2

34th Annual Conference of the Military Testing Association

PROCEEDINGS
Volume I

THE MILITARY TESTING ASSOCIATION

MTA

Prepared and Coordinated by

Navy Personnel Research and Development Center
San Diego, California

26 - 29 October 1992

34th Annual Conference of the Military Testing Association

Hosted by the

Navy Personnel Research and Development Center
53335 Ryne Road
San Diego, California 92152-7250

26 - 29 October 1992

Conference Committee

Commanding Officer, NPRDC	*CAPT Thomas Finley*
Chairs, MTA-92 Conference	*Drew Sands* *John Ellis*
Chair, Program Committee	*John Ellis*
Chair, Registration Committee	*Margie Sands*
Chair, Hospitality Committee	*John Ellis*
Chair, Publication Committee	*Drew Sands* *John Ellis*

Acknowledgments

The members of the MTA-1992 Conference Committee wish to express their appreciation for the expertise and dedication of the following individuals. Each person had an important role in the success of the 34th Annual Conference of the Military Testing Association.

Audiovisual Support

Jim Julius

Database

Margie Sands
Anthy Dunlap
Christina Reese

Finance

John Ellis
Drew Sands

Hospitality

John Ellis
Kathy Ellis

Program Organization

Margie Sands

Program Production

John Ellis
Margie Sands
Ruth Ireland
Mely Leano
Marci Barrineau

Proceedings Publication

Drew Sands
John Ellis
Marci Barrineau
Carmen Fendelman

Registration

Margie Sands
Anthy Dunlap
Christina Reese
Peggy Laone

Session Chairs

Herb Baker
Mike Cowen
Ronna Dillon
John Ellis
John Folchi
Paul Foley
Dennis Gettman
Alice Gerb
Janet Held
Rebecca Hetter
Jerry Laabs
Gerald Larson
Reynaldo Monzon
Robert Morrison
Randolph Park
Shelley Perry
Josephine Randel
Malcolm Ree
Dave Robertson
Carol Robinson
Ellie Robinson
Hendrick Ruck
Michael Rumsey
Drew Sands
Dan Segall
Steve Sellman
George Seymour
Wallace Sinaiko
Mannie Somer
Friedrich Steege
Hervey Stern
Walt Thode
Tom Trent
Wolfgang Weber
John Welsh
Lauress Wise
Martin Wiskoff

Foreword

The Proceedings of the 34th Annual Conference of the Military Testing Association document the presentations given during the Conference paper and panel sessions. The papers present a broad range of topics by contributors from the military, private industry, and the educational communities. It should be noted that the papers reflect the opinions of the authors and do not necessarily reflect the official policy of any military service, institution, or government.

Contents

Conference Opening

Keynote Address

Session 1

Session 2

Session 3

Session 4

Session 5

Session 6

Session 7

Session 8

Session 9

Session 10

Session 11

Session 12

Session 13

Session 14

Session 15

Session 16

Session 17

Session 18

Session 19

Session 20

Session 21

Session 22

Session 23

Conference Opening

The following is a summary of the *Welcome* speech given by Dr. Sorenson, Navy Personnel Research and Development Center's Technical Director (Acting), at the 34th Annual Conference of the Military Testing Association.

Good morning, ladies and gentlemen of the Military Testing Association. The Navy Personnel Research and Development Center and I as its Technical Director are very pleased and proud to be a part of this 34th annual conference.

The purpose of MTA and your meeting this week is important to your individual organizations and to the military as a whole.

Profound changes are taking place that will affect the military and, hence, the MTA. Changes in roles and missions, in technology, and in personnel.

Change in Roles and Missions—The Service's management initiative, Project Reliance, is the most comprehensive restructuring effort involving the technology base in over 40 years. DoD challenged the services to create a new approach to science and technology management that would increase efficiency and reduce overlap in the research, development, test, and evaluation activities. The Armed Services Training and Personnel Systems Science and Technology Evaluation and Management Committee (TAPSTEM) is the recognized integrating mechanism responsible for compliance with Reliance objectives in Manpower and Personnel and Training Systems technology areas. The goals of the TAPSTEM committee are to facilitate management coordination, improve information exchange, and accomplish training and personnel systems science and technology activities pertinent to the changing missions of the Army, Navy, Marine Corps, and Air Force. These role and mission changes were ignited by the greatly reduced threat of blue-water naval conflict and increased focus on limited "come-as-you-are" regional crises, including military interventions like the Panama invasion, Desert Storm, the Los Angeles riots, and Hurricane Andrew.

Change in Technology—Computing and communications technologies are advancing at a rapid pace. Computer networks are a fact of life and these "information technologies" are central to our daily operations. The pace will not slow as information technologies continue to advance like a speeding train.

Change in Personnel—From 1990-1997, civilian end strength will be cut 25%. (By necessity, overhead expenses must shrink.)

These changes demand a reevaluation of our operations: what are we doing, why are we doing it, how are we doing it, and who are we doing it for? These changes provide you with an opportunity to actively reshape your programs.

You need to identify your customers and their needs. You need to leverage technology to increase productivity. You need to manage more cost-effectively. And you need to leverage your MTA connections to share capabilities and resources among all organizations--not just within your own.

I challenge you to seize the opportunity. Use your expertise to educate and train; to inject quality into every step of the military personnel process; to increase productivity and efficiency; and to meet the needs of your customers.

I appreciate very much the opportunity to speak with you this morning and to help launch your 34th annual conference. I wish you the best of good debate and discussion, and look forward to the report of your deliberations.

Into the 21st Century:
The Changing Face of Military Manpower

by

W. S. Sellman

Office of the Assistant Secretary of Defense
(Force Management and Personnel)

Introduction

In September 1968, I attended my first MTA conference in San Antonio. I was a relatively new Air Force Captain at the time, and MTA was a very big event for me. It was the first conference at which I presented a professional paper and perhaps more importantly, I served as the transportation and logistics officer for the conference. In the later regard, my wife had just given birth to our second daughter and was only recently returned home from the hospital. Of course, I was at the hotel most of the time coordinating the pick-up of conferees at the airport, arranging bus transportation to Hemisfair, obtaining tooth paste and shaving cream for conferees and doing just about anything my commander needed to have done. Today, some 25 years later, whenever my wife feels the need to punish me, she recalls the 1968 MTA. There are other foibles that I have committed over the years that she uses to beat upon me but the '68 MTA with me at the hotel and her home with the two children is something that has never left her consciousness.

Over the years, it has been my privilege to attend 17 MTAs in places like French Lick, Indiana; Lake Geneva, Wisconsin; and Gulf Shores, Alabama, not to mention New York City, Washington, DC, San Antonio, San Diego, Oklahoma City, and of course Munich. As is true with all of you who have attended MTAs, I have many wonderful and rewarding memories. In fact, if you want to know about one of them find Marty Wiskoff and ask him about the time that the Germans gave us awards for outstanding service to MTA. I also have had the opportunity to present many papers at MTA conferences and to ghost-write three keynote speeches for my Air Force bosses--1973, 1977 and 1984; but, I must confess that never in my wildest dreams did I ever believe that someday I would be standing here before you, presenting the MTA keynote on my own behalf. So it is a particular honor and pleasure for me to be here to discuss what I hope is a vision of tomorrow's military and what that might mean for personnel testing.

Let me begin by saying that these remarks are my own speculations and do not reflect the official position of the Department of Defense, and in the years to come I may, in fact, even deny having said them. It depends on how well things work out. Second, anyone who tries to predict the future is foolish; weathermen have the best track records and everyone knows about their levels of precision in terms of their predictions. Nevertheless, it is fun to try to be a futurist, so here we go.

The last three years have been interesting times for military manpower. First, we started with the military drawdown. Then, after one year of drawdown, came Operations Desert Shield and Desert Storm and that entailed a call-up of the Reserves, stop loss, renewed recruiting, and an increase in military end strength. Desert Storm was followed by a resumption of the drawdown--voluntary separation pays, Selected Early Retirement Boards, and wide-spread fear of reductions in force (RIFs). I believe the next 10 years will be even more exciting and challenging as we move toward the 21st century.

Most aspects of the 1990s are as yet unknown, but I believe that there are three things in 1992 that we know for sure: (1) the military will continue to get smaller, (2) the Defense budget also will continue to shrink, and (3) there will be a changing role for military personnel. So for the next few minutes, I would like to muse about the future composition of the military, and the changing role of the military in American society.

Composition of Future Military

Today's military is the best in the history of our country; aptitude levels and educational attainment for both new recruits and the career force have never been higher. This year, Fiscal Year 1992, new recruits will be composed of 99 percent high school diploma graduates and 75 percent of new recruits with diplomas will score at or above the 50th percentile on the Armed Forces Qualification Test. As you might suspect, the Services will do everything in their power to try to preserve that quality, and they probably can accomplish that goal given reasonable resources.

As you also know, Congress is under enormous pressure to reduce the deficit, and with the end of the Cold War, money for military manpower will become ever more scarce. For example, since Fiscal Year 1989, recruiting resources are down by 20 percent. Advertising is down by 55 percent and there was talk this year of prohibiting the Services from television advertising. The number of production recruiters is down by 10 percent with another 10 percent scheduled for reduction over the next two years. We have cut the number of recruiting offices around the country by over 20 percent. Large budget cuts can drive down recruit quality, and cuts in pay, reenlistment bonuses, and other benefits can have an negative impact on the career force.

Will this happen in the decade ahead? Both President Bush and Governor Clinton are pledged to maintain a strong military and to avoid the low quality, hollow force of the late 1970s. I believe that we will continue to have a high quality force in terms of aptitude and education. The size of the force will both require it and allow it. No matter who wins the election, I believe the force will continue to shrink to levels somewhere between 1.3 million and 1.4 million members as opposed to the 1.6 million which is now commonly reported. This will push us from a force of specialists to one more of generalists. People will be asked to perform a wider variety of tasks with less direct supervision. High quality people will be needed, and as the civilian youth population begins to grow, the selection ratio (vacancies compared to available manpower) will become more favorable.

Of course, the changing demographics of the youth population may negatively affect that situation. In the years to come, there will be more minorities, more women, and more aliens in the labor force. Many minorities will come from high schools with marginal programs. So unless education reform works, minority high school graduates may lack the basic skills needed for entry level positions. This could pose problems not only for the military but for industry as well. However, vigilance on the part of the Department, the Congress, and the Administration should help avoid serious declines in force quality. In the long run, sufficient resources are the key; with them we are okay; cut too deep and there could be a crisis in military manpower.

In the future, the Services also will continue to seek diversity in their military personnel. It will be of continuing importance that the Services reflect the composition of the society they protect. During Operations Desert Shield and Desert Storm, much was said in the media about overrepresentation of minorities and that they would pay a disproportional burden if we went to war. Fortunately, casualties in the Persian Gulf never reached predicted levels. However, with the return to force drawdown, critics alleged that those cutbacks would have an adverse impact on minorities. Obviously, this is a "dammed if you do, dammed if you don't" situation.

Following Desert Storm, there was a serious decline in the propensity of Black youth to enlist. We talked to recruiters from

inner cities and other areas with heavy minority representation and there seemed to be three reasons for lowered propensity:

- **Black leaders (ministers, teachers, etc.) were still concerned about the "burden" problem.**

- **Black entertainers had released antimilitary music (particularly rap music).**

- **Many youth believed that with the downsizing there were no longer opportunities within the military.**

During the early part of Fiscal Year 1992, the percentage of new recruits who were Black did decline, but that proportion has stabilized at about 17 percent, as compared with about 14 percent Blacks in the civilian youth population. Analyses also have shown that minorities in the career force are not really being disproportionately hurt by drawdown activities. I believe that throughout the 1990s minorities will continue to enter service (both officers and enlisted) at levels slightly above their percentage in the youth population. A smaller military will still have opportunities for rewarding careers for capable young people.

Women in the military, and particularly in combat, also has been and will continue to be a defining issue. The President's Commission on the Assignment of Women in the Armed Forces has been considering the role of military women for the past year. Its report is due to the President on November 15. The Commission has heard extensive testimony from people within and outside the military on all aspects of women in combat. It also has sponsored surveys of military members, general/flag officers, and the American public regarding their attitudes about women in combat, and on November 2-3, the Commission will decide what it will recommend to the President. My thoughts on this issue are that women should be allowed to hold any job within the military for which they qualify. Standards should be set relative to performance and both women and men should meet those standards.

My prediction is that the Commission will endorse women to fly combat aircraft and to serve on combatant ships, but the Commission will not endorse women in ground combat. Even though the burden of proof lies with those against women in combat, the issues of strength/stamina, unit cohesion, pregnancy, men's protective attitudes towards women, potential torture of women POWs, etc. are still of considerable concern. Yet, by the year 2000, I believe these issues will be resolved (either politically or through research) and that women will be allowed to participate in any occupation for which they qualify. I guess my egalitarian tendencies are showing; by the new century people will be judged on their skills, abilities, and potential — not membership in a group.

Role of Military in American Society

Now, let me spend a few minutes talking about the future role of the military in society. On June 23, Senator Sam Nunn gave a speech on the floor of the Senate in which he recommended that military personnel be used to assist civilian efforts in critical domestic needs. In the speech, Senator Nunn suggested several ways the military could help local communities:

- Military personnel as role models--hard working, disciplined men and women who command respect can serve as a powerful force among our young people.

- Rehabilitation and renewal of community facilities--schools, public housing, recreational facilities, roads, and bridges are in need of repair.

- Summer outreach programs for disadvantaged children who need help with basic skills.

- Medical transport--military helicopters could be used to carry seriously ill or injured people to hospitals.

- Nutrition programs--military personnel could assist state and local welfare agencies in distribution of surplus food.

In his speech, Senator Nunn also stated three principles (or constraints) under which this type of program would work:

- It could not interfere with the military mission.

- It could not compete with private or other government efforts.

- It could not be used as justification for additional resources.

Of course, the Services have always been involved in such activities, but they have never been centralized within the Department. For example:

- JROTC - there are nearly 1,600 units in high schools around the country emphasizing citizenship, leadership, and self-esteem. DoD provides uniforms, equipment, and

shares the cost of retired military instructors. We now are planning to expand the number of JROTC units to about 3,500 and pay more of the costs for schools in disadvantaged neighborhoods.

- **Service academies have conducted educational programs on campus in areas of science, math, and computers.**
- **Military personnel have long served as role models--military people already volunteer their time in scouts, churches, community organizations, and charities, etc.**
- **The military supports the homeless with excess materials (blankets, cots, food) to local organizations.**
- **The military is involved in law enforcement through counter narcotic efforts.**
- **The military already has medevac programs.**
- **And of course, we were involved in emergency response and disaster relief with the recent experiences with Hurricane Andrew.**

As Senator Nunn promised in his speech, he included a new "Civil-Military Cooperative Action Program" in the Fiscal Year 1993 DoD Authorization Act. The President signed that bill into law Friday night (October 23, 1992) at 5:30. The law directs the Secretary of Defense to set up programs to use skills, capabilities, and resources of the Armed Forces to meet domestic needs of the United States. The objectives of the program, now codified are:

- To enhance individual and unit training and morale of military personnel through meaningful community involvement.
- To encourage cooperation between civilian and military sectors of society in addressing domestic needs.
- To advance equal opportunity.
- To enrich the civilian economy through education, training, and transfer of technology.
- To improve environmental, economic, and social conditions.
- To provide opportunities for disadvantaged citizens.

To accomplish these objectives, the Secretary of Defense shall encourage establishment of advisory councils at regional, state, and local levels to coordinate projects and activities. Membership will include military officials and representatives from local and state agencies, and representatives from civic/social service organizations, business and labor. But, you do not have to be an oracle to see this sort of formalized program is the wave of the future, especially in the aftermath of Hurricane Andrew. What I do not know is how this will play in terms of the development of national service programs.

Four years ago, Senator Nunn and Congressman Dave McCurdy (Democrat-Oklahoma) introduced national service legislation. This was part of the Democratic Leadership Council's (DLC) agenda. Governor Clinton was chairman of the DLC at the time. Today with the downsizing, about 100,000 young people who could have enlisted three or four years ago will be denied that opportunity. What will happen to those 100,000 young people? They are good young people. They are bright. They are well-educated, and only because we are going to have a smaller military force will they be denied entrance. I believe that they will displace people who are in entry level positions in the civilian sector. In turn, that displacement will ultimately affect even more disadvantaged young people. Young people coming from Appalachia, from the inner-cities, who will no longer have the opportunity to compete for entry level jobs. I hope that the United States will not throw these young people away. Perhaps, national service is a concept whose time has come, and perhaps this civilian-military cooperation initiative is Sam Nunn's national service program to be run by the Department of Defense.

One last aspect of the military role in society that I would like to cover pertains to a garrisoned versus deployed force. With a smaller force and less Defense dollars, there will be fewer overseas assignments and less frequent reassignments within the United States. It is likely that people will be assigned to the same base for longer periods of time, maybe even up to eight years or more. The issue will be whether the force looks outward into the community or looks inward to itself, as it did during World War II.

I believe military personnel will become more and more a part of the communities in which their bases are located. They will become more involved in local off-base activities, will shop in civilian grocery stores and department stores, will live among non-military neighbors, and send their children to community schools. This will lead to a reinforcement of the concept of citizen-soldier, if that concept needs any reinforcing. Military people and their families will hear the opinions and concerns of their neighbors, can explain military positions on different issues or situations (i.e., why we don't send troops to Bosnia) to those same neighbors, and everyone benefits because of better understanding and a lessening of the "them and us" concept. I also believe the military will develop a greater sense of belonging, of social integration, a sense of identify, of caring about those people with whom they live and ultimately defend and protect.

In sum, I believe that the military of the 21st century has a real chance to be even better than the military of today. It will be composed of smart, well educated people of diverse backgrounds who will be judged by their individual attributes. They not only will be well trained and ready for all contingencies, but also will make their communities better places to live through personal involvement and service. Tomorrow's military is one which will deserve our honor and pride.

Implications for Testing

Let me close with some observations about how this changing military will affect military personnel testing because after all, this is the Military Testing Association. When I began this morning, I mentioned that I had previously been involved in the preparation of three MTA keynote speeches. I would like to go back to 1973 and 1977 and see how some of the points in those speeches relate to our evolving military. In 1973, then Lt General John Roberts, Deputy Chief of Staff for Personnel, said that tests of the future should focus more on the individual--on his or her unique talents and desires and that tests should maximize the opportunity for choice--both by the individual and by his or her employer. Then in 1977, Major General Herb Emanuel, Vice Commander of the Air Force Military Personnel Center gave examples of research topics that addressed General Roberts' principles. Two of the research topics that were discussed by General Emanuel, computerized adaptive testing and prediction of first-term enlisted attrition, will be discussed in some depth at this MTA.

I still believe that the 1973 principles stated almost 20 years ago are relevant. Our military of the future will be required to operate equipment and systems of ever increasing sophistication and with fewer resources--money and people. We, as testing experts and personnel managers, must develop and use selection and classification procedures that more than ever will maximize performance and allow us to get the most from our increasingly scarce dollars. In the case of computerized testing, we must be able to demonstrate to Congress that we can sufficiently improve how we do business to amortize the costs of buying the hardware. At this conference, you will hear about innovative ways to do enlistment testing, including the possibility of contracting outside the Department with companies to purchase computer systems and to administer the tests for us. That would solve the problem of buying equipment that is no longer state-of-the-art before it is delivered. We also continue to struggle with attrition management and the use of education credentials in the enlistment screening process. This afternoon, you will have the opportunity to learn the latest on the politics and progress of this seemingly intractable area.

In the final analysis, the message that I would like to leave with you this morning is this. Even though the military of the future will be smaller, more mobile, required to be more proficient in a wider variety of tasks, and perhaps somewhat different in its social composition, it will still be a military of people. We must select the best available men and women, and place them into jobs for which they are qualified, and we must never forget those people are individuals, with skills, abilities, aspirations, motivations--i.e. talents that are unique to them. We must maximize their future contribution to the military, but we must always treat them with dignity. In the future, we will need innovative testing research and methods more than ever. Let us join together in the knowledge that there is excitement and opportunity to improve the management of the most precious asset entrusted to any organization--its people. And let us do it wisely.

Computer-Based Testing R&D at the Navy Personnel Research and Development Center: An Overview

W. A. Sands*
Director, Personnel Systems Research Department
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

ASVAB

The Armed Services Vocational Aptitude Battery (ASVAB) is a multiple aptitude test battery taken by all applicants for enlistment into the U. S. military services. This conventionally-administered, paper-and-pencil test battery (P&P-ASVAB) involves eight power tests and two speeded tests. The power tests are General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Auto and Shop Information, Mathematics Knowledge, Mechanical Comprehension, and Electronics Information. The speeded tests are Numerical Operations and Coding Speed. The results of these tests are used for both initial eligibility determination and subsequent classification into entry-level training.

The U.S. Military Entrance Processing Command administers the ASVAB under two programs: the Enlistment Testing Program and the Student Testing Program. In the Enlistment Testing Program, ASVAB is administered to over 800,000 applicants each year, in approximately 70 Military Entrance Processing Stations (MEPS) and approximately 970 Mobile Examining Test Sites (METS) nationwide. In the Student Testing Program, ASVAB is administered in over 15,000 schools to over 1,000,000 students annually.

* The opinions expressed in this paper are those of the author, are not official, and do not necessarily reflect those of the Navy Department.

CAT-ASVAB

The U. S. Department of Defense has a Joint-Service Program to develop, test, and evaluate a Computerized Adaptive Testing version of the battery (CAT-ASVAB). The Department of Navy was designated as the Executive Agent, with the Navy as the Lead Service. The Navy Personnel Research and Development Center (NPRDC) was designated the Lead R&D Laboratory. The Air Force was assigned responsibility for the development of the large test item banks needed for CAT-ASVAB. The Army was assigned responsibility for the procurement, deployment, and implementation of the full-scale operational testing system.

CBT R&D at NPRDC

This symposium includes six papers that describe some of the research and development in Computer-Based Testing (CBT) conducted by the Navy Personnel Research and Development Center. This R&D has included studies emphasizing psychometrics and studies involving operational implementation.

The first paper, written by R. D. Hetter, D. O Segall, and B. Bloxom, addresses the issue of medium of administration in item calibration. Historically, new test items have been calibrated by administering them to large numbers of examinees in a paper-and-pencil format. The development of the microcomputer-based CAT-ASVAB system has opened the possibility of collecting item calibration information using a computer-based test administration. The purpose of this study was to determine if there was any difference between items calibrated on data collected under a paper-and-pencil vs. a computer-based administration, when the items were used in a CAT-ASVAB administration. Results indicated that the medium of administration had no effect on the reliability of the CAT-ASVAB scores.

The second paper, written by K. E. Moreno and D. O. Segall, examines the measurement precision of CAT-ASVAB. This study involved two aspects: (1) a comparison of the alternate form reliability of CAT-ASVAB and P&P-ASVAB, and (2) a comparison of (a) the correlations between CAT-ASVAB scores and corresponding P&P-ASVAB scores and (b) the correlations between corresponding tests in two forms of P&P-ASVAB. Results indicated that seven of the ten CAT-ASVAB tests were significantly more reliable than the corresponding P&P-ASVAB tests, while there was no difference on the other three tests. Finally, the correlations between CAT-ASVAB tests and P&P-ASVAB tests were as high as the correlations between two forms of P&P-ASVAB.

The third paper, written by D. O. Segall, addresses equating, a central issue in evaluating CAT-ASVAB as a potential replacement for the P&P-ASVAB. Equating tests involves placing the scores from two instruments on the same metric, so that their scores are interchangeable. The established approach to equating forms of ASVAB has involved two stages: (1) a provisional equating, based upon data collected under non-operational conditions, and (2) an operational equating, based upon data collected under operational conditions. If the results are in close agreement, this increases confidence that the equating is appropriate. If the results differ, this raises questions concerning the data representativeness, collection procedures, and analyses. This paper discusses procedures for equating CAT-ASVAB to P&P-ASVAB.

The fourth paper, written by G. E. Larson and D. L. Alderton, examines the reliabilities and practice effects for the tests in the Enhanced Computer-Administered Test (ECAT) battery. The ECAT tests are promising experimental tests that are possible candidates for incorporation into ASVAB at some time in the future. Results showed practice effects for all the experimental tests except Assembling Objects. Reliabilities for the ECAT tests ranged from .75 to .91, a range comparable to existing ASVAB tests. Females obtained lower scores on all the psychomotor tests and some of the spatial tests. Finally, it was noted that many of the tests had significant correlations with grade point average in civilian schools, suggesting that the tests may have validity for military training school selection.

The fifth paper, written by J. H. Wolfe and D. O. Alderton, examines the incremental validity of six experimental tests for predicting performance criteria in nine Navy training schools. These tests included measures of working memory, spatial ability, reasoning, and perceptual speed. Results showed that the experimental tests improved prediction over the standard ASVAB tests for five of the nine schools.

The last paper describes the Operational Test and Evaluation (OT&E) of the Computer-Based Testing battery. This battery includes the operational administration of CAT-ASVAB and the experimental administration of ECAT. The purpose of the OT&E is to gather information needed for a nationwide implementation. This OT&E is currently on-going in three locations: San Diego California, Los Angeles California, and Jackson Mississippi. Two additional locations will be involved in the future: Baltimore Maryland and Denver Colorado. Preliminary results from the initial locations are very favorable. The test administrators have experienced few problems running the CBT system and, in general, prefer it to P&P-ASVAB. In addition, this generally very positive attitude has been expressed by both examinees and recruiting personnel.

ITEM CALIBRATION MEDIUM EFFECT ON CAT SCORES

Rebecca D. Hetter and Daniel O. Segall
Navy Personnel Research and Development Center

Bruce M. Bloxom[1]
Department of Defense Manpower Data Center

INTRODUCTION

The Navy Personnel Research and Development Center is conducting research to design and evaluate a computerized adaptive test (CAT) as a potential replacement for the paper-and-pencil Armed Services Vocational Aptitude Battery (ASVAB). In support of this effort, the Accelerated CAT-ASVAB Program (ACAP) is evaluating item pools specifically developed for computerized adaptive testing.

Background

An important question in the development of item pools for computerized adaptive tests is whether data for calibrating items should be collected by a paper-and-pencil (P&P) or a computer administration of the items. Even though research shows that computerized adaptive tests with P&P item calibrations can have validities comparable to conventional P&P tests (Moreno, Segall and Kieckhaefer, 1985, p. 29-33), there is continuing uncertainty about how much less than optimal these computerized adaptive tests might be.

The concern about medium of administration (MOA) in item calibration is that item parameters for some types of items (e.g., items with long paragraphs or with graphics) may differ between computer and P&P administrations. This could result in less-than-optimal item selection and score estimation in adaptive tests. If P&P administrations do not yield precise enough calibrations, items must be administered by computer during calibration just as they are during testing.

Objective

The purpose of this MOA study is to evaluate the effect on adaptive scores of using a P&P calibration. Specifically, to what extent do adaptive scores obtained with computer-administered items and a P&P calibration correspond to adaptive scores obtained with computer-administered items and a computer calibration? If there is a lack of correspondence, is it greater than would be found by chance, i.e., greater than would be found with the use of another computer calibration?

METHOD

Fixed blocks of items were administered by computer to one group of examinees and by P&P to a second group. These data were used to obtain computer-based and P&P-based calibrations of the items. Each calibration was then used to estimate item response theory adaptive scores (thetas) for a third group of examinees who had received the items by computer. The effect of medium of administration was assessed by comparative analyses of the thetas using the alternative calibrations.

Subjects

The subjects were Navy recruits who were randomly assigned to one of three groups. Data were collected for 2955 examinees, 989 in Computer-Group 1, 978 in the P&P Group, and 988 in Computer Group 2. These sample sizes provide enough data for independent calibrations, since Hulin et al (1983, p. 101-110) simulation results suggest that substantially larger samples produce little improvement in the precision of item characteristic curves and scores, given the number of items (40) used in these calibrations.

Testing was conducted at a Recruit Training Center in San Diego, California. ASVAB scores of record were obtained for nearly all of the recruits and were used to assess whether the groups were

[1] The opinions expressed in this paper are those of the authors, are not official, and do not necessarily represent those of the Department of the Navy or the Department of Defense.

comparable in ability levels.

Items

The items were taken from pools specifically developed in support of CAT-ASVAB by Prestwood, Vale, Massey, and Welsh (1985). Forty items from each of four ASVAB content areas (general science, arithmetic reasoning, word knowledge, and shop information) were administered by computer to Groups 1 and 3, and by P&P to Group 2. The items were conventionally administered in order of ascending difficulty, using the difficulties obtained by Prestwood et al (1985). The three groups received the same items with the same instructions and practice problems, in the same order and with the same time limits. Although only four of the 11 CAT-ASVAB subtests were included in this study, subtest order was the same as in the CAT-ASVAB. Time limits were prorated from 95 percent completion times for the same content areas in ACAP, with the addition of 10 percent to allow for a higher completion rate. Subtest order and time limits were as follows: GS (19 min), AR (63 min), WK (16 min), and SI (17 min). The total time was 115 minutes.

The 40 items included 34 high-usage items (usage obtained from ACAP simulation studies) and six "seeds" (not-scored items administered for the purpose of gathering data for on-line calibration research). The booklet format was the same used in the original P&P calibration by Prestwood et al (1985), and the computer format was the same used in ACAP. Practice problems and instructions were also as in ACAP.

Item Calibrations

Item response theory parameter estimates based on the three-parameter logistic model (Birnbaum, 1968) were obtained in separate calibrations for Computer Group 1 (calibration C1) and for the P&P Group (calibration C2). The data sets on which the calibrations are based are labelled U1 and U2, correspondingly. The calibrations were performed with LOGIST6 (Wingersky, Barton, & Lord, 1982) a computer program that uses a joint maximum-likelihood approach. Data set U3 from Computer Group 2 was not used in the calibrations. The design with the corresponding notation is summarized in Table 1.

Table 1

Calibration Design

Group	Medium	Data Set/ Item Responses	Item Parameters/ Calibrations
1	Computer	U1	C1
2	P&P	U2	C2
3	Computer	U3	n.a.

Scores

For each recruit in Group 3 two ability scores were computed: T1 and T2 (see Table 2). Both scores were based on U3 responses. T1 scores were calculated using the computer-based item parameters (C1). T2 scores were calculated using the P&P-based item parameters (C2). T1 and T2 were adaptive scores, computed as described below using only 10 of the 40 responses from a given examinee.

Adaptive Scores. To compute the adaptive thetas (T1 and T2), 10-item adaptive tests were simulated using actual examinee responses. Owen's Bayesian scoring (Owen, 1975) was used throughout the test to update the ability estimate, and a bayesian modal estimate was computed at the end of the test to obtain the final score. Items were selected from information tables on the basis of maximum information. (An information table consists of lists of items by ability level. Within each list, all the items in the pool - 40 in this case - are arranged in descending order of the values of their information functions computed at that ability level. This study used 37 ability levels equally spaced along the [-2.25, +2.25] interval).

Table 2 summarizes the method used for computing the theta scores used in the analyses.

ASVAB Scores. ASVAB subtest scores for the four content areas of interest were obtained from the records for most of the examinees. The scores included General Science (GS), Arithmetic Reasoning (AR) Word Knowledge (WK), Auto Shop (AS), plus the Armed Forces Qualification Test (AFQT) composite. It should be noted here that the ASVAB's Auto Shop subtest covers two content areas: Auto

Table 2

Computation of Theta Scores				
Calibration Parameters	Response Set	Scoring Method	Test Length	Theta
C1 (Computer Group 1)	U3	Adaptive	10 Items	T1
C2 (Paper-and-Pencil)	U3	Adaptive	10 Items	T2

Information and Shop Information, whereas in the CAT-ASVAB each area constitutes a separate subtest. Since only Shop Information (SI) was administered in this study, ASVAB-AS was compared to MOA-SI.

RESULTS

Calibration Samples

Two cases had fewer than 10 valid responses (not-reached greater than 30) in Subtests WK and SI, Computer Group 2, and LOGIST omitted them from the calibrations. The cases were consequently eliminated from all subsequent analyses of WK and SI, Computer Group 2. Final sample sizes were 989 for Computer-Group 1, 978 for the P&P Group, 988 for GS & AR in Computer-Group 2, and 986 for WK & SI in Computer-Group 2.

Analyses using ASVAB scores of record were performed to determine whether the three groups were comparable in examinee ability. Results clearly indicated that there are no differences among the groups.[2]

Reliability Analysis

A design was developed to assess the effect of calibration medium on test reliabilities. The statistical model and the LISREL specifications are described below. It should be noted that these tests would assess overall effect across the four content areas simultaneously; if a significant effect was found, further analyses would be required to attribute the error to specific subtests.

Statistical Model. Let's assume that the observed theta values, $\hat{\theta}$, have three components: true ability level θ, measurement error ε, and random error due to calibration δ. Then,

$$\hat{\theta} = \lambda\ (\theta + \varepsilon) + \delta$$

$$\hat{\theta} = \lambda\ \xi + \delta$$

where $\xi = \theta + \varepsilon$, the true ability plus the error of measurement and λ is a scale factor. Then the basic measurement model can be described by the eight equations listed below. Also listed next to each equation is the corresponding subtest score and the data sets used to compute it.

Equation	$\hat{\theta}$	Responses	Item Parameters
$\hat{\theta}_1 = \lambda_1\ \xi_1 + \delta_1$	T1-GS	U3	Computer
$\hat{\theta}_2 = \lambda_2\ \xi_2 + \delta_2$	T1-AR	U3	Computer
$\hat{\theta}_3 = \lambda_3\ \xi_3 + \delta_3$	T1-WK	U3	Computer
$\hat{\theta}_4 = \lambda_4\ \xi_4 + \delta_4$	T1-SI	U3	Computer
$\hat{\theta}_5 = \lambda_5\ \xi_5 + \delta_5$	T2-GS	U3	P&P
$\hat{\theta}_6 = \lambda_6\ \xi_6 + \delta_6$	T2-AR	U3	P&P
$\hat{\theta}_7 = \lambda_7\ \xi_7 + \delta_7$	T2-WK	U3	P&P
$\theta_8 = \lambda_8\ \xi_8 + \delta_8$	T2-SI	U3	P&P

[2] Results are available by request.

The statistical tests available for selecting the best-fitting model consist of: (1) estimating the model in which certain parameters are set to be equal, and (2) estimating a less constrained model. The test consists of assessing the statistical significance of the improvement in fit going from the more constrained model to the less constrained model. If the more constrained model fits the data as well as the less constrained model (i.e., within sampling error limits), then one may conclude that the constraints do not seriously erode the fit of the model.

In this case, one model is specified such that the calibration errors of the pseudo-true test scores are constrained to be equal for the computer-based and the P&P item parameters; another model is specified such that these calibration errors are free to vary between the two media of administration. If the constrained model provides just as good a fit as the free model, then constraining calibration errors to be equal across item-parameter sets does not erode the fit of the model to the data, and one would conclude that the calibration errors of the ability scores were equal for computer and P&P item-parameters.

According to the model, the variance-covariance matrix Σ among the observed scores has the form:

$$\Sigma = \Lambda_X \Phi \Lambda_X' + \theta_\delta$$

where Λ_X is a diagonal matrix with standard deviations in the diagonal, θ_δ is a diagonal matrix of variances attributable to calibration error, and Φ is the attenuated correlation matrix among the ability values ξ. Notice that the matrix Φ is attenuated from one source of error only, that source attributable to the calibration; Φ is not attenuated with respect to measurement error. The fixed and estimated parameters of this model are displayed in Table 3.

Notice in Table 3 that the correlation of a subtest with itself (across media) is equal to one, and that the correlations between same-name subtests are assumed to be equal both across and within calibration media, e.g., for *AR* and *GS*,

$$r\,(AR_1,GS_1) = r\,(AR_2,GS_2) = r\,(AR_1,GS_2) = r\,(AR_2,GS_1)$$

Table 3

Correlation and Variance Matrices in the Model

	Correlation Matrix Φ							
	T1 (Computer)				**T2 (P&P)**			
	GS-1	AR-2	WK-3	SI-4	GS-5	AR-6	WK-7	SI-8
GS-1	1.0							
AR-2	(2,1)	1.0						
WK-3	(3,1)	(3,2)	1.0					
SI-4	(4,1)	(4,2)	(4,3)	1.0				
GS-5	1.0	(2,1)	(3,1)	(4,1)	1.0			
AR-6	(2,1)	1.0	(3,2)	(4,2)	(2,1)	1.0		
WK-7	(3,1)	(3,2)	1.0	(4,3)	(3,1)	(3,2)	1.0	
SI-8	(4,1)	(4,2)	(4,3)	1.0	(4,1)	(4,2)	(4,3)	1.0
	Variance Matrix θ_δ							
θ_δ	(1,1)	(2,2)	(3,3)	(4,4)	(5,5)	(6,6)	(7,7)	(8,8)

In the correlation matrix Φ, the numbers in parentheses refer to the subtest scores used to compute the correlation, e.g., (2,1) represents the correlation between T1 scores from subtests AR and GS (AR-2, GS-1). The correlation between T2 subtests AR and GS should be (6,5); however it is represented by (2,1) because under the model they are assumed to be equal.

In the variance matrix θ_δ, the numbers represent variances, e.g., (7,7) represents the variance of the T2 scores from subtest WK (WK-7, WK-7).

LISREL Models. To test the model fit, two models were specified and corresponding LISREL runs were performed. In Model 1, the variances of errors due to calibration (θ_δ) were free to vary between the two media of administration. In Model 2, the variances of errors due to calibration were constrained to be equal for same-name subtests.

The Φ constraints displayed in Table 4 were imposed for both Model 1 and Model 2.

The LISREL output yields a chi-square statistic that is a measure of how much Σ differs from S, that is, of how well the model fits the data. The difference in the chi-squares from the two models is also a chi-square with *df* equal to the difference in *df* from the two models. If it is not significant, then the data satisfies/fits the model independently of the calibration errors. That is, errors due to calibration across media for same-name subtests are equal.

The LISREL specifications for Model 1 were:

(1) Lambda-X = Diagonal Matrix, Free
(2) PHI = Symmetrical Matrix, Free
(3) Theta-Delta = Diagonal Matrix, Free

The LISREL specifications for Model 2 were:

(1) Lambda-X = Diagonal Matrix, Free
(2) PHI = Symmetrical Matrix, Free
(3) Theta-Delta (TD) = Diagonal Matrix with Constraints
(4) TD Constraint #1: All off-diagonal θ_δ fixed at zero, and
(5) TD Constraint #2: θ_δ $(COMPUTER) = \theta_\delta$ $(P\&P)$, that is,
$\theta_\delta(1,1) = \theta_\delta(5,5)$; $\theta_\delta(2,2) = \theta_\delta(6,6)$; $\theta_\delta(3,3) = \theta_\delta(7,7)$; and $\theta_\delta(4,4) = \theta_\delta(8,8)$.

Table 4 presents goodness-of-fit statistics for these models. The likelihood ratio chi-square value of the model in Model 1 was 14.07 with 14 *df*. The result is not statistically significant, indicating that the model adequately explains the observed covariance matrices. Results for Model 2 show a chi-square value of 19.57 with 18 *df*, which is also not statistically significant. The difference in chi-squares between Model 1 and Model 2 is itself distributed as a chi-square with *df* equal to the difference in *df* from Model 1 and Model 2. This value (5.50 with 4 *df*) was not significant, indicating that allowing the error term to be free does not change the fit of the model.

Table 4

Test for Equality of Reliabilities

Model	χ^2	df	Prob	Gof	Adj-Gof	RMSR
1. θ_δ = free	14.07	14	0.445	0.996	0.991	0.002
2. $\theta_\delta(COMPUTER) = \theta_\delta(P\&P)$	19.57	18	0.358	0.995	0.990	0.003
3. 2–1	5.50	4	*n.s.*			

Prob = Probability
Gof = Goodness of Fit
RMSR = Root Mean Square Residuals

CONCLUSIONS AND RECOMMENDATIONS

Results of the reliability analyses indicate that random errors due to calibration have equivalent variance across different media. These findings indicate that the use of item parameters obtained in a P&P calibration will not affect the reliability of CAT-ASVAB test scores and clearly support the use of the P&P parameters of the current CAT-ASVAB item pool, an important concern of the ACAP program.

REFERENCES

Birnbaum, A. (1968). Some latent-trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Cohen, J. & Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory: Applications to Psychological Measurement.* Homewood, IL: Dow Jones-Irwin.

Joreskborg, K.G. & Sorbom, D. (1986). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods.* Mooresville, IN: Scientific Software, Inc.

Moreno, K.E., Segall, D.O., & Kieckhaefer, W.F. (1985). A Validity Study of the Computerized Adaptive Testing Version of the Armed Services Vocational Aptitude Battery. *Proceedings of the 27th Annual Conference of the Military Testing Association.* San Diego, CA: Navy Personnel Research and Development Center.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association, 70,* 351-356.

Prestwood, J. S., Vale, C. D., Massey, R. H., & Welsh, J. R. (1985). *Armed Services Vocational Aptitude Battery: Development of an adaptive item pool.* (AFHRL-TR-85-19). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton, NJ: Educational Testing Service.

CAT-ASVAB Precision

Kathleen E. Moreno and Daniel O. Segall
Navy Personnel Research and Development Center
San Diego, CA 92152

INTRODUCTION

The Navy Personnel Research and Development Center (NPRDC), as lead military laboratory for the Computerized Adaptive Testing (CAT) project, is responsible for the development and evaluation of a CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). One part of the evaluation process involved conducting an empirical study to assess the precision of the CAT-ASVAB system intended for operational use.

BACKGROUND

NPRDC has conducted various studies which provide information concerning the precision of CAT-ASVAB and the relationship between the CAT-ASVAB and P&P-ASVAB subtests. Some of these studies have used an experimental version of the CAT-ASVAB system to collect empirical data, while others have used simulated data.

Empirical Studies. As part of a joint-service study on the validity of CAT-ASVAB, recruits in each of the four services were administered an experimental version of the CAT-ASVAB and those P&P-ASVAB subtests used in computing a recruit's school composite score. These data were used to compute cross-medium correlation coefficients. Correlations between CAT-ASVAB subtest scores and corresponding P&P-ASVAB subtest scores were as high as correlations between two alternate forms of the P&P-ASVAB. The results of these analyses indicate that the experimental CAT-ASVAB is measuring the same types of abilities that are measured by the P&P-ASVAB. Indirectly, the results indicate that the reliability of the CAT-ASVAB subtests is as high as that of the P&P-ASVAB subtests. (Moreno, Segall, & Kieckhaefer, 1985)

In 1987, NPRDC conducted an alternate forms study at the Recruit Training Center (RTC), San Diego, which provides more direct information about the reliability of the CAT-ASVAB item pools. This study was conducted using preliminary item pools and Apple /// hardware. The purpose of the study was to collect empirical data that could be used in a preliminary evaluation of the ACAP item pools. Navy recruits were randomly assigned to one of three groups. Group 1 received two forms of the CAT-ASVAB, Group 2 received two forms of the P&P-ASVAB, and Group 3 received one form of the CAT-ASVAB and one form of the P&P-ASVAB. Results of this study showed that the CAT-ASVAB subtest reliabilities, with the exception of Coding Speed, either equaled or exceeded P&P-ASVAB subtests in terms of alternate forms reliability. Further investigation of Coding Speed showed that if a number correct score were used instead of a rate score, the differences in reliabilities disappeared. Cross-medium analyses indicated that CAT-ASVAB and P&P-ASVAB are measuring the same constructs. (Day & Kieckhaefer, 1987)

Simulation Studies. In addition to empirical studies, NPRDC has conducted numerous simulation studies. As part of these studies, NPRDC has examined several measures of precision for the ACAP item pools (Segall, 1987). One measure examined the conditional precision of both the CAT-ASVAB and the P&P-ASVAB. Score information functions were computed for each content area and form of the CAT-ASVAB, and for one form of the P&P-ASVAB (9A). The CAT-ASVAB score information functions equaled or exceeded the P&P-ASVAB information across ability levels, for all 18 pools examined.

Simulated test-retest reliabilities were computed to assess the unconditional precision of both the CAT-ASVAB and P&P-ASVAB (Hetter & Segall, 1986). Again, the simulated CAT-ASVAB reliabilities matched or exceeded the corresponding P&P-ASVAB reliabilities.

On the whole, the studies that have been conducted to date have provided us with valuable information concerning both CAT, in general, and, more specifically, the CAT-ASVAB item pools. However, an empirical study of the precision of the CAT-ASVAB "system", final items pools adminstered on the Hewlett Packard Integral, needed to be conducted prior to operational use of the system.

PURPOSE

The primary purpose of the study described in this paper was to compare CAT-ASVAB subtest and composite scores to corresponding P&P-ASVAB subtest and composite scores. A secondary purpose was to (1) compare the correlations between CAT-ASVAB and corresponding P&P-ASVAB subtests to those between two forms of P&P-ASVAB.

METHOD

Examinees. Navy recruits stationed at the Recruit Training Center in San Diego served as examinees in this study. The total number of examinees tested was 2,090. There were 1,057 in the CAT-ASVAB group and 1,033 in the P&P-ASVAB group. A large percentage of these subjects did not have complete data because they did not return for the second test. After eliminating cases with incomplete data the sample sizes were 744 for CAT-ASVAB and 726 for P&P-ASVAB.

Design. This study used an equivalent groups design. Examinees were randomly assigned to one of two groups. Group 1 was administered form 1 of the CAT-ASVAB, followed by form 2 of the CAT-ASVAB. Group 2 was administered form 9B of the P&P-ASVAB, followed by form 10B of the P&P-ASVAB. There was an interval of five weeks between when an examinee took the first test and when he took the second test. This interval was constant for all examinees. There was no counter-balancing of forms. Comparisons between CAT-ASVAB and P&P-ASVAB reliabilities should be unaffected by the order of administration of the forms. In addition, counter-balancing would have been administratively more difficult.

Tests. The two forms of the CAT-ASVAB were forms 01C and 02C, developed for operational use. The two forms of the P&P-ASVAB were forms 9B and 10B. These P&P-ASVAB forms were chosen for this study because NPRDC has item parameters available for these forms that are on the same scale as the CAT-ASVAB item parameters. This will permit the results of this study to be compared to results from simulation studies conducted at NPRDC.

Procedure. Upon arrival, all examinees were given general instructions explaining the experimental testing in which they were participating, and signed a privacy act statement allowing use of the data for research purposes. Examinees were then seated in the appropriate room (CAT-ASVAB or P&P-ASVAB), based on a random assignment list. CAT-ASVAB was administered using Hewlett Packard Personal Computers, following the same procedures developed for operational implementation. The P&P-ASVAB was administered following procedures outlined in the ASVAB Test Administrator's Manual. At the conclusion of testing, test administrators collected the following data from the examinees' DD Form 1966: population group, ethnic group, date of birth, education, operational ASVAB test form, operational ASVAB subtest scores, and date of enlistment.

Scores. The scores used for those examinees taking the CAT-ASVAB were the equated CAT-ASVAB raw scores (unless otherwise indicated in the table). The tables obtained from the CAT-ASVAB score equating data collection were used to obtain these scores (Segall, 1990). The scores used for the P&P-ASVAB were the number right for the subtests. Composite scores were obtained in the usual method.

Data Editing. A data editing procedure which compared post-enlistment (non-operational) scores to pre-enlistment (operational) scores was used to eliminate "unmotivated" examinees (Segall, 1990). After editing, the sample size was 723 for the CAT-ASVAB group and 706 for the P&P-ASVAB group.

Evaluation of Equivalent Groups. The equivalency of the two groups was checked by (1) comparing the two groups on race and years of education, and (2) comparing the distribution of operational subtest scores for the two groups.

Precision Analyses. Alternate forms correlations were computed for each of the CAT-ASVAB and P&P-ASVAB subtests and the service composites. Fisher's z transformation was used to evaluate the difference between CAT-ASVAB and P&P-ASVAB reliabilities, at the subtest level.

Cross-medium Correlations. Correlations were computed between the CAT-ASVAB subtest scores and the corresponding operational ASVAB subtest scores. Correlations between non-operational P&P-ASVAB and the corresponding operational P&P-ASVAB subtests were also computed.

RESULTS

Evaluation of Equivalent Groups. Table 1 shows the distribution of race and years in school for each of the two groups. As shown, these distributions are not significantly different. For race, however, the value is larger than expected. When this test was rerun eliminating the "other" category, the P-value was .80. For each of the subtests, a K-S test was conducted to evaluate the difference between the score distributions for the two groups. There were no significant differences.

Table 1

Frequency of Race and Education by Group[a]

	CAT-ASVAB	P&P-ASVAB	χ^2	P-value
Race			9.14	.058
Whites	541	539		
Blacks	141	134		
Asian	27	34		
Am. Ind.	6	6		
Other	24	8		
Education			1.39	.85
< 10 yrs	22	21		
10 yrs	27	34		
11 yrs	70	73		
12 yrs	565	537		
> 12 yrs	55	56		

[a] Note. This was run on the unedited data set. There were 10 cases with missing data.

Precision Analyses. As shown in Table 2, the alternate form correlations for the CAT-ASVAB subtests either matched or exceeded those of corresponding P&P-ASVAB subtests. This table also shows the alternate form correlations for the CAT-ASVAB subtests, using Owen's Bayesian estimate of theta and the final ability estimate (penalized mode for the power tests, rate score for the speeded tests). In addition, a comparison of CAT-ASVAB composite scores with P&P-ASVAB composite scores showed that for all 32 composites, CAT-ASVAB scores were higher than corresponding P&P-ASVAB scores. The values for the Armed Forces Qualification Test composite and the Verbal composite, used by all services, are shown in Table 2.

Table 2

Alternate Forms Correlations for Subtests

Content Area	Owens	Final	Equated Raw	Raw
GS	.85	.85	.84*	.73
AR	.82	.80	.82**	.77
WK	.86	.86	.83	.82
PC	.59	.58	.54	.48
NO		.84	.81*	.71
CS		.76	.78	.75
AS	.85	.85	.89*	.77
	.78	.78		
MK	.88	.87	.88*	.82
MC	.75	.75	.75**	.70
EI	.77	.75	.73*	.65
			Standard Scores	
AFQT			.92	.89
VE			.82	.80

Note. The scores for the speeded tests in column 2 of the table are rate scores. Those reliabilities with one asterisk were significantly higher, $p < .01$, those with two asterisks were significantly higher, $p < .05$.

Cross-medium Correlations. Table 3 shows the cross-medium correlations between CAT-ASVAB subtests and operational P&P-ASVAB subtests and between non-operational P&P-ASVAB subtests and operational P&P-ASVAB subtests.

Table 3

Correlations with Operational P&P-ASVAB

Content	CAT-ASVAB		P&P-ASVAB	
Area	Form 1	Form 2	Form 9B	Form 10B
General Science	.83	.82	.79	.73
Arithmetic Reasoning	.81	.75	.76	.72
Word Knowledge	.83	.81	.81	.78
Paragraph Comp	.54	.43	.48	.38
Numerical Operations	.60	.60	.65	.56
Coding Speed	.57	.54	.65	.62
Auto & Shop Info	.83	.83	.76	.74
Math Knowledge	.85	.83	.83	.80
Mechanical Comp	.69	.64	.66	.65
Electronics Info	.73	.72	.66	.65

SUMMARY

The results of this study show that seven of the ten CAT-ASVAB subtests are significantly higher in reliability than the corresponding P&P-ASVAB subtests. The other three subtests are as reliable as the corresponding P&P-ASVAB subtests. In addition, CAT-ASVAB subtests correlate as highly with pre-enlistment P&P-ASVAB as do alternate forms of the P&P-ASVAB. These findings indicate that, from a psychometric standpoint, CAT-ASVAB can be used as a replacement for the P&P-ASVAB.

REFERENCES

Day, L. E., & Kieckhaefer, W. F. (1987). *A Comparison of CAT and ASVAB Alternate Forms Reliability* (Contract N66001-83-D-0343 Deliverable) San Diego, CA: RGI, Incorporated.

Hetter, R. D., & Segall, D. O. (1986, November). *Relative precision of paper-and-pencil and computerized adaptive tests*. Paper presented at the 28th annual meeting of the Military Testing Association, Mystic, Connecticut.

Moreno, K. E., Segall, D. O., & Kieckhaeffer, W. F. (1985). *A validity study of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery.* Proceedings of the 27th Annual Conference of the Military Testing Association (pp. 29-33). San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O. (1990). *Score Equating Development Analyses of the CAT-ASVAB* (Draft Technical Report). San Diego, CA: Navy Personnel Research and Development Center.

Segall, D. O. (1987). *ACAP Item Pools: Analysis and recomendations*. Unpublished manuscript.

Equating of the Computerized Adaptive Testing Version of the ASVAB

Daniel O. Segall

Navy Personnel Research and Development Center

This paper describes an equating of the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). This study is part of a larger effort to field an operational CAT-ASVAB for use in the selection and classification of military applicants. During early stages of implementation, both CAT-ASVAB and the paper-and-pencil version (P&P-ASVAB) will be in operational use. Some applicants will be accessed using scores from the CAT-ASVAB, while others will be enlisted using their scores obtained on the P&P-ASVAB. In order for the scores on the two versions to be exchangeable, an equivalence relation (or equating) between CAT-ASVAB and P&P-ASVAB must be obtained. The primary objective of this equating is to provide a transformation of CAT-ASVAB scores that will preserve the flow rates currently associated with the P&P-ASVAB. Consequently, the P&P-ASVAB and equated CAT-ASVAB will possess identical subtest and composite distributions. This property will allow the two versions to be used interchangeably, without effecting flow-rates into the military, or into various occupational specialties. The CAT-ASVAB will be equated in two phases: (1) Score Equating Development (SED) phase and (2) Score Equating Verification (SEV) phase. Although data collection for both phases have been completed, the study described in this paper covers the first of these phases.

The purpose of the study described here is to provide an equating of the CAT-ASVAB that could be used at a later date to select and classify military applicants. During this study, both CAT-ASVAB and P&P-ASVAB were given non-operationally to equivalent groups. The tests were non-operational in the sense that scores on these tests had no impact on the applicants' eligibility for the military. The equating obtained from this phase was however used operationally during a recent data collection effort, SEV. During this recent (SEV) phase, all applicants were administered one test only; either an operational CAT-ASVAB or an operational P&P-ASVAB. Both versions used in the equating impacted applicants eligibility for military service. Data for this recent data collection effort are being analyzed and will be applied to future CAT-ASVAB applicants.

CAT-ASVAB MEASUREMENT PROPERTIES

Subtest Composition

The CAT-ASVAB and P&P-ASVAB share many common features. In an attempt to design CAT-ASVAB as a substitute, many of the fundamental characteristics of the ASVAB were left unchanged (e.g. subtest content, multiple choice item types, exposure rates, etc.). However, the nature of a computerized adaptive test imposes many striking differences. The P&P-ASVAB is composed of 10 subtests which span verbal, mathematical, technical, and clerical content areas. The CAT-ASVAB is composed of 11 subtests. These subtests correspond in content to the P&P-ASVAB subtests, with one exception. The Auto and Shop Information content areas are measured by two CAT-ASVAB subtests (Auto-Information and Shop-Information) rather than a single test (Auto/Shop) as given by the P&P-ASVAB. Each CAT-ASVAB subtest is a fixed number of items, and has a fixed maximum time-limit. On average, CAT-ASVAB subtests contain 40-percent fewer items than their P&P counterparts.

Item Selection

The CAT-ASVAB has two unique forms, each form drawing items from a different item pool. For the adaptive tests, items are selected using maximum IRT information. After each item, the ability estimate is updated using Owens (1975) estimator. A new item is selected from among the most informative items at that ability level, subject to the constraints imposed by the item exposure control algorithm (Sympson & Hetter, 1985). This algorithm reduces the exposure rate of certain highly informative items, while increasing the exposure rate for other items. The result is an upper ceiling on the exposure among the test items. The exposure-rate of the pools most informative items was targeted at 1/3 for subtests contained within the AFQT and 2/3 for the remaining power tests. Considering that CAT-ASVAB has two forms, the effective exposure rate for the most informative items is comparable to the P&P-ASVAB. General Science has one additional constraint imposed on item selection; items are balanced among life, physical, and chemistry items. Items are selected from among these content areas in the sequence (P L P L P L P L P L P L P L C) where P = physical, L = life, and C = chemistry. That is, the first item administered is a physical science item, the second is a life science item, ..., and the last item is a chemistry item. During the adaptive sequence, each item is selected from among the most informative items within the appropriate content area. However, this selection within a given content area is contingent on the constraints imposed by the exposure control algorithm.

Power Subtest Scoring

The CAT-ASVAB uses Owens (1975) bayesian procedure to update the ability estimate after each item. This updated ability estimate is used to select items for administration. Consequently, provisional ability estimates are obtained after responding to each item. A final Owens estimate can be obtained by updating the estimate with the response to the final test item. However, the Owens estimate, as a final score has one undesirable feature: it depends on the order in which the items were administered. Consequently, it is possible for two examinees to receive the same items, provide the same responses, but receive different final Owens ability estimates. This could occur if the two examinees received the items in different sequences. Although this event is relatively unlikely, a decision was made to use a final estimator that is order independent. The mode of the posterior distribution (bayesian mode) is used at the conclusion of each power test to provide a final ability estimate. This estimator is unaffected by the order of item administration, and provides slightly greater precision than the Owens estimator.

Speeded Subtest Scoring

The two speeded subtests CS and NO are scored using the rate score where the geometric mean of screen times is divided into the proportion of correct responses. The proportion-correct in the numerator of the rate score is corrected for chance guessing. Without this feature, an applicant could receive a very high score by pressing any key quickly, without reading the items. Such an examinee would receive a low proportion correct, but a high rate score because of the fast responding. Correcting the score for chance guessing eliminates the advantage associated with fast random responding.

Scoring Incomplete Tests

There is one property of the bayesian modal estimator (BME) that could be problematic in the context of incomplete tests. As with bayesian estimators in general, the BME contains a bias that draws the estimate toward the mean of the prior. This bias is inversely related to test-length. That is, the bias is larger for short adaptive tests, and smaller for long adaptive tests. A low ability examinee could use this property to his/her advantage. If allowed, a low ability examinee could receive a score at or slightly below the mean by answering one or two items. Even if the items were answered incorrectly, the strong positive bias would push the estimator up towards the mean of the prior. Consequently, a below average applicant could use this strategy to increase their score: just answer the minimum number of items allowed. To discourage the use of this strategy, a penalty procedure was developed for use in scoring incomplete tests (Segall, 1988).

The penalty procedure used in CAT-ASVAB provides a final score that is equivalent (in expectation) to the score obtained by guessing at random on the unfinished items. The size of the penalty depends on the number of unanswered items, the particular subtest, and the provisional ability level.

SCORE EQUATING DEVELOPMENT

Data Collection Design

Data from 8040 applicants were collected from six geographically disperse regions within the continental Unites States. Within each region is a Military Entrance Processing Station (MEPS), and associated with each MEPS are a number (between 3 and 16) of Mobile Examining Team (MET) sites. Each of these MEPS and MET sites were included in the data collection for about 7 to 12 weeks. The six regions were selected to provide a representative and diverse sample of military applicants. These six locations taken together were expected to provide a nationally representative sample with respect to AFQT, Race and Gender. Each applicant was randomly assigned to one of three groups.

Each group was assigned a different non-operational ASVAB. Examinees in one group were assigned to a non-operational P&P-ASVAB (Version 15C). Examinees in the other two groups were assigned either Form 1 or 2 of the CAT-ASVAB. In addition to taking a non-operational battery (either P&P-ASVAB or CAT-ASVAB) each applicant was administered an operational P&P-ASVAB for enlistment purposes. Scores from this operational test were retained for the analysis of sample characteristics, and to screen unmotivated applicants from the equating.

Smoothing and Equating

The objective of equipercentile equating is to provide a transformation that will match CAT-ASVAB score distributions with P&P-ASVAB distributions. This transformation, which is applied to CAT-ASVAB, would allow scores on the two ASVAB versions to be used interchangeably without disrupting applicant flow rates. One method for estimating this transformation involves the use of the two empirical cumulative distribution functions (CDF's). Scores on CAT-ASVAB and P&P-ASVAB could be equated by matching the empirical proportion scoring at or below observed score levels. However, this transformation is subject to random sampling errors contained in the CDF's. It is a universal belief that the precision of the equating transformation can be improved by smoothing either: (a) the equating transformation, or (b) the two empirical distributions which form the equating transformation. A considerable amount of controversy has arisen however on the choice of a decision rule for specifying the amount or degree of smoothing. One primary objective of the CAT-ASVAB equating was to use smoothing procedures that provide an acceptable trade-off between random and systematic error. In this study smoothing was performed on each distribution (CAT-ASVAB and P&P-ASVAB) separately. These smoothed distributions were used to specify the equipercentile transformation.

One additional concern arises over the shape of the equating transformation in the lower score range, where data are usually sparse. Typically, most equating procedures provide a transformation that is either undefined or poorly defined over this lower range. This problem was overcome here by fitting logistic tails to the lower portion of the smoothed density functions. These tails achieved two desirable results. First, the distributions were extended to encompass the entire lower range, thus defining the equating transformation over this entire range. Second, by pre-specifying the fit-point of the tail, the distribution (and consequently the equating transformation) above that point is left unaltered by the tail. Consequently the tail-fitting procedure altered the equating only over a pre-specified lower range; the equating transformation above that range was unaltered.

P&P-ASVAB Smoothing

The procedure used to smooth the P&P-ASVAB, developed by Segall (1988), estimates the smoothest density that deviates from the observed density by a specified amount. The observed and smoothed densities deviate from one another by an amount to be expected from sampling error.

The two parameter logistic cumulative density function (CDF) was used to specify density values for the lower tail of the discrete distributions. The logistic CDF provides a close approximation to the normal CDF and is often used as a substitute since it provides mathematically tractable expressions for both the density and distribution functions. Although the function is usually used to define a continuous CDF, it is used here to define a discrete density. Two constraints were placed on the logistic function. The first constraint assures that there is a smooth fit of the logistic tail to the estimated density. This is accomplished by constraining the last bin of the tail to equal the estimated value of the smoothed solution. The second constraint assures that the proportion contained in the logistic tail will equal the proportion contained in the tail of the smoothed solution (about .05). It follows from this constraint that together, the logistic tail and the upper portion of the smoothed solution will define a density (i.e. sum to 1). Once the above constraints are imposed, values for the two parameters of the logistic CDF can be obtained. These values can be derived through an iterative numerical procedure.

Figure 1 displays the smoothed solution and the fitted tail for one example from among the ten subtests of the P&P-ASVAB 15C. The empirical proportions for each bin are indicated by the height of the bar. The smoothed (or fitted) density values are indicated by the small circles joined by the dotted lines. The point at which the tail was joined to the smoothed solution is indicated by an arrow.

Smoothing CAT-ASVAB Distributions

The procedure used to smooth the CAT-ASVAB distributions (Kronmal and Tarter, 1968) provides a fourier estimate of the density function using trigonometric functions. In order to obtain a useful density estimate, it is necessary to smooth the series by truncating it at some point. Kronmal and Tarter provide expressions which relate the Mean Integrated Square Error (MISE) of the fourier estimator to the sample fourier coefficients. These MISE expressions are used to specify a truncation point for the series, making it possible to specify an optimal number of terms in the series.

The logistic CDF was also used here to smooth the lower portion of the fourier estimate where data are sparse. This tail fitting involved the following steps. First, the proportion contained in the tail was specified according to the proportion contained in the tail of the corresponding discrete (P&P-ASVAB) distribution (about .05). A second constraint assured that the density value of the logistic tail at the join point equaled the density of the fourier estimate. This provided a continuous transition between the fourier estimate and the logistic tail. Once the above constraints are imposed, values of the two logistic parameters can be obtained through an iterative numerical procedure.

Figure 2 displays the smooth fourier estimate and the fitted tail for one of 20 CAT-ASVAB distributions (Since Form 1 and Form 2 were smoothed separately, 20 estimates were performed in total). The empirical histogram for the CAT-ASVAB distribution is indicated by the height of each bar. The smoothed (or fitted) density function is displayed by the dotted curve. The fitted logistic tail is indicated by a solid bullet.

Figure 1
Arithmetic Reasoning
P&P-ASVAB

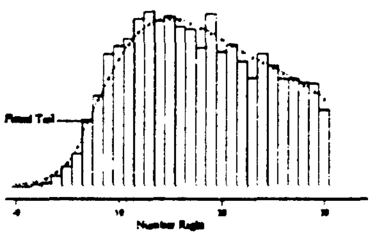
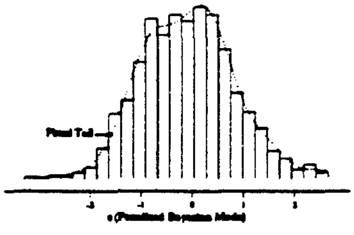


Figure 2
Electronics Information
CAT-ASVAB



Equating Transformations

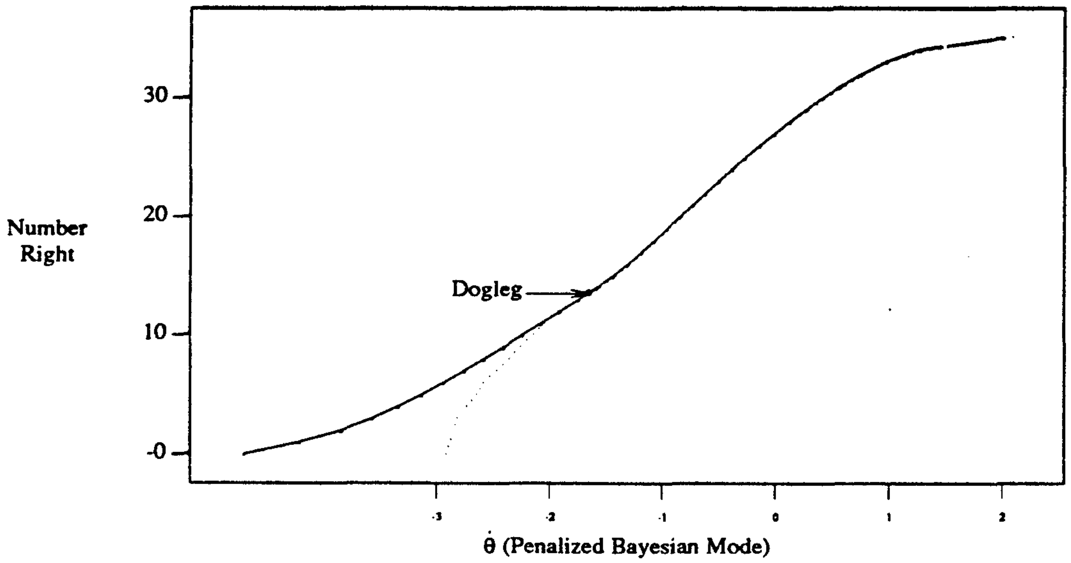
The smoothed distributions were used to specify the equipercentile transformation for the CAT-ASVAB. There were a total of 20 equatings, one for each content area of each CAT-ASVAB form. For each P&P-ASVAB number-right score, an interval of the continuous CAT-ASVAB scores that contained the same estimated proportion was obtained. Figure 3 compares the equating function based on two smoothed distributions with the function based on two empirical unsmoothed distribution. The Word-Knowledge content area is shown. The smoothed function is indicated by the circles joined by solid lines. The dogleg portion of the function (that portion effected by the tail fitting procedure) is indicated by a large bullet. The unsmoothed transformation is indicated by the dotted function. For both the smoothed and unsmoothed transformation, each number-right (on the y-axis) is plotted against the midpoint of the CAT-ASVAB score interval (on the x-axis). The agreement between the smoothed and unsmoothed functions is very high above the dogleg portion. Notice that the tail appears to provide a smooth extrapolation of the equating function over the lower range, and in no way effects the agreement of the function above the dogleg portion. Also notice that the dogleg provides a monotonic increasing function for mapping CAT-ASVAB scores into number-right.

Composite Equating

Equating the CAT-ASVAB to the P&P-ASVAB involves matching subtest distributions using an equipercentile method. This distribution matching provides a transformation of the CAT-ASVAB ability estimates to number-right equivalents. Once this transformation is specified for each subtest, raw-score equivalents can be computed. These raw-score equivalents provide the basis for the computation of service specific composites, as well as the AFQT and Verbal composites. One concern is that the distributions of CAT-ASVAB composites will differ systematically from P&P-ASVAB composite distributions. This difference could be caused by differences in subtest reliabilities or content.

Sums of (equated) subtest standard scores were computed for the 29 service composites and for the AFQT. The VE composite was also computed from the sum of Word-Knowledge and Paragraph-Comprehension raw-scores. After these sums were obtained, the appropriate scale conversion was applied to place each composite score on the metric used for classification decisions by the services.

Figure 3. Equating Transformation: Word Knowledge



Each CAT-ASVAB composite distribution (Form 1 and 2) was compared to the corresponding P&P-ASVAB composite distribution. Two different methods were used to examine the significance of the differences. First, the Kolmogorov-Smirnov two-sample test (KS) was used to detect overall differences between Form 1 and P&P-ASVAB, and between Form 2 and P&P-ASVAB. Since this test is not highly sensitive to differences of a specific nature (e.g. differences in variances), an *F*-test was also used to assess the differences between Form 1 and P&P-ASVAB variances, and between Form 2 and P&P-ASVAB variances. Both significance tests were performed on all 31 composites. Of the 62 comparisons tested using the K-S tests, none were significant at the .01 level. Three of the 62 variance comparisons were significant at the .01 level.

These significance tests are generally indicative of no differences between CAT-ASVAB and P&P-ASVAB composite score distributions. The three significant differences that were identified are likely due to Type I errors that occur when a large number of comparisons are made. In this study, over 124 comparisons were made. Finding at least three significant differences (at the .01 level) is a highly probable occurrence even when no true differences exist between the composite distributions. Even though these significant differences are probably due to chance (Type I errors), it is prudent to examine the consequence of not equating these composites, under the assumption that these observed differences are real. That is, suppose these observed differences in composite distributions were treated as true differences; what consequence would this difference have on flow rates? This issue was investigated for the Navy EG composite which displayed the largest variance difference between 15C and CAT-ASVAB. The training schools that select on EG all happen to employ a cut-score of "96". The proportion of applicants scoring at or above 96 on each of the CAT-ASVAB forms and 15C was examined. If the observed sample differences are treated as true differences, then 2%-3% additional CAT-ASVAB applicants would qualify for schools using the Navy EG composite. This difference is relatively small.

SUMMARY

The next major analysis effort for the CAT-ASVAB project is Score Equating Verification. The data collected from this portion of the study will be used to "re-equate" the CAT-ASVAB. One key psychometric question will be: how much does the equating based on the operational data differ from the current equating based on non-operational scores? Preliminary analyses indicate that this difference is small. This outcome suggests that the methods and procedures used in this study will provide a proven framework for future CAT equatings.

References

Kronmal, R. and Tarter, M., (1968). *The estimation of probability densities and cumulatives by Fourier series methods*. Journal of the American Statistical Association, 69, 925-952.

Owen, R. (1975). *A Bayesian sequential procedure for quantal response in the context of adaptive mental testing*. Journal of the American Statistical Association, 70, 351-356.

Segall, D. O., (1988). In minutes of the January, 1988 meeting of the CAT-ASVAB Technical Committee.

Sympson, J. B. and Hetter, R. D., *Controlling item exposure rates in computerized adaptive tests*. In Proceedings of the 27th Annual conference of the Military Testing Association, 1985.

RELIABILITIES AND PRACTICE EFFECTS FOR THE ENHANCED COMPUTER-ADMINISTERED TEST (ECAT) BATTERY

Gerald E. Larson
David L. Alderton

Navy Personnel Research and Development Center
San Diego, CA 92152-6800

The Office of the Assistant Secretary of Defense for Military Manpower and Personnel Policy formed the Test Advisory Selection Panel (TASP) in December 1989 and gave it the responsibility to assemble, from the available experimental aptitude tests, an optimal battery for validation. The TASP considered a number of factors, including reliability, uniqueness, and possible bias, and selected a set of tests referred to as the Enhanced Computer-Administered Test (ECAT) battery. A test/retest analysis of ECAT is documented in the present report.

METHOD

Subjects - Military scheduling considerations made recruit testing impractical for the current research. Thus, high school and junior college students in the San Diego vicinity were recruited as subjects, with the restrictions that subjects must be between the ages of 16 and 26, with the total sample having no more than 35% females and no less than 60% caucasians (to ensure comparability between the sample and military recruits). As an incentive to participate in the study, each subject was paid $70.00. Three hundred and thirteen subjects (223 males, 90 females) completed both test sessions. They averaged 19.3 years-of-age, with a standard deviation of 2.8. The ethnic breakdown was as follows: 73% Caucasian, 10% Hispanic, 6% Asian, 4% Filipino, 3% African-American, 4% "Other."

APTITUDE TESTS

Each subject completed an approximately 3-hour battery of 10 computerized tests, presented on Hewlett-Packard Integral microcomputers operating under UNIX™. Nine of the 10 tests comprised the actual ECAT battery. The tenth, "Perceptual Speed," was included as a supplemental measure. Tests 8-10 below (Target Identification, One-hand Tracking, and Two-hand Tracking) used a custom built "response pedestal" with response buttons, sliders, and a joy stick.

1. Integrating Details - A complex 40 item spatial problem solving test. Each item consists of two separate screens. The first screen contains from 2 to 6 regular geometric puzzle pieces that must be mentally brought together to form a completed object. Having connected all of the puzzle pieces, the individual must remember the final object, then press a response key indicating that she/he is ready. Once the key is pressed, the puzzle pieces are replaced by a new screen with a single completed object. The subject must indicate if the completed object shown is a product of the original puzzle pieces. There are three dependent measures for each trial; time spent studying the puzzle pieces, time spent deciding if the completed form is valid, and response accuracy.

2. Mental Counters - Mental Counters is a 40 item working memory test. Each screen contains three horizontal lines, arrayed left to right. Each line represents a counter with an initial value of zero. During an item, boxes appear sequentially, one at a time, either above or below one of the three lines. If a box appears above a line, the value for that counter is incremented by +1. If a box appears below a line, that counter is decremented by -1. On each trial either 5 or 7 boxes appear. The boxes appear at one of two rates, either one every 1.33 seconds or one every .75 seconds. The subject must make a series of rapid calculations and select, from a four-alternative multiple choice menu, the set of correct final counter values. Number of correct responses is used as the summary score.

3. Sequential Memory - Sequential Memory is another complex test of working memory. Each item consists of three to five horizontally arrayed dots on the screen. Each dot is given a numerical value; these must be memorized. The item is then presented in a series of 5 to 7 "calls" to the dots; where each call is announced by briefly turning one of the dots into an "X." The person must report the digit string that corresponds to the order that the dots were "called." In the second half of the test, after all the calls for an item have been made, the examinee is told to translate each number in the ordered number list into a different number and then type in the new ordered list. There are 10 items in the first half of the test and 25 in the second half of the test. The dependent variable is the proportion of digits correctly reported by the examinee.

The opinions expressed are those of the authors, are not official, and do not necessarily reflect the views of the Department of the Navy.

4. Spatial Reasoning - A figural inductive reasoning (or series extrapolation) test. Items use a combination of geometric forms and arbitrary figures presented in a series of four frames. The subject's task is to induce the transformation rule controlling the series and then select one of five alternatives that correctly completes the series. The dependent variable is number correct across the 30 items. There is a 12 minute time limit.

5. Perceptual Speed - Perceptual Speed (Alderton, 1990) is a clerical/perceptual speed test. Each item consists of two side-by-side symbol strings of the same length. The examinee's task is to determine whether the two symbol strings are identical, and to make these judgements as rapidly as possible while maintaining 90% accuracy. Symbol string length is systematically varied from 1 to 7 elements. The test is divided into 3 subtests based on string content: Numbers (56 items), letters (56 items), or abstract stick figures (60 items). Each item type (number of elements X symbol type) has a minimum and maximum response time bracket associated with it. If an examinee responds too quickly or too slowly she/he is warned to slow down or speed up. Cumulative accuracy is retained and used in feedback after every 10-14 items. To control for speed/accuracy tradeoffs, the examinee is warned to slow down if accuracy drops below 85% or to speed up if accuracy goes above 95%. The primary dependent variable is the average rate score across the three subtests.

6. Assembling Objects - A spatial construction test. Each item consists of a frame with several (2-6) separate elements. The subject's task is to choose, from four alternatives, the answer that correctly represents how the elements should be connected. There are 32 items in the test. The first 15 items are semi-mechanical items with labels indicating how the elements should be connected. The final 17 items in the test consist of small jigsaw puzzles with no labels showing how the puzzle pieces are to be connected, but only one of the four answer choices includes all of the puzzle elements. The dependent variable is the number of correct items solved in 16 minutes.

7. Spatial Orientation - A spatial perspective test. Each item consists of an environmental view, such as a bridge over a river or a farm house. In each view the horizon is apparent. These views are rotated away from the "natural" horizon in a frame. At the bottom of the frame is a circle with a dot on the perimeter. The subject's task is to rotate the frame around the view until it corresponds with the natural horizon of the view and determine where the dot on the circle would be located. This information is then used to select which of 5 alternatives correctly shows where the dot would be on the circle (following the rotation). The score is the number of items (of 24) solved correctly.

The next 3 tests use the ECAT response pedestal to input responses.

8. Target Identification - A hybrid test combining aspects of choice reaction time and spatial mental rotation tests. Each item consists of a figure in the top half of the screen and three alternative figures in the bottom half of the screen. The correct answer is the alternative (at screen bottom) that represents the same object as the standard, even though the standard may be distorted (e.g., rotated, shrunken, or both) relative to the answer choice. (Answer choices are always presented in a "natural" upright position) The examinee's task is to select the correct alternative as rapidly as possible. The figures are schematic line drawings of simple objects, such as trucks, helicopters, and tanks. Before each item the subject is required to hold down 4 "home" buttons, two on the left and two on the right. While all four buttons are simultaneously depressed the item is presented. As soon as the examinee decides upon an answer, either hand may be used to press the button (on the top of the pedestal) that corresponds to the selected alternative. As soon as any of the four "home" buttons are released the alternatives are masked (blacked out). The dependent variable is the average correct decision time where decision time is defined as the time between item presentation and "home" button release. There are 36 items administered with a maximum 7 minute total test time.

9. One-Hand Tracking - A psychomotor test that uses a response pedestal. Each item begins with a "path" on the computer screen. The path is simply a contiguous string of lighted screen pixels. The path goes up/down and/or right/left, parallel with the sides of the screen and makes only 90 degree turns. At one end of the path is a diamond indicating the path's termination point. Starting at the other end is a box that travels forward along the path. The subject moves a joy-stick that controls the movement of a "cross-hair." The subject's task is to keep the cross-hair on the moving box. Items vary in terms of the length of the path which is inversely related to the speed at which the box moves (total item duration is thus constant). For each item, the "score" is the average absolute Cartesian pixel distance between the cross-hair and the moving box (a distance reading is taken every 50 msec during the item). There are 18 items. The dependent variable for the test is the average of the 18 item scores.

10. Two-Hand Tracking - Another psychomotor test that has exactly the same structure and task constraints as One-Hand Tracking described above. The only difference is that movement of the cross-hair is controlled by two slide potentiometers. One of the slides controls the horizontal (left/right) movement of the cross-hair while the second slide controls the vertical (up/down) motion of the cross hair. One hand must be used for each slide control. The slides are arranged such that the horizontal slide's physical

movement is right and left while the vertical slide's physical movement is up and down. Number of items, test scoring, and final test score are the same as above.

RESULTS

Prior to the main analyses the data were trimmed to eliminate subjects who scored 10% or more below chance on the power tests. Also, subjects were eliminated if their scores declined 50% or more from session one to session two, or if the score for either session lagged four standard deviations below the sample mean. Finally, speed test scores were discarded if accuracy was below 70%. These data editing rules were designed to eliminate unmotivated or severely impaired examinees. Upon implementing these rules, the proportion of subjects excluded from the analyses ranged from a high of 6% on Assembling Objects and Mental Counters to a low of .3% on One-hand Tracking.

Practice Effects. Descriptive statistics and practice effects for the remaining subjects are shown in Table 1. As can be seen, practice effects (reflecting improved performance) were significant for all tests except Assembling Objects. Given the relative novelty of the experimental measures, some improvement with practice was to be expected. In many cases, however, improvements were of little practical importance despite statistical significance. For example, note the slight (less than one tenth of a standard deviation) though significant gain for the Integrating Details test. In general, score gains were greatest for speeded and/or psychomotor tests (especially Two-hand Tracking) and it is therefore this category of measures which should be the focus of concern for issues such as practice and coaching.

Insert Table 1 about here

Reliabilities. Test reliabilities are shown in Table 2. Retest reliabilities range from .75 to .91, with a median of .81. These figures compare favorably with ASVAB retest reliabilities, which range from .63 to .88, with a median of .79 (Wolfe, in preparation). Internal consistency estimates are also quite acceptable, ranging from .78 to .97 across both sessions. In general, reliabilities were somewhat higher for speeded and/or psychomotor tests that for power tests. Since as noted above the former also showed the greatest practice effects, one may infer that practice caused an upward shift in the psychomotor score distribution without a substantial reordering of individual ranks.

Insert Table 2 about here

Gender Effects. Table 3 shows test performance as a function of gender. Females scored significantly below males on five of the ten tests; two of these tests were spatial in nature (Integrating Details and Spatial Orientation) and three were psychomotor (Target ID, One- and Two-hand Tracking). To provide a better context for these findings, it should be noted that there were no gender differences in academic standing (i.e., grade point average) within the sample, nor were there differences on the ECAT reasoning and working memory tests. Therefore, there is no reason to believe that the gender effects reflect underlying general intelligence differences rather than specific spatial and psychomotor differences.

Insert Table 3 about here

DISCUSSION AND CONCLUSIONS

Results from our test/retest administration of the ECAT battery are, for the most part, highly encouraging. Test reliabilities are at least as good as those for the operational ASVAB. Of concern in the present study are the significant practice effects observed for nearly all tests, and the female score deficit observed on some spatial tests and all psychomotor tests. These issues should be addressed by follow-on research prior to operational use of ECAT tests. An important question is whether adding more practice items at the beginning of the tests can stabilize performance prior to the administration of operational items. With regard to gender differences, follow-on analyses must include actual military criterion performance measures. For example, if females under-perform (relative to males) on criterion as well as predictor measures then the latter deficit does not reflect test bias. A finding of gender equivalence on the criteria would, however, suggest that the test is a biased predictor and that alternative tests or administration formats must be sought.

TABLE 1
DESCRIPTIVE STATISTICS AND PRACTICE EFFECTS

	SESSION1		SESSION2				
					t		2-tail
	Mean1	SD1	Mean2	SD2	Value	df	Prob.
VARIABLE							
PSRATE	.709	.089	.753	.101	-12.89	308	.000
SEQMEM	.707	.140	.761	.141	-10.76	307	.000
SP_REAS	.692	.199	.733	.175	-5.22	295	.000
INTEGRATE	.773	.132	.784	.128	-2.24	306	.026
ASSEMBLE	.673	.214	.686	.211	-1.77	292	.079
ORIENT	.530	.258	.628	.256	-9.21	294	.000
COUNTERS	.781	.160	.795	.183	-2.04	292	.042
TARGETID	1.66	.568	1.37	.504	14.99	310	.000
TRACK1	2913	432	2777	475	9.35	312	.000
TRACK2	3863	531	3549	619	21.03	309	.000

TABLE 2
TEST RELIABILITIES

VARIABLE	SESSION1 ALPHA	SESSION2 ALPHA	RETEST RELIABILITY
PSRATE	.95[1]	.94[1]	.86
SEQMEM	.88	.89	.81
SP_REAS	.87	.86	.75
INTEGRATE	.79	.78	.79
ASSEMBLE	.87	.89	.83
ORIENT	.89	.90	.75
COUNTERS	.89	.91	.79
TARGETID	.97[1]	.97[1]	.80
TRACK1	.97[1]	.97[1]	.84
TRACK2	.97[1]	.97[1]	.91

[1] Split-half reliabilities

TABLE 3

GENDER DIFFERENCES IN TEST PERFORMANCE

	SESSION1				SESSION2				GENDER EFFECT		GENDER X SESSION	
	Males		Females		Males		Females					
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	F	Prob.	F	Prob.
PSRATE	.707	.092	.716	.082	.750	.104	.760	.092	.76	NS	.07	NS
SEQMEM	.707	.146	.708	.125	.753	.149	.780	.119	.73	NS	5.42	.02
SP_REAS	.700	.209	.674	.175	.734	.185	.731	.151	.41	NS	1.84	NS
INTEGRATE	.791	.130	.729	.128	.791	.131	.766	.121	8.17	.00	12.00	.00
ASSEMBLE	.684	.218	.648	.203	.699	.209	.658	.214	2.22	NS	.07	NS
ORIENT	.565	.264	.444	.221	.662	.257	.545	.234	15.78	.00	.03	NS
COUNTERS	.794	.153	.750	.172	.799	.186	.783	.176	2.18	NS	3.74	NS
TARGETID	1.56	.548	1.91	.536	1.26	.447	1.62	.547	34.69	.00	.00	NS
TRACK1	2778	378	3247	375	2648	418	3096	457	91.75	.00	.39	NS
TRACK2	3670	466	4334	359	3339	537	4063	493	142.1	.00	3.36	NS

Females:	Ns range from 86 to 90
Males:	Ns range from 205 to 223

Navy Incremental Validity Study of New Predictors*

John H. Wolfe and David L. Alderton

Navy Personnel Research and Development Center, San Diego, CA 92152-6800

Paper-and-pencil testing of mental aptitudes reached a plateau of predictive validity during and shortly after World War II. During the last decade, the development of personal computers has offered the opportunity to administer new types of tests of memory, reaction time, spatial ability and psychomotor skills. The military services recognized this opportunity and initiated projects to develop new tests and find whether they improved validity when added to the existing selection tests, called the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB can be administered in either a paper-and-pencil mode (P&P-ASVAB) or a computerized adaptive testing mode (CAT-ASVAB).

The ASVAB is heavily weighted towards crystallized academic skills. It lacks any direct measure of spatial ability or perceptual speed. Working memory, which Kyllonen and Christal (1990) and others believe to be the basis for fluid intelligence, is difficult to test in a paper-and-pencil mode. The ASVAB also lacks an abstract reasoning test.

The purpose of the present study is to determine whether the validity of the ASVAB can be improved by supplementing it with a battery of new computerized tests of working memory, spatial ability, and perceptual speed.

Method

Tests

Table 1 lists the tests that were administered.

Table 1
Tests in the Navy Validity Study

Working Memory
Sequential Memory (Larson & Alderton, 1990)
Mental Counters (Larson, Merritt, & Williams, 1988)
Spatial Visualization
Integrating Details (Alderton, 1989)
Space Perception (from ASVAB Form 6)
Perceptual Speed
Overall Rate Score on three new Subtests (Alderton, 1990)
Reasoning
Figural Reasoning (from Project A)
ASVAB (Alternate Forms, Afternoon Testing)
P&P ASVAB for 1/2 Examinees
CAT-ASVAB for 1/2 Examinees

* The opinions expressed in this paper are those of the authors, are not official and do not necessarily reflect the views of the Navy Department.

Subjects

The examinees were Navy recruits at the Great Lakes Recruit Training Center who were scheduled for technical training at one of nine Class "A" schools, as shown in Table 2. 4989 recruits were tested, of which 3997 graduated. This number was further reduced to 3356 by eliminating cases with missing data on the dates of their testing or other factors.

Table 2
Schools for Navy Validity Study

Abbrev.	Title	Tested	Enrolled	Graduated
AD	Aviation Machinist's Mate	136	125	115
AMS	Aviation Structural Mechanic - Structures	122	115	104
AO	Aviation Ordnanceman	128	125	117
AV	Avionics Total, consisting of:	368	330	294
	Aviation Electronics Technician	241	213	186
	Aviation Fire Control Technician	80	72	66
	Aviation Antisubmarine Warfare Technician	47	45	42
BT/MM	Boiler Tech/Machinist, consisting of:	1169	988	935
	Boiler Technician	427	353	335
	Machinist's Mate	742	635	600
GMG	Gunner's Mate - Phase I	447	427	398
HM	Hospitalman	782	832*	628
HT	Hull Maintenance Technician	454	418	391
OS	Operations Specialist	1155	1109	1015
	Unassigned	228	0	0
	Total	4989	4469	3997

* 87 initially unassigned cases were sent to HM school.

Testing Schedule

Testing was conducted from June 1989 through February 1990. The examinees were tested in their second week of training, immediately after classification into different occupational specialties. All examinees received approximately three hours of experimental cognitive tests in the morning. In the afternoon, they were randomly assigned to either P&P ASVAB or CAT-ASVAB testing.

Instructions

Examinees were first given written and oral information about their rights under the Privacy Act, told that the testing was for research purposes, would be kept confidential, and would have no effect on their careers, and then asked to sign a statement giving permission to be tested under these conditions. Of course, these instructions eliminated much of the incentive to perform well on the tests.

Criteria

School performance data were obtained from a variety of sources. Final School Grades (FSG) were readily obtained for each school from existing records, often in computerized data bases. The same sources provided internal consistency reliability estimates for FSGs. An exception was the Avionics school, where no reliability estimate could be obtained for FSG, which had a mean of 99.2 out of a possible 100, with a standard deviation of only 1.78. Avionics FSG was omitted from the subsequent analyses.

In addition, every effort was made to obtain records of hands-on practical laboratory exercises. In most cases, these turned out to be simple pass-fail marks, with everyone passing. In three schools -

Aviation Machinist, Avionics, and Hull Technician, meaningful practical criteria were available. These were factor analyzed, and the factor pattern guided the construction of composites of unit-weighted criterion variables.

Communalities were used as estimates of the reliabilities for the components of the composites, and then reliabilities of the composites were computed using the standard formulas for the correlation of sums. Table 3 shows reliabilities and corrections for range restriction.

Table 3
Characteristics of School Performance Criteria

						Uncorrected		Corrected	
School	Criterion	N	Mean	Min	Max	Std. Dev.	r_{xx}	Std. Dev.	R_{xx}
AD	FSG	92	87.2	78.5	96.5	4.49	.950	6.95	.979
AMS	FSG	89	81.3	72.9	92.3	3.82	.900	6.87	.969
	LAB	89	84.3	76.7	94.8	3.75	.606	4.75	.755
AO	FSG	94	82.2	69.9	[illegible]	6.56	.880	8.32	.925
AV	LAB1	226	94.6	66.4	100.0	4.77	.512	4.90	.536
	LAB2	226	93.8	85.0	98.9	2.30	.412	2.73	.582
	LAB1+LAB2	226	94.1	81.5	99.1	2.82	.617	3.08	.678
BT/MM	FSG	811	86.0	75.1	99.9	5.23	.810	6.09	.860
GMG	FSG	324	86.0	73.0	98.7	4.79	.920	6.04	.950
HM	FSG	491	82.1	73.5	94.6	3.92	.930	4.84	.954
HT	FSG	322	90.5	82.6	97.4	3.11	.910	3.55	.931
	QUIZES	322	91.4	80.9	98.0	3.34	.819	4.26	.889
	LAB1	322	94.0	86.7	96.9	1.58	.788	1.68	.812
	LAB2	322	97.5	91.8	99.6	1.08	.438	1.10	.459
	LAB1+LAB2	322	95.4	90.8	97.5	1.08	.753	1.13	.775
OS	FSG	907	88.3	74.8	98.1	4.40	.900	5.67	.940

Hypothesis Testing

In a study of this kind, hundreds, or even thousands of hypotheses could be tested. In order to control the Type I error associated with multiple significance tests, a hierarchical approach was used (Cohen & Cohen, 1983, p. 172). First, a single hypothesis for the whole study is tested, then hypotheses for each school, hypotheses for each new predictor, and finally hypotheses for school x new predictor combination.

The multiple correlation of all ten ASVAB tests was computed for each criterion for each school.[1] Next, the multiple correlation of the ASVAB plus four composite predictors with each criterion was computed. For each criterion, the probability associated with the difference was determined from the F-distribution with degrees of freedom = 4 and N - 15, where

$$F_{4,N-15} = \frac{\Delta R^2}{1 - R^2_{ASVAB+CTB}} \cdot \frac{N-15}{4}.$$

These probabilities were combined into a single number that represents the probability that the new predictors have no incremental validity in any school. For each school, only one criterion was chosen for inclusion in the aggregate probability. The combined probability is given by the chi-square distribution of $\sum_{i=1}^{Schools}(-2\log P_i)$ with $2\times Schools$ degrees of freedom (Fisher, 1932).

[1] In a separate analysis, Wolfe (1992) found no significant difference between the validities of CAT-ASVAB and P&P-ASVAB, so results were pooled here.

Schools: If the global null hypothesis is rejected, the previously computed probability values for each school are used to decide if the results for that school are significant.

New Predictors: If the global null hypothesis is rejected, probability values are computed for adding only one new predictor to the ASVAB for each school and each new predictor. The results are accumulated across schools, using the Fisher chi-square method described above. This yields a probability value for each new predictor for the whole study. However, the probabilities for the new predictors are not independent, as they are for schools.

In a similar manner, probabilities are computed for *deleting* one predictor from the complete battery of ASVAB plus all new predictors. This p-value is used to test whether a given new predictor is redundant with respect to the other new predictors.

Predictor × School: If a given school and a given predictor separately show significant incremental validity, then the previously computed joint probability of using that predictor in that school is used to test the hypothesis that adding that one predictor to the ASVAB improves validity for that school.

Estimating the Magnitude of Validity Increments

All hypothesis testing was based on uncorrected correlations. To estimate the magnitude of the validity increments, several corrections were applied at various stages of the analysis. Lawley's (1943) range restriction corrections were applied to the correlation matrix of predictors and criteria, using all ten preenlistment ASVAB tests as explicitly selected variables. Multiple correlations based on either corrected or uncorrected correlations were "shrunken" to estimate population values, using the Wherry formula.[2]

Finally, the multiple correlations were corrected for criterion unreliability, using range-corrected reliabilities.

Results

Global Statistical Significance

The probability for the null hypothesis of no incremental validity in any school turned out to be 3×10^{-10}.

Schools

Although five of the nine schools show significant results, the largest gains occur in three schools: Aviation Ordnanceman, Avionics, and Hull Technician, where the validity increments exceed .05. Table 4 shows the results.

Predictors

We looked at the validity increments associated with adding only one new test to the regression equation with all ten ASVAB tests. Combining probability values across samples, we found that every new predictor had significant incremental validity when added to the ASVAB.

We also looked at the effect of deleting one predictor from the augmented battery of ASVAB plus new tests. Results show that the Spatial composite, Mental Counters, and Sequential memory tests have unique predictive ability not measured by other tests in the battery. Either ASVAB-6 Space Perception or Integrating Details could be deleted from the battery without significant effect, but not both. Figural Reasoning or Perceptual Speed could be deleted from the battery without significant effect if the other tests remained.

[2] Because negative correlations differences were replaced by zeros, the mean correlation difference may be larger than the difference in the mean correlations.

Table 4
Incremental Validities over Post-enlistment ASVAB

			Uncorrected				Fully Corrected		
School	Criterion	N	R_{ASVAB}	$R_{ASVAB+CTB}$	Effect†	$P(F_{4,N-15})$	R_{ASVAB}	ΔR	Percent Gain
AD	FSG	92	0.500	0.515	0.020	8.168×10^{-1}	0.796	0.000	0.0
AMS	LAB1	89	0.306	0.401	0.080	2.197×10^{-1}	0.639	0.017	2.6
AMS	FSG	89	0.404	0.417	0.013	9.116×10^{-1}	0.833	0.000	0.0
AO	FSG	94	0.554	0.643	0.182	9.727×10^{-3}	0.733	0.052	7.1 **
AV	LAB1	226	0.338	0.396	0.051	3.313×10^{-2}	0.423	0.061	14.5 *
	LAB2	226	0.383	0.450	0.070	6.326×10^{-3}	0.784	0.036	4.6 **
	LAB1+LAB2	226	0.400	0.468	0.075	4.034×10^{-3}	0.600	0.051	8.5 **
BTMM	FSG	811	0.494	0.498	0.006	3.498×10^{-1}	0.708	0.000	0.1
GMG	FSG	324	0.539	0.562	0.036	2.692×10^{-2}	0.751	0.008	1.1 *
HM	FSG	491	0.547	0.551	0.006	5.726×10^{-1}	0.745	0.000	0.0
HT	FSG	322	0.413	0.428	0.015	3.213×10^{-1}	0.602	0.002	0.3
	QUIZES	322	0.525	0.547	0.033	4.167×10^{-2}	0.776	0.008	1.0 *
	LAB1	322	0.312	0.371	0.047	6.658×10^{-3}	0.444	0.043	9.7 **
	LAB2	322	0.243	0.306	0.038	2.272×10^{-2}	0.355	0.059	16.7 *
	LAB1+LAB2	322	0.336	0.410	0.066	5.497×10^{-4}	0.449	0.063	13.9 **
OS	FSG	907	0.457	0.492	0.045	6.297×10^{-8}	0.736	0.015	2.1 **

Notes:

†Effect $= \dfrac{\Delta R^2}{1-R^2_{ASVAB+CTB}}$. * $p < .05$ ** $p < .01$

As shown in Table 5, significant validity increments greater than .02 occur in three schools - Aviation Ordnanceman, Avionics, and Hull Technician. The two spatial tests improve validity for all three schools. Of the two working memory tests, Mental Counters increases validity in Hull Technician lab, while Sequential Memory is involved in Avionics lab. Figural Reasoning also has incremental validity in Avionics lab.

Table 5
Fully Corrected Incremental Validities over Post-enlistment ASVAB

School	Criterion	Mental Counters	Sequential Memory	ASVAB-6 Space	Integrating Details	Perceptual Speed	Figural Reasoning
AO	FSG	.000	.000	.037**	.033*	.003	.000
AV	LAB1	.000	.074**	.000	.024	.008	.055**
	LAB2	.001	.000	.012*	.016*	.018*	.017*
	LAB1+LAB2	.001	.026*	.009	.023*	.017*	.038**
GMG	FSG	.002	.001	.000	.002	.007**	.001
HT	QUIZES	.002	.000	.000	.002	.004*	.001
	LAB1	.039**	.000	.013	.028**	.002	.008
	LAB2	.005	.000	.038*	.000	.021	.000
	LAB1+LAB2	.043**	.000	.026**	.028**	.009	.002
OS	FSG	.009**	.006**	.001	.006**	.000	.010**

* $p < .05$ ** $p < .01$

Discussion

One fact that emerged from the study was that the ASVAB is a remarkably good predictor of Navy training school grades. When corrected for restriction in range and criterion unreliability, most multiple correlations are in the mid .70's. There is little room for improvement, and the incremental validities of the new predictors are generally low for predicting grades. In contrast, laboratory performance criteria are not so well predicted by the ASVAB, and here the new predictors have their greatest incremental validities, up to 16.7%. The ASVAB is best at measuring academic aptitude, or "book learning" ability. Laboratory or shop work may require more fluid intelligence, spatial ability, and working memory, which the new predictors measure. The practical skills may be more important for subsequent job performance than the academic learning measured with written tests. Thus the utility of new predictors for selecting personnel may be better estimated from their incremental validities for predicting lab criteria than for grades.

Schmidt, Hunter, and Dunn (1987) estimated that a 3% improvement in the average validity of the ASVAB could produce an annual utility increase of $83 million for the Navy. In the present study, the incremental validity averaged 2% over all schools, including some with zero improvement. This increase translates into a $55 million improvement in utility for the Navy, and at least three times that for all of the military services combined.

The predictors used in this study were chosen for exploratory research to determine whether the constructs of spatial ability, working memory, and perceptual speed could improve prediction of school performance. They are not necessarily the optimum enhancements to the ASVAB. The battery omits other important ability measures, such as psychomotor ability. The tests themselves could be psychometrically engineered for higher reliabilities or adaptive administration. Thus further research might be able to double or triple the incremental validities found here, especially if lab or shop criteria were used.

An especially important finding was that the working memory tests have unique predictive power not redundant with other new predictors or the ASVAB. The working memory tests require computer administration. Hence the ASVAB cannot be improved beyond a certain point without becoming computerized.

References

Alderton, D. L. (1989). *Development aned evaluation of Integrating Details: A complex spatial problem solving test* TR 89-6. San Diego: Navy Personnel Research and Development Center.

Alderton, D. L. (1990). *Revisiting a cognitive framework for test design: Applications for a computerized perceptual speed test.* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences* (2nd ed.). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Fisher, R. (1932). *Statistical methods for research workers.* (4th ed.) London: Oliver & Boyd.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 12,* 131-147.

Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence, 14,* 309-325.

Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: Some implications of task complexity. *Intelligence, 12,* 131-147.

Lawley, D. (1943). A note on Karl Pearson's selection formulae. *Royal Society of Edinburgh Proceedings, Section A, 62,* B 28-30.

Schmidt, F., Hunter, J., & Dunn, W. (1987, September). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB)* TCN 86-698. (Contract No. DAAL03-86-D-0001, DO 0053). Research Triangle Park, NC: U. S. Army Research Office.

Wolfe, J. (1992). *Validity equivalence of computerized adaptive testing and conventional administration of the Armed Services Vocational Aptitude Battery for predicting training performance in nine Navy technical schools.* Manuscript submitted for publication.

THE OPERATIONAL TEST AND EVALUATION OF COMPUTER-BASED TESTING

by

E. R. Wilbur, K. E. Moreno,
and R. D. Hetter[1]
Personnel Systems Department
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

Research laboratories for the military services have been conducting research in the area of Computer-Based Testing (CBT) for more than a decade. Psychometric studies have been very encouraging. Research on the Computerized Adaptive Testing version of the ASVAB (CAT-ASVAB) has shown that most CAT-ASVAB tests are significantly more reliable than corresponding Paper-and Pencil ASVAB (P&P-ASVAB) tests despite the shorter length of CAT-ASVAB (Moreno & Segall, in press). Results from a Navy validity study on new CBTs demonstrate that, for the prediction of school performance, some of these tests provide incremental validity over P&P-ASVAB (Wolfe, 1991).

Results of psychometric studies provide information necessary to evaluate CBT in terms of improved accuracy. They do not, however, provide all of the information necessary to decide whether CBT should be implemented. Two additional factors, the associated costs of implementation and the method in which the test is implemented, or the concept of operation, may be the most critical elements in an implementation decision.

Cost of implementation is associated with psychometric characteristics of a test, for example, in the reduction of school failure rates and maximization of on-the-job performance. However, it is also influenced by the concept of operation. This is particularly true for CBT. The number of machines needed to implement CBT, one of the most influential factors in determining implementation costs, varies drastically with the concept of operation and test siting strategy. Concept of operation also influences costs such as recruiter time and travel, applicant travel, and test administrator time and travel. In addition, the concept of operation may impact other issues of concern to the military services such as test security, accession flow rates, and processing capacity.

[1] The opinions expressed in this paper are those of the authors, are not official, and do not necessarily represent those of the Department of the Navy.

A Joint-Service effort is underway to evaluate concepts of operation for future ASVAB (Weiss, 1991). Several concepts are being evaluated, including CBT. As part of this effort, an Operational Test and Evaluation (OT&E) of CBT is being conducted. The purpose of this OT&E is to collect information that will be used in evaluating alternative concepts of operation and the costs and benefits of the various concepts.

APPROACH

Test Sites

OT&E data collection will be conducted in four Military Entrance Processing Stations (MEPS), San Diego, Jackson, Baltimore, and Denver. It will also include data collection in some of the Mobile Examining Team Sites (METS) associated with these four MEPS. At these sites, CAT-ASVAB will be administered to all military applicants and CAT-ASVAB test scores will be the scores of record. Toward the latter part of OT&E, selected Enhanced Computer Administered Tests (ECAT) may be added to CAT-ASVAB. ECAT will be administered as a non-operational test and scores will be used for research purposes.

Data Collection

During OT&E, data will be collected on factors influencing implementation decisions for CBT. The methods used for data collection will be CAT-ASVAB test administration, questionnaires administered to recruiters, applicants, and MEPS personnel, on-site observation, and interviews with MEPS personnel. The types of information to be collected are as follows:

1. Impact of change in operational procedures. The government will be able to test different operational scenarios such as a variable-start scenario where examinees arrive and begin testing at successive intervals. The effect on MEPS personnel, recruiters, and applicants will be evaluated.

2. Equipment needs. The amount of equipment required under certain concepts of operation will be studied. For example, the number of machines required under a variable-start scenario at OT&E sites can be used to estimate machine requirements nationwide under this type of scenario.

3. Test Administrator training and performance. During OT&E, test administrators currently giving P&P-ASVAB will administer CBT. A training program designed for test administrators will be developed, evaluated, and revised based on this limited trial.

4. Logistics. Associated logistical concerns such as equipment setup, takedown, and maintenance over an extended period of time will be assessed.

5. User acceptance studies. Public relations concerns will be identified and appropriate procedures and materials developed to address these problems.

6. Security issues. Extended operational data collection will allow the government to assess procedures for identifying potential security problems. It will also allow evaluation of the effectiveness of item exposure control.

7. Pilot testing of ECAT. Applicant data will be used to evaluate test instructions, item difficulty, and test time limits and to provide an improved basis for decisions regarding operational implementation of ECAT.

8. Item/test functioning. Classical and item response theory statistics computed from a broad, heterogeneous sample will be used to examine item functioning across ability levels and subpopulations such as gender or ethnic group.

9. Motivation studies. A large sample of motivated applicants will provide valuable data for examining issues in appropriateness measurement and deliberate-failure.

10. CAT-ASVAB retest data. Extended periods of testing in multiple locations will provide data on CAT-ASVAB retest examinees. These data will be used to compare the expected change in CAT-ASVAB and P&P-ASVAB scores as a function of retesting.

RESULTS

Data collection for OT&E is currently in progress at three sites. Data collection began on June 1, 1992 in San Diego, California, August 2, 1992 in Los Angeles, California, and September 1, 1992 in Jackson, Mississippi. Original plans for OT&E did not include testing in Los Angeles. However, when the MEPS burned during the Los Angeles riots, CAT-ASVAB was installed with equipment added to telecommunicate scores to the San Diego MEPS for processing. Data collected in Los Angeles will be included in the OT&E database.

Data collection in Baltimore, Maryland, will begin on February 1, 1993 and data collection in Denver, Colorado is currently scheduled to begin on May 1, 1993. MET site testing is still in the planning stages.

Preliminary results have been obtained for some of the issues being studied. These results are primarily based on site observations. Some data are available from applicant questionnaires that have been administered.

Test Administrator Training

The training program for military test administrators has been very successful. Test administrators met all course objectives and required minimal on-the-job training. Observation of performance on the job has shown that test administrators can satisfactorily operate the system. However, it has become obvious that due to the high turnover in test administrators and scheduling conflicts, "group-administered" classroom training is not the appropriate method. For these reasons, a computer-based training program using an intelligent tutoring system will be used in Baltimore and Denver for training CAT-ASVAB test administrators.

Flexible Start Option

All OT&E sites are currently using a flexible start option. Each MEPS has established an arrival window during which applicants may arrive and begin the test. Recruiters and applicants find flexible start reduces scheduling problems and makes is easier for applicants to attend testing sessions. MEPS personnel were initially concerned about the flexible start option because it differed from the traditional group administration mode. They have found, however, that the procedure works well.

Recruiter Reactions

Recruiter reactions are overall very positive but do vary by MEPS. Recruiters in San Diego have expressed concern over the differences between CAT-ASVAB and P&P-ASVAB. They find it difficult to understand how a test with 16 items can provide a number-correct score of 35. This demonstrates the need for public relations materials. In fact, recent briefings provided to recruiters on CAT-ASVAB have greatly reduced concerns.

These same concerns have not been found in Los Angeles or Jackson. Recruiters in these areas are highly enthusiastic about CAT-ASVAB. The benefits of reduced CAT-ASVAB test lengths and receiving immediate scores appear to outweigh any other concerns. In fact, some recruiters in these areas travel a considerable distance to bring applicants to the MEPS for CAT-ASVAB testing rather than to a nearby MET site for P&P-ASVAB.

Test Administrator Reactions

Test administrator reactions to CAT-ASVAB are much more positive than expected. Test administrators in Jackson and San Diego prefer CAT-ASVAB to P&P-ASVAB. According to Jackson test administrators, even with flexible start, CAT-ASVAB conserves time both in test administration and processing of test results. Los Angeles test administrators provide printed scores immediately to recruiters thereby reducing the number of unqualified applicants sent to San Diego for testing. Test administrators at all sites have expressed appreciation for the computer-administered instructions and timing. In addition, because examinee responses are scored by the computer, there are no answer sheets to scan.

Applicant Reactions

A questionnaire to assess applicant reactions to both P&P-ASVAB and CAT-

ASVAB has been administered at OT&E sites.[2] Preliminary analyses from San Diego data show differences between CAT-ASVAB and P&P-ASVAB examinee responses on several questionnaire items. CAT-ASVAB examinees are less tired at the end of the test, feel less pressured, are better able to concentrate during the test, and feel the test was shorter. CAT-ASVAB examinees also perceive the test as being slightly more difficult, feel slightly worse about taking the test and are slightly more worried. While the results are statistically significant, the magnitude of the difference is small and attempts at interpretation should consider the adaptive nature of CAT-ASVAB. Traditionally, high ability examinees have been accustomed to answering most or all items on a test correctly. In CAT-ASVAB, a correct response is followed by an item of higher difficulty which could lead to the perception of a more difficult test and of poorer performance. Finally, the questionnaire contains a comment section. Most applicants who choose to comment, and especially those who previously tested on P&P-ASVAB, express a clear preference for CAT-ASVAB.

Other Observations

Observations of field operations during OT&E have revealed design changes that could improve CAT-ASVAB software and hardware. For example, test administrators currently must hand-write and keystroke applicant information on required processing forms even though that information has been previously entered in the CAT-ASVAB system. Designing and printing the forms at the Test Administrator Station will eliminate this duplication of effort. Oth:r beneficial software changes will be to streamline the "stand-alone" mode of operation, revise interactive screen dialogues for ease and speed of comprehension, and to enter additional applicant information in the system to further reduce the work load of test administrators.

OT&E has also shown that for nationwide implementation of CAT-ASVAB, the hardware system must be more portable. During OT&E, MEPS and MET site installations are considered nearly permanent operations. In other words, equipment at these sites remains installed for the duration of testing. However, during nationwide implementation, for MET sites in particular, equipment may need to be set up for each testing session, necessitating a much more portable design than currently in use.

SUMMARY

To date, OT&E has provided information that will be very useful in evaluating concepts of operation for CBT. In addition, the OT&E data will be valuable in designing and fielding a system for nationwide implementation, if such a decision is made. OT&E has shown that CBT meets the needs of recruiters, applicants, and MEPS personnel.

[2] Questionnaires were administered to CAT-ASVAB applicants only at Los Angeles.

REFERENCES

Moreno, K. E. and Segall, D. O. (In press). Alternate form reliability and cross-correlational analyses of the CAT-ASVAB. San Diego, CA: Navy Personnel Research and Development Center.

Wise, L. (1991). ASVAB overview: Process, content, and evaluation factors. Unpublished manuscript.

Wolfe, J. (1991, July). Navy validity study of new predictors: Preliminary findings. Paper presented at the meeting of the Defense Advisory Committee on Military Personnel Testing, Monterey, California.

THE MOTIVATION TO LEARN: ADVANCED TECHNOLOGIES FOR DISTANCE LEARNING

INTRODUCTION

Dennis J. Gettman
Armstrong Laboratory
Brooks AFB, Texas

For more than 40 years distance learning has opened the door of educational opportunity to thousands who, without distance instruction, would simply not have received it, or would have traveled great distances for the opportunity to learn. Television technology has served this effort well. Direct transmission, satellite, microwave, and cable have made analog television the dominant paradigm for distance learning. Paper-based correspondence and more recently distributed computer software programs continue to present inexpensive alternatives for distance learners.

There have been problems, however. Broadcast channels are expensiva and inflexible. Receive sites must be permanent facilities containing a great deal of special equipment including satellite dishes and private line phone service for interaction. Distribution of materials and collection of homework and tests are often conducted through slow mail services, delaying student feedback and grades. Finally, most traditional approaches rely on a "talking head" to deliver instruction in imitation of a live classroom, though there is little of the immediacy and impact of live performance that can be translated through television. Such systems work, but only when participating students are highly motivated to perform and succeed. Thus, traditional distance learning creates a media-poor environment, devoid of immediate feedback, student involvement, one-on-one interaction, and useful exchange of materials.

I believe that the problems we will encounter for distance learning in the future are more psychological than they are technological. I see great disparity between how potential users/managers perceive distance learning and what could actually be put to use. The paradigm is shifting and we may not be shifting with it fast enough to take full advantage of the exciting changes in communications technology. Conceptually, multimedia distance learning connects students and teachers through computers to create a media-rich environment for collaborative education and exploration. The paradigm shift should occur in how we envision the blending of technologies, the proper mix, under appropriate circumstances for the purpose of stimulating motivation and thus, enhancing the learning process. Further challenges will involve cost and production. How expensive will the implementation of these amazing technologies be and will it be worth it? What are the most efficient and effective ways of producing lessons for multimedia distance learning courses? What types of courses best lend themselves or even require multimedia? What should the teachers do? What should the students do? How can we best take advantage of technology for education and training?

Obviously there are many questions. The following papers presented in the distance learning panel at this MTA conference yield hope in finding answers to many of the questions. The real answer is that there are a range of answers which can be implemented when appropriate. For example, as you will see, paper-based correspondence has been providing instruction to motivated students for over 40 years. Certainly improvements can and are being made all the time, but this type of distributed education is inexpensive and works with the right students. On the other hand, the future holds much promise for the use of very sophisticated multimedia formats including television, DVI technology, collaborative groupware, and even intelligent tutoring at a distance. The question for distance learning users of the future should not be which technology for which course, but rather, which combination of technologies will work best to stimulate and motivate the learner.

EMPIRICAL COMPARISON OF ALTERNATIVE INSTRUCTIONAL TV TECHNOLOGIES

Henry Simpson
Navy Personnel Research & Development Center
San Diego, California

An experiment was conducted to compare the training effectiveness and user acceptance of live instruction and six different alternative Instructional TV (ITV) technologies: multi-channel 2-way video with 2-way audio, single-channel 2-way video with 2-way audio, 1-way video with 2-way audio, 1-way video with intermittent 2-way audio, and audiographics. Findings were that in comparison with live instruction ITV in several different forms was effective both in terms of student performance and student and instructor acceptance. The most successful ITV technologies were those allowing continuous 2-way audio communication between classrooms with either 2-way or 1-way video. Using 2-way video does not appear to improve student performance as compared to 1-way video, but instructors prefer 2-way video and students expressed the desire to see their cohorts in other classes, which requires 2-way video. Student test performance was poorer with ITV systems that restricted remote students' ability to converse with or see the instructor and the performance decrement was evident in both local and remote classrooms. This difference may be accounted for by the additional interactivity possible with ITV systems allowing continuous 2-way audio communication. Evidence also suggests that student acceptance of partially-interactive ITV technologies was lower than with fully-interactive ITV. Similar results would be expected with videotaped instruction. Students adapted to compensate for the video and audio shortcomings of ITV technologies. The most serious shortcoming of the simulated ITV technologies was audio. Additional work needs to be done to refine the audio systems and procedures used in VTT.

The research was performed under the sponsorship of the Office of Naval Technology. The opinions expressed are those of the author, are not official, and do not necessarily reflect the views of the Navy department.

Collaborative Instructional Development Environment: A Stage for the AIDA

Robert G. Main & Andrew S. Wilson

INTRODUCTION

The Air Force has identified a goal concomitant with the development of the Advanced Instructional Design Advisor (AIDA) to create a multi-use instructional design workstation that will provide designers/developers the power to locally produce instructional materials ranging from graphics to video to computer-based interactive training. The Collaborative Instructional Development Environment (CIDE) Workstation is a set of functional specifications for hardware, software and network communications to operate on the Air Force Workstation III and IV. The specifications were developed from the responses of a Delphi committee asked to evaluate potential components of such a system in terms of their most common tasks and development efforts. The purpose of this research is to ascertain not only the computer-based media development tools required for effective instructional design by the Air Force, but to explore the type of collaborative support environment that will make available to instructional developers the expertise necessary to produce curriculum materials that fully exploit the power of media to motivate as well as teach.

DISCUSSION OF THE PROBLEM

Well-developed, appropriate media enhances instructional quality (Johnston, 1987). However, its use and effectiveness are hampered by:

- High cost
- Inability to determine appropriate media
- Lack of production expertise
- Substantial lead time for production
- Communications problems

While an Advanced Instructional Design Advisor (AIDA) can aid in selecting the appropriate media for an instructional activity, the problems listed represent serious impediments to the instructional developer in incorporating that media into the lesson. An experienced live instructor can often compensate for problems with instructional media, but the trend toward electronic delivery of instruction demands highly effective stand-alone media that communicates and motivates (Winn, 1987). Thus, there is a pressing need to empower instructional designers/developers (IDD's) with the tools to create powerful and dynamic instructional aids both for standup instruction and for computer-based instruction and distance learning.

The software and hardware tools now exist to create a multimedia production workstation that will allow development of a wide array of instructional media from documents to animation to interactive digital video. However, it is unrealistic to believe that instructional developers, often subject matter experts in a particular area of training, will also have the skills to make the best use of the powerful tools such a workstation will provide. Therefore, how to empower IDD's to produce and manipulate a variety of instructional media without placing on them the burden of requisite expertise in what are typically specialist areas has become a primary concern. For example, providing a graphics software package that is easy to use does not endow the user with the talent to generate visuals that are creative and powerful in their ability to facilitate the transfer of knowledge and also grab the learner's attention and hold his or her interest.

AIDA offers a partial solution. AIDA can assist an IDD in selecting an appropriate medium to accomplish specific instructional objectives. It will guide developers through many complex design decisions and help them clarify goals and directions. But AIDA will not be able to review the aesthetics of media development. It will not be able to examine a proposed media solution for clarity of purpose and execution. Though it will be able to point out potential errors, it will not be able to answer specific questions regarding the easiest way to solve a particular communication problem; nor can AIDA provide years of media production experience to a neophyte designer.

So while AIDA provides less experienced IDD's important guidance and assistance in instructional design decisions, it cannot give comprehensive direction and evaluation in media production. They need both the valuable guidance of AIDA and a support infrastructure for technical and creative assistance in making the best use of the increasingly complex tools at their disposal and for producing the media that will most effectively support the instructional objectives.

THE COLLABORATIVE INSTRUCTIONAL DEVELOPMENT ENVIRONMENT (CIDE)

The increasing demand for multimedia instructional materials has created the need for a collaborative instructional development environment that includes not only instructional design and subject matter expertise, but media development specialists as well. The addition of a communications network to the computer platform on which designers/developers will be running both AIDA and the designer-specified hardware and software for media production can create a collaborative open network for sharing expertise in mentor/apprentice relationships between developers and special interest advice conferences led by media specialists. Providing on-demand assistance should solve most media production problems and provide a logical path for the instructional developer to travel from the recommendations of the AIDA to the finished lesson material required by the instructional design.

The inherent versatility of an open network brings a number of added benefits to Air Force instructional design organizations including:

- Centralized Multimedia Education and Training Archive (META) of reusable instructional media for easy updates, adaptation and integration.
- Ability to search local and remote archives for appropriate existing media.
- Organizational gateway for send/receive faxing and E-mail.
- Centralized electronic media publishing.
- Automated document flow for formative and summative evaluation.
- Improved task management and group coordination.

This study will address the following questions pertinent to a collaborative instructional development environment:

1. Is electronic collaboration effective?
2. What are the instructional media needs of Air Force instructional developers and how are they currently being met?

3. **Can a collaborative instructional development environment be created with current commercial-off-the-shelf technology?**

ASSUMPTIONS

That electronic media provide effective enhancement to instructional strategies in terms of improved cognition, information organization and integration, and learner motivation has been generally accepted since the 1968 research of Chu and Schramm (1979). Therefore, this study does not address the use of electronic media in improving the effectiveness of instruction. The term "effective" implies a judgment relative to a standard. This has customarily meant comparison of some electronic medium relative to face-to-face instruction without electronic media. "Medium" is defined by the American Heritage Dictionary as "an agency, such as a person, object or quality, by means of which something is accomplished, conveyed or transferred." The electronic media are simply carriers of information for presentation to the learner. The potential of a medium for transfer of knowledge and skills has more to do with how the information is packaged and accessed by the learner than it does with the characteristics of the medium itself (Johnston, 1987).

Information encoded as print differs very little whether it is presented by video, CRT or in a handout. The spoken word may be carried by the instructor's voice in person, on audio tape or by video. The assumption is made that learning does occur with electronic media and that the way the information is packaged, accessed and presented can have great effect on the learning process. Given this assumption, the problem becomes how to provide the tools and the expertise to permit the instructional developer to have choices in the selection of an appropriate medium and the most effective packaging of that instruction for optimal learning. An important consideration in the design and development process is the "mindware", a term coined by Salomon (1985) that refers to the mindset a learner brings to the instructional process. It includes the learner's propensities and associations with media. Main (1992) has attempted to systematize these motivational factors by generating a model of instructional design which integrates the affective domain into the curriculum development process. Although empirical data is sparse, the attractiveness of electronic media is evident in the amount of leisure time spent with television and electronic games by both children and adults. Though there is some ongoing discussion in the literature over how much and what types of media are appropriate to various tasks (see Friedman, Polson and Spector, 1991), we will not attempt to explore these issues here since they are more the province of an AIDA, SME's and IDD's than they are dependent on a CIDE.

The very strength of a collaborative instructional development environment is the versatility to produce any and all types of media as deemed suitable to a particular task by the cadre of professionals employed for ISD. Whether the media need is a published document, graphic slide, computer-based animation, or edited videotape, the personal computer has matured sufficiently as a platform to produce professional quality products. We state here that the PC is sufficiently robust to simultaneously support a panoply of media production tools, an AIDA, and a collaborative network to form electronic work groups.

Electronic work groups as discussed here are the result of computer-mediated communication (CMC). Analog communication networks are not considered because of their relatively high cost when used for media exchange. Computer-based communication systems range from simple electronic mail to voice mail, interactive chat forums, desktop video conferencing, document transfer and shared screen editing. Basic CMC systems have gained enormous popularity over the last ten years as exemplified by the rise in usage of public network services such as MCI Mail, Prodigy and America On-Line, and the worldwide research collaboration taking place on Internet, NSFnet and Bitnet. For millions of people, checking their e-mail messages or logging into an interactive chat conference has become a daily ritual. Internet, a consortium of universities and research institutions, expects to have more than two million participants by 1995 (*Communications Week*, July 6, 1992, p. 1).

These services are active collaborative communities. Over one thousand special interest forums on the Internet, for example, allow asynchronous discussion of social, technological and scholarly issues ranging from animal psychology to quantum physics. The participants represent a huge knowledge base. By actively sharing information they multiply their individual skills and abilities.

IS ELECTRONIC COLLABORATION EFFECTIVE?

The literature on collaborative work is rich with examples of increased employee productivity. Though one study of Air Force cadets found that highly competitive people perform best when given individual rewards (Porter, Bird, & Wunder, 1990), the majority of researchers have reported improvement in performance on complex tasks by collaborative groups (Bassin, 1988; Blaye, Light, Joiner, & Sheldon 1991; Johnson, Maruyama, Johnson, Nelson & Skon, 1981; Katz, Kochan & Weber, 1985). In their excellent literature review, Tjosvold & Tsao (1989) state:

> Considerable research, including field experiments, indicate that people in cooperation compared to those in competition exchange resources, assist each other, and manage conflicts constructively so that they are all successful (p. 189).

Citing the findings of Johnson et. al. (1981) they continue, "as they work cooperatively, employees explore issues and make successful decisions, and are more productive especially on complex tasks that benefit from sharing information" (p. 189).

According to Bassin (1988), "It's not the gifted individuals who make peak performance possible as much as the dynamics of belief, collaboration and support" (p. 64). Bassin believes cooperative work groups are effective because of the resources of individual members, diversity of ideas, emotional support, mutual motivation and increased job satisfaction. He feels that isolated employees are at a fundamental disadvantage, unable to grasp how their work output fits into the overall performance of the organization. Collaborative teams solve these problems. According to Tjosvold et. al. (1989):

> In cooperation, people believe their goals are positively linked; one's goal attainment helps others reach their goals. Alternatively, mistrust, individual tasks, and win/lose rewards induce competition. Competitors believe their goals are negatively correlated so that one's goal attainment makes it more difficult for others to attain their goals (p. 189).

Finally, members of cooperative work groups report increased job satisfaction and organizational loyalty (Finholt, Sproull & Kiesler, 1990; Tjosvold et. al., 1989; Bassin, 1988; Sproull and Kiesler, 1986; Tjosvold, Andrews and Jones, 1983; Johnson et. al., 1981). A reduced sense of isolation, greater understanding of organizational objectives, emotional support and social interaction all seem to play important roles. Sproull and Keisler (1986) point out that cooperative work groups often use electronic mail to provide a productive outlet for natural desires for sociability and organizational attachment. "People like to be sociable at work. A technology that makes it easy to be sociable—be it a water fountain, coffee pot, telephone, or EMS [electronic messaging system]—will be used for sociability" (p. 1151). Likewise, Tjosvold, et. al. (1983) suggest that cooperative interaction strengthens morale, commitment to the organization and productivity. The positive experiences of working together lead employees to believe they have gained a great deal from the employer; and teamwork binds them to each other and to the organization.

It is generally felt that members of a cooperative work group benefit from the strengths of the talented individuals of whom it is comprised. (Bassin, 1988). Technical expertise and design experience more readily cross the organizational lines in an ad hoc cooperative group, providing just-in-time support for mission-critical objectives (Finholt, et. al., 1990). Employees participate in organizational goals and enjoy increased productivity and greater job satisfaction. But traditional methods of forming and maintaining ad hoc work groups such as face-to-face meetings may cost an organization a great deal in terms of travel, time to distribute materials, and time required to meet and to schedule more meetings, especially amongst geographically remote participants (Finholt et. al., 1990).

The literature provides sufficient evidence that electronic collaboration is effective. There is a note of caution. Changing communication patterns and protocols also changes organizational culture. These issues are not addressed in this paper but considerable empirical evidence is available and should be examined before establishing capabilities (see for example Lea and Spears, 1991; Dubrovsky, Kiesler and Sethna, 1991; Smilowitz, Compton and Flint, 1989; Sproull et. al., 1986).

WHAT ARE THE INSTRUCTIONAL MEDIA NEEDS OF AIR FORCE IDD'S AND HOW ARE THEY NOW BEING MET?

A Delphi group of expert Air Force IDD's was used to determine the instructional media needs of the Air Force and what tools would most enrich the collaborative instructional development environment.

The methodology for this study involved a review of state-of-the-art technologies in the field of personal computers and desktop workstations, media production tools and communication software. The evaluation was limited to commercial-off-the-shelf applications or products being beta tested for commercial release. An examination of the trade publications in personal computing, desktop publishing, digital photography, graphic design, video production and communication networking was used to establish a taxonomy of available products and services that appeared to have value for instructional development. A trip was made to the InfoMart in Dallas to see some of the candidate technologies demonstrated.

The information from the technology review was used to generate a list of 42 product categories divided into three areas of instructional design/development: 1) Instructional materials development, 2) Management and 3) Collaboration. To provide a rational method for evaluating the importance of the functions represented by these product categories, a combination of the Delphi methodology and the Kepner-Tregoe rational decision model (1965) was used.

The Kepn r-Tregoe rational decision model uses the technique of determining what are the essential outcomes and what are the desirable outcomes for any decision situation. It is widely used in the evaluation of competing systems because it provides a quantifiable method for comparing products with a variety of disparate features. The "must" category of features must be met by all candidate systems or they are dropped from further consideration. Those features deemed desirable but not essential are labelled "wants" and are assigned weights (usually by a panel of users). The evaluation is made by experts who test the system's ability on each item.

For our panel of experts we elected to use Air Force IDD's from a number of organizations that would reflect a variety of instructional development needs from standup courses to computer-based technical training. We opted for this approach over using consultants or academics because of the importance of user involvement in system design.

Systems development theory (Kling, 1991; Conner, 1985; Boar, 1984) and practical field experience (Kyng, 1991; Perin, 1991) both indicate that potential users of a system must be involved during the early stages of design. Kyng (1991) advocates a doctrine of "mutual learning" where designers teach users about the technological possibilities while users instruct designers in the task specifics of their work. Perin (1991) discusses the problems created when systems are mandated for unwilling users. Computer systems that extend the abilities of subordinates, and especially those that may create informal social fields among them, may threaten managers. "The challenge is to create computer support that acknowledges, if not incorporates these realities, rather than presuming the technology will by itself reform or obliterate them" (p. 81).

Therefore, while expert consultants might easily specify an extremely competent design system in terms of the prevalent ISD models and perceived needs of IDD's, there is no certainty that such a system will be readily adopted by Air Force IDD's. For these reasons, we assembled a Delphi panel of experienced users to assist in developing the functional design requirements for the CIDE. We were assisted in identifying expert IDD's by Lt. Sheila Robinson HQ ATC/TTDD and by Maj. Richard O'Neal HQ ATC/XPCR at Randolph Air Force Base.

Delphi is a technique developed by the RAND Corporation to be used in technical forecasting or to achieve consensus amongst a group of experts without undue influence (halo effect) by prestigious individuals (Tersine and Riggs, 1976). For this study, a Delphi group of ten experienced Air Force IDD's was selected. Their combined experience totals 108 years in curriculum development. The participants are expert practitioners rather than a representative sample of Air Force instructional developers. A survey by Walsh, Yee, Grozier, Gibson & Young (1992) of 256 Air Force personnel involved in developing computer-based instruction (CBI) found the average experience of the IDD's to be just 20 months. Participants were selected for this panel because of their knowledge and experience with Air Force instructional design and development, not because they represented typical IDD's.

Eight members of the panel of experts are male and two are female. Seven are civilian employees of the Air Force and three are career military personnel. The level of sophistication of the group is quite high. One participant is a manager of an instructional development group. Although not involved in the actual design of instruction at this time, he has more than seven years of prior experience in training development. He is included because his managerial responsibilities include the design and development of all types of instruction from traditional classroom to CBI. Eight of the members are experienced in designing and developing CBI and three of them do this exclusively. Five panel members are involved with the design of standup training using static media aids and six develop dynamic media for their instructional programs. Seven of the designers have at least some experience in multimedia CBI development. This level of expertise and experience in the field makes this group well qualified to offer expert evaluations concerning the functional requirements and desired features for a collaborative workstation to improve both the productivity and quality of Air Force instructional design and development.

INSTRUMENTATION AND EVALUATION

A structured questionnaire was distributed to the panelists in which they were asked to specify the percentages of instruction created using different types of media both within their organization and Air Force-wide. We then asked them to tell us how much instruction using each type of media they thought would be most appropriate for use by their organization and the Air Force.

They were presented with the list of 42 candidate technologies developed from the trade journals, literature review and vendor presentations. They were told their expertise was being solicited to assist in determining the design features of a collaborative instructional development workstation. They were asked to evaluate each technology category to determine if they felt it was essential (a "must") to quality curriculum development. If the technology was judged not to be essential, the panelists were asked to place a value on its worth (0=valueless to 20=nearly essential) to an instructional developer.

Finally, they were asked about media they can and cannot presently develop in-house and their collaborative relationships with other designers and subject matter experts.

We sought to answer five basic questions that bear directly on the functional specifications of the CIDE:

- What types of instructional media are presently being developed?
- What types of instructional media would IDD's prefer to develop if they had more resources?
- By whom is various media now developed (IDD's, non-training agencies, contractors)?
- What kind of collaboration is necessary to the development of effective media for instruction?
- What technologies are perceived as essential to IDD; which are desirable, and which are unnecessary?

The data gathered was averaged and used to rank order potential technologies that could be included in the CIDE. Using the Kepner-Tregoe (1965) decisioning system, we were able to determine which technologies constitute the necessities of the system and which the niceties. Using this "rational" decisioning system helps forestall the desire to add every available technology under the assumption that if we provide it to designers they will learn to want it and use it— the "Field of Dreams" approach.

THE TECHNOLOGIES

Our only constraint (self-imposed) was that all software and hardware technologies specified for the design of the CIDE should be compatible with the Air Force Workstation III and IV. Our intent in this is not only to reduce eventual development costs and to work with a computing platform that has already been approved and implemented by the Air Force, but also to ensure compatibility with the Advanced Instructional Design Advisor being developed for Air Force ISD (Hickey, Spector and Muraida, 1992).

Technological feasibility was determined through review of computer trade publications and an on-site visit to Dallas' InfoMart. While specific software and hardware selections will require further study and additional input from potential users, there will be a discussion below of critical technologies that match the user requirements determined by the Delphi panel. The research and development paradigm is to establish a rational ordering of functional requirements and assess the status of commercial-off-the-shelf (COTS) tools available to meet those requirements.

FINDINGS

Our hypothesis holds that Air Force IDD's are probably designing more standup instruction and more instruction with static media than they would prefer. If true, we would suspect it is because of an inability to design more dynamic curriculum materials stemming from a variety of reasons ranging from lack of skills to lack of equipment to insufficient time. To test this theory, we asked the Delphi panel for their best estimate of the quantities of instructional media of various types being produced by them, their organizations, and their best estimate of the media type's use Air Force-wide. Summaries of their responses are contained in figures 1-8.

As we postulated, individual Air Force IDD's generally feel they are developing more instruction without media, or instruction that is dominated by static media than they would prefer (Figure 1). The mean portion of curriculum hours developed as standup instruction with no media was estimated by our panel to be 40 percent within their own organizations and 32 percent overall for the Air Force. They believed a more suitable amount of this type of instruction would be about 25 percent.

The participants also indicated they would prefer to see less instruction supported by static media such as slides, overhead transparencies, etc. (Figure 2). They estimated instruction with static media accounted for almost 50 percent of the hours of instruction produced in their organization and nearly 60 percent Air Force-wide. Their preference was that approximately one-third of the instructional hours be standup instruction supported by static media.

The use of dynamic media for instruction shows an opposite result, i.e., the participants would like to use dynamic media more than it is being used now (Figure 3). Participants would like to increase their organization's use of dynamic media from 30 to nearly 40 percent of instruction designed, and would like to see it account for one-third of total Air Force instruction.

As we suspected, print-based media is still the most widely used medium (Figure 4). More than 90 percent of instructional hours are supported by some printed materials in the form of student handouts and workbooks. The effective penetration of desktop publishing and familiarity of nearly all instructional designers with paper-based production certainly facilitates its ubiquity. Nevertheless, panelists feel that the amount could be reduced somewhat without damage to the instructional process.

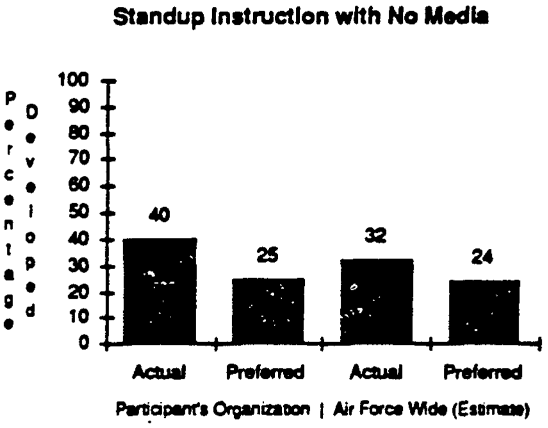


Figure 1

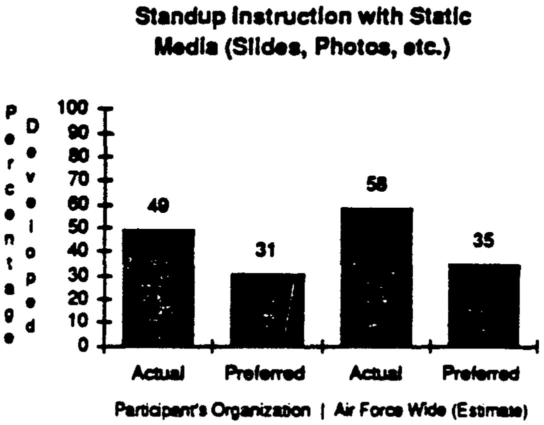


Figure 2

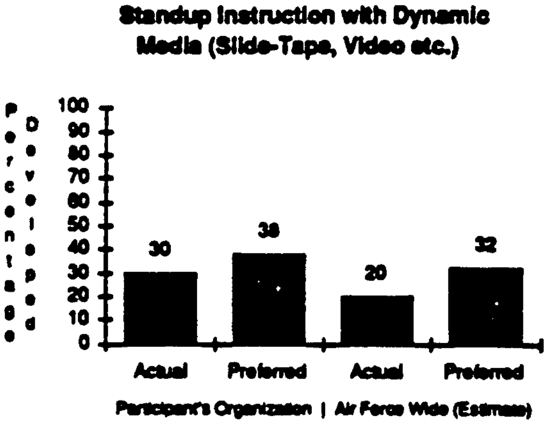


Figure 3

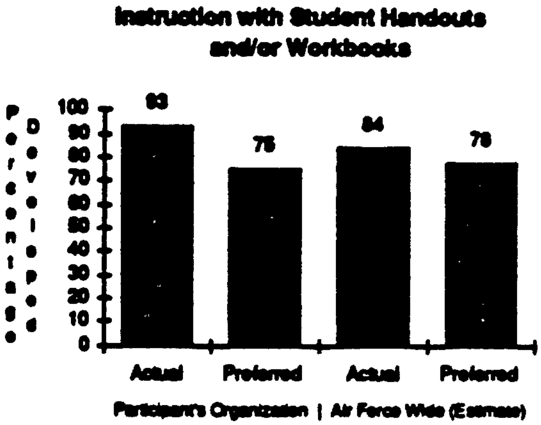


Figure 4

While the percentage of instruction presently developed for computer-based delivery is low (18 percent within our panelist's organizations and less than 10 percent estimated Air Force-wide), most participants would like to see CBI use increased greatly (Figure 5). Although one panelist charged exclusively with CBI development feels that CBI should only account for 10 percent of all instruction, other participants felt the amount of CBI desired should be nearly one-third of Air Force-wide instruction.

Every respondent wants to see more multimedia CBI developed for Air Force instruction (Figure 6). At present, one-third of all CBI being developed within participants' organizations consists of dynamic multimedia, along with an estimated 25 percent of such CBI Air Force-wide. But panelists believe that the amount should be pushed above 50 percent, that is, more than half of all computer-based instruction should be multimedia. This suggests a clear need for interactive dynamic multimedia. However, as participant comments allude, many of the tools required to develop motivational multimedia are presently unavailable in the field.

Air Force IDD's also indicate an interest in developing more instruction for distance learning applications (Figure 7). By their estimate a scant two percent of Air Force instruction now constitutes distance learning. However, they believe as much as

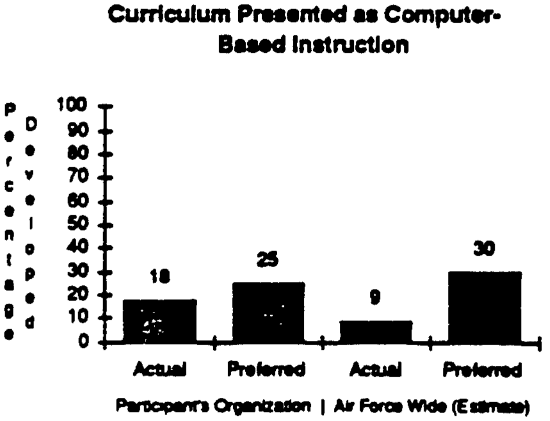


Figure 5

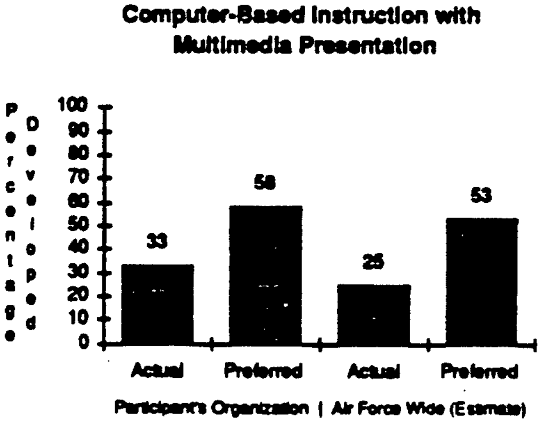


Figure 6

27 percent of their instruction has distance learning applications, and see a potential for 22 percent of total Air Force instruction to be delivered remotely.

Finally, panelists feel that too great an emphasis is now placed on objective examinations. They show a clear preference for developing performance-based evaluations both within their own organizations and Air Force-wide (Figure 8). The data suggests most of the experts want performance-based evaluations to supplement rather than totally replace objective exams. Given the nature of many of the tasks for which Air Force IDD's develop instructional systems, it seems likely that objective measures may often be insufficient. Many skills-based tasks can only be effectively evaluated by proficient performance. Objective exams are generally easier to develop, administer and evaluate than performance-based tests, suggesting that IDD's may benefit from tools that help them develop more innovative evaluation measures.

In summary, the experienced Delphi panel would like to develop less standup instruction that is unsupported by media or has only static media. Although they would like to reduce their reliance on traditional text-oriented student handouts and workbooks, they still want print support for three-fourths of their instruction. They would like to increase the use of dynamic media as a support for standup instruction and they would like to increase CBI and in particular the use of CBI that includes multimedia presentation. They would also like to increase the amount of courses offered through distance learning systems and they would like to implement more performance-based evaluations.

MEDIA PRODUCTION METHODS

Most of our Delphi panelists (seven of nine responding) contract the final development of graphics. Reasons cited for this range from not violating a base-negotiated contract by developing graphics in-house, to a respondent who cites excessive time and effort spent to produce non-professional looking graphics. Only two of nine IDD's report they develop their own final graphics. Two report they are starting to develop more internally. One of those cited slow turnaround by contractors.

All respondents' organizations contract for printing services, though six of nine provide camera-ready copy, indicating the penetration of desktop publishing. One panelist reports that their organization is beginning to desktop publish and hope to soon provide camera-ready copy; one panelist reports that contractor turnaround is, "not very fast". Two panelists report that printing services are handled by contractors, but do not specify who provides camera-ready copy.

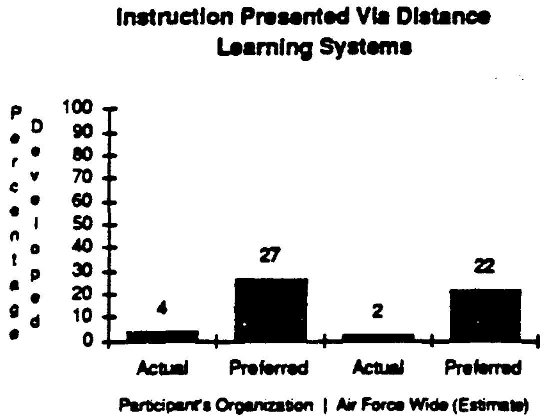


Figure 7

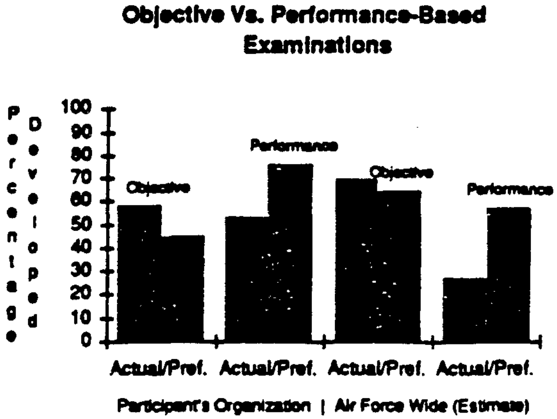


Figure 8

Five of seven respondents use on-base contractors for all photography. Two of seven provide their own photos. Similarly, five of nine respondents use on-base contractors for all video footage, while two of nine produce their own internally. Of these two respondents, one is tasked with CBI development, the other is a manager whose organization develops primarily textual media and team training. Two of nine share the task of video development with contractors. Three of seven respondents contract for slide/tape program production, two produce in-house and two report no slide/tape productions. Three of eight panelists report audio production is provided by on-base contract while two develop in-house. Three panelists state audio production services are not available (even though computer-based audio production tools are highly developed and inexpensive).

ESSENTIAL AND DESIRED DEVELOPMENT TOOLS

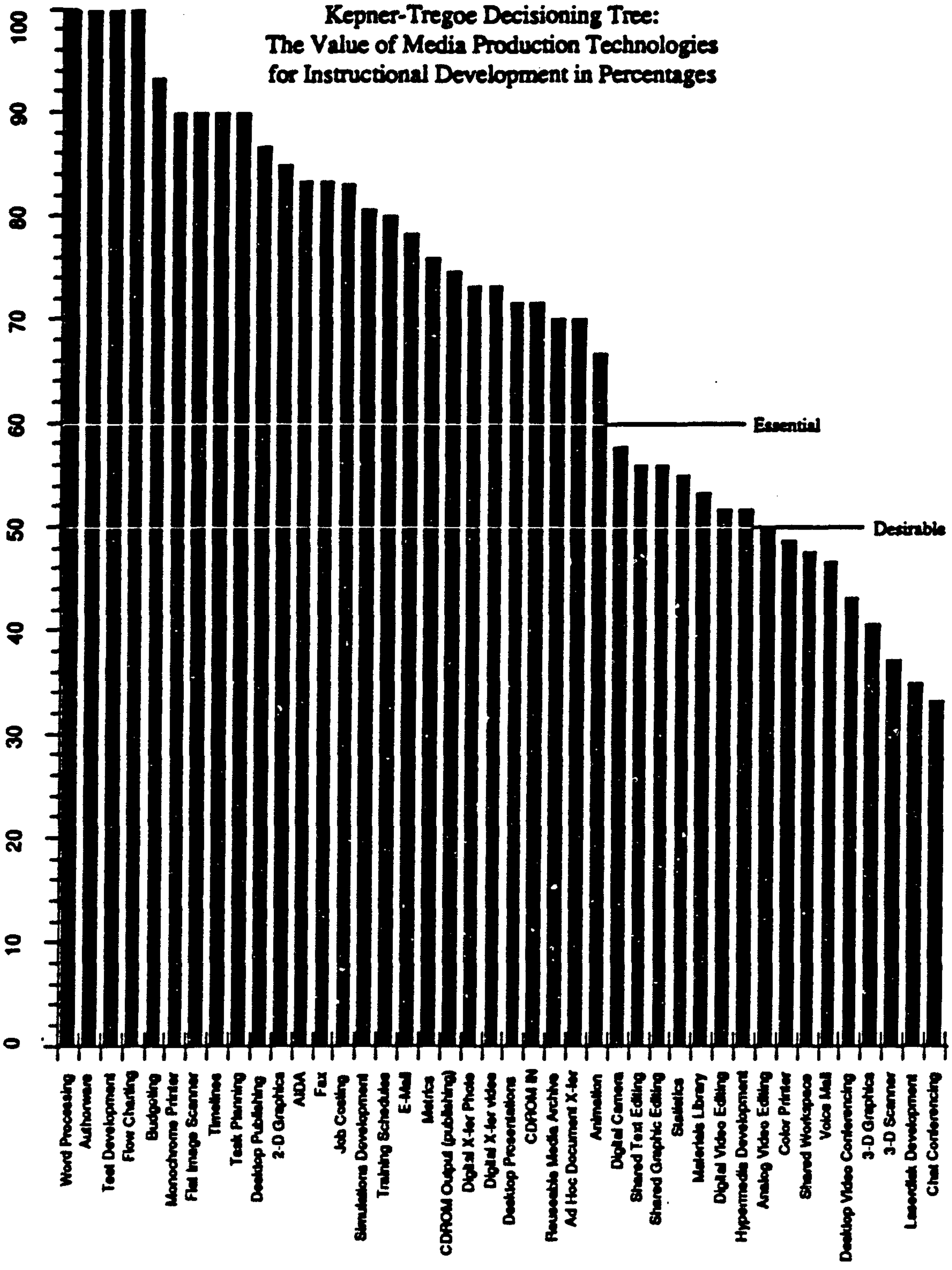
Respondents were asked to determine with a yes or no vote whether a variety of computer-based media development tools were essential to ISD. Where they voted no, they were asked to determine the usefulness of the tool for ISD on a scale of 0-20. Each "yes" vote is valued at 30 points, while each "no" vote is valued at its given weight. These data are totaled, divided by 30 and used to create the Kepner-Tregoe decision tree shown in Figure 9. Tools that receive 60 percent of possible points (180 points of the 300 points possible) are considered to be essential to ISD and, therefore, to the CIDE workstation; those totaling 50 percent (or 150 points) are deemed highly desirable; others below 50 percent are considered useful in proportion to their weights. Note that no tool received a weight below 30 percent (90 points), and at least two panelists considered any given tool absolutely essential (yes votes). Thus, all of these tools should probably be available as add-on features to the workstation to support the task needs of particular designers.

Clearly Air Force instructional designers and developers would like to be producing more dynamic and motivational media for standup instruction, computer-based instruction and for distance learning applications. A variety of reasons for the present lack of media are suggested in the panelists' comments on their organizations' current arrangements for final media production.

One respondent states that the existing base-negotiated graphics contract legally prevents them from developing graphics in-house. One is simply lacking sufficient equipment. Three others cite training and poor final quality of in-house work due to "seldom used but technically difficult skills". Four of the respondents bemoan long turnaround time for most contracted media, while three others cite low quality in contractor-developed media due to poor communication or insufficient familiarity with the subject matter. Despite these problems, all parties indicated the need to use more dynamic and motivational instructional media.

The literature cited earlier suggests that many of these problems could be remedied by the installation of a collaborative network. Creativity and expertise hurdles can be surmounted by special interest groups and just-in-time technical support. Communications problems with contractors can be circumvented with more timely collaborative sessions. With the appropriate tools, more preparatory development work can take place in-house even if prior agreements stipulate that contractors must produce final media. And with the right tools, designers will have the freedom to explore more creative, dynamic and motivational media solutions to instructional problems.

From the Kepner-Tregoe decisioning tree it is fairly obvious that IDD's themselves recognize this. Of course there was unanimous agreement for word processors, authorware, flowcharts and test development tools— the staples of the trade. But not



Kepner-Tregoe Decisioning Tree:
The Value of Media Production Technologies
for Instructional Development in Percentages
100
90
80
70
60
50
40
30
20
10
0
Essential
Desirable
Word Processing
Authorware
Text Development
Flow Charting
Budgeting
Monochrome Printer
Flat Image Scanner
Timelines
Task Planning
Desktop Publishing
2-D Graphics
AIDA
Fax
Job Costing
Simulations Development
Training Schedules
E-Mail
Metrics
CDROM Output (publishing)
Digital X-fer Photo
Digital X-fer video
Desktop Presentations
CDROM IN
Reuseable Media Archive
Ad Hoc Document X-fer
Animation
Digital Camera
Shared Text Editing
Shared Graphic Editing
Statistics
Materials Library
Digital Video Editing
Hypermedia Development
Analog Video Editing
Color Printer
Shared Workspace
Voice Mail
Desktop Video Conferencing
3-D Graphics
3-D Scanner
Laserdisk Development
Chat Conferencing

Figure 9

surprisingly, there was extremely strong interest in desktop publishing, two-dimensional graphics, image scanners, simulations development, digital and analog video editing, video and photography transfer, and CDROM input and publishing. These are tools to create a media-rich instructional environment. They are not simple to use, requiring technical proficiency and creativity, but IDD's understand their value to the development of highly motivational, dynamic media.

The data indicate, as well, that they understand the value of collaboration with peers and specialists to quality ISD. Every respondent reported routine collaboration with a subject matter expert. Four participants agreed with the panelist who stated, "[it is]...impossible to develop any kind of quality training without an SME". In addition, four of nine cite frequent collaboration with other IDD's for ideas, and evaluation; five of nine report collaboration with graphics specialists; and three of nine report contact with technical experts and media specialists: "without this link, our product would not get out".

Presently, the bulk of this collaboration is conducted face-to-face or over the telephone. Where media production and graphic design specialists are concerned, considerably more work is conducted face-to-face than by any other method. Yet, most of these experts cite time and communication factors as primary impediments to more extensive use of motivational media. Clearly a well-implemented collaborative network that connects IDD's, SME's and media production specialists—including contractors—would solve most of the aforementioned problems, streamline collaborative processes by eliminating many face-to-face meetings, and encourage the creative development of more effective, dynamic and motivational media.

Finally, the experience of the Navy cited by Cantor (1988) suggests that centralized archival of reusable, adaptable media would save time and money, while encouraging IDD's to make use of the best stock media available. This type of networked archive would make media produced by the most highly skilled and talented producers readily available to IDD's throughout the Air Force's widely distributed ISD agencies.

CAN A COLLABORATIVE INSTRUCTIONAL DESIGN ENVIRONMENT BE CREATED WITH CURRENT COMMERCIAL-OFF-THE-SHELF TECHNOLOGY?

The answer to this questions is a qualified yes. Every item on the list of 42 functions identified as essential or desirable for at least some IDD's are available right now. The open architecture and communication networking capabilities are present. What cannot be answered by this study is what integration software, degree of data interchange standards and communication data speeds are necessary and available for implementation of every function. To answer this question definitively will require additional study, prototype development and beta testing which is strongly recommended. A beta test would create a field laboratory environment that could be useful in answering a variety of research questions regarding process and task procedures for optimal use of the collaborative instructional development environment. The constraining factors for such a system are not hardware and software.

SUMMARY

As indicated throughout the study, the mere provision of development tools to IDD's does not empower them individually to design efficiently or effectively. Instructional Systems Design (ISD) is a complex task requiring variously the input of IDD's, SME's, graphic designers, video production specialists, technical writers and CBI programmers. While increased attention to the user interface has created a generation of media production software much more friendly to the average user, true mastery of any such tool is the result of both practice with the software and a thorough understanding of the knowledge domain to be represented and the traditions and techniques native to the media being developed. The proliferation of desktop publishing provides an illustrative example. While user-friendly programs like PageMaker and Ventura Publisher brought computer-based publishing to everyone, they did not communicate the traditions of the typographer's art. As a result, the average quality of typeset materials now varies a great deal. For expert typesetters, the software tools were a productivity boon, allowing them to create high-quality typeset pages more quickly and less expensively than ever before. However, in the hands of the non-typesetter, they allowed only the quick production of readable but inelegant pages devoid of ligatures, gendered quotes, and properly kerned letter pairs. To quote Bob Krejci, "There is no one-person authoring tool that can produce the kind of product that an experienced staff of designers, subject matter experts, artists, and programmers can develop... There is no 'Van Gogh in a spray can' product" (1992).

This is not to suggest that there is not merit in providing IDD's with the wide variety of tools required to develop affective instructional media. On the contrary, while questions remain as to the amount and design of potent media for various instructional objectives and within specific domains of knowledge, there is no doubt that visual and auditory media can be an appropriate and highly effective means of organizing and presenting some information (Gildea, Miller and Wurtenberg, 1990). Analysis of specific appropriate forms and applications of graphics have been begun with respect to the development of an Automated Instructional Design Advisor (Friedman, et. al., 1991). While an AIDA may well include guidance towards appropriate applications and designs for graphic instructional media, the goal of a development workstation should be to provide robust fully-featured design tools. Whether a screw or nail is appropriate is the decision of the carpenter and architect. A good toolbox contains both driver and hammer. For example, the graphic design and imaging segments of the PC market have matured sufficiently that there are a variety of extremely competent software packages presently available for the development of fine art (e.g., Fractal Design Painter), line art (e.g., Corel Draw, Adobe Illustrator) CAD (e.g., AutoCad, Eazy Cad), solid model rendering (e.g., AutoCad, Renderman) and photo image processing (e.g., Image-In, Photo Finish). While the Air Force may choose to standardize on one or more of these packages, the CIDE workstation will be designed to accommodate one or all of them in ad hoc arrangements to support the task at hand. However, any such integration of new tools requires an equal commitment to technical and creative support services to aid in the transfer of sufficient expertise to make the tools useful. The creation of ad hoc collaborative work groups can address the problem of providing technical development assistance to less experienced IDD's and those without sufficient expertise in development of specific types of media. Simultaneously, collaboration should improve worker efficiency and job satisfaction as well as creating a community of expertise and a professional growth environment for the motivated IDD.

The superimposition of a collaborative open network on the instructional development environment will permit ad hoc working arrangements for mentor/apprentice relationships, creative and technical consulting support and multiple problem solving perspectives for individual IDD's as well as bringing other benefits to improve both ISD and IDD performance. A collaborative open network among instructional developers may include technologies as commonplace as fax, e-mail and voice mail or those as esoteric as ISDN-based digital document transfer and two-way video conferencing. Fundamentally, it constitutes the creation of open communication channels that engender the formation of ad hoc work groups and technical interest forums, allow the transfer of documents and resource materials, and provide an easily accessed, non-threatening means to seek technical help and creative assistance.

Building an open network will allow both synchronous (live chat forums or video conferences) and asynchronous (e-mail, fax) communications amongst IDD's. But perhaps just as valuable, it will enable document transfer for evaluation purposes, scheduling and coordination, and sharing of valuable resources such as adaptable existing media and a centralized archive for some materials. Perez (1992) in an exploration of traits of the expert training developer discovered that senior designers developing instruction through a team approach, "...developed formal conventions and guidelines to insure the uniform execution of the instructional design" (p. 13). CMC can support this kind of control and coordination as well as increasing efficiency in the group design process. Cantor (1988) reported that an automated curriculum design environment developed for the Navy that included archiving of boilerplate text and graphics as a shared resource reduced time spent on repetitive work by allowing incorporation and adaptation of existing materials. He cited an aggregate reduction in ISD time-on-task from between 45 and 66 percent.

The study by Walsh et. al. (1992) of Air Force CBT developers reinforces the value a collaborative instructional development environment could provide. They found 78 percent of CBI development team members were inexperienced in CBI design and development. More than one in four of the development team members felt the team's activities were not well coordinated and that communication between team members was unclear and ineffective.

A formal job/task analysis was not performed in one of every three CBI development efforts. About one in four CBI instructional developers relied on previous course materials or analyses. Forty percent used learning objectives modified from previous lessons. Only one-half of developers indicated a media analysis was performed as part of the CBI design process, and when an analysis was performed a subject matter expert was used over two-thirds of the time. Of the 253 CBI designers surveyed, not one indicated a media specialist was involved in the media analysis process. Of development team members, just over ten percent were described as media experts and they were all graphic artists. For the CBI projects, 95 percent contained some graphic components (icons, charts, tables, diagrams, maps, equipment, human figures, even animation); 48 percent had audio content (bells, beeps, tunes as rewards, signals, music, engines and verbal commands, questions, etc.); and 45 percent included some still or motion video (for identification of equipment, body parts, panels, etc., procedures and interpersonal and communication skills). The most commonly cited reasons for not using multimedia were lack of capability, not enough time and not being trained for development. There was no mention at all of interactive media applications.

Walsh's survey of practicing computer-based instructional developers strongly indicates that Air Force CBI development could benefit substantially from the availability of a collaborative instructional development environment as outlined in this study. The data from the Delphi group indicates non-CBI instructional development needs a similar capability.

Providing computer-based media development tools in a collaborative environment should streamline many work processes and stimulate interactive evaluation of instructional components. Tessmer and Wedman (1992) discovered that the most common reason cited by professional IDD's for skipping an ISD activity was not lack of money or experience but lack of time. They state, "A means of 'cutting corners while controlling risk' in ID/D [ISD] project[s] needs to be developed" (p. 16). We believe the CIDE can assist in cutting those corners.

CONCLUSIONS

The Air Force should explore implementation of a collaborative instructional development environment workstation. All the technologies involved are readily available in COTS packages. Many of the individual media development technologies such as flat bed scanning, graphic design, desktop publishing, authorware and laser printing are already in place in some ISD organizations. Issues to be resolved include:

1) the bandwidth and best method for ad hoc networking and multi-network management (e.g., simple ethernet, ISDN, FDDI, ATM, SNMP, etc.);
2) whether members of the collaborative community can benefit from broadband communications technologies such as desktop video teleconferencing, voice mail and groupware (shared graphic and text editing);
3) what kinds of materials should be centrally stored in a Multimedia Education and Training Archive (META) and whether those archives should be maintained within each ISD organization or by on-base visual information (VI) agencies such as Combat Camera, or both.
4) strategies for perusing, indexing and previewing contents of the META, including Boolean search techniques and indexed multimedia information retrieval.
5) determining any potential adverse consequences of installing a CMC network and developing strategies to offset them.

With the development of AIDA, the firm adoption of a standardized computer platform, and the maturation of computer-based media production tools ranging from graphic design to desktop digital video editing, the time is ripe for implementation of an instructional development environment with standard tools, central archiving, and a collaborative network for exchange of creative and technical support, reusable media, and formative evaluation. Changes to organizational culture, while predictable, will remain minor and manageable. In reality they will most likely contribute to increased employee satisfaction and organizational loyalty. Without question, such a system will contribute to the development of more effective, motivational instruction.

Bibliography

Bassin, M. (1988). Teamwork at General Foods: New and improved. *Personnel Journal*, 67, 5 62-70.

Blaye, A., Light, P., Joiner, R. and Sheldon, S. (1991). Collaboration as a facilitator of planning and problem solving on a computer-based task. *British Journal of Developmental Psychology*, 9, 471-483.

Boar, B. H. (1984). *Application Prototyping*. New York: John Wiley & Sons, Inc.

Cantor, J. A. (1988). An automated curriculum development process for Navy technical training. *Journal of Instructional Development*, 11, 4, 3-11.

Chu, G. C., and Schramm, W. (1979). *Learning from Television— What the Research Says, 4th Ed.* (first edition released in 1968), Washington, D.C.: National Association of Educational Broadcasters.

Communications Week (1992). Intel board to vote on IP fix. *Communications Week*, #410, July 6, 1992, p. 1.

Conner, D (1985). *Information system specification and design road map.*Englewood Cliffs, NJ, Prentice-Hall.

Dubrovsky, V.J., Kiesler, S. and Sethna, B.N. (1991) The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups. *Human-Computer Interaction*, 6, 119-146.

Finholt, T., Sproull, L., and Kiesler, S. (1990) Communication and performance in ad hoc task groups. In, Galegher, J., Kraut, R. E., & Egido, C (Eds), *Intellectual Teamwork, Social and Technological Foundations of Cooperative Work.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Friedman, A., Polson, M. C., and Spector, J. M. (1991). Designing an advanced instructional design advisor: Incorporating visual materials and other research issues. USAF Document # AL-TP-1991-0017-Vol-4 HRTC, Armstrong Laboratories, Brooks, AFB, TX.

Gildea, P. M., Miller, G. A. and Wurtenberg, C. L. (1990). Contextual enrichment by videodisc. In: Cognition, Education, Multimedia, Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.

Hickey, A. E., Spector, J. M. and Muraida, D. J. (1992). Design specifications for an Advanced Instructional Design Advisor (AIDA). USAF Document AL-TR-1991-0085-Vol-2 HRTC, Armstrong Laboratories, Brooks, AFB, TX.

Hiltz, S.R. and Turoff, M. (1978). *The Network Nation: Human Communication via Computer*. Reading, MA: Addison-Wesley.

Johnson, D. W., Maruyama, G., Johnson, R. T., Nelson, D. and Skon, S. (1981). 'Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis', *Psychological Bulletin*, 89, 47-62.

Johnston, J. (1987). *Electronic Learning from Audiotape to Videodisc*. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Katz, H. C., Kochan, T. A. and Weber, M. R. (1985). 'Assessing the effects of industrial relations systems and efforts to improve the quality of working life on organizational effectiveness', *Academy of Management Journal*, 28, 509-526.

Kepner, C. H. and Tregoe, B. B. (1965). *The Rational Manager*. New York, McGraw Hill.

Kling, R. (1991). Cooperation, coordindation and control in computer-supported work. *Communications of the ACM*, 34, 12, 83-88.

Krejci, B. (1992) Response to a letter to the editor. *Instruction Delivery Systems*, 6, 4, p.5.

Kyng, M. (1991). Designing for cooperation: Cooperating in design. *Communications of the ACM*, 34, 12, 65-75.

Lea, M. and Spears, R. (1991). Computer-mediated communication, de-individuation and group decision-making. *International Journal of Man-Machine Studies*, 34, 283-301.

Linde, C. (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. *Language and Society*, 17, 375-399.

Main, R. (1992). Integrating the affective domain into the instructional design process. Brooks A. F. B., Texas, Armstrong Laboratories Report AL-TP-1992-0004, HRTC, Armstrong Laboratories, Brooks, AFB, TX.

Monge, P. R. and Kirste, K.K. (1980). Measuring proximity in human organization. *Social Psychology Quarterly*, 43, 1, 110-115.

Perez, R. S. (In Press). Modelling the expert training developer. In Advanced Training Technologies Applied to Training Design. (Eds.) R.J. Seidel & P. Chatelier, Plenum Press.

Perin, C. (1991). Electronic Social Fields in Bureaucracies. *Communications of the ACM*, 34, 12, 75-82.

Porter, D. B., Bird, M. E., & Wunder, A. (1990). Competition, cooperation, satisfaction and the performance of complex tasks among Air Force cadets. *Current Psychology: Research & Reviews*, 9, 4, 347-354.

Ridgeway, C.L., Berger, J. and Smith, L., (1985). Nonverbal cues and status: An expectation states approach. *American Journal of Sociology*, 90, 5, 955-979.

Salomon, G. (1985). Information technologies: What you see is not (always) what you get. *Educational Psychologist*, 20, 4, 207-216.

Smilowitz, M., Compton, D. C., & Flint, L. (1989). The effects of Computer Mediated Communication on an individual's judgment: A study based on the methods of Asch's social influence experiment. *Computers In Human Behavior*, 4, 311-321.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science*, 32, 1492-1512.

Tersine, R. J., and Riggs, W. E. (1976). The Delphi technique: A long-range planning tool. *Business Horizons*, April, 51-56.

Tessmer, M., & Wedman, J. (1992). The practice of instructional design: a survey of what designers do, don't do and why they don't do it. San Francisco, CA: paper presented at the annual meeting of the American Educational Research Association.

Tjosvold, D. & Tsao, Y. (1989). Productive organizational collaboration: The role of values and cooperation. *Journal of Organizational Behavior*, 10, 189-195.

Tjosvold, D., Andrews I. R. and Jones, H. (1983). 'Cooperative and competitive relationships between superiors and subordinates', *Human Relations*, 36, 1111-1124.

Walsh, W. J., Yee, P. J., Grozier, S. A., Gibson, E. G., and Young, S. A. (1992). *A survey of Air Force computer-based training (CBT) planning, selection, and implementation issues*. AL-TP-1991-0059, HRTC, Armstrong Laboratories, Brooks, AFB, TX.

Winn, W. (1987). Instructional design and intelligent systems: Shifts in the designer's decision-making role. *Instructional Science*, 16,59-77.

ISDN: BACKBONE FOR AN INSTRUCTIONAL DEVELOPMENT, ARCHIVAL AND DELIVERY ENVIRONMENT

Andrew S. Wilson
California State University Chico
Chico, California

California State University, in partnership with Pacific Bell and AT&T, has undertaken an ongoing evaluation of the effectiveness of Integrated Services Digital Network (ISDN) technology as a delivery channel for multimedia distance education and collaborative work group support.

ISDN comprises a method of dividing existing twisted pair copper wire in the public switched network to create substantial bandwidth for transmission of digital data. A standard telephone provisioned for Basic-rate ISDN is divided into three channels: two 64K-bits per second (bps) bearer channels (B channels) and a 16Kbps data channel (D channel) used for packet mode oriented signalling. The resulting 144Kbps data pipeline can be used to carry voice, data, and compressed video. Broadband ISDN (B-ISDN) is based in a T1 line and provides 23 B channels and one 64Kbps D channel for total bandwidth of more than 1.5Mbps.

The potential of ISDN for distance education resides in its relatively low cost and flexibility. A single basic rate line costs only twice as much as a standard telephone line, yet it can move 15 times as much data as a normal line with a 9600 baud modem. Moreover, because it uses the public switched network, ISDN distance learning networks can be configured in ad hoc relationships on-demand with the ease of placing a phone call. Receive sites can be added to the instructional network by simply ordering a new phone line, and all the equipment required is commercial-off-the-shelf technology.

In the first two phases of CSUCs multimedia distance learning trials, two basic rate ISDN lines were used. One provided connectivity between computer-based platforms at the delivery and receive sites and permitted private voice calls. The second line was used to provide two-way real-time interactive video and audio using two B channels (112Kbps). The results suggest that while two-way compressed video at 112Kbps was of sufficient quality to recognize intensity available. While highly compressed video is adequate to recognize a speaker, it does not carry sufficient information to permit expression recognition or clear transmission of board work or real-time video media with which the students are not already familiar.

One potential solution to this problem would be to employ more basic rate ISDN lines of a single broadband primary rate ISDN line to increase available bandwidth for two-way video transmission significantly to 384kbps or above. Multiband multiplexers can combine 56Kbps, 64Kbps and 384Kbps channels in ad hoc configurations for bandwidth-on-demand ranging from 56Kbps to 2.04Mbps. Higher bandwidths can transmit broadcast quality two-way video. Though this is significantly more expensive than using two basic rate channels, it is considerably cheaper and more flexible than traditional broadcast technologies.

Another option to be explored is transmission of digital rather than analog video. Digital video can be compressed more readily and with less total breakdown or "mosaicing" of the image since it does not have to pass through analog-to-digital and digital-to-analog converters. In addition, some digital video compression schemes such as Intel's DVI may be transmitted without significant information loss at data rates as low as 156Kbps, making digital transmission of full-motion two-way video and full-duplex audio over basic-rate ISDN feasible.

The forthcoming implementation of National ISDN One will integrate the isolated pockets of ISDN service into a seamless network with the potential to carry multimedia data in ad hoc configurations at a relatively low cost. this technology may have implications for entertainment, file transfer, shopping and information services. It certainly has broad potential for both educational delivery and cooperative work group support.

MOTIVATIONAL APPEAL IN PAPER-BASED CORRESPONDENCE COURSES

REID J. MATHERNE
Extension Course Institute
Gunter AFB, Alabama

Texts of distance education or correspondence courses are instructional texts. Their purpose is not only to communicate knowledge in an effective way, but to lead the student to an adequate understanding of the knowledge presented through a structured study process. The design of distance texts, therefore, is an instructional design process.

The U.S. Air Force has been publishing paper-based correspondence courses or texts for 41 years. I suppose it is a fair question to ask why the Air Force uses such a medium to train its personnel. Well, it is obvious that distance study programs are designed to be replicable over large numbers of individuals and are usually print-based. Consequently, there are economies of scale which tend to yield cost and efficiency benefits. Portability is also an obvious advantage. Students can carry their printed courses on the flight line, in the field, and just about anywhere; furthermore, an appointed time to study is not required. However, the most probable reason the Air Force selects print-based distance learning programs to train airmen relates to time, space, and budgetary constraints. The genius of distance study is that students can stay where they are, perform their jobs, and still attain Air Force educational requirements peculiar to their career fields.

The Air Force considers instructional system design, good writing, and properly developed textual material as basic ingredients for any text. Because of the distance aspect of correspondence courses, where students enrolled do not have the advantage of a classroom instructor who explains, demonstrates, and illustrates, inherent teaching techniques created through motivational appeal become equally important. Drawing from research this paper addresses some of those techniques employed by the U.S. Air Force Extension Course Institute (ECI).

A format for motivational appeal in a text is the attitudinal structure of the book. It contains those friendly, verbal gestures (Moore, 1985) from the affective domain that unlock the doors of cognition in distance learners. It charts a path for learning; it is user friendly.

When a distance learner first picks up a book, the visual message on the cover should be appealing. The cover design should have enough eye appeal to stimulate the student to open the text and begin to work. Thematic concepts that will occur throughout a text may first appear on the cover. For example, in figure 1 the airplane on the cover of an ECI text is a subtle thematic cue that connotes the ideas of 'motion' and 'moving ahead,' always to the right.

Another motivational feature a good correspondence text should contain is a menu page. Some people call it a table of contents, but it really is more than that. Figure 2 shows that like a computer menu, the menu page can be used by the student to select any unit for further examination. In this case, a student who wants to know more about the Digestive System in Unit 1 just turns to page 1-25. The unit menu tells the student that the unit has ten major sections containing twenty six lessons. By reading the menu, the student can get a good idea of the scope of the text and how each unit fits into the gestalt of the text (Hesse, 1985). Furthermore, the menu should help the student find material without reading the entire text.

Along with the menu page, ECI considers clear and concise directions which leave no questions at all in the student's mind an equally important motivational feature. Notice the directions in figure 3. Along with the text Preface is a note with a brief, simple set of instructions.

Still another feature of formatting for appeal is that of controlling the overall appearance of the printed text by judicious use of typography, white space, and figure placement. In other words, the concern here is the designer's efforts to achieve a balance of visual imagery to print media (Adams and Fuchs, 1985). Striking the right balance in line length, type size, and interline spacing ensures reader comfort. It also affects the readibility of the material. It ultimately affects the success of the book, because line length, type size, and spacing determine the cost of the number of pages. ECI attempts to be careful not to sacrifice comfort and readibility for less money. If a book is a success with students, money is saved in the long run.

There has to be logic used when making decisions about the columns of text material, another important motivational feature. Column is just another way of saying line length. In this regard, ECI takes into consideration its audience.

Columns are our foundation or our structural skeleton. The basic one-, two-, or three columns are the traditional arrangement, tested over time and found to be the most efficient in terms of space utilization (allowing the minimum for left and right margins). For the most part ECI uses the two-column page with the one-column introduction at the beginning of its units (figure 4).

Unit introductions are somewhat larger type, but do not cover the complete two column width. Psychologically, this is good (for us), because it sets the introduction apart and says to the student, "This is only an introduction; read only to see what is going to happen." It is a warmup and establishes comfort. Introductions provide the "mind set" for what is to come and to establish comfort with the material. At ECI, introductory material is not tested in either our unit review exercises or in our proctored course examinations.

Text outside margins need not be more than the standard 1/2 inch if using space economically. The inside margin (or gutter margin) varies. The inside margin should be determined by how the book is fastened together. A "bound" book requires at least a one-inch margin so the book does not have to be broken in order to reveal type. A book that will be placed in a ring binder requires a 5/8 inch margin. If a book is printed in multi-page forms, collated, folded, trimmed and saddle-stitched (magazine-fashion), the gutter margin can be less than 1/2 inch; it is kinder on the reader, however, to make it at least 3/4 inch.

Margins determine the width of the "live matter area" -- the printable part of the page that we subdivide into columns. The precise width of the live matter area allows us to devise the geometry of the page. ECI approaches the width of the live matter in terms of what is the best design structure for our purpose. Let's assume our live matter area is 41 picas (pica being 10 spaces per inch). 41 is a magical number in that it is easily divisible into several column arrangements. The 41 pica-wide page can actually be broken into at least five differing arrangements. To reiterate, ECI uses primarily the two column arrangement of 20 picas with one 1-pica gutter. However, we are presently experimenting with a one column arrangement of 41 picas.

The geometry of column width merely determines the size of spaces to be filled with type. Of course, it is the type that matters, not the space into which it is poured. ECI's goal is to proportion the type so subtly that readers will read their texts with such facility that the fact that they are working is not noticeable.

Another very important motivational feature in ECI courses is type size. The most commonly used type sizes run from 6 to 12 point (Hamilton, 1980). While a standard 10-point size medium-weight serif type of font is used for texts at ECI (figure 5), other sizes, weights, and type styles are used in headings. The numbered section heading is displayed in a bold sans-serif letter with initial caps that command attention. Adequate white space is used again to separate elements of the text. The learning statement then follows in a lowercase demi-bold sans-serif font diminished in size from the section head. Also, there is further subordination in the text flagged by a boldfaced variation of the basic 10-point text font.

In the use of white space, the leading or interline spacing must also be considered. A proper balance of leading to typography can enhance readability and comprehension. Minimum leading causes the type to appear smaller (Jones & Taylor, 1985).

Another important motivational feature which affects "the look" of a text is the manner in which it is deployed. Deployment here refers to how text is displayed through paragraphing, tables, lists, etc. Text deployment facilitates learning in three ways: by identifying the type of information available; by identifying how it is to be used; and, by providing a visual model of what is to be done and allowing the 'earner to process the material visually (Gropper, 1988).

A numbered list signifies a sequence of steps, while a bulleted list signifies a series of discrete facts. One typeface may signify definitions, another examples. Matrices may signify the association of facts and their referents.

Some deployments also create a look that can announce how its content is to be processed. Paragraphs announce that a conventional reading sequence will be sufficient (left to right and top to bottom). Lists announce that its items are to be taken one at a time and in order. Adjacent material announces "read in this direction" or "compare these two items: or 'if this--do that."

Text deployments also provide a visual model. The listing of procedural steps, if-then displays, or adjacent items to be distinguished not only invite the practice of verbal responses. By visually modeling the relevant performance, they invite processing in a visual mode. Learners see elements of the performance to be learned laid out in ways that stimulate the performance required on their jobs. It invites visual processing of that performance. Prompted by the look of text deployments learners may be prompted to visualize: units of information; itemization; divisions between units of information; links between them; sequence to be observed; or routes to follow. In its text development process ECI is challenged daily with finding text deployments that elicit all three of the ways of facilitating learning, matching them to conditions for which they might be relevant, and stepping back to assess the benefits.

ECI judiciously uses nonverbal as well as verbal cues as an effective means to aid students in identifying, organizing, and interpreting the most important content in a text. Nonverbal cues include: highlighting (italics, boldface, box, etc.); underlining; asterisks; parenthesis; brackets; exclamation points; and white space (horizontal and vertical). Verbal cues include: short signal phrases; adjunct questions; instructional objectives or topical statements; advance organizers; outlines; headings; and margin notes. Again, all of these cues are used sparingly; overuse diminishes the importance of the really important.

The last motivational feature I want to address is the use of graphics in distance education texts. ECI attempts to place figures conveniently for reference where the student is introduced to the illustrated concept in the text to ensure the educational adequacy of the page (Figure 6). Proper placement also demands that the figure be located on the page to create a pleasing visual balance and not to distract from the text (Figure 7). Again, ECI practices restraint in its use of graphics. The ready availability of sophisticated graphic tools can easily overwhelm many users. Some pitfalls ECI attempts to avoid include: using too many boxes; using too many fancy borders; using photos that are not cropped; using erratic caption placement; using too much space between visuals; and overprinting text and graphics.

In summary, Extension Course Institute's philosophy is that the format of correspondence course texts should complement its instructional design to the student's interests, attitudes, and aesthetics. Basically, texts should incorporate design features that include student direction; that make appropriate use of typography, white space and leading; and that balance page appearance with figure placement. Such texts should have appealing covers and use menu pages. The overall visual appeal of the text should be motivational and stimulating.

References:

Adams, D.M. & Fuchs, M. New digitized literacies: mixing visual media, the humanities, print, and computer-based technology, Educational Technology, 1985, 25 (5), 16-18.

Glynn, S.M., Britton, B.K., & Tillman, M.H. Typographical Cues in Text: Management of the Reader's Attention, The Technology of Text II, 1985, Educational Technology Publications, Inc., Englewood Cliffs, N.J., 8, 192-207.

Gropper, G.L. How Text Displays Add Value to Text Content, Educational Technology, 1988, 28 (4), 15-21.

Hamilton, R.B. Course design and layout. In M. Lambert and S. Welch (Eds.) Home Study Course Development Handbook. Washington: National Home Study Council, 1980.

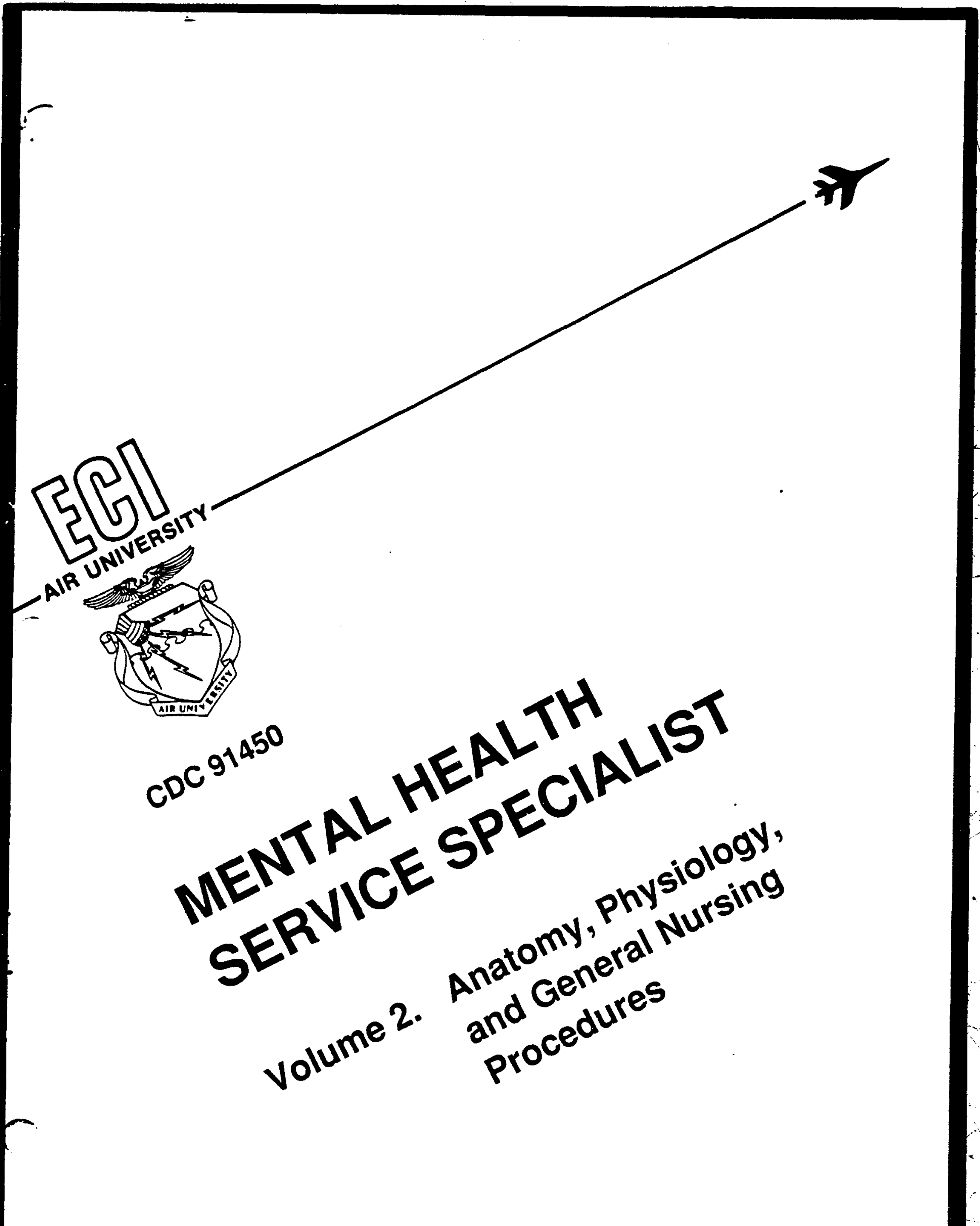
Hesse, C.W. A format proposal. Paper presented at Interservice Correspondence Exchange conference, Oklahoma City, November 14, 1985.

Jones, B.S. & Taylor, R.B. Talking paper on proposed new CDC format. Extension Course Institute. Gunter AFB, AL. 1985.

Merrill, P.F. & Bunderson, C.V. Preliminary Guidelines for Employing Graphics in Instruction, Journal of Instructional Development, 1981, Vol. 4, No. 4.

Moore, M. Adult learning at a distance. Paper presented at Effective Teaching at a Distance Conference, University of Wisconsin, Madison, August 5, 1985.

White, J.V. Good design: a balancing act, Art World, 1989, January/February, 22-28.



ECI

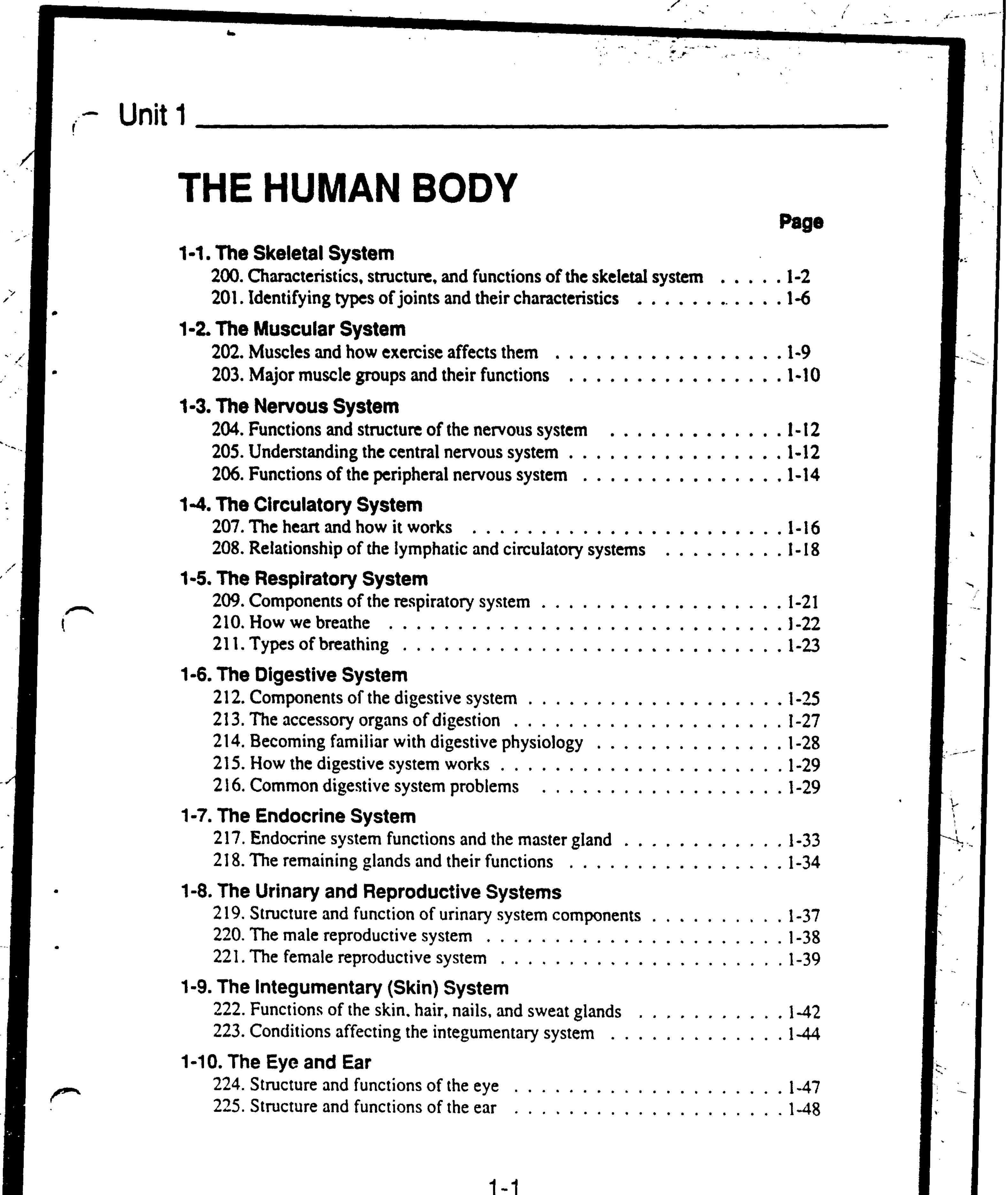
AIR UNIVERSITY

CDC 91450

MENTAL HEALTH
SERVICE SPECIALIST

Volume 2. Anatomy, Physiology,
and General Nursing
Procedures

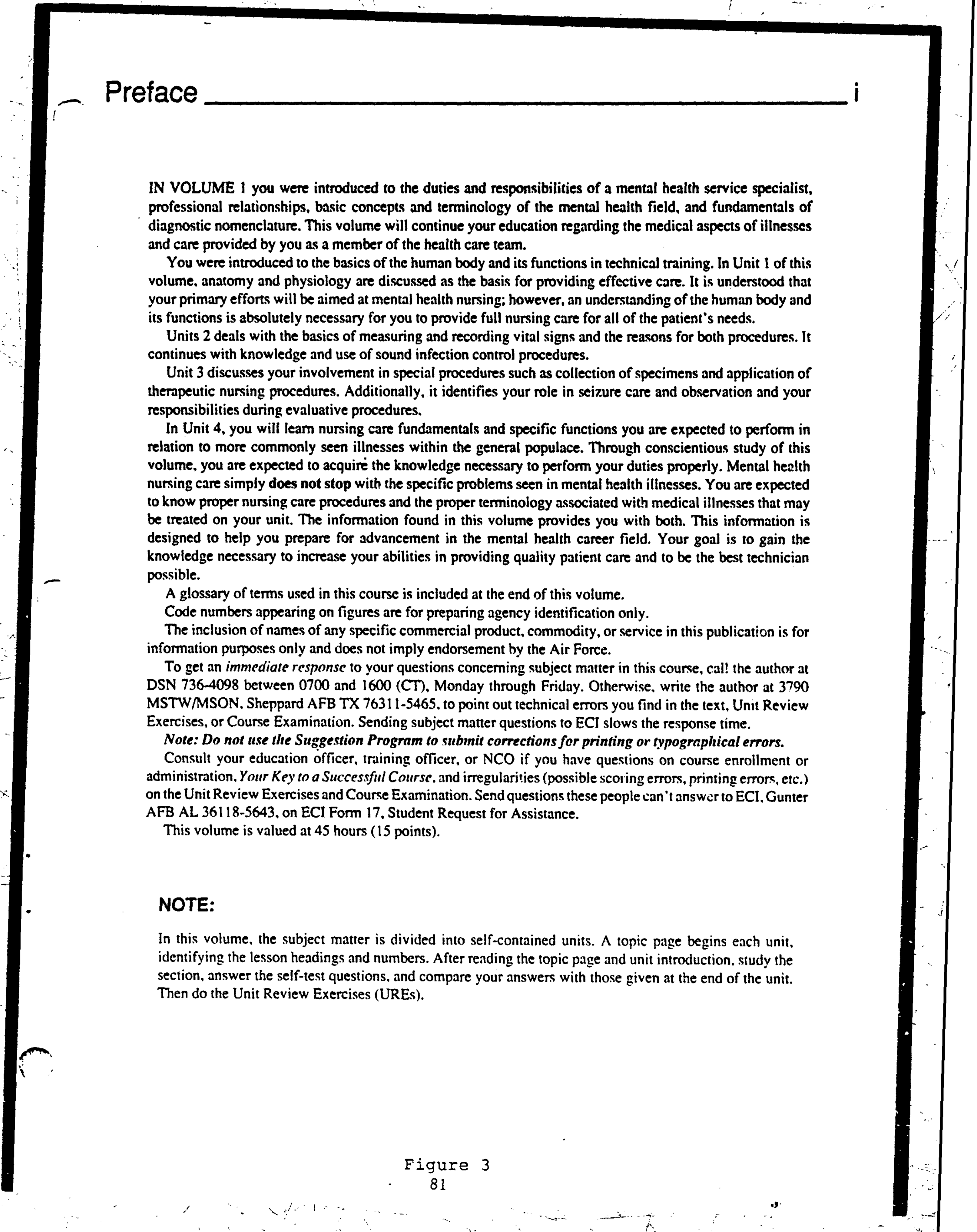
Figure 1



Unit 1

THE HUMAN BODY

Figure 2



Preface

i

IN VOLUME 1 you were introduced to the duties and responsibilities of a mental health service specialist, professional relationships, basic concepts and terminology of the mental health field, and fundamentals of diagnostic nomenclature. This volume will continue your education regarding the medical aspects of illnesses and care provided by you as a member of the health care team.

You were introduced to the basics of the human body and its functions in technical training. In Unit 1 of this volume, anatomy and physiology are discussed as the basis for providing effective care. It is understood that your primary efforts will be aimed at mental health nursing; however, an understanding of the human body and its functions is absolutely necessary for you to provide full nursing care for all of the patient's needs.

Units 2 deals with the basics of measuring and recording vital signs and the reasons for both procedures. It continues with knowledge and use of sound infection control procedures.

Unit 3 discusses your involvement in special procedures such as collection of specimens and application of therapeutic nursing procedures. Additionally, it identifies your role in seizure care and observation and your responsibilities during evaluative procedures.

In Unit 4, you will learn nursing care fundamentals and specific functions you are expected to perform in relation to more commonly seen illnesses within the general populace. Through conscientious study of this volume, you are expected to acquire the knowledge necessary to perform your duties properly. Mental health nursing care simply does not stop with the specific problems seen in mental health illnesses. You are expected to know proper nursing care procedures and the proper terminology associated with medical illnesses that may be treated on your unit. The information found in this volume provides you with both. This information is designed to help you prepare for advancement in the mental health career field. Your goal is to gain the knowledge necessary to increase your abilities in providing quality patient care and to be the best technician possible.

A glossary of terms used in this course is included at the end of this volume.

Code numbers appearing on figures are for preparing agency identification only.

The inclusion of names of any specific commercial product, commodity, or service in this publication is for information purposes only and does not imply endorsement by the Air Force.

To get an *immediate response* to your questions concerning subject matter in this course, call the author at DSN 736-4098 between 0700 and 1600 (CT), Monday through Friday. Otherwise, write the author at 3790 MSTW/MSON, Sheppard AFB TX 76311-5465, to point out technical errors you find in the text, Unit Review Exercises, or Course Examination. Sending subject matter questions to ECI slows the response time.

Note: Do not use the Suggestion Program to submit corrections for printing or typographical errors.

Consult your education officer, training officer, or NCO if you have questions on course enrollment or administration, *Your Key to a Successful Course*, and irregularities (possible scoring errors, printing errors, etc.) on the Unit Review Exercises and Course Examination. Send questions these people can't answer to ECI, Gunter AFB AL 36118-5643, on ECI Form 17, Student Request for Assistance.

This volume is valued at 45 hours (15 points).

NOTE:

In this volume, the subject matter is divided into self-contained units. A topic page begins each unit, identifying the lesson headings and numbers. After reading the topic page and unit introduction, study the section, answer the self-test questions, and compare your answers with those given at the end of the unit. Then do the Unit Review Exercises (UREs).

Figure 3

The dictionary defines fundamental as "forming or serving as an essential component of a system or structure: basic". In this unit you will learn about the basic essential components necessary in the health care system provided to promote our patients' health and well-being. Specific areas of care aimed at patient comfort are covered. We will, once again, review your responsibility as a mental health service specialist health care provider. As it pertains to providing nursing care, one of your primary jobs is to provide quality basic health care and response to patients' physiological needs. You are the one responsible for providing the majority of this care while working the inpatient mental health units (including Alcoholism Rehabilitation Centers). With these points in mind, let's look at your expected role.

2-1. The Technician's Role on the Health Care Team

A patient can be described as anyone who receives help or health care from a doctor, nurse, social worker, or any other health team member. Providing good health care means you provide a service to people; you are not providing medical care. That responsibility belongs to doctors, physician assistants, and others licensed to do so. As part of the mental health care team (see fig. 2-1.), you function as an assistant to the physician and nurse to help hospitalized patients throughout their time on the unit. You are part of a team of professionals and subprofessionals working with one goal in mind: to help the patient get well and return to duty!

As noted in Volume 1, the major stimulus responsible for better patient care in the United States military was conflict related. Can you guess what it was? If you said the Civil War (1861-1865), you're absolutely correct. Guess who was responsible for providing most of the fundamental care for soldiers struck down in battle? If you said nurses and doctors, you're wrong! There were doctors here and there in some of the more fortunate units. Depending on how you look at it, you might say that some of those units were unfortunate, considering the lack of training and equipment these doctors and surgeons had. There were very few, if any, assigned nurses in the field of battle. The majority of nursing care in the field was provided by fellow soldiers; buddy care was one of the names attached to this response. These (buddies) could be considered the first medical service specialists or mental health specialists. Due to the lack of available medications, a great deal of relaxation therapeutic techniques were probably used unknowingly.

The military, in time, recognized the need for better care of its sick and wounded and progressively took steps to remedy the lack of organized care. By 1901, the Army Nurse Corps was established. In 1908, the Navy established its Nurse Corps. The Air Force, a much younger service, didn't formally organize its medical service until 1949. Through these developments came the establishment of the corpsmen or medical technician roles to support the military medical programs and to provide the fundamental care once provided by the "buddy" in the field. Now that you're somewhat familiar with how the whole idea of technician use got started, let's move on to the fundamental care you are expected to perform as a mental health service specialist.

226. Obtaining physiological measurements

As part of the mental health care team (fig. 2-1), you will be asked to obtain physiological measurements. Physiology pertains to the biological science that deals with the essentials of life processes. Since it has to do with living organisms and is one of the processes that keeps all organisms going, it most likely can be *measured* or compared against a "norm" or normal expected quantity or quality. The vital signs are a perfect example. As you may know, the vital signs include the patient's temperature, pulse, respiration, and blood pressure. They are called *vital signs* because of the significance of the data they offer which may indicate the condition of the patient physically or mentally. Vital comes from the Latin word *vita*, which means life. And chances are if you can't get a measurement of any one of these four vital signs, the patient is having a serious problem.

Most of these measurements vary within certain limits during a 24-hour period. There are areas that may affect the vital signs such as sleep, exercise, noise, weather, medications, stress, illness, or fear. The affect of these areas is dependent upon the individual's personality, specific illness, emotional state, or body chemistry. The basic skills for taking a patient's vital signs are simple but should not be taken for granted. Accuracy is an absolute must. The vital signs are the quickest, most accurate way of determining a patient's medical condition and needs. Changes in how the body is functioning and how emotions are affecting the patient are often reflected in the vital signs. The vital signs often reflect even minor changes in a patient's health or emotional state. Keeping this in mind, let's look at the four vital sign areas and discuss the methods used to obtain the data necessary from each area.

Temperature or Body Temperature. The body's tissue, cells, and other organisms function at their best in a narrow

Figure 4

There may be times when it will be necessary to admit or transfer patients with medical or surgical problems to your unit. This is usually done because the patient's behavior is such that the staff on the medical or surgical unit is unable to provide the total specialized care needed. In these cases, the need for emotional and psychiatric support is placed in a higher priority. The opposite is true when it becomes necessary to transfer one of your patients to a medical or surgical unit for specialized care because of a serious medical or surgical problem. In addition, one of your patients may develop a medical or surgical problem that the physician will want to treat. In this unit, we cover some of the medical and surgical illnesses you may encounter and some general nursing care required for these patients.

All patients, whether they are mentally ill or have medical or surgical problems, have some of the same or similar needs. In this section we look at some of the general problems and needs of the medical/surgical patients. Some of the material may appear to overlap previous units. Review is often necessary for subject matter applying to the proper care of patients. Keep in mind that these patients are actually physically ill and require special care and considerations from you.

4-1. Respiratory Disorders

Respiratory disorders cause a greater amount of sickness within the military than any other illness. How many times have you heard, "He has the flu," or "She's in the hospital with pneumonia." Illnesses such as influenza and pneumonia are common causes of respiratory distress and require a different type of nursing care. In this section, we will look at the symptoms of respiratory disorders and the nursing care of these disorders.

Do you recall the definitions of pulmonary and respiration? Simply put, pulmonary means "of or pertaining to the lungs". Respiratory means "the act of breathing", taking in oxygen, giving off carbon dioxide. Both terms will be used in this section; they are often interchanged with each other. The key is to remember whether it is a pulmonary or respiratory disorder. It's getting in the way of breathing, and that's not good. Let's start by looking at some of the symptoms of respiratory disorders.

240. Recognizing the symptoms of pulmonary/respiratory disorders

Most patients with pulmonary disorders have several symptoms in common and, therefore, have problems and needs that are common. Among these common symptoms are cough, pain, and dyspnea. The cough can be much the same for a patient with pneumonia or tuberculosis. For instance, the patient with pneumonia has an annoying cough accompanied by a rusty-colored or blood-streaked sputum. One with tuberculosis m**a**y present the same symptom. It is only from other symptoms and diagnostic tests that the correct diagnosis is made.

Chest Pain. Chest pain may result from many causes and it has several meanings. It may be caused from tissue destruction, diseased chest muscles, irritation to nerves, or inflammation of pleural tissues. No two patients react to pain in the same way. Where some patients tolerate pain and do not complain, others complain loudly. For this reason, it is best to observe and report the pain according to the following pattern:

- When did the pain start?
- What is the nature of it? Is it sharp, burning, or knifelike?
- Is it severe or mild?
- Where is it located (right upper chest, lower left chest, etc.)?
- Does it come and go?

Dyspnea. Dyspnea, or labored breathing, is another common symptom among patients with pulmonary disorders. It can be caused from an obstruction, congestion, or any inflammatory condition of the respiratory tract. Dyspnea may be treated by removing the cause and administering oxygen. Signs of dyspnea include cyanosis, increased respiration, and restlessness.

Problems in respiratory functioning can be caused by illness and conditions that affect ventilation. This in turn affects oxygen transport. Three primary factors affecting respiration are hyperventilation, hypoventilation, and hypoxia.

Hyperventilation. This term is used to refer to a state of ventilation in excess of what is required to maintain normal levels of carbon dioxide in the body. The harm doesn't come from taking in too much oxygen. Instead, it prevents exhaling too much carbon dioxide. A condition known as respiratory alkalosis can occur in a hyperventilating patient. This means there is a loss of acid and an increase in the pH factor which ultimately could affect the entire body functioning.

This problem of breathing too quickly can be caused by severe stress, anxiety, injury, or infection of the respiratory center in the medulla. Hyperventilation also causes diminished blood flow to major organs because of low carbon

Figure 5

1-4. The Circulatory System

The circulatory system involves the cardiovascular and lymphatic processes. Patients with disorders of this system are very often seriously ill, and the nursing care they receive is a crucial and vital part of their treatment. These disorders are subject to abrupt and sudden changes and require close observation. They are different from most other body system disorders in that the circulatory system is vital to all other body systems and functions. It is possible to carry on without a kidney or lung, but without a heart, life stops.

The circulatory system includes all structures concerned with the transportation and distribution of body fluids from one region of the body to another. It includes a cardiovascular system and a lymphatic system. The cardiovascular system consists of the heart, arteries, arterioles, veins, venules, and capillaries. The lymphatic system is made up of lymph capillaries, lymph vessels, and lymph ducts. We will consider each of these systems separately.

207. The heart and how it works

The Cardiovascular System. The cardiovascular system comprises the heart and blood vessels. The heart propels blood through the blood vessels—a system of closed tubes composed of arteries, capillaries, and veins.

The heart. The heart (fig. 1-12) is a hollow, muscular pump that lies in the middle portion of the chest (mediastinum) slightly to the left side. It is flanked by the right and left lungs. Protection is provided anteriorly by the sternum and posteriorly by the vertebral column. The heart is said to be about the size of a man's fist, cone-shaped, with the apex directed downward and to the left. It is contained within a fibrous sac called the *pericardium*. The serous inner layer of the pericardium normally allows free cardiac motion. This pericardium has two layers: the visceral and the parietal pericardium. The space between these two layers contains pericardial fluid which prevents external trauma from being transmitted directly to the heart.

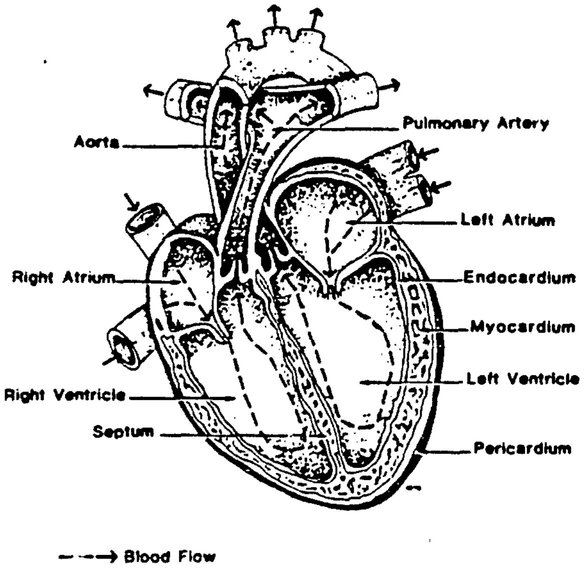


Figure 1-12. The heart.

The heart is composed of three separate layers of tissue. These are the epicardium, the myocardium, and the endocardium. The outer layer, the epicardium, covers the surface of the heart. The endocardium, or innermost layer, is a thin layer of tissue that lines the inside of the heart and covers the cardiac valves.

The interior of the heart is divided into a right and left portion by a septum (a dividing wall or partition). In each half there is an upper chamber (the atrium) which receives blood from the veins and a lower chamber (the ventricle) which receives blood from the atrium and pumps it out into the arteries. The openings between the chambers on each side of the heart are supplied with one-way valves which prevent backward flow of the blood. The valve on the right is the tricuspid valve; the one on the left is the bicuspid or mitral valve. The outlets of the lower chambers of the heart (ventricles) are supplied with similar valves. The pulmonary valve is at the origin of the pulmonary artery; the aortic valve is at the origin of the aorta, which is the largest artery.

Physiologically, the heart acts as two separate pumps. The right side receives deoxygenated blood into the atrium from the various regions of the body. Then, the ventricle pumps it into the lungs. There it receives a fresh supply of oxygen and gives off carbon dioxide. This phase is called *pulmonary circulation*. The left side of the heart receives the oxygenated blood from the lungs into the atrium, and the ventricle pumps it into all regions of the body through the arteries. This phase is the systematic circulation.

Each contraction of the heart is followed by limited relaxation. The cardiac muscle never completely relaxes but always maintains a degree of tone. Contraction of the heart is the systole, the period of work; relaxation of the heart is the diastole, the period of rest. A complete cardiac cycle is the time from the onset of one contraction or heartbeat to the onset of the next.

During the cycle, the blood going through the valves and chambers of the heart produces sounds. These sounds can best be imitated vocally by the syllables "lubb" and "dubb," separated by a brief pause. Conveyed to the ears of the physician by a stethoscope, variation in the rhythm, intensity, or character of these sounds furnishes valuable clues for detecting cardiac disease.

The heart has its own system of blood vessels: the right and left coronary arteries which arise from the aorta as soon as it leaves the heart, and the coronary sinus which collects the venous blood from the heart and empties it directly into the right atrium.

Ficure 6

muscles are also attached to it. The various surfaces and processes enable the vertebrae to move upon one another.

There are seven cervical vertebrae in the neck. The first is the atlas, which supports the head. The second is the axis upon which the head turns. These are the only vertebrae with names; all of the others are numbered. There are 12 thoracic vertebrae in the posterior chest region. They articulate with the ribs. There are five lumbar vertebrae. The sacrum articulates on each side with the hip bone and coccyx, forming the posterior wall of the pelvis.

Between the vertebrae, from the second to the sacrum, are intervertebral discs. They act as shock absorbers for the vertebral column. They contain an elastic, pulpy substance called *nucleus pulposus*.

The thorax. The thorax (fig. 1-4) is a cone-shaped, bony cage formed by the 12 thoracic vertebrae, the ribs that terminate anteriorly (remember all that terminology in Volume 1?) in articulating cartilages, and the sternum. It houses the heart and lungs. There are 12 ribs on either side of the thorax, extending from the first through the twelfth thoracic vertebra. Ribs are identified by number and by the side of the body on which they are located. The first seven pairs are considered true ribs because they are attached to the thoracic vertebrae and to the sternum. The remaining five pairs are considered false ribs because they do not articulate directly with the sternum. The sternum, commonly called the *breastbone*, is long and flat and is located at the midanterior part of the thoracic cage. It protects the heart, lungs, and greater vessels. The sternum consists of three portions: the manubrium, the body, and the xiphoid process.

The upper extremities. Each upper extremity, shown in figure 1-5, consists of the clavicle, shoulder, arm, forearm, wrist, and hand. Both extremities together total 64 separate bones.

The clavicle, or collar bone, forms the anterior part of the shoulder girdle. It lies horizontally just above the first rib. It is attached to the scapula and the sternum. Because of its anterior location, it is often fractured as a result of falls.

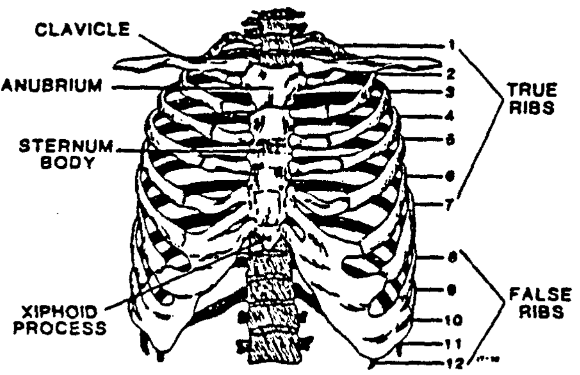


Figure 1-4. Bony thorax, anterior view.

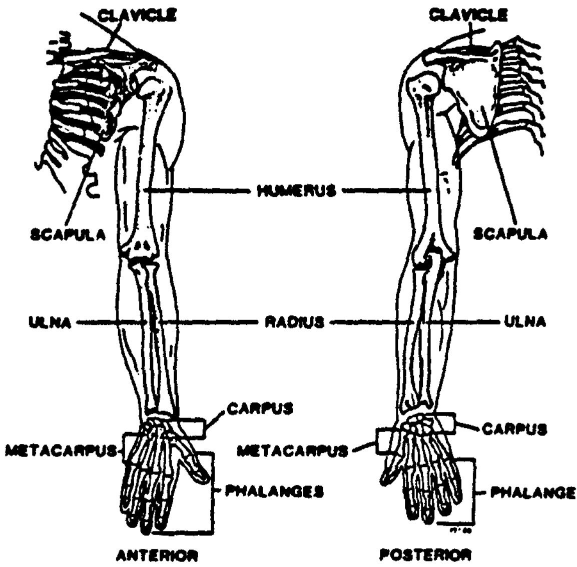


Figure 1-5. Upper extremities.

The scapula (shoulder bone) is a triangular-shaped bone lying in the upper part of the back. It forms the posterior portion of the shoulder girdle and a part of the shoulder joint.

The humerus is the bone of the upper arm and is classified as a long bone. It articulates with the shoulder girdle to form the shoulder joint and the bones of the forearm to form the elbow joint.

The bones of the forearm are the ulna and the radius. From an anatomical position (palms facing forward), the radius is on the lateral or thumb side with the ulna on the medial or little finger side. The ulna and radius articulate at their proximal ends (nearer the center of the body) with the humerus, and at their distal ends (farthest from the center of the body) with some of the carpal bones.

The wrist and hand are made up of 27 separate bones. The wrist contains 8 small bones called *carpals*. The hand consists of 5 metacarpal bones and 14 phalanges. The metacarpal bones are numbered 1 to 5 to correspond to the 5 fingers; the thumb is the first finger. The phalanges are the small bones of the fingers.

The lower extremities. The bones that make up the lower extremities (fig. 1-6) are the innominate, femur, patella, tibia and fibula, tarsals, metatarsals, and the phalanges.

The innominate, or hipbone, is composed of three parts—the ilium, ischium, and pubis—that are firmly united into one bone. The two hipbones, together with the sacrum and coccyx, posteriorly form the pelvic girdle. This girdle forms a deep basin that protects the organs of the lower abdomen, the bladder, lower bowel, and reproductive organs.

The femur, or thigh bone, is the longest bone in the body. The proximal end is rounded and has a head that fits into the acetabulum. It also has a neck, the part of the femur most frequently fractured.

Please write your response to unit self-test questions and then check the text answers at the end of the unit.

SELF-TEST QUESTIONS

207. The heart and how it works

1. Match each term listed under column B with its the correct description in column A. Some responses may be used more than once.

Column A

____(1) A hollow, muscular pump in the middle portion of the chest.

____(2) A fibrous sac containing arteries.

____(3) The layers of the heart.

____(4) Upper chambers that receive blood from the veins and lower chambers which pump blood into the arteries.

____(5) Acts as two separate pumps.

____(6) Receives oxygenated blood from the lungs.

____(7) The time from the onset of one contraction or heartbeat to the onset of the next.

____(8) Carries the blood to the heart muscle.

____(9) Increases the activity of the heart as a pump.

____(10) Strong elastic tubes, constructed to withstand the high pressure placed upon their walls when the heart pumps blood to the body.

____(11) Located at the ends of the arterioles and are so small that red blood cells pass through in single file.

____(12) The contracting force of adjacent muscles and the action of the diaphragm aid in the forward propulsion of blood through them.

Column B

a. The heart.
b. Right and left coronary arteries.
c. Veins.
d. Pericardium.
e. Right and left atria; right and left ventricles.
f. Epicardium, myocardium, endocardium.
g. Arteries.
h. Sympathetic nervous system.
i. Left atrium.
j. Capillaries.
k. Complete cardiac cycle.

208. Relationship of the lymphatic and circulatory systems

1. What is the relationship of the lymphatic system to the circulatory system?

2. What is lymph, and how is it formed?

3. Describe lymph nodes, state another name for them, and tell where they are located.

4. State the two functions of the lymph nodes.

5. What role does the spleen play in the lymphatic system?

Figure 8

ADVANCED TECHNOLOGJES FOR DISTANCE LEARNING

CONCLUDING REMARKS

J. Michael Spector
Senior Scientist, Instructional Design Branch
Armstrong Laboratory
Brooks AFB, Texas

The focus of this symposium on distance learning has been on how emerging technologies can be used to enhance interactivity and learning in distance settings. Our basic assumptions are (1) that media use or misuse can make a difference in learning outcomes; and (2) that there is a some relationship between interactivity of the instruction and learning in structured situations. The debate in the literature with regard to the first assumption is extensive, although we believe that there is sufficient evidence to continue testing the hypothesis that media use can affect learning outcomes (cf., Kozma, 1991; Kulik & Kulik, 1991). The evidence connecting interactivity and learning has also been studied extensively, although the connections are not as clearly established as we would like (cf., Hannafin, 1991; Riel, 1990).

The most obvious and immediate conclusion of this discussion is that there is a paucity of research concerning systematic methods that can be used to optimize learning in distance settings. Because there is great potential for distance learning, both in terms of learning outcomes and cost savings, there is a need to conduct additional studies, to establish an ongoing research program, and to develop government and industry standards for distance learning.

Reid Matherne of the Air Force Extension Course Institute (ECI) provided a report on how distance learning support materials might be optimized. If we adopt the well-established principle that instructional materials should be designed with the delivery setting clearly in mind, then it is certainly the case that this area deserves careful examination. ECI's principle is that supporting material should fit easily into the format of the delivery setting. In addition, such materials should contain directions and white space for student notetaking during a live distance delivery. The real challenges occur as the delivery setting becomes more dynamic and interactive. Our expectation is that supporting materials will need to be computer-based in order to support the collaborative development and delivery of instruction that is likely to increase in the next decade.

Henry Simpson of the Navy Personnel Research & Development Center has conducted some empirical studies comparing live instruction with a variety of instructional distance technologies (one/two way audio/video). His findings were that one-way video is about as effective as two-way video in spite of strong preferences for the latter on the part of both learners and instructors.

Professor Robert Main of California State University Chico proposed a collaborative environment for the development of computer-based instructional materials for distance learning. This is an interesting variation on the need for learner collaborations in distance settings. As media become increasingly complex and expertise becomes more sparsely distributed, the need for collaborative instructional developments at a distance becomes more significant.

Andrew Wilson, also of California State University, provided an extensive evaluation of the capabilities of the Integrated Services Digital Network (ISDN) to support both distance development and delivery of multimedia computer-based instructional materials. His discussion emphasizes the fact that as the technologies become more affordable the need to optimize and to standardize becomes more critical.

In conclusion, we can honestly say that we have only scratched the surface with regard to issues related to the future of distance learning. What appears worthy of immediate research and development are (1) an exploration of the nature of interactivity and its relationship to learning; (2) the significance of collaborative environments for development and delivery; and (3) how supporting materials can be designed to best support a variety of distance settings.

References

Hannafin, M. J. (1991). Effects of elaboration strategies on learning and depth of processing during computer-based instruction. *Journal of Computer-Based Instruction*, 18(3), 77-82.

Kozma, R. B. (1991). Learning with media. *Review of Educatinal Pesearch*, 61(2), 179-212.

Kulik, C. C. & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, 7, 75-94.

Riel, M. (1990). Cooperative learning across classrooms in electronic learning circles. *Instructional Science*, 19, 445-466.

SENIOR MILITARY EDUCATION IN THE U.S. ARMY: PERSPECTIVES FROM GRADUATES

Glenda Y. Nogami
U.S. Army War College

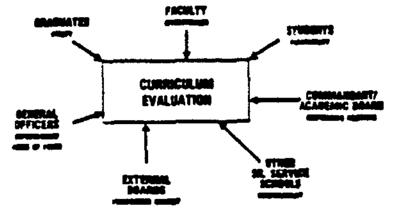
The U.S. Army War College (USAWC) is the senior service college of the Army. USAWC is charged with the responsibility to prepare officers and civilians for "senior leadership responsibilities in a strategic environment during peace and war," as well as to "study the role of landpower, as part of a joint or combined force, in support of the U.S. national military strategy" (USAWC Curriculum Pamphlet, Academic Year 1991).

The USAWC offers a Military Education Level-1 "degree" (MEL-1). This is the highest military education level designator, and is required in many of the higher level positions. The MEL-1 degree provides broad education at the senior, strategic level. Currently, 96.7% of all General Officers in the tri-services and 99.2% of all Army General Officers are MEL-1 graduates (General Officer Management Office, 1990).

USAWC Curriculum Evaluation Model

Over the years, many individuals and groups have influenced the curriculum. For any school, there are many myriad groups and individuals who seek to influence any curriculum. Some of these are mandated by law, others because of tradition, because of expertise, because they are the recipients of our educational process, or simply because of interest. In seeking to gain an overall perspective on the various points of view of the constituency groups, USAWC has developed a comprehensive curriculum evaluation model. The model recognizes that there are at least seven important constituency groups: (1) current students, (2) current faculty, (3) graduates, (4) general officers, (5) other senior service schools, (6) mid-career officers - prior to entry at USAWC, and (7) external boards of inquiry and evaluation (Nogami, 1990). Although all groups provide information on all facets of the curriculum, each group's primary contribution is unique (Figure 1).

CURRICULUM EVALUATION



Students and faculty are best able to evaluate the current curriculum. The students provide evaluations on individual courses and on the complete Academic Year. The quality of course materials and instruction is also rated, but primarily the data is indicative of what they think will be useful to them and what they enjoyed. Faculty are best able to judge course and lesson content and the effectiveness of different teaching methods. They are responsible for selecting and preparing course materials, developing effective presentation methods, teaching and evaluating student performance.

The real test of whether USAWC is teaching the skills and knowledge necessary for senior leadership can best be determined by USAWC graduates and General Officers.

Graduates should be better able to successfully perform their duties because of the skills and knowledge imparted or developed at USAWC. The input from graduates is a validity check: did USAWC prepare them for their assignments? (Nogami and Davis, 1989).

The separate groups of General Officers (GOs) and graduates will overlap with time. General Officers have an overview of the many job related requirements of colonels and GO's. They offer at least two important groups of information: consensual validation (to the graduates' input) and identification of new GO skills that will be required in the future - helping to make USAWC more pro-active in curriculum planning and implementation.

With the emphasis on jointness and cooperation, it is imperative that USAWC have (at least) a comparable program with the other services. The level of instruction and the information presented should be appropriate and on a par with education from other senior service schools. This will ensure both a more effective joint service environment, as well as help ensure that the best officers desire to come to USAWC as faculty and students.

External boards of inquiry or evaluation come from various sources: Congress (e.g., Skelton, 1988; GAO, 1991), the American Council on Education, DOD and DA Commissions & Panels (e.g., Haines, 1966; Joint Professional Military Education Panel, 1990; etc.). These boards provide information on the comparability of the USAWC curriculum to other curricula - both military and civilian. In the context of total military education, these boards put the USAWC experience in perspective - as a strong link in the chain of military education. Their primary contribution involves "how others see USAWC."

USAWC Biennial Survey of Graduates

The purpose of the USAWC is to prepare the Army senior leaders for future positions. In operational terms, this means preparing officers for the 5-7 years of service following USAWC. During these years, it can be assumed that the graduate will fill 2-5 different assignments, as diverse as brigade commander to National Security Council or Joint Chiefs of Staff adviser. Skills and knowledge important to any one position may not directly apply to another position. To be able to identify specific job requirements, the USAWC has instituted a biennial survey of graduates beginning in 1990. This is a longitudinal survey which will follow individuals over time. This will allow USAWC to identify skills and topics that are useful, not just in the job the respondent is currently holding, but in all positions s/he has filled.

METHODOLOGY

Respondents

Although the students/graduates are from all the sister services, civilian federal agencies, and other countries, USAWC's primary audience is Army officers (Regular Army, Army Reserves, and Army National Guard). The criteria for inclusion for the survey were: (1) Army officer, (2) graduate of USAWC (USAWC MEL-1), (3) MEL-1 from AY83 to AY89, (4) not on retired status. MILPERCEN provided current rank and addresses were provided for each of the officers who met the above criteria.

In April 1990, survey packets were sent to each of the 1,570 eligible officers. Two months after the first mailing of the survey, a reminder postcard was sent to the nonrespondents. Surveys received by 30 September 1990 - five months after the initial mailings - were included in the analyses.

Survey Instrument

The Biennial Survey is designed to provide USAWC information which will help to "evaluate the relevance of the curriculum and help the College plan for future needs and long term educational objectives" (letter from the Commandant, 1990). The survey is designed to have two parts: one, a core set of questions about the curriculum; and two, a more changeable set on issues of immediate concern. These are questions that refer to specific, time sensitive topics which may have little or no applicability in a longitudinal study.

RESULTS

A total of 1,110 completed surveys were received. This represents a 71% response rate (1,110 divided by 1,570 sent). Due to space limitations, suffice it to say that 2/3 of the respondents were graduates of the USAWC resident program, 2/3 were in the Regular Army, slightly over 50% were in the Combat Arms, 20% each in the Support and Service Support Branches, and every Academic Year 1983-89 was well represented.

The respondent population (current rank) is predominantly colonels (O-6). As expected, the more recent year groups have a larger percentage of LTC(P) and LTC than earlier year groups for all USAWC MEL-1 Programs. General officers made up 23% of the 1983 year group, and only 1% of the 1989 class. USAWC graduates were more likely to be commissioned through ROTC or OCS than USMA (79% vs. 10%).

Curriculum Topics

Eighty-four percent (84%) of the respondents agreed with the statement that the "USAWC curriculum covered the right subjects for my professional development." Respondents were asked to rate the utility of forty separate curriculum topics and programs for their current position on a 5-point rating scale (from 1 = not at all to 5 = very greatly). Because the respondents spanned seven years with different curricula, a sixth response - "not applicable" - was available. Table 1 shows the average rating for each of the curriculum topics. All "not applicable" responses were excluded. All topics are above the mid-point of the scale (mid-point = 2.5). More than 80% of the officers rated each of these topics as being useful in their current assignment. Specific analyses for assignment, job, branch, component (Active vs Reserve), and year of graduation are available in the complete report (Nogami, Colestock, and Phoenix, 1991).

Educational Objectives

The graduates are expected to "enhance the effectiveness of the U.S. Army," and USAWC is charged with preparing them to "meet the full range of responsibilities and challenges (they) will encounter as a senior leader(s)" (Curriculum Pamphlet, Academic Year 1990, USAWC). This has been translated into seventeen objectives of the USAWC academic program. Graduates were asked to indicate how well these objectives were accomplished for them on a 5-point scale (1 = not at all to 5 = very greatly). The table below shows the overall mean rating for each objective. Judging by the mean ratings the graduates have given, all of the objectives have been well accomplished, all ratings above the mid-point, with the vast majority floating in the 4.0 range (on a 5.0 scale).

UTILITY OF CURRICULUM TOPICS

	Mean
The Senior Leader	
Senior leadership competencies	3.9
Ethics and values of the senior leader	4.0
Strategic and operational decision making	3.8
Self assessments (e.g., M-B personal/pref)	3.6
Command in war	3.3
Strategic vision	3.7
Human dimension of combat	3.3
War, National Policy and Strategy	
Theory and nature of war	3.4
Elements of national power	3.7
Domestic environments on national security policy	3.7
Global environments on national security policy	3.7
Formulating/analyzing national security policy	3.4
Formulating/analyzing national military strategy	3.5
Historical assessment of national strateegy	3.3
Strategic/theater nuclear concepts/issues	3.1
Regional and global strategic appraisals	3.6
Implementing National Military Strategy	
Operational continuum (spectrum of conflict)	3.3
Formulating joint military strategy and doctrine	3.1
Joint operations planning system (JOPS)	2.9
Planning, programming, budgeting system (PPBS)	3.4
Joint strateegic planning system (JSPS)	2.8
Structure & capabilities of military forces	3.4
Army develops, resources, sustains, mobilizes	3.5
Planning and execution of strategy	3.3
Process of mid-range policy formulation	3.1
Theater planning at Unified Command level	2.8
Security assistance	2.8
Operational art	3.0
Organization & functions of non-military agencies	3.0
Risk assessment	3.2
Other/Complementary Curriculum Topics and Programs	
Effective oral communication	3.8
Effective written communication	4.0
Word processing & computer skills	3.3
Assessing your general health & fitness	3.8
Type A/B and stress management	3.7
Military studies program (MSP)	3.3
Military families program	3.2
Advanced courses program	3.6
TV and media workshop	3.0
Military history - lessons learned	3.5

Suggestions for Additional Curriculum Topics

"What skills/knowledge do you feel senior officers will need in the next five to ten years that should be added to the curriculum?" This open-ended question elicited numerous topics and reflected some thought on how the Army and the global situation in which the Army operates might develop. As the curriculum has evolved some of the suggestions have already been incorporated. Indeed, many of the suggestions are not for "new" courses or topics, but topics for additional or "continued" emphasis. The suggestions reflect their perception of the kind of world in which the Army must operate.

Table 24. Educational Objectives

How well has USAWC prepared you to:	Mean
1. Set an ethical climate in your service/organization	3.8
2. Be physically fit	3.4
3. Be mentally fit	3.9
4. Deal with problems which have no clear cut solutions	3.8
5. Be an innovator/initiator of policy	3.7
6. Succeed in positions of broad scope & responsibility	4.0
7. Assess/plan for the future while executing in the present	3.8
8. Think conceptually	3.9
9. Think critically	3.9
10. Work in a strategic environment	3.9
11. Understand the role of the military in a democratic society	4.1
12. Be adept in the development and use of military forces to achieve national objectives	3.9
13. Advise the National Command Authorities on the use of military forces	3.6
14. Make better decisions and give better advice	4.0
15. Provide a frame of reference which recognizes the complexity of the issues but also provides the perspective to work through them to find solutions	3.9
16. Serve in an orga ization involving joint forces	3.5
17. Serve in an organization involving combined or coalition forces	3.3

Joint From the graduates' comments, future graduates will operate in a "joint environment." To optimally operate in it, the Army, Air Force, Navy, Marines, as well as the Reserve Components must work together. Force reductions and budget cuts are perceived to "place a tremendous burden on all services to cooperate with each other to maintain our military readiness." Teaching "jointness" has to be more than information on the "joint system." It means learning a style of responding to problems in which branch of service does not enter into the equation, where it is the "U.S. military, not Army to meet the threats" (or other service).

Graduates anticipate that the "peacetime uses" of the military will include more involvement in domestic issues (i.e., drug interdiction, "natural disaster recovery" and crisis management). Effectively coordinating military, reserve and civilian agencies will be a challenge. In addition to understanding how each agency is organized and operates, it will be important to "know the rules, opportunities, and pitfalls" for involvement.

Changing national priorities will necessitate working smarter within stricter environmental constraints. Installation and resource management will increasingly be concerned with environmental issues. The impact of training exercises on the environment, as well as disposal of hazardous waste materials, will be issues that installation commanders and resource managers must handle.

Economic problems within this country will necessitate working with fewer materiel and personnel resources - i.e., resource management. Planning, programming, budgeting (PPBS) acquisitions, then managing the resulting dollars and people means knowing the system and how to manipulate it. The problem of "shrinking dollars and people" is not perceived to be a "short term problem," but one that will be with us for a long time. "Downsizing" the Army while maintaining "training readiness" for a credible force to "execute" military strategy will require effective resource and personnel management skills.

New alliances and a decrease in superpower tension, will require an increased emphasis on international relationships and cultures. The threats are more diverse and from all regions. The probability of more "low intensity conflicts to include terrorism" will require new thinking and strategies. Knowledge of other cultures and languages, understanding of treaties, negotiations, and international diplomacy will be increasingly important for understanding new friends and foes. Allied, combined, and coalition operations will be more common. An increased emphasis on the needs of third world countries and regions may require nation building and security assistance skills, as well as understanding the armed forces in these countries.

"How to think," with an innovative, open-minded approach will be needed to deal with accelerating change. As the global situation changes, when ambiguity prevails, and information is incomplete, the officer must be able to adapt and be flexible. Analysis, vision, strategy, as well as conceptual and creative abilities, are required of the senior officer. Innovation and creativity require the ability to "shed the rigid rules of doctrine and be visionary in approach." Utilizing decision support systems and computerized information systems may make the tasks easier to accomplish by making "fast and voluminous" information more manageable if senior leaders are able to capitalize on them.

Communication - both verbal and written, already very important, will become even more so. Competing for national resources will require the Army to present a cogent case to Congress and the public. The senior leader's ability to communicate will greatly influence Congress' and - through the media - the public's reaction to the Army. Communication must be accomplished both up and down the chain of command - to soldiers and their families.

As important as the above are, warfighting is still the major skill required of the senior Army leader. In order to effectively use land forces, he must be extremely knowledgeable about warfighting across the spectrum of conflict, operational planning, and contingency operations.

REFERENCES

General Accounting Office. (1991). Army: Status of Recommendations on Officers' Professional Military Education (GAO/NSLAD-91-121BR). Washington, D.C.: U.S. Government Printing Office.

Haines Board (1966). Report of the Army Board to Review Army Officer Schools: Washington, D.C.: Department of the Army.

Nogami, G. Y. and Davis, R. J. (1989) Report on the Survey of USAWC Graduates from Academic Years 1983 - 1987. Carlisle, PA: U.S. Army War College.

Pappas, G. S. (1979) Prudens Futuri: The US Army War College 1901-1967. Carlisle, PA: The Alumni Association of the US Army War College.

Skelton Commission (1988). Hearing on Professional Military Education. Washington, D. C.: Panel on Military Education, Committee on Armed Services.

USAWC Curriculum Pamphlet, Academic Year 1991 (1990). Carlisle, PA: U.S. Army War College.

DEVELOPING LEADERSHIP SKILLS IN ARMY UNITS

Joan Harman
U.S. Army Research Institute
for the Behavioral and Social Sciences

Introduction

During the span of years when the U.S. Army Research Institute (ARI) has been carrying out research for the Army, no topic has been of more consistent importance than that of leadership. The Army has undergone vast changes in structure and philosophy during the present century and, with each change, leadership has been a central critical issue. Not only has Army leadership evolved as an internal phenomenon, but also the Army has demonstrated widespread leadership for American society as a whole. Two striking examples of such successful demonstrations are the racial integration of the Army after World War II and Project 100,000 that resulted in the mainstreaming of citizens who had suffered education decrements.

In recent years, ARI's leadership research has been sponsored by the Center for Army Leadership (CAL) at Fort Leavenworth. The study that is addressed by this paper concerns the policies, practices, and tools used by unit commanders to develop leadership skills in junior leaders (specifically, newly commissioned 2nd Lieutenants and newly promoted Staff Sergeants). CAL's interest is in a closer reconciliation between institutional leadership training and training in operational settings that will constitute an integrated leader development system. To contribute to this development, ARI researchers traveled to Army posts and interviewed unit leaders about orienting junior leaders to unit practices, assessing leadership skills, providing feedback, developing leadership skills, and improving leadership development in units.

Procedure

Subjects

97 Battalion Commanders and Command Sergeants Major, Company Commanders and First Sergeants in 10 battalions at 5 Army posts (Forts Carson, Ord, Polk, Riley, and Stewart) were interviewed. These leaders were members of light infantry, mechanized infantry and armor battalions.

Method

Researchers developed and pilot tested interview guides for Battalion Commanders, Command Sergeants Major, Company Commanders and First Sergeants. Using the revised guides, each interviewer both took notes on responses and tape recorded them with the consent of the persons interviewed.

Results

Orientation

The Army uses fairly extensive procedures to integrate newly assigned soldiers into units. When a unit is notified about a soldier who will be reporting for assignment, it sends an information packet and a welcome letter to the soldier. When the soldier arrives at the post, he is inbriefed by his chain of command during which he is apprised of the standards and expectations of his commanders. He is provided with a sponsor who is an experienced peer and who will guide him through resolution of personal and professional problems. Unit commanders express special concern about the need for soldiers to get their families settled so that their minds will be free to concentrate on learning their new jobs. Frequently, newly assigned soldiers are granted 10 days temporary duty to assist with family settlement. Battalion commanders address groups of newly assigned soldiers and give them the history of the battalion and their command philosophy.

It's important to note, however, that units carry out these orientation practices when they are able to do so. In some cases, units receive no notice before soldiers report for assignment and, conversely, units send out welcome packets to soldiers who are later assigned elsewhere. Furthermore, when units are preparing for major training events, very little time is avai'able for any other activity. At the very least, however, newly assigned soldiers are inbriefed by their commanders and learn about standards and expectations.

Assessment

Interviewees were asked how early they start to appraise the leadership skills of newly assigned 2nd Lieutenants and Staff Sergeants. Most responded that they make a great effort to avoid passing judgement on any soldier until they are able to observe his performance both in garrison and in the field. However, a soldier's appearance, his attitude and his ability to communicate during inbriefings can provide some information about his potential. In the case of Staff Sergeants, the records they bring to inbriefings show their past assignments and testing records.

Interviewees were asked what criteria they use to assess leadership skills after they have sufficient opportunities to observe a soldier's performance. The following responses are rank ordered by frequency of mention:

personal appearance (this includes physical condition)
interactions with troops, peers, superiors
technical and tactical competence
motivation/eagerness
teaching/training
ability to communicate
ability to care for soldiers

counseling skills
initiative
leading from the front
unit success
integrity
attentive/respectful troops
background
leading by example
decisiveness/problem solving
confidence
dirty leadership (out from behind the desk leadership)
military bearing
responsibility
Skill Development Test scores
Officer Evaluation Report/NCO Evaluation Report criteria
ability to delegate
ability to plan/prioritize
education level
aggressiveness/boldness
motivating troops

The following responses were given by one or two interviewees: dependability, handling pressure, commitment to duty, open mindedness, judgement, pride, caring for families, maturity, punctuality, loyalty, courage, assessment by subordinates, goal setting, and caring about the mission.

When asked why they use these particular standards, the most frequent response was that these senior leaders developed the standards as a result of career experience during which they honed their judgement about characteristics of effective leaders. They also referred to standards set forth in leadership manuals by the Army; e.g., FM 22-100 and FM 22-101.

Feedback

Interviewees were asked about their methods for communicating to newly assigned soldiers how effectively these soldiers are performing as leaders. The formal ways of providing feedback include written counseling (most often every 30 days for NCOs) and quarterly Officer Evaluation Reports and NCO Evaluation Reports. After Action Reviews are conducted after every training exercise to review performance and are another way of communicating about the progress of leadership development. Also, some units carry out quarterly squad assessments.

The senior leaders who were interviewed seemed to consider the more informal methods of providing feedback to be the most effective. Most commonly this is on-the-spot, event-driven counseling, both positive and negative, given during senior leaders' day-to-day movements among the troops. More than one interviewee reported that most soldiers regard formal counseling as primarily negative, so that daily contact provides the best opportunity to give encouragement and guidance. Another means of providing feedback is a result of reviewing junior leaders' counseling statements to track and guide their developing skills as counselors. However, formal counseling is needed in order for soldiers to receive letters of concern, bars to promotion, bars to reenlistment, reductions in rank, disciplinary action, etc.

Leadership Development

All of the counseling opportunities described above contribute heavily to developing leadership skills. Other contributors include programs focused on leadership such as Officer Professional Development and NCO Professional Development. Another common program is Sergeant's Time. This program schedules a half day's time each week to allow NCOs to train soldiers in whatever areas they judge necessary. Senior leaders also consider education opportunities to have value for developing leaders. These opportunities include job-relatcd classes in military schools, correspondence courses, college courses, and courses offered as part of the NCO Education System.

Commissioned Officers are given extra duties, such as those of Safety Officer and Drug and Alcohol Control Officer, to expand their leadership skills. Developing leaders are also given one-level-up training and assignments are rotated to provide broad exposure to diverse tasks. In cases in which senior leaders identify areas in which developing leaders need extra support, they may pair the soldier with a strong, experienced peer, or assign the soldier to teach classes in areas identified to need improvement.

Improving Leadership Development

Although, overwhelmingly, the leaders interviewed were well satisfied with their units' leadership development practices, they responded readily when we asked them how they would improve these practices if they had complete freedom. The responses most frequently given were:

more training time and resources
more NCO Education System slots and other education opportunities
fewer training distractors
more field exercises
more unit autonomy to control training.

The following options were mentioned by one or two respondents:

more force-on-force training
more NCOs
improved Officer Basic Course instructors and curricula
better quality soldiers
better promotion systems
more training areas
formalized leader development practices

classes that cover battle-focused training, caring for soldiers, problem solving and job responsibilities

more Officer Development Programs and NCO Development Programs as well as improved classes in planning, conducting training and counseling

more specific feedback during counseling

more involvement of NCOs in military bearing, appearance and discipline

integration of individual development in leader training manuals

job-specific leader training programs

keeping new lieutenants in jobs for 3 years

placing more trust in officers and communicating more with subordinates

making junior leaders responsible for company combat operations

providing more mentoring

giving extra rewards to soldiers in leadership positions

making it easier to discharge unsatisfactory soldiers.

Discussion

If one examines the long list of leadership criteria that were gathered during this study, it seems clearly unreasonable to expect any individual to encompass every one of those excellent characteristics and abilities. However, the list communicates several things of significance to both our sponsor and developing leaders. First, it demonstrates areas where more sophisticated leaders place higher priorities and, by their emphasis, suggests areas in which shortfalls may occur most frequently. Second, it alerts developing leaders to characteristics and abilities that are more highly valued, as well as to a real need to become familiar with their commanders' preferences. Third, the listing includes trainable items (technical/tactical competence) together with items that are very much more difficult to train (integrity, boldness). In terms of including critical leadership skills in Army curricula, this is a consideration to be carefully weighed.

The wish list concerning improving leader development that respondents provided also contains significant information. It informs developing leaders about the limitations and frustrations experienced by their superior officers and units. It also signals the Army about ways to improve leadership skills development in unit settings.

References

Headquarters Department of the Army (1990). Military Leadership. Field Manual 22-100, Washington, D. C.

Headquarters Department of the Army (1985). Leadership Counseling. Field Manual 22-101, Washington, D. C.

Leadership Training for Cohesion and Motivation

Guy L. Siebold and Twila J. Lindsay

U.S. Army Research Institute for the
Behavioral and Social Sciences

ADDRESS CORRESPONDENCE TO:

Commander
U.S. Army Research Institute
ATTN: PERI-RL, Dr. Guy L. Siebold
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Commercial: (703) 274-8293
AUTOVON: 284-8293
FAX: (703) 274-8578
Home: (703) 548-0587

Presented at the 34th Annual Conference of the Military Testing Association, Del Mar, CA 26-29 October 1992.

Leadership Training for Cohesion and Motivation

Guy L. Siebold and Twila J. Lindsay

U.S. Army Research Institute for the
Behavioral and Social Sciences

Recent research has demonstrated that the levels of soldier cohesion and motivation are of critical importance to small unit performance and highly predictive of small unit performance, as measured by external observers, in extended field training exercises (e.g., Siebold and Lindsay, 1991; Siebold, in press). Significant correlations between soldier motivation and small unit performance, under the condition of strong leadership, have reached $r = .93$. Nonetheless, small units differ substantially in their levels of soldier cohesion and motivation, as well as in their performance. Further, median unit levels of cohesion and motivation in most unit samples are not overly strong (usually between 3 and 3.5 on a 5 point scale). Major demographic factors (e.g., soldier's racial/ethnic group, superior's racial/ethnic group, marital status, living on/off post, or having dependents) do not appear to be significantly correlated with soldier cohesion or motivation. Time in the unit, however, does appear to be negatively and non-linearly related to cohesion and motivation. The authors have concluded that the main determiners of small unit cohesion and motivation are small group processes and leadership from the chain of command.

Given the impact of leadership and leadership-influenced small group processes on soldier cohesion and motivation, the authors investigated the training which small unit leaders reported they had on building, assessing, and sustaining cohesion and motivation among their subordinates. The purpose of the investigation was to find out how leaders evaluated their training, what kind of training was considered the best, and how much of a perceived need there was for stronger training on motivating subordinates and building unit cohesion

Method

Sample. The sample consisted of small unit leaders in a light infantry battalion, an airborne battalion, and a mechanized infantry battalion. Specifically, respondents were 58 squad leaders, 19 platoon sergeants, 17 platoon leaders, and 8 company commanders from the three battalions for a total N = 102. Only 36 of the 58 squad leaders had completed the Basic NCO Course; only 14 of the 19 platoon sergeants had completed the Advanced NCO Course; 16 of the 17 platoon leaders had completed the Officer Basic Course; and 6 of the 8 company commanders had completed the Officer Advanced Course. About one half of the squad leaders and platoon sergeants had airborne training, regardless of battalion. Almost all the platoon leaders and company commanders received both ranger and airborne training.

Procedure. Researchers met with intact platoon leadership teams (i.e., the squad leaders, platoon sergeant, and platoon leader from the same platoon as a group). Company commanders were interviewed individually. Meetings took place in dayroom or office settings and usually lasted a little over one hour. At the beginning of each session, each leader in the session was given a brief questionnaire (example at Appendix A) so that he could describe his formal training and his evaluation of his training (formal or informal) on motivating subordinates, building cohesion, and assessing their levels in units. After the questionnaire administration, the researchers conducted interviews on the same topics and on how to improve leadership team motivation and cohesion.

Results

Questionnaires. The questionnaire data showed that the small unit leaders rated their formal school training on motivating their subordinates and on building cohesion among their subordinates as generally effective, with response means of 6.7 (*SD* = 1.8) and 6.0 (*SD* = 1.8) respectively, using a 10 point rating scale. The leaders rated their on-the-job training while in units a little higher in effectiveness, with the response mean for motivating subordinates at 7.2 (*SD* = 1.8) and for building cohesion among their subordinates at 6.4 (*SD* = 2.1). The leaders rated the effectiveness of their overall training at assessing the level of unit motivation at 6.8 (*SD* = 1.9) and at assessing the level of unit cohesion at 6.1 (*SD* = 1.8). On the other hand, the leaders indicated that it would be of high value to them to be able to better assess the levels of motivation and cohesion among their subordinates, with response means of 8.2 (*SD* = 1.5) and 8.1 (*SD* = 1.8) respectively.

There were no major differences in the leaders' ratings if broken out by the leaders' battalion (i.e., no battalion effect). However, there were some discernible patterns in the ratings if broken out by leader duty position (see Table 1). Platoon leaders and company commanders rated their on-the-job training in units as more effective than their NCOs did. Leaders in all positions (but platoon leaders especially) indicated that the ability to better assess the levels of motivation and cohesion in their subordinates would be of high value to them. Most leaders rated their school and unit training concerning motivation as more effective than their training concerning cohesion.

Interviews. In the interviews, the leaders expressed a less positive view of the effectiveness of Army schools in training them to motivate subordinates or build unit cohesion. The leaders typically stated that the content of the course material was too general, too little, and usually too underemphasized to be effective. They wanted more "meat" in the courses and a focus on what would help them in future assignments. Some of the courses presented a synopsis of theories (e.g., of motivation), but the courses did not sufficiently integrate the theories with

Table 1

Means for Motivation and Cohesion Training Scales by Duty Position

	Duty Position			
Motivation and Cohesion Training Scales	SL	PS	PL	CC
Rate the effectiveness of the training you received in formal Army Schooling on--				
motivating subordinates	6.6	6.6	7.4	6.4
building cohesion	5.9	5.2	6.8	6.5
Rate the effectiveness of on-the-job training you have received while serving in units on--				
motivating subordinates	6.9	6.8	7.9	8.6
building cohesion	6.2	5.7	7.2	8.3
Rate the effectiveness of the training you have received over your career on--				
assessing level of unit motivation	6.7	7.1	6.7	6.4
assessing level of unit cohesion	6.0	5.8	6.8	6.4
Of how much value to you would it be to be able to better assess--				
level of motivation in subordinates	8.0	8.1	8.8	8.1
level of cohesion among subordinates	7.9	8.1	8.9	8.1

Note. Duty positions were: SL (squad leader), PS (platoon sergeant), PL (platoon leader), and CC (company commander). Responses were based on a 1 (low) to 10 (high) rating scale. Most *SD*s were about 2.0, + or - .5. Most ranges were about 7 scale points, + or - 2.

the day to day problems the leaders would face in their units. On the other hand, most officers raved about the effectiveness of the (special) training they received in Ranger School (and to some extent Airborne School) in providing the experiences which facilitated the leaders' understanding of the importance and value of soldier motivation and cohesion. But the value was in the experience, not in the direct school instruction concerning motivation or cohesion. It is probably fair to summarize the comments of the leaders NOT by stating that they felt the school training was ineffective BUT by stating that they felt the school training was incomplete and not sufficiently focused on their upcoming needs as leaders. And the complete segregation, for example, of almost all formal school training by rank inhibited mutual understanding and shared views on training concerning soldier motivation and cohesion. Some leaders suggested pairing platoon leaders and platoon sergeants, who would serve together in their next assignments, during the Officer Basic and Advanced NCO Courses at Fort Benning, GA so that they could take joint subcourses together with integrated training materials.

In the interviews (as also reflected in the questionnaire responses), the leaders stated that their best training for motivating their subordinates and building cohesion came from their experiences in troop units. Most leaders reported that they benefited a great deal from other leaders who acted as (positive or negative) role models or occasionally as mentors. The leaders also reported that they benefited a lot from informal peer group discussions, WHEN they had time and facilities for them. Some reported that higher leaders (e.g., first sergeants or company commanders) in their units had tried to initiate some within-unit-leadership training for their NCOs and Lieutenants. However, usually the implementation of the within-unit training was overridden by training schedule problems, turbulence, and various details or taskings. Overall, it is probably fair to summarize the comments of the leaders by stating that the majority of leadership training on motivating subordinates and building unit cohesion comes from leaders informally teaching leaders and from the experiences of being a leader in a troop unit. Needless to say, the quality and quantity of this training is uneven and not integrated across leaders. It is also fair to say that while many leaders would like to be able to better manage and assess motivation and cohesion in their units, they do not view these areas as the source of their greatest challenges or difficulties (e.g., compared to turbulence, command climate problems, and coordination with the higher echelons).

Discussion

Most leaders know that soldier motivation and unit cohesion are important. Recent research has documented their extensive importance to small unit (field training excercise) performance. The current research effort and the substantial variation in unit cohesion and motivation suggest that leaders need to be better trained and assisted to improve and sustain soldier motivation

and unit cohesion. The simple question is how can the Army do this in an effective and cost-efficient way?

While the question is simple, the answer is complex, as is leader training and development. The first step is that leaders and appropriate Army institutions must recognize that improvement in these areas is needed and should be assigned some priority within the many issues and problems facing the military services today. As part of that recognition, leaders should understand that the capability to enhance soldier motivation and small unit cohesion through better leadership training is an opportunity to greatly enhance unit performance without a substantial increase in funds, resources, or personnel. Improvements in these areas can be part of the answer to maintaining needed capabilities within the context of downsizing and budget reductions.

Potential fixes to improve leadership skill in motivation and cohesion are many and not necessarily difficult or costly. Leadership training (e.g., in the competency areas of "Teaching & Counseling", "Soldier-Team Development", "Professional Ethics", "Supervision", and "Communications") should strongly emphasize and integrate aspects impacting upon motivating subordinates and building cohesion. Field manuals, pamphlets, training circulars, and military qualification standards and materials can provide more focused and relevant information and ideas in their next updated versions. Formal Army courses (e.g., Primary Leader Development Course, Basic NCO Course, Advanced NCO Course, Officer Basic Course, Officer Advanced Course, and the Battalion and Brigade Pre-Command Courses) and more specialized training courses can be examined individually and as a structured system to improve their substance and integration (i.e., do they have meat in them, and do they reflect a consistent, sequential and progressive, coherent picture?). High level unit leaders need to consider the impact of their decisions, taskings, communications, and command climate on lower level motivation and cohesion and get real feedback on the impact. In addition, these leaders need to consider 1) implementing a stronger and more organized system of mentoring, 2) establishing a system for getting periodic objective feedback from subordinate leaders about their units (simple, cost-effective tools are available for this), and 3) facilitating leaders working with leaders to enhance unit cohesion, soldier motivation, and leadership.

References

Siebold, Guy L. (in press). "The Relation Between Soldier Motivation, Leadership, and Small Unit Performance." In H. F. O'Neil, Jr. and M. Drillings (Eds.), Motivation: Research and Theory. Hillsdale, NJ: Lawrence Erlbaum.

Siebold, Guy L. And Twila J. Lindsay (1991). "Correlations Among Ratings of Platoon Performance." Proceedings of the 33rd Annual Conference of the Military Testing Association, San Antonio, TX. (28-31 October); pp. 67-72.

APPENDIX A
LEADER TRAINING & DEVELOPMENT SURVEY--SOLDIER-TEAM DEVELOPMENT

Circle your duty position: Company Commander Platoon Leader
Platoon Sergeant Squad Leader

Circle the formal Army schooling (resident or correspondence course) you have received (circle all that apply):

BNCOC, ANCOC Ranger, Airborne OBC, OAC, CAS3

Circle the Army Publications (addressing motivating subordinates and building cohesion) which you have studied (at any time):

FM 22-8 Unit Cohesion
FM 22-100 Military Leadership
FM 22-102 Soldier Team Development
STP 21-II-MQS (MQS II) Manual of Common Tasks
Pam 350-2 Training, Developing and Maintaining Unit Cohesion

For the questions below, use a scale from 1 (lowest) to 10 (highest), with 5 being average. Circle your response.

1. Rate the effectiveness of the training you received in formal Army schooling

--on motivating subordinates:
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

--on building cohesion among subordinates:
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

2. Rate the effectiveness of on-the-job training you have received while serving in units

--on motivating subordinates:
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

--on building cohesion among subordinates:
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

3. Rate the effectiveness of the training you have received over your career

--on assessing the level of unit motivation:
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

--on assessing the level of unit cohesion:
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

4. Of how much value to you would it be--to be able to better assess

--the level of motivation in your subordinates?
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

--the level of cohesion among your subordinates?
(Very Low) 1 2 3 4 5 6 7 8 9 10 (Very High)

"NEGATIVE" LEADERSHIP AS REPORTED REASON FOR LEAVING MILITARY SERVICE[1]

Mary Sue Hay and Trueman R. Tremble, Jr.
U.S. Army Research Institute
for the Behavioral and Social Sciences

It is not uncommon to find that soldiers exiting military service frequently choose an item like "quality of leadership" when given a list of possible reasons for leaving. One recent interpretation (McGee, 1992) suggested that such a finding indicates a need for better leader development. That is, the leadership education and training system is not adequately preparing individual leaders to perform their leadership functions. While this interpretation may be correct, endorsements of single items are open to multiple interpretations. It is, of course, for this reason that scales composed of a single item, or only a few, are less reliable than longer scales, where each individual item is presumed to tap into a slightly different aspect of a larger construct.

Thus, endorsement of a single item such as "quality of leadership" could indicate dissatisfaction specifically with the individual's immediate supervisor, a broader dissatisfaction with the leadership of the post, or even a very diffuse dissatisfaction with the total Army, since leaders represent the larger service to their subordinates. This ambiguity is similar to the traditional problem of determining leadership effectiveness--for example, whether to measure leadership behaviors or group outcomes. The research described here explores the meaning of such a single-item endorsement.

Method

Source of Data

The Army Career Transitions Survey (ACTS) is an exit survey for separating and retiring active duty personnel, administered as soldiers go through out-processing at the Army's 40 major Transition Points. The basic ACTS questionnaire is a single-sheet form, consisting of a few demographic items and 45 satisfaction items which are rated on a four-point scale (Very Satisfied, Satisfied, Dissatisfied, and Very Dissatisfied). These 45 items cover various aspects of Army life, including such considerations as job challenge and fulfillment, housing, benefits, support and recreational facilities, health care, family services, and job assignments. In addition to rating each of the 45 items in terms of satisfaction, soldiers select the one item which made them first think about leaving the Army.

[1] The views, opinions, and findings in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other authorized documents.

One additional page can be added to the basic ACTS questionnaire in order to provide data on topics of special interest. Issues arising from Operation Desert Shield/Storm (ODS/S)--including its impact on Reserve/Guard enlistment intentions, advice to others on joining the military, and overall satisfaction with the Army--were addressed by the supplemental questions discussed in this paper.

Administration of the ACTS began in April, 1991 and is continuing on an indefinite basis. Data used in the analyses reported here were collected during the period from April through December, 1991, and represent 13,275 completed questionnaires.

Analyses

Because it is likely that perceptions of the Army vary somewhat depending on the reason for separation (e.g., retirement vs. involuntary separation), as well as differing by Military Personnel Category (e.g., Officers vs. Enlisted), the sample was limited to enlisted soldiers who were leaving the Army after having fully completed their expected terms of service. This provided a relatively homogeneous sample of 9172 soldiers.

The 45 satisfaction items were factor analyzed, using principal components analysis. Principal components analysis is particularly appropriate for an initial pass through the data, where the general objectives are data reduction and interpretation (Johnson & Wichern, 1982). Respondents who indicated that an item was "Not Applicable" were dropped from the analysis, as were respondents with missing data on any of the 45 items, thus reducing the sample size for the principal components analysis to 1103. However, this sample size should be more than adequate for a factor analysis of 45 items.

Five factors were retained, based on a minimum eigenvalue of 1.0 and a scree test. To gain more interpretable results, the factor structure was rotated using an Equamax (orthogonal) method. The Equamax rotation was selected because it is considered to be particularly good at breaking up the large first factor that sometimes results from principal components analysis (Cattell, 1978).

Based on the rotated factor pattern, five scales were constructed from the items associated with the five factors. Items with loadings of less than .50 for a factor were completely eliminated, and items with cross-loadings were eliminated unless there was a difference in loadings of at least .10. In those few cases, the item was assigned to the factor with the higher loading.

Scale scores were calculated for each respondent, and were correlated with items from the ACTS supplement which asked about soldiers' advice to potential military enlistees, likelihood of joining a Reserve or Guard unit, overall satisfaction with the

Army, and satisfaction with their leader's technical competence and concern for soldiers. Sample sizes for the correlational analyses ranged from 8257 to 8981.

Results

The rotated factor pattern is shown in Table 1, with a description of each item, as well as our subjective name for each factor. The five factors together explained 61% of the variance, with each factor of the rotated factor structure accounting for approximately 12% of the total variance. This "evenness" of the factors in explaining the variance is chiefly because of the Equamax rotation, which tends to balance out the factors. The fourth factor, "Leadership/Leadership Climate," is the primary focus here. This factor explained 11.96% of the total variance.

Correlations of the scale scores (derived from the factor analysis) with seven relevant items from the ACTS supplement are shown in Table 2. Both scales and items were scored in positive directions. That is, the higher the scale score, the higher the satisfaction, and the higher the item score, the higher the degree of positive advice, likelihood of joining a Reserve/Guard unit, or satisfaction with the Army and with leaders' technical competence and concern for soldiers' welfare.

Note that all correlations in Table 2 are positive. This means that the higher the satisfaction, the greater the likelihood that the soldier would give favorable advice, join a unit, be satisfied with the Army experience, or believe that Army leaders were technically competent and concerned about their soldiers' welfare. Note also that the Leadership/Leadership Climate scale is the strongest correlate of six of the seven items in Table 2. It is also the second strongest correlate of overall satisfaction with the Army experience, surpassed only by Professional Development.

Discussion

"Quality of Leadership and Management" has consistently received one of the lowest satisfaction ratings of the 45 items in the ACTS questionnaire (McGee, 1992). In addition, this item has been one of the most frequently selected when soldiers choose the reason which first made them think of leaving the Army. The question, of course, has always been: What, exactly, do these soldiers mean what they think of "leadership"? This research provides some insight into that question.

Based on our results, it appears that soldiers really do mean a unique constellation of leadership behaviors, rather than some diffuse dissatisfaction with the Army. The first four items on the Leadership factor--Quality of Leadership and Management, Supervisor Competence, Respect from Superiors, and Recognition for Accomplishments--reflect leadership characteristics which are primarily, if not exclusively, under the control of the

Table 1

Rotated Factor Pattern

Factor					
Army Environment	Living Conditions	Family Care	Leadership/ Leadership Climate	Professional Development	Questionnaire Item
64	21	30	23	24	Availability of Army Housing
63	24	27	29	15	Amount of Family Separation
60	8	21	30	34	Promotion/Advancement Opportunity
59	17	36	18	20	Special Pay (Bonuses)
56	15	37	28	20	Amount of Basic Pay
19	**66**	18	21	22	Amount of Overseas Duty
33	**60**	27	13	28	Quality of Government Housing
22	**60**	24	21	34	Stateside Living Conditions
10	**59**	18	30	14	Geographic Location of Job
31	**58**	25	21	22	Number of PCS Relocations
18	**55**	27	24	17	Overseas Living Conditions
23	10	**80**	26	12	Quality of Family Health Care
28	16	**73**	23	15	Quality of Military Health Care
39	10	**68**	25	15	Availability of Family Health Care
12	38	**65**	20	30	Quality of Family Service Centers
10	41	**63**	17	29	Dependent Facilities/Schools
24	7	23	**73**	22	Quality of Leadership/Management
35	10	18	**67**	24	Supervisor Competence
7	9	17	**62**	39	Amount of Respect from Superiors
35	10	15	**59**	26	Recognition for Accomplishments
12	42	22	**57**	17	Amount of Regulations/Discipline
39	31	17	**51**	15	Amount of Personnel to do Work
16	33	27	**50**	27	Number of Quick Response Tasks
25	21	19	29	**68**	Use of Skills/Training on Job
36	26	15	24	**66**	Assignment to Leadership Jobs
-6	27	23	36	**59**	Level of Job Fulfillment
49	15	22	25	**59**	Assign. with Tech/Prof Development
44	11	19	34	**56**	Control over Job Assignments

Note. Loadings are multiplied by 100 and rounded to the nearest integer.

individual's immediate superiors. The other three items--Amount of Regulations/Discipline, Amount of Personnel to do Work, and Number of Quick Response Tasks--are probably less directly under the supervisor's control. We believe, however, that these items are associated with leadership behaviors. That is, these are aspects of the soldier's job where the superior can have some impact, if not absolute dominion, because it is the immediate leader who transmits these facets of the Army to the soldier.

Table 2

Correlations of Scale Scores with Items from ACTS Supplement

	Scale				
	Army Environ.	Living Cond.	Family Care	Leader-ship	Prof. Devel.
Advice to friend on seeing a military recruiter	.28	.24	.22	.35	.32
Advice on joining the Army	.24	.21	.18	.32	.29
How likely to join Army Reserve Unit	.11	.08	.09	.14	.12
How likely to join Army National Guard Unit	.06	.05	.09	.10	.08
Overall satisfaction with the Army experience	.30	.29	.26	.45	.47
Belief in Army leaders' technical competence	.31	.19	.22	.55	.38
Belief in Army leaders' concern about soldier's welfare	.32	.21	.24	.53	.38

Note. All correlations are significant at *p* < .001.

It is worth noting that we explored a number of factor extraction and rotation methods, with varying numbers of factors, before finally selecting the one reported here as having the most interpretable results. These methods included principal factors extraction, as well as principal components analysis, and both orthogonal (Varimax, Equamax) and oblique rotations (Promax, Harris-Kaiser). In all cases, we found that the first four items were consistently grouped together and that this cluster of items always loaded first on a factor. The tendency of these four items to cluster together, regardless of the specific factor solution, suggests a certain stability in the meaning of "leadership."

The stability and validity of the Leadership factor is also supported by the finding that satisfaction with leadership was most strongly correlated with the belief that the individual's Army leaders were technically competent and concerned with soldiers' welfare. These are usually considered the two major dimensions of leadership and a relationship with "quality of leadership" is precisely what we would expect to find.

Overall, these findings suggest that these leadership behaviors may have a strong influence on soldiers' perceptions of the Army. The correlational analyses show a clear trend in which satisfaction with leadership is more strongly related than any of the other factors to several items which have implications for the Army.

Satisfaction with leadership tends to be associated more closely than other factors with the extent to which veterans will prove to be valuable ambassadors for recruiting, by their advice to potential recruits. This holds true, also, for the relationship between satisfaction with leadership and the likelihood of joining an Army Reserves or National Guard unit after separation, but the relationships are much weaker in this area.

Most interesting, perhaps, is the finding that soldiers' ratings of overall satisfaction with their Army experience tend to be more closely related to the Leadership factor than to other factors emphasizing such basics as pay, family benefits, and living conditions. Only satisfaction with professional development opportunities is more strongly associated with overall satisfaction, and the edge is very slight.

Given these findings, it appears that emphasizing the development of high-quality leadership skills, in both technical and affective domains, could well prove beneficial to the Army. Satisfaction with leadership seems to be a major part of veterans' overall perceptions of the Army, satisfaction with their Army experience, and future behaviors.

References

McGee, M. (1992, June). The fundamental disconnection in Army's leadership assessments. Army, p. 18.

Cattell, R. B. (1978). The scientific use of factor analysis in behavioral and life sciences. New York: Plenum Press.

Johnson, R. A., & Wichern, D. W. (1982). Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice-Hall, Inc.

The Naval Junior ROTC Leadership Academy

By

Donald E. Dorin, Ph.D.
NJROTC Cadet Education
Chief of Naval Education and Training
NAS Pensacola, Florida

Shortly after noon some participants begin arriving by private automobiles, a number of others show up in school mini-vans, but most of them disembark from naval C-9 aircraft. During the month of June, this scenario is being played at several naval bases throughout the country. Its the time of the year when 900 Navy Junior Reserve Officer Training Corps (NJROTC) cadets from 226 NJROTC units throughout the nation give up seven days of their high school summer vacation to attend the Naval Junior ROTC Leadership Academy.

The NJROTC Leadership Academy is like the NJROTC itself, a special program--one of self-development and opportunity for young men and women who are to become cadet leaders in their respective high school NJROTC units. Cadets selected for the Academy engage in an exciting and challenging course of study, with the principal goal of developing citizenship, leadership, teamwork, and high standards of personal appearance.

For the past several years the Naval Junior ROTC Leadership Academy training has been held at Pensacola, Norfolk, Annapolis, Great Lakes, Corpus Christi, San Diego and Alameda. Participating cadets understand they will remain in a controlled atmosphere during their stay, although some limited free time may be available. Cadets are informed that the purpose of the program is to prepare them for leadership roles in their respective units. All NJROTC cadets are thoroughly counseled on the physical and disciplinary demands of this accelerated training. They are made aware of what is expected of them at the Leadership Academy before they decide to attend, and told there will be very little time for socializing and recreation.

Academy instructional staff are solicited from area NJROTC units on a voluntary basis. Efforts are made to provide for a balanced instructional staff in an attempt to avoid having "too many Chiefs (NJROTC naval science instructors) and not enough Indians (associate naval science instructors)." Personnel selected have a minimum of 2 years as an instructor in the NJROTC program.

Instructors selected for Leadership Academy are provided with a copy of an Academy Administrative Manual, Instructor Guide, Cadet Guidebook, and a proposed training schedule. With these documents available, instructors have ample time to completely familiarize themselves with the subject matter, the instructional references, and the training aids. Each instructor annotates and customizes their lesson guides for ease of use.

The nature of the Academy requires instructors and escorts to set appearance and behavior standards that cadets can only hope to emulate. At no time is profanity used by instructors when addressing, correcting, or inspecting cadets. Also, instructor's uniforms are expected to be impeccable at all times, and worn correctly. Attention to detail is essential.

During the seven days the cadets attend the Leadership Academy, they attend classes and are tested on their ability to compete in a stringent physical and mental environment. Classes consist of subjects on physical fitness, practical leadership, advanced military drill, orienteering, service etiquette and social manners, self-awareness and NJROTC subject areas which familiarize the cadets with the many duties in the administration of an NJROTC unit.

Cadets completing the entire leadership syllabus in a highly successful manner are awarded a distinguishing silver shoulder cord. Those cadets who, for a variety of reasons fail to meet minimum requirements in any major instructional area, may receive a certificate of completion but they will not receive the silver cord. Cadets who arrive physically unfit, are unmotivated, or project an unacceptable attitude, will not receive the silver cord either.

Cadets selected for the Leadership Academy must meet the following criteria:

1. Be a second or third year male or female cadet expected to return and complete the junior and/or senior year in the unit, or be an outstanding first year student slated for a leadership position in the unit.

2. Have no record of disciplinary problems in the unit or school.

3. Have a high school grade point average of 2.5 or higher on a 4.0 scale.

4. Be highly motivated and well trained in the basics of military drill.

5. Demonstrate to the unit's naval science instructor (NSI) an aptitude for the NJROTC that will ensure success in a military training environment.

6. Be in good physical condition as determined by participation in school physical activities. A cadet must have earned the Physical Fitness ribbon before selection.

7. Having demonstrated to the NSI, no more than 6 weeks prior to the Academy, that the cadet can perform the required number of sit-ups and push-ups, and achieve the required time in the 1.5 mile run.

Al] exercise requirements are expected to be completed on the second full day of the Academy. Failure to meet minimum times and numbers may result in a cadet's immediate departure from the Leadership Academy at his/her own expense. Anyone not passing the PT requirements on the second day is brought before a review board of at least three Academy staff members to consider extenuating circumstances. The board determines the proper course of action. Cadets failing to complete the qualification requirements do not receive the silver cord awarded at graduation.

Naval science instructors are expected to test their cadets to assure that minimum requirements can be achieved prior to Academy departure. Instructors must also be careful not to bring or send cadets who may have a medical or physical fitness problem which would prevent participation. Cadets are required to complete medical and dental accident insurance data on the NJROTC Standard Release Form. The release form is rescreened upon arrival. Should screening indicate that a cadet is not fit for training or the insurance data is lacking, he/she is not accepted. Particular attention is paid to cases where the cadet is taking medication that could limit participation.

Personnel and barracks inspections, and proper grooming standards play an important part in the Academy training. Male cadets have NJROTC regulation haircuts, female cadets are not allowed to wear make-up, and NJROTC uniforms are required for the entire period of training. Uniforms, quarters, bunks, lockers, and personal gear are inspected at least once a day. Cadets are expected to improve their appearance and give attention to detail as the course progresses. With the number of cadets involved, it is mandatory that personal hygiene is of prime importance. All cadets shower, and shave (if necessary) daily. Failure to meet standards results in a reduction in overall grades for the Academy.

When conducting the Academy, certain leadership activities are mandatory, and certain activities are optional, depending upon base locale and available facilities. The following items are mandatory and have priority when preparing Academy training schedules:

1. The physical training requirements for completion of the Academy.

2. The full practical leadership series including:

   a. Leadership characteristics demonstrated by use of VHS videocassette dramatic vignettes.
   b. Service etiquette and social manners.
   c. The leadership field event (orienteering).
   d. Platoon drill and ceremonies.
   e. Obstacle course.

There are other military leadership activities which can be scheduled, time permitting, but are optional and considered secondary to the aforementioned primary practical leadership training. The NJROTC Leadership Academy staff members are guided by the Academy mission and objectives in all Academy matters. The four main objectives for the Leadership Academy are:

1. To promote habits of orderliness and precision, and to develop respect respect for constituted authority.

2. To challenge and motivate cadets to push toward their physical and intellectual limits. Cadets will continually be called upon to meet high standards of personal appearance, self-discipline, and meticulous attention to detail.

3. To instill a high degree of personal honor, self-reliance, and confidence in each cadet by presenting a military environment in which cadets will be forced to rely upon themselves and their shipmates to study, work, and learn.

4. To enhance the basic attitude, knowledge and skills required to practice the art of leadership.

Because the main objective of the Leadership Academy is to prepare selected NJROTC cadets for leadership roles, a significant amount of time is spent teaching the cadets certain leadership characteristics which aren't necessarily available nor normally taught at their home unit. The Academy instructors have lesson guides which outline 22 specific leadership characteristics to be presented and discussed in the classroom. The individual lesson guides have been developed in a style that provides for a logical lecture and discussion sequence of content, and seeks to elicit active verbal participation by the Leadership Academy cadets. Included with each lesson outline is a brief VHS videotape vignette designed to be viewed as a supplement to the classroom instructor's lecture. It is intended that with the presentation of the classroom lecture and the showing of the vignette, no more than twenty to thirty minutes of classroom time be given to an individual leadership characteristic. The 22 cadet leadership characteristic outlines and videotape vignettes are as follows:

Leadership Topic

1. Leadership: Making Things Happen Through People.
2. Issuing Orders to Subordinates.
3. Providing Effective Feedback to Seniors and Subordinates.
4. Developing Loyalty: A Two-way Street.
5. Establishing Goals.
6. Disciplining Inappropriate Behavior.
7. Developing a Positive Mental Attitude.
8. Delegating Authority.
9. Making a Decision.
10. Maintaining Integrity.
11. Taking the Initiative.
12. Managing Time and its Importance.
13. Developing Economical Working Habits.
14. Being Physically Fit.
15. Rewarding Accomplishments.
16. Communicating and Keeping and Open Mind.
17. Planning for the Unexpected.
18. Having Self Discipline and Relationships with Subordinates.
19. Overcoming Bias and Prejudice.
20. Maintaining Safety and the Minimization of Accidents.
21. Developing Teamwork and Coordinating Operations.
22. Monitoring the Plan.

The Instructor Guide and corresponding videotape vignettes have been developed in a sequential format for ease of use. The individual lessons are divided into several sections: (1) the first three sections are lecture/discussion topics which the instructor presents to the class allowing 2-3 minutes for each lecture/discussion topic; (2) the fourth section is a discussion topic that takes about 2 minutes to set a frame of reference from which the cadets are to view the brief videotape; (3) the fifth section on the outline uses no more than 6 minutes for the actual classroom showing of the leadership vignette and a short period of time in which the cadets discuss the leadership characteristic and the vignette among themselves; (4) the sixth section of the lesson takes 10 minutes or less for the cadets to summarize the leadership style exhibited by the cadet(s) in the vignette; and (5) the final section takes about 3 minutes for the instructor to summarize the unit topic and the learning outcomes from this lesson. The Instructor Guide is not designed as a verbatim script to be read by the instructor to the cadets, but rather as a basic outline for the instructor to follow in his/her lecture or discussion of the leadership characteristic. Most instructors paraphrase the wording contained in the written outline, and expand on the topic with personal experience and past training.

Instructors begin the academic portion of the Leadership Academy by establishing contact with the cadets. They explain how the lessons and the videotapes will be presented. Some of the vignettes dramatize a positive example of the leadership characteristic, some of them dramatize a negative example, and a few present both a positive and negative characteristic. They explain the value of learning the various leadership characteristics and their effect on helping the cadets to assist the NSI in the running of the NJROTC unit. Cadets are informed that they will be expected to exhibit this kind of behavior should he or she be placed in a leadership position in the coming school year. The class is divided into sections of 5 to 7 cadets each, depending upon the number of cadets in the class, and told that each group will have a spokesperson who will summarize their group's general consensus of opinion on the leadership style exhibited by the cadet(s) in the videotape vignette. To prevent one cadet from being the only speaker, a different cadet serves as the spokesperson for each leadership characteristic.

Cadets also spend a portion of their time at the Leadership Academy studying the survival skill of orienteering. Learning to find their way in unfamiliar countryside is that aspect of the leadership training whereby the cadets combine their total talents of common sense, intelligence and physical abilities to solve a land navigation problem and learn the skills of teamwork. Cadets are divided into groups of 5 or 6 members, briefed on the grid system of map construction and use, shown the general layout of the particular course, explained the procedure for reading the location of an object, then sent out in different directions alternately. Instructors stress that in an orienteering situation, teamwork is essential. All team members must share all their skills for the benefit of the team's accomplishment.

Attending classes where service etiquette and social manners are taught is a popular part of the leadership training with the cadets. Instructors emphasize that good manners are important to get along in our society, but are essential if one is to get ahead in life. Cadets study the importance of making a good first impression, and the importance of personal cleanliness and hygiene. They learn the value of keeping timely appointments, and that promptness and responsibility go together. Instructors explain to the cadets how to conduct themselves with decorum at all times. To teach the customs, courtesies, and etiquette relevant to a formal sit down dinner, instructors use the base mess hall to demonstrate to the cadets the proper dinner manners, eating habits and table settings.

On the average two sessions a day at the NJROTC Leadership Academy are devoted to military drill. In addition to these scheduled sessions, opportunities to teach drill are available when cadets move to and from classes, meals and various activities. Drill instructors seize every opportunity to accomplish the goals and objectives for military drill. The purpose of any training at the Leadership Academy is of course "Leadership," and all drill activities emphasize those purposes of military drill that particularly aid in the development of leadership techniques.

The objectives for drill training is to enhance the individual cadet's opportunity to gain confidence in his/her leadership ability. Cadets are taught the necessary skills to lead other cadets in military drill. Emphasis is placed on instruction in proper methods of giving commands, such as correct voice inflection and projection, the timing and use of proper cadence, and the training in giving remedial instruction to other cadets. During the week all cadets are provided the opportunity to lead other cadets in squad and platoon close order drill in order to gain the confidence and assurance necessary to assist in instruction of junior cadets in their respective units.

The schedule for military drill sessions is "open ended," that is, there are no strict expectations of progress for each session. Although the cadets all have at least a year of experience in drill, their skills will vary and some additional instruction in basics may be necessary at the beginning of the training cycle.

Drill competitions are scheduled often between platoons to add to the competitive spirit and pride in the individual platoon. Instructions for these competitions are issued to drill instructors early in the training cycle and drill instructors are encouraged to allow the platoons to practice in order to be at their best.

An important aspect of the training at the Leadership Academy includes verbal and psychological stress imposed on the cadets by the Academy instructors. This does not include profanity or other display of verbal abuse. This helps instill within the cadets, self-discipline, subordination, and the ability to work towards important Academy goals. Despite the numerous upsetting distractions caused by this psychological stress, it contributes towards building cadet self-confidence, and plays a part in giving cadets who complete the training, a feeling of having succeeded at a difficult series of tasks.

Leadership Academy graduates are not allowed to return to their units with the notion that yelling, harassment, and the imposition of stress constitute good leadership. Cadets in leadership positions who try these methods on subordinate cadets, create problems for all concerned. Therefore, at the end of the Leadership Academy, a special effort is made which emphasizes to the graduates the purpose of the stress, the shouting, and the petty abuse they experienced during the Academy. A final review of the leadership principles and concepts discussed throughout the Academy reinforces the idea that a good leader does not lead by frequent intimidation, and that this is is not the way a good leader performs.

Formal graduation exercises conducted at the end of training are open to the public and attended by high ranking dignitaries. The public perception of how well a military group is trained is often based on the success of their performance in military drill at these functions. Therefore training time is set aside for the practice of the graduation pass-in-review.

The Naval Junior ROTC Leadership Academy is a physically and mentally demanding but very rewarding experience. Those cadets who come well prepared receive a tremendous benefit from this program. Over the past several years the numbers of cadets receiving leadership training has continued to increase. Graduates of the Leadership Academy return to their units well prepared to assume their NJROTC responsibilities as cadet leaders.

LEADERSHIP RESEARCH AT THE AIR FORCE HUMAN RESOURCES DIRECTORATE

Thomas W. Watson
Leasley K. Besetsny

Armstrong Laboratory

Leadership is an important issue in the United States military services. The Army is the lead service in leadership research and has been involved in this type of research for many years, for the effective functioning of cohesive, motivated soldiers, working as a team, is critical to their mission. However, the other services are increasing their interest in leadership research, and interservice collaboration is becoming more common. This paper addresses leadership research at the Air Force Armstrong Laboratory's Human Resources Directorate, the personnel research arm of the Air Force. However, it should be noted that the United States Air Force Academy also conducts leadership research (USAFA, undated).

The need for Air Force leadership research is underscored by Scott (1987), who indicated that dissatisfaction with leadership is often a cause of pilot retention problems. Recent surveys administered by HQ AFMPC and HQ USAF also report considerable dissatisfaction with senior leadership and concerns that leaders may not be sufficiently sensitive to people or their needs.

The Leadership Effectiveness Assessment Profile (LEAP)

As an initial leadership research project, begun in the late 1980's, the Human Resources Directorate developed and field-tested the Leadership Effectiveness Assessment Profile (LEAP), a biographical inventory designed to measure leadership and management potential, ability to function well in team situations, propensity for commitment to the Air Force, and related attributes. Initially, both officer and enlisted versions of the LEAP were planned. The officer version underwent initial development (Appel, Grubb, Shermis, Watson, & Cole, 1990) and iterative field testing and refinement (Appel, Quintana, Cole, Shermis, Grubb, Watson, & Headley-Goode, in press). The enlisted version underwent taxonomic and item pool development, but field testing and refinement await additional evidence of the utility of the officer LEAP (Appel, Grubb, Elder, Leamon, Watson, & Earles, 1991). The officer LEAP underwent six iterations to date, as summarized in Table 1. Sample sizes ranged from small (61) to moderate (673). Cadets or junior officers were used as subjects.

Table 1

Overview of LEAP Field Testing

LEAP Version	Population Sampled	Sample Size	Location	Type of Administration	Type of Feedback
0-1	Junior officers	61	Randolph, Brooks AFBs	One-on-one oral	Face-to-face
0-2A	Junior officers	71	Keesler AFB	Small group	Focus groups
0-2B (ROTC)	1990 ROTC summer cadets	345	Lackland AFB	Large group	Questionnaire
0-2B (OTS)	OTS cadets	72	Lackland AFB	Large group	Questionnaire

0-2C (OTS)	OTS cadets	156	Lackland AFB	Large group	Questionnaire
0-2D (ROTC)	1991 ROTC summer cadets	673	Lackland, Lowry, McConnell, Plattsburgh, Vandenberg AFBs	Large group	None

In its most recent form, the LEAP is composed of 12 scales measuring leadership constructs, plus a faking detection scale. Scales range from 8 to 23 items and the total number of items in the latest version is 120. The various scales investigated are identified in Table 2. During LEAP field testing, problems were sometimes encountered in obtaining sufficient subjects and adequate criteria. However, we were quite successful in obtaining rich feedback from subjects for instrument improvement, especially during the first and second field tests in which one-on-one or small group feedback was obtained. Thus, respondents became co-developers of the instrument. LEAP field testing was moderately successful, despite the obstacles faced. As shown in Table 2, test-retest reliability data collected on the last three versions generally improved with successive versions. However, even for version 0-2D using the empirical key, reliability data for four scales fell below .50.

Table 2

Test-Retest Reliability for LEAP 0-2B, 0-2C, and 0-2D

Scale	0-2B (n=72)	0-2C (n=156)	0-2D (n=430)	0-2D[a] (n=263)
Total LEAP Score	.64	.69	.73	.71
Transformational Leadership	.60	.46	.65	.46
Transactional Leadership	.15	.20	.48	.48
Decision Making Ability	.57	.55	.63	.67
Giving and Seeking Information	.48	.67	.54	.66
Team Player Orientation	.45	.70	.61	.54
Self-Sufficiency Orientation	.71	.58	.63	.49
Physical Fitness Factors	.49	.80	.71	.63
Institutional Commitment	.31	.59	.67	.66
Occupational Commitment	.31	.47	--	--
Persistence to Excellence	.80	.83	.81	.78
Toleration of Adversity	.47	.65	.63	.64
Socialized Power	--	--	.58	.58
Retention Propensity	--	--	.79	.66
Quantity of Work Alternatives	.81	.84	--	--
Quality of Work Alternatives	.46	.82	--	--
Faking Detection	--	--	.65	.43
Time Interval (in months)	1	2	1	1

[a] Unlike the other three data sets, these results are based on the ordinal empirical key rather than the rational key.

Validation was accomplished during the latest field test using ROTC field training performance (FTP) scores and responses to a 19-dimension Air Force Peer Rating Form (AFPRF) created to parallel the LEAP dimensions. Relationships between the LEAP scales and AFPRF dimensions were not strong. However, when the ordinal empiricai key rather than the rational key was used, each of the LEAP scales correlated significantly with FTP except Transactional Leadership. In addition, the ordinal empirical key. was successful in predicting 27% of the overall criterion variance. These data are presented in Tables 3 and 4. The LEAP was far more effective than the Air Force Officer Qualifying Test (AFOQT) in predicting the FPT criterion. The AFOQT alone accounted for only 4% of the explained variance. Together they contributed 30%.

Table 3

LEAP Scales Validated Against Field Training Performance (FTP) Score

LEAP Scales	Rational Key[a]	Ordinal Key[b]
LEAP TOTAL	.11	.45***
Transformational Leadership	.03	.21***
Transactional Leadership	.05	.04
Decision Making Ability	.05	.22***
Giving and Seeking Information	.07	.21***
Team-Player Orientation	.10	.15**
Self-Sufficiency Orientation	.07	.25***
Physical Fitness Factors	.19	.35**
Institutional Commitment	.05	.14**
Persistence to Excellence	.11	.25***
Toleration of Adversity	.10	.10*
Socialized Power	.01	.22***
Retention Propensity	-.04	.08*

[a] *n* = 382. [b] *n* = 263.
**p* < .05. ***p* < .01. ****p* < .001

Table 4

Regression Analysis Predicting FTP Score Based on Ordinal Key[a]

			Cumulative		
Step	Variable Entered	R^2	R^2	F	P
1	Physical Fitness	.12	.12	44.18	.0001
2	Persistence to Excellence	.04	.16	16.46	.0001
3	Socialized Power	.03	.19	11.76	.0007
4	Self-Sufficiency Orientation	.03	.22	10.36	.0014
5	Retention Propensity	.02	.24	6.23	.0131
6	Giving and Seeking Information	.01	.25	6.02	.0147
7	Decision Making Ability	.01	.26	4.55	.0336
8	Transformational Leadership	.01	.27	3.11	.0790

[a] *n* = 263.

Proposed Medical Treatment Facility Leadership Research

During July 1992, the Wilford Hall USAF Medical Center Commander requested our assistance in the identification, recruitment and preparation of medical treatment facility commanders. He had been asked by the Air Force Surgeon General to chair an ad hoc committee to address these issues. The senior author met with the committee on 21 September 1992 to gain a better understanding of the issues and will report back to the committee in February 1993 on an initial survey of physicians.

A Physician Leadership Survey has been developed and will be reviewed by the Surgeon General's committee, and finalized during November 1992. This survey focuses on incentives and disincentives for command, including future incentives which could be used to interest physicians in assuming command positions. The survey also addresses physicians' willingness to pursue management training and asks general attitudinal questions about the role transition from direct care to command.

We will propose to the committee that a multimethod approach be used. We will recommend that interviews be conducted with physicians, individually and in small groups, at southwestern United States locations to complement the survey approach. We believe the problem is so multifaceted that these plans constitute only the early stages of a longitudinal investigation into a variety of issues. We may be able to draw upon the fine work already performed by the Army in Executive Leadership (Jacobs & Jaques 1987, 1990, 1991).

Future Directorate Leadership Research

Medical executive leadership research will be pursued vigorously. Also, work will need to continue on the LEAP if it is to be prepared for operational use. Additional instrumentation for identifying leadership potential at the point of entry may also be needed, but would not be limited to biodata. However, recent collaboration with the other services, and our own review of the leadership literature, suggests that our leadership program should be expanded and refocused. In addition, it should emphasize increased cooperation within and across the services, and with academia.

Leadership is an extremely complex, multidetermined phenomenon. As the integrating conceptual framework provided by Yukl (Van Fleet & Yukl, 1986; Yukl, 1989a, 1989b) attests, leadership involves a myriad of factors including leader and member characteristics; skills; behaviors; personal and other sources of power; and intervening, situational and end-result variables. Therefore, it follows that leadership should not be divorced from the organizational processes of which it is a part (Hosking & Morley, 1988; Jacobs & Jaques, 1991). Also, as Yukl (1989b) and Rast (1991) explain, there is an emerging trend toward decreasing the focus on formal leaders and acknowledging that leadership is a *shared, multidirectional process* between members embedded in organizational or other social systems. Thus, leadership must be studied in context: in the context of the developing self and one's relationship to the external world (Forsythe, 1992, Kegan, 1982; Kegan & Lahey, 1984; Kuhnert & Lewis, 1987; Roberts, 1987); in the context of interactions with others (Jacobs, 1979; Katz & Kahn, 1978; Likert, 1967); in the context of the increasingly complex cognitive demands placed on leaders as they ascend to senior executive levels (Jacobs & Jaques, 1987, 1990, 1991: Sashkin & Fulmer, 1988); and in the context of the politics, structure, culture and climate of organizations (Hosking & Morley, 1988; Jacobs & Jaques, 1991; Jaques, 1976; Schein 1985; Yukl, 1989a, 1989b). Leadership as a shared social process is consistent with recent advice from our former Chief Scientist that we return to studying organizations, not simply personnel (Howell, 1992). It is also consistent with the concept of empowerment so commonly referred to in the "total quality" movement, which is a modern manifestation of the participative management premises of 30 years ago (Howell, 1992: Yukl, 1989b). As Graham (1988) emphasized, "Fostering follower *autonomy* is the hallmark of effective leadership" (p. 73).

In planning future leadership research, we will take into account the complexities discussed above. We will study leadership in the actual organizational environments in which it occurs. Alternatively, we will study leadership in training settings which provide special opportunities for assessment. In addition, we may examine leadership in leadership laboratories in which real-world conditions are simulated under more controlled conditions. Emphasis will be placed on how teams function effectively.

Since methods for studying leadership have their limitations (Schriesheim & Kerr 1977; Yukl, 1989b), multiple methods will be used and new methods explored. Surveys and individual or small group interviews will be used in many situations, but not exclusively. Formal and informal leadership will be addressed, and vertical and horizontal communication and linkage patterns will be examined. Direct or indirect leadership at multiple organizational levels will be studied to identify specific role demands at these different levels and the requisite skills, abilities and attributes needed for effective performance. Often neglected developmental, cognitive and situational factors will be given explicit consideration. Research will also seek to identify what is the most appropriate approach to leadership in different situations or for different types of organizational members. Traditional conceptions of leadership may be counterproductive in those instances where substitutes for leadership exist (Howell & Dorfman, 1986; Kerr & Jermier, 1978), or when work teams are self-managed (Manz & Sims, 1987, 1989).

REFERENCES

Appel, V.H., Grubb, P.D., Shermis M.D., Watson, T.W., & Cole R. W. (1990). The Leadership Effectiveness Assessment Profile (LEAP): Initial officer prototype development (AFHRL-TR-90-19, AD-B146 272L). Brooks AFB, TX: Manpower and Personnel Research Division, Air Force Human Resources Laboratory.

Appel, V.H., Grubb, P.D., Elder, E.D., Leamon, R.E., Watson, T.W., & Earles, J.A. (1991). Leadership effectiveness profile (LEAP): Organizational taxonomy enlisted pool development (AL-TP-1991-0025, AD-A239 973). Brooks AFB, TX: Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division.

Appel, V.H., Quintana, C.M., Cole, R.W., Shermis, M.D., Grubb, P.D., Watson, T.W., & Headley-Goode, A. (in press). Leadership Effectiveness Assessment Profile (LEAP): Officer instrument field testing and refinement (AL-TR-1992-105). Brooks AFB, TX: Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division.

Forsythe, G.B. (Spring 1992). The preparation of strategic leaders. Parameters, XXII (1), 38-49.

Graham, J.W. (1988). Commentary: Transformational leadership: Fostering follower autonomy, not automatic followership. In J.G. Hunt, B.R. Baliga, H.P. Dachler, & C.A. Schriesheim (Eds.), Emerging leadership vistas. Lexington, MA: Lexington Books.

Hosking, D.M., & Morley, I.E. (1988). The skills of leadership. In J.G. Hunt, B.R. Baliga, H.P. Dachler, & C.A. Schriesheim (Eds.), Emerging leadership vistas. Lexington, MA: Lexington Books.

Howell, J.P., & Dortman, P. W. (1986). Leadership and substitutes for leadership among professional and nonprofessional workers. The Journal of Applied Behavioral Science, 22, 29-46.

Howell, W.C. (1992, July). Scince Notes. Unpublished status report. Brooks AFB, TX: Armstrong Laboratory, Human Resources Directorate.

Jacobs, T.O. (1979). Leadership and exchange in formal organizations (USAOECS RB 22-101-1). Fort Ord, CA: US Army Organizational Effectiveness Center and School.

Jacobs, T.O., & Jaques, E. (1987). Leadership in complex systems. In J. Zeidner (Ed.), Human productivity enhancement: Volume 2. Organizations, personnel and decision making. New York: Praeger.

Jacobs, T.O., & Jaques, E. (1990). Military executive leadership. In K.E. Clark and M.B. Clark (Eds.), Measures of leadership. West Orange, NJ: Leadership Library of America.

Jacobs, T.O., & Jaques, E. (1991). Executive leadership. In R. Gal & A.D. Mangelsdorff (Eds.), Handbook of military leadership. New York: John Wiley & Sons.

Jaques, E. (1976). General theory of bureaucracy. New York: Halsted Press.

Katz, D., & Kahn, R.L. (1978). The social psychology of organizations (2nd ed.). New York: Wiley.

Kegan, R. (1982). The evolving self. Cambridge, MA: Harvard University Press.

Kegan, R., & Lahey, L.L. (1984). Adult leadership and adult development: A constructionist view. In B. Kellerman (Ed.), Leadership: Multidisciplinary perspectives. Englewood Cliffs, NJ: Prentice-Hall.

Kerr, S., & Jermier, J.M. (1978). Substitutes for leadership: Their meaning and measurement. Organizationational Behavior and Human Performance, 22, 375-403.

Kuhnert, K.W., & Lewis, P. (1987). Transactional and transformational leadership: A constructive, developmental analysis. Academy of Management Review, 12, 648, 657.

Likert, R. (1967). The human organization. New York: McGraw-Hill.

Manz, C.C., & Sims, H.P., Jr. (1987). Leading workers to lead themselves: The external leadership of self-managing workteams. Administrative Science Quarterly, 32, 106-128.

Manz, C.C., & Sims, H.P., Jr. (1989). Super leadership: Leading others to lead themselves. New York: Simon & Schuster.

Roberts, N.C. (1987). Leadership from a developmental perspective. Paper presented at the "Command Climate: Focus on Leadership Research" Conference, Center for Army Leadership, Kansas City, MO.

Rost, J.C. (1991). Leadership for the twenty-first century. New York: Praeger.

Sashkin, M., & Fulmer, R.M. (1988). Toward an organizational theory of leadership. In J.G. Hunt, B.R. Baliga, H.P. Dachler, & C.A. Scriesheim (Eds.), Emerging leadership vistas. Lexington, MA: Lexington Books.

Schein, E.H. (1985). Organizational culture and leadership. San Francisco: Jossey-Bass.

Schriesheim, C.A., & Kerr, S. (1977). Theories and measures of leadership: A critical appraisal of current and future directions. In J.G. Hunt and L.L. Larson (Eds.), Leadership: The cutting edge. Carbondale, IL: Southern Illinois University Press.

Scott, L.M., Schoen, R.J., & Blaine, C.L. (1987). Job attitudes and commitment of Air Force pilots. Unpublished manuscript, F.J. Seiler Research Laboratory, USAFA, Colorado.

USAFA (undated). USAF Academy leadership studies. Colorado Springs, CO: USAFA.

Van Fleet, D.D., & 'ukl, G.A. (1986). Military leadership: An organizational behavior perspective. Greenwich CT: JAI Press.

Yukl, G.A. (1989a). Leadership in organizations (2nd ed.). Englewood Cliffs NJ: Prentice-Hall.

Yukl, G.A. (1989b). Managerial leadership: A Review of theory and research. Journal of Management, 15, 251-189.

Discussant Remarks
Reducing Attrition: No More Tiers?

by

W. S. Sellman
Office of the Assistant Secretary of Defense
(Force Management and Personnel)

In 1977, the Office of the Secretary of Defense and the Office of Naval Research conducted a 3-day conference on attrition screening. Service researchers and policy makers attended to learn what might be done to control first-term attrition. Shortly after this conference, the Office of the Secretary of Defense issued a policy memorandum in which it directed the Services to maintain a specific level of attrition. Not amazingly, within about 3 months the Services had achieved that level. Consequently, it should be noted that while we are engaged in research to predict attrition from variables that could become part of the personnel screening system, attrition itself is managed.

There are many officials within the Services who believe that a given level of attrition is appropriate; no matter the quality of people entering service, there will always be some who will not adapt. After the 1977 conference, two different ways of controlling attrition were reviewed. One was a selection and classification approach. The other was organizational development--changing the environment which exists within the military that leads to attrition.

Today, 15 years later, attrition is still at about the same level that it was in 1977. Individuals with a regular high school diploma attrite at a rate of about 20 percent; non-graduates and people with alternative education credentials attrite at about a 40 percent rate. This means, as Jim Kinney pointed out, that out of every 10 non-graduate and alternative credential applicants, six would be successful. Because we cannot identify those six reliably, there has been a tendency to exclude the entire group. In essence, we are throwing away six out of 10 young people who could be successful members of the military, if given the opportunity. That is our actuarial approach.

Today's world is more complicated than in 1977, as described in Janice Laurence's paper. The 1987 revision to the education tiers based on actual attrition rates of recruits with different credentials held great promise for reducing turnover and saving millions of dollars by avoiding costs of recruiting, training and equipping replacements. Nevertheless, because of political pressure, the education tiers used in 1992 are not the empirically based tiers originally established in 1987. Home schoolers should be in Tier 1 and adult diploma holders should in be Tier 2. I have no doubt that an attempt to return adult program graduates to Tier 2 would lead to intensive lobbying and congressional intervention.

In a way, my office, the Directorate for Accession Policy in the Office of the Secretary of Defense, is to blame for this conundrum. At the end of each quarter, we issue a recruiting press release. In it, we report to the Congress and the American public the aptitude levels and educational attainment of new recruits. We also report these statistics to Congress as indices of recruit quality. As a result, the Services have developed a score card mentality. They compare themselves on the basis of these measures. They program their recruiting resources on the basis of recruit quality, and defend their budget requests to Congress on the basis of recruit quality.

Attrition statistics are not reported to Congress, nor are they shown in the press releases. The Services only pay lip service to attrition when justifying why they recruit high school diploma graduates. The Services would rather look good and recruit people with diplomas than they would like to control attrition. According to the General Accounting Office (GAO), every time someone leaves service prematurely, it costs $18,000 to recruit, train and equip a replacement. Today, we are recruiting almost no one who does not have a high school diploma and yet attrition rates have remained constant. Attrition looks like it did 10 years ago, and even 15 years ago. The Services prefer to recruit diploma graduates and justify that recruiting strategy on the basis of controlling attrition. Yet somehow, more graduates never gets translated into reduced attrition. That is really a serious problem.

Now, a couple more random thoughts. When the GAO was asked by Congressman Thomas Sawyer to review the EBIS study, it was clear they had received their marching orders to discredit the research--this despite the fact that the GAO had previously published three reports urging recruitment of high school diploma graduates to control attrition. The GAO also had in 1982 issued a report urging the Services to implement a biodata questionnaire, such as the Armed Services Applicant Profile. This time the GAO clearly had their instructions from Congressman Sawyer and Senator John Glenn. We were told education credentials were not good enough because high school diplomas correlate with attrition only about .22. If we refused to place adult diploma holders in Tier 1, Congress would have a reason, this low correlation, to outlaw the use of education credentials in enlistment screening. In fact, congressional staffers threatened that if we did not move adult graduates back to Tier 1 that they would codify in law language that would prohibit screening on the basis of education credentials.

Finally, the Navy is to be commended for its work on the Compensatory Screening Model. I wish they were using it with all applicants, not just nongraduates. It took considerable courage to implement CSM at all, because the Navy's nongraduate rate (using a model that will probably result in a more GEDs) will rise. In 1977, the Air Force implemented a similar compensatory screening system called IMAGE, and it worked. It lowered attrition and expanded the manpower pool. Yet, after about 6 months, IMAGE was discontinued because the Air Force's nongraduate rate was increasing. Never mind that IMAGE would save money though reduced attrition, the Air Force was sensitive about having more nongraduates enter service. I hope the Navy will be able to stand fast against this kind of pressure.

I also would like to see a biodata questionnaire used with CSM--at least to try it operationally to see if coaching and faking really are insoluble problems. The Services are opposed to such a questionnaire because they do not want to turn away a single high school diploma graduate who is otherwise qualified for enlistment. So, I doubt if we will ever truly know if biodata questionnaires will work as part of an actual personnel selection system.

In sum, we have come a long ways in the last 15 years in developing a viable and valid attrition screening system, but we still have further to go. The research described today is impressive and should be pursued vigorously. Perhaps as a result, maybe at some future MTA conference we can announce that the attrition screening problem (both political and technical) has been solved. Not likely, but you never know.

Education Standards in the Military: The Way They Were, Are, and Will Be[1]

Janice H. Laurence
Human Resources Research Organization

If you've heard it once, you've heard it a bazillion times: The personal characteristic most related to completion of a term of enlistment is education credential. Thus, to screen for first-term attrition, the Military Services consider a prospective recruit's educational background--not for the content but for the credential. Since the 1960s, high school graduation status has been used to predict who is likely to finish an obligated term. Decades of research have corroborated the finding that high school diploma holders have almost an 80 percent chance of completing the first three years of service whereas the corresponding figure for nongraduates is between 50 and 60 percent. Screening on the basis of high school diploma status was a reliable and practical way to reduce attrition and its associated costs. However, the 1970s and 1980s witnessed a proliferation of secondary school credentials which led to uncertainties and inconsistencies in attrition screening policies. Different Services handled the recruitment and enlistment of persons with alternative credentials (e.g., General Educational Development or GED and other test-based equivalencies, adult education credentials, diplomas from non-accredited schools, and home study diplomas) differently, depending upon their interpretation of why the diploma was the single best predictor of service adaptability.

Based upon empirical evidence that the attrition rates for GED holders were higher than other high school graduates (HSG) and more similar to non-high school graduates (NHSG), in 1975, the Department of Defense (DoD) formally modified the HSG definition for enlistment purposes. A three tier system was adopted with a new acronym--HSDG (high school diploma graduate)--coined for applicants with the real McCoy while equivalency holders continued with the HSG designation which was only a small step up from NHSG in terms of enlistment preference.

Because empirical attrition data were lacking for alternative credentials other than the GED and the Services didn't know why the diploma was a good predictor of adaptation to the military, the three-tier categorization system used "more rhyme than reason" in sorting out newfangled credentials. As a result of rather elusive and capricious criteria, consistency within, let alone between Services regarding the treatment of secondary level credentials was notably absent until the mid to late 1980s. Although none of the Services "preferred" GEDs, persons with other test-based equivalencies were enlisted as HSDGs by the Air Force. On the other hand, other equivalencies were apparently less revered than even the GED by the Army and Navy. And, although the Air Force considered correspondence school graduates as HSDGs, the Army and Navy did not. Although the Marine Corps was in agreement with the Air Force regarding correspondence credentials, the Corps' position seemed askew of their "seat time" theory that guided the enlistment preference of attendance certificate holders. Credential categorization wasn't static across time either. The Air Force reversed its preference for attendance and completion certificates in 1983 and by 1981 only the Air Force hung on to an accreditation criterion for considering a high school diploma as such.

The Politics of Categorizing Education Credentials

Mounting public criticism of education enlistment policies became a major problem. Word that applicants from nontraditional schools or those holding diploma alternatives were welcome by some Services but not others spread to concerned parents, educational institutions issuing the credentials in question, and ultimately to Congress' ears. Interested parties wanted to know why it was harder for holders of certain credentials to enlist in their Service of choice.

[1] Excerpted from Laurence, J.H. (in press). Education standards and military selection: From the beginning. In T. Trent & J.H. Laurence (Eds.). Adaptability screening for the armed forces. Washington, DC: Office of the Assistant Secretary of Defense (Force Management and Personnel). The full text contains relevant citations.

There was more at stake for nonpreferred groups than having to score higher than bona fide high school graduates on the enlistment screening test, the Armed Forces Qualification Test (AFQT). Depending upon applicant supply and demand, at times the Services turn away alternative credential holders regardless of their AFQT scores in pursuit of quality goals defined jointly by aptitude and education credential. Tied to entry denial, certain credentials and educational programs were being denigrated inadvertently. The military's selection policies regarding education are often misinterpreted as meaning that regular diploma holders are preferred because they are more skilled and able or just plain smarter. The setting of higher aptitude minimums for nongraduates is meant to limit those without a "real" diploma, if and when they are enlisted, to individuals who can be expected to perform relatively well in training and on the job. Despite this explanation of Service education policies, it is difficult for most people to disassociate the meaning of education credentials from cognitive ability or achievement level. In other words, it is difficult to disentangle the credential's schooling signal from its perseverance predictiveness. Even recruiters have contributed to the confusion because they haven't been able to explain, appropriately, the reasons for turning away those not considered diploma graduates.

Initially, in the hopes of protecting the reputations of their programs, alternative credentialing program executives lobbied the Services and DoD officials directly, attesting to the academic rigor of their curricula and extolling the virtues of their graduates. However, spokespersons did not remain limited to program Pooh Bahs but increasingly included members of Congress. One of the first to write DoD, in October of 1981, was the Honorable Bill Nichols (D-AL) as he pushed for HSDG status for graduates of nonaccredited schools. Six months later, this same House member, who was the Chairman of the Military Personnel and Compensation Subcommittee on Armed Services, went up the chain of command and sent a dispatch to then Secretary of Defense Caspar W. Weinberger as he championed the GED cause. He recognized that attrition rates were higher for GED equivalency holders but his bone of contention was that the Services' overly broad education policies disqualified many would-be-successful recruits. He encouraged DoD to adopt new screening criteria based on individual attributes as a replacement for the three broad credential categories. The Congressman continued his correspondence with the Secretary of Defense. His next letter, in December of 1982, again addressed the plight of the unaccredited Christian schools, and the individuals who were being adversely affected. Chairman Nichols was informed that DoD would base its decision on the results of newly initiated research which promised to shed light on the ever-growing credential controversy.

Toward a Unified Solution

DoD was indeed planning to answer the mail. The Office of the Assistant Secretary of Defense for Manpower, Installations and Logistics convened a Joint-Service Education Credentials Working Group and in March of 1982 sponsored the development and administration of the Educational and Biographical Information Survey (EBIS).

The EBIS was administered to more than 34,000 applicants for enlistment and 40,000 new recruits between February and June of 1983. A great deal of biographical data was available through the EBIS, but most pertinent to solving the credential conundrum were items that elicited information regarding respondents' education credentials and types of schools attended.

Linking EBIS data with Defense's automated personnel files made it possible to track attrition status for more specific credential types. Unfortunately, it would take a few years for the data to mature and yield stable est mates of attrition. EBIS applicants had to be given time to enter service and, along with the recruit sample, needed time to demonstrate adaptability or the lack thereof.

Meanwhile, Back on the Hill

As DoD and the Services awaited an empirical basis for categorizing alternative credentials into enlistment priority groups, loyalists kept those cards and letters coming. The most persistent battle over the status of the Christian Schools was waged by Kindness -- Congressman Thomas Kindness (R-Ohio) that is. The Congressman began with a missive to Secretary of the Air Force, Verne Orr, on February 17, 1983. He urged the Air Force to resolve the matter as had the Army and Navy by allowing graduates of nonaccredited Christian schools to be enlisted as diploma graduates regardless of accreditation status. The Air Force's response was that the jury was still out; there were as yet no data on which to base a decision.

Kindness was not discouraged. He wrote back to Orr on March 25, 1983, somewhat piqued that the Air Force had not promptly acceded. If the Air Force wasn't going to be "practical and sensible," then he threatened to go directly to the Secretary of Defense and the White House!

The correspondence continued and escalated. As promised, Representative Kindness sent a letter directly to Weinberger wherein he accused the Air Force of "footdragging" on the issue of non-certified religious high school graduates. After a series of internal Defense coordination memoranda, the Honorable Caspar Weinberger sent a reply to the Honorable Thomas Kindness on May 16, 1983. Over Weinberger's signature, Defense explained education enlistment policies yet again. Kindness was told that the Air Force chose to be more cautious than its fellow Services in amending its position on the accreditation issue while awaiting the EBIS results. The Secretary stated that he would not direct the Air Force to alter its enlistment policies at that time. The Air Force had won the battle, but the war went on.

One key aspect of the dialogue between Congress and Defense is the difference in perspectives. Congressional arguments were, to a great extent, on behalf of the individual while DoD's logic was tied to group statistics. Another way of viewing this difference of opinion is that the contrast was between the individual (Congress' view) and the institutional perspective (DoD's view). Congress was voicing concern that individuals who wanted to serve in the military and who might be successful were denied entry because they were part of a "group." DoD was trying to accept quality recruits and in so doing reduce attrition and its associated costs, efficiently. Both stands are valid but often they are not completely reconcilable.

Early snapshots (e.g., after 3 months in service) indicated that private school graduates were doing as well as public school graduates; graduates from non-accredited schools were roughly on par with those from accredited schools; and graduates of church-related schools had rates similar to non-church school graduates. These early data coupled with the continuing barrage of "congressionals" led to an ahead-of-schedule policy change. By FY 1985, the Joint-Service Group members (including the Air Force spokesperson) agreed to formally drop the accreditation requirement.

The GED lobbyists also kept the heat on during the course of the EBIS study but the initial results were not so promising in their case. Nonetheless, the GED Testing Service and DoD had a running dialogue going. They pushed for DoD to measure the individual attributes associated with attrition directly rather than using group data. And they expressed outrage that recruiters were discrediting the credential and in many cases not allowing those with a GED to try to qualify for service entry.

The Services were alerted to the above-mentioned recruiting injustices by the Deputy Assistant Secretary of Defense through a memo to the Service Assistant Secretaries. Recruiters were told to allow a GED holder to take the ASVAB and were instructed how to explain the reason for rejecting such an applicant without denigrating the credential. But in the interim, as well as in the end, there was no triumph for the GED cause.

While awaiting the final EBIS results, the Working Group met periodically to establish precise, Service-common definitions of education credentials. The new definitions were far less ambiguous and reflected the proliferation of alternative certificates. For example, the definition of a high school diploma graduate included that the applicant had attended and completed a 12-year/day program of classroom instruction and possessed a locally issued diploma.

The Unified Solution

Because of the relentless politicking over the fate of many of the alternative credentials, DoD settled on a 30-month attrition criterion rather than the customary 36 months. EBIS analyses indicated that alternative credential holders and those with no credential did not, on average, adapt to military life as well as those with regular or traditional diplomas. That is, their attrition rates were shown to be higher--in most cases more than 50 percent higher--than those found for traditional diploma graduates--regardless of whether the diploma was issued from a public or private school.

Not only were the attrition rates for equivalency credential holders and adult education diploma holders much higher than for regular diploma graduates, but their rates were closer to the No Credential group than to the High School Graduate group. And, although there were Service differences in overall attrition rates, the relationships among the credential groups held across Services. Furthermore, an examination of behavioral and demographic variables was shown not to mitigate the attrition differences. Attrition was not simply spuriously related to credential.

Armed with actual data on the adaptability of recruits with non-regular diplomas, the Services now had an empirical basis for categorizing the myriad of credentials into enlistment priority groups. A March 24, 1987 memorandum from Chapman B. Cox, the Assistant Secretary of Defense (Force Management and Personnel), to the Service Manpower Assistant Secretaries established "new policy for determining the educational enlistment status of individuals applying for military enlistment." There remained three tiers but the composition had changed. Furthermore, for the first time, all Services were categorizing credentials in the same way. They were still free to determine their own enlistment standards and enlistment priority for each tier but a correspondence school diploma was committed to Tier 2 regardless of Service. And, a new acronym was coined--ACH (for alternative credential holder)--to replace HSG. On October 1, 1987, the new policy would be binding for all.

The Problem With the Solution

It would be an understatement to say that the transition to the new education enlistment policies did not proceed smoothly. This time, the war was declared by the Adult Education constituency, with Ohio and California on the front lines. Adult education graduates had enjoyed top "high school diploma graduate" status by the Army and Navy--the largest Services--until the credential categories were revised.

Shortly after being warned of the impending policies, letters were directed to the Secretary of Defense and his assistants. In one two-page letter from the Principal of Akron Evening High School, addressed to the Principal Director of Military Manpower and Personnel Policy, the policy decision was described as unfounded, appalling, based on flawed expertise, whimsical assumptions and questionable data, blatantly skewed, a perilously prejudicial position which was nearly clandestinely disseminated, and had not one redeeming feature. There was no need to read between the lines here; this fellow hated the prospective policy.

Adult education lobbyists were relentless. They had a strategy that included sending letters to Pentagon officials, soliciting the help of federal and state legislators, getting news coverage, trying to discredit DoD's data, and if possible taking a case or two to the courts through the American Civil Liberties Union. As the president of the California Council for Adult Education put it: "The more "hell" ...we can raise, the better." Fortunately for Defense, this advocate for adult education admitted that the attrition statistics could not be repudiated.

The fact that the policy was indeed implemented did not deter lobbying efforts on behalf of adult diploma holders. The adult education agents followed through on their threats. Much to DoD's surprise and dismay, the assault on education credential standards was coming on strong from this newly disenfranchised group. No longer was this just a continual GED problem.

Representative Thomas Sawyer (D-OH) went so far as to introduce a bill on June 17, 1987 before the House Armed Services Committee (HASC) to halt the implementation. The HASC wanted a delay so that the General Accounting Office could evaluate the study on which the new policy was based. The policy went ahead as did the GAO investigation--with Senator John Glenn (D-OH) now taking the lead. The Tier 2 treatment of adult education graduates sent Glenn into orbit. The GAO levied many lame criticisms but supporting arguments fell on deaf ears. A bold letter from HASC Representatives Les Aspin (D-WI, Chairman) and Sawyer in February of 1988 asked not for a temporary reprieve for adult diploma holders but for their assignment to the preferred tier. In their letter, these Congressmen went so far as to say, if you do this "there will be no reason to continue the GAO report on this matter." Aspin and Sawyer also called for the Services to accelerate their alternative approach to adaptability screening so that individual rather than group characteristics would drive selection decisions.

The proverbial straw that broke the camel's back was a similar letter by Senator Glenn, Chairman of the Subcommittee on Manpower and Personnel of the Senate Committee on Armed Services. Shortly thereafter, on February 29, 1988 a memo went out to the Service Assistant Secretaries for Manpower from the Honorable Grant S. Green, Jr., Assistant Secretary of Defense (Force Management an Personnel) stating that it was "best to honor the congressional request." Given the strong pressure applied and the fact that only a handful of adult education diploma holders would be expected to seek enlistment, DoD weighed the potential costs and benefits and decided to yield. Effective starting April 1, 1988 for FY 1989 accessions-to-be waiting in the DEP, adult diploma graduates would be part of Tier 1 at least until data from the new credential coding system, which was being tracked, might indicate otherwise.

Education Credential Tiers Today

Except for the recategorization of adult diploma holders, the policies devised by the Joint-Service Education Credential Working Group are in force today. Though such policies are alive, they are not necessarily well.

As promised, DoD continued to examine the attrition rates of the various credential holders now that a common and precise coding system was in place. The latest attrition data compiled were for the combined FY 1988 - FY 1989 cohorts after 24 months of service. Each time the conclusions were the same. Adult diploma holders didn't seem to belong in Tier 1 and neither did those without a traditional diploma who completed just one semester of college. In contrast, home schoolers looked as if they should have been "upped" a tier.

Disaffected groups continued to vie for Tier 1 status. Though adult education lobbyists were silenced, old and new voices were raised. The GED contingent renewed its protest. Congress has been making inquiries of late as well, not only on behalf of GED constituents (e.g., an April 7, 1992 letter from Senator Lloyd Bentsen, D-TX) but to address the concerns of Correspondence School graduates and, more forcefully, the Home School lobby. In the case of the former, it seems that correspondence school is the way to go in remote sites such as Alaska and public servants such as Senator Ted Stevens (R-AK) are concerned about not closing the military career path to such youth.

Even more noise has been made recently about those who are legally educated at home. Senator Jesse Helms together with House members Joel Hefley, William Dannemeyer, William Dickinson, and James Sensenbrenner wrote a low-key letter to Assistant Secretary of Defense (Force Management and Personnel) Christopher Jehn on October 11, 1991. Subsequent letters from Congressman Frank R. Wolfe and Senator Bob Dole sent in April of 1992 urged a Tier 1 status for home schoolers.

In response to recent queries, DoD asked for patience. Rather than moving education credentials around and so adjusting tiers, the whole system may be replaced. Again, that's maybe.

A Replacement for Education Credential Tiers?

Fearing that external political pressure would serve to put the kibosh on the three-tier system, DoD had continued to investigate alternative means of adaptability screening. The Services developed a biographical questionnaire called the Armed Services Applicant Profile (ASAP) which comprised multiple-choice items pertaining to background characteristics and behaviors such as high school academics and work history. ASAP was administered from December 1984 through February 1985 to over 120,000 military applicants across all four Services and ultimately had a good showing, relative to credential tiers, for predicting attrition. Regardless of the apparent power of the ASAP over the existing system, many technical, practical, and political concerns impeded progress toward operational implementation. Among the issues raised by Service researchers and policymakers was the potential for recruiter coaching and applicant faking of this self-reported inventory of items. In response to this concern over possible compromise, portions of the Army's Assessment of Background and Life Experiences (ABLE) which were devised to detect faking were recommended to supplement the ASAP. The ABLE was a temperament inventory developed by the Army just subsequent to the ASAP and designed to predict more than just attrition (e.g., leadership potential). With the Army's push toward ABLE and the ASAP progress to date, a combined instrument (with a shortened ASAP and portions of ABLE) known as the Adaptability Screening Profile (ASP) was pilot tested toward the end of 1988.

Worried that the validity of such a biodata inventory would not stand up over time, DoD came up with an alternative to the alternative. Instead of supplanting education credential tiers with ASAP or ASP alone, the idea was to develop a compensatory screening model (CSM). The CSM approach would consider an applicant's credential (actually the specific credential rather than grouped into tiers) along with other characteristics routinely gathered in the enlistment process (e.g., age, marital status, aptitudes) and maybe even ASAP (or ASP) score (or maybe not, given technical and practical concerns). A person would be evaluated on the basis of all such information and, as the name implies, the compensatory screening model would enable one attribute to compensate for another. Though preliminary research has been done, the fate of the CSM is also uncertain. Credential tiers are efficient and with revamping could be made a bit more effective. The CSM, on the other hand, would involve more extended applicant processing time not to mention being a bigger burden on other resources. Though the payoff in terms of attrition reduction looks good in theory, there is less certainty about its value in practice, particularly in light of concerns over biodata, which is the CSM's strongest element (at least in a research setting).

The future of adaptability screening awaits decision. Under consideration now are three options: 1) revise the current three-tier system; 2) adopt a compensatory screening model (but hold the biographical inventory) in its stead; or 3) adopt the CSM with the ASP included. The Services have rendered their opinions in favor of option number 1. The CSM approach seems popular only within the Navy and even for this Service the scope is limited. The Services seem to be saying, "Why bother?" After all, recruiting is pretty good right now and to replace the tiers with CSM might make quality look less rosy than it is.

The Navy at least, perhaps owing to its more stubborn attrition and recruiting problems, is willing to accept a change in the status quo, at least a little. In July 1992 the Navy initiated an operational test of its own version of a CSM for those who do not hold a regular high school diploma. Though high school graduates are exempt, the adaptability of a small percentage of alternative credential holders and those without any secondary school documentation is being evaluated on the basis of variables like specific credential, years of schooling, age, employment status, AFQT, moral waiver status, and military youth program participation. After two years the Navy will decide whether the CSM is worth it. And who knows, the other Services just might come on board. Of course, it is possible that the Navy could throw the CSM overboard before its trial period is scheduled to end.

The ultimate fate of education standards is uncertain. The options, together with the Services predilections, are being weighed by the Assistant Secretary of Defense for Force Management and Personnel--The Honorable Christopher Jehn. Will the tier system survive in one form or another? It seems safe to predict that tiers will remain, but compositional changes will be made. From the data alone, adult education credentials and transcripts attesting to the completion of one semester of college should be moved to Tier 2 and home study diplomas should go into Tier 1. However, apart from the empirical data, evidence from the past suggests that adjusting tiers will not pass unnoticed. Opposition is likely to be fierce.

Which will it be? What will the future hold for adaptability screening? To partially answer this, an odd twist to an old adage seems apropos: "The less things change, the more they remain the same."

A Biodata-Based Compensatory Screening Model[1]

James R. McBride
Human Resources Research Organization

This paper summarizes the work of the Compensatory Screening Model (CSM) development project, in which the Navy Personnel Research and Development Center developed attrition-screening models as possible replacements for current enlistment eligibility criteria.

The CSM Project resulted from a conflict between personnel screening procedures used by the Services, and advocates of alternatives to the traditional high school diploma -- alternatives such as correspondence school, adult education, home study, and educational equivalency diplomas, among others. Advocates of alternative educational credentials argue that more than half of enlistees holding such credentials have been successful in completing their terms of enlisted service.

The CSM project sought alternative criteria that could open up enlistment eligibility to individuals with alternative education credentials, without detriment to first-term attrition rates. This would be accomplished by allowing positive indicants in an applicant's record to compensate for an education credential with a relatively high attrition risk. This paper describes the development and evaluation of logistic regression models, embodying this compensatory feature, for predicting successful completion of a term of service. The predictor variables included the applicants' age, aptitude test scores, educational attainment, and number of dependents, as well as scores on the ASAP (Trent et al., 1990), a biodata instrument designed to predict attrition risk. Joint Service models were developed, as well as separate models for each of the Services.

Each of the logistic regression models is fully specified by the estimated values of its regression parameters. The estimated regression parameters for six models are listed in Table 1. Four of the models are Service-specific equations, while two are Joint Service models. Among the latter, the model labeled "DoD1" makes no distinction among the four Services; model "DoD2" includes indicator variables for the Services.

Inspection of the DoD 1 model regression parameters indicates that all predictor variables except the dependents variable had statistically significant regression weights; this was true for DoD 2 as well. In the latter model, branch of Service was significant for the Air Force and Marine Corps, but not for the Navy.

Looking at the Service-specific models shows less consistency across models in terms of which variables' regression parameters are significant. Only the ASAP score was significant in all four of these models. The education scale variable was significant in three of the Services' models, but not in the Air Force model. The aptitude variable was significant in all Services except the Marine Corps. The pattern for the age and dependents variables was inconsistent from one Service to another.

Effectiveness of the CSM Models

The effectiveness of the alternative CSM models for predicting 24-month Service completion was evaluated in two ways: (a) by means of cross-validated correlation coefficients; and (b) by actual count of the relative frequency of service completion.

Cross-validated correlations. For each of the alternative CSM models, the correlation of the model with the criterion was calculated in a cross-validation sample of 27,068 cases. The results are shown in the column labeled "Accessions Sample Correlation" in Table 2. These correlations summarize the strength of the linear relationship between Service completion and the composite defined by each model.

Table 1

The Six "First-Wave" Compensatory Models: Logistic Regression Weights for Predicting Completion

	Army	Navy	USAF	USMC	DoD.1	DoD.2
Constant	-7.996*	-7.475*	-4.803*	-7.516*	-7.370*	-7.666*
	(.951)	(1.40)	(2.18)	(2.61)	(.704)	(.711)
ASAP Score	.039*	.048*	.058*	.034*	.043*	.043*
	(.002)	(.004)	(.004)	(.005)	(.002)	(.002)
Education Scale	.679*	.471*	-.725	.809*	.560*	.588*
	(.065)	(.089)	(.623)	(.186)	(.048)	(.050)
Math Knowledge Std Score	.014*	.028*	.028*	.011	.020*	.020*
	(.003)	(.004)	(.005)	(.006)	(.002)	(.002)
Age Scale	5.514*	2.294	-2.989	5.605	3.531*	3.889*
	(1.11)	(1.64)	(2.63)	(3.17)	(.834)	(.840)
Dependents (0 or 1)	-.099	.060	.317*	-.140	.020	.006
	(.066)	(.129)	(.121)	(.205)	(.051)	(.051)
Navy Adjustment						-.055
						(.040)
USAF Adjustment						-.129*
						(.043)
USMC Adjustment						-.229*
						(.051)

Notes: * indicates significant regression weight (p < .05).
() indicates standard error of the regression weight.

The strength of those relationships in the applicant population is of greater interest, but unfortunately cannot be observed, because criterion data are not available for anyone but accessions. The population correlation can be inferred, however, by applying an adjustment for range restriction to the observed correlations' values. This was done by applying a multivariate correction for the effects of selection, using the MVCOR program (Sympson & Candell, 1983). These "corrected" correlations are also listed in Table 2, labeled "Applicant Population Correlation."

Among the models developed in the all-DoD sample, the uncorrected correlations were approximately .24 for the three models that included ASAP in their equations and developed optimal weights.

The Service-specific models had uncorrected correlations ranging from about .17 to about .27, with the largest correlation observed in the Navy sample, and the smallest observed in the Marine Corps sample.

The corrected correlations showed a similar pattern of relative magnitudes. The DoD sample estimates ranged from .22 to .27; the Service-specific sample estimates ranged from .19 (the Marine Corps) to .31 (the Navy).

Table 2
Cross-validated correlations of
CSM score with 24-month completion

	Sample Size	Accessions Sample Correlation (Observed)	Applicant Population Correlation (Inferred)
First-Wave CSM Equations			
DoD 1	27,068	.239	.27
DoD 2	27,068	.237	.27
Army	12,134	.247	.28
Navy	5,880	.270	.31
Air Force	6,075	.196	.22
Marine Corps	2,979	.171	.19

Squared correlations express the proportion of variance accounted for by the various models. For this purpose, we prefer to use the inferred population correlations, since the models are intended to be applied in that population. Proportions of variance accounted for by the Service-specific models ranged from 3.6 per cent (for the Marine Corps model) to 9.6 per cent (for the Navy model). DoD-wide, the proportion ranged from 4.8 per cent (for the model without ASAP) to 7.3 per cent (for all three full DoD models).

Discussion

In this section, we will interpret and qualify the results reported above. The discussion will address three topics: (a) the predictor variables included in the alternative CSM models, (b) Service-specific differences implied in the models, and (c) what difference the use of compensatory models might make in personnel selection.

The Predictor Variables. Three of the six predictor variables made the largest contributions to predicting service completion: ASAP, education, and aptitude. Of these, ASAP clearly stood out: In addition to being the predictor variable that made the largest contribution to the equation, ASAP was the only predictor variable that was statistically significant in every CSM model that included it. Based on the data analyzed here, ASAP appears to be the single most powerful predictor of 24-month completion.

For the analyses reported here, education credential was rescaled from a categorical variable to a graded quantitative variable adjusted for the effects of most other predictor variables. This "education scale" variable made the second largest contribution to the compensatory screening models, in every instance except the Navy and Air Force-specific CSM equations. Based on the results obtained here, as well as the historical importance of education as the best predictor of service completion, education clearly has a role in any compensatory screening system.

Cognitive aptitude, as measured by the ASVAB Mathematics Knowledge standard score, made the third greatest contribution overall to the DoD equations, and was significant in every Service-specific CSM equation except that for the Marine Corps.

Age, as measured by the age scale variable, made the fourth largest contribution to the DoD prediction equations. The size of that contribution was small, however, and it was significant in only one of the four Service-specific equations -- the Army.

The dependents variable was significant in only one of the six first-wave equations -- that of the Air Force -- and its contribution to that model was relatively small. We are inclined to dismiss the present dependents variable as a useful predictor.

Service-specific Differences. The last predictor variable to be discussed here is branch of Service. Branch of Service makes a difference in the relationship of the predictor variables to the probability of service completion. These differences could be taken into account in two different ways. The first is to include branch of Service as a predictor variable in a Joint-Services CSM equation, as was done in the case of the DoD 2 model. The other is to develop and use Service-specific compensatory model equations; this was also done in the preceding analyses.

What Difference Do the Compensatory Models Make? Aside from statistical significance, is there any reason to believe that the use of compensatory models would improve personnel screening? The fact that all of the alternative models were correlated with service completion suggests that the answer is "yes." However, the magnitude of the correlations was modest: .17 to .24, uncorrected; approximately .22 to .31, corrected for range restriction. These numbers imply that a selection procedure that achieves them will be more effective in controlling attrition than no procedure at all; they do not, however, indicate whether such procedures might improve over present selection practices. To answer that question, we would need to have attrition data on the 1985 applicants who were not enlisted.

Because we do not have these data, we will resort to Taylor and Russell tables, which allow the user to estimate the average level of employee productivity, if a selection procedure with specified degree of validity were employed. Table 3 illustrates the effect that valid selection procedures can be expected to have on service completion rates. The table displays expected completion rates as a function of two aspects of the situation: the selection ratio -- the percentage of applicants selected -- and the validity of the selection procedure.

The table shows that completion rates can be expected to increase as (a) the validity of the selection procedure increases, and (b) the selection ratio decreases. The implication is that an organization that wants to improve its personnel attrition situation can do so by either or both of two means: Hire a smaller proportion of applicants, and/or use a selection procedure with higher validity than the current one. The former method can be achieved by either hiring fewer people or attracting a larger number of applicants. In most cases neither of these is practical; one results in personnel shortages, the other may entail unacceptable recruiting costs. If reducing the selection ratio is not an alternative, the organization can introduce a more valid selection procedure, or tolerate personnel attrition.

It should be of interest to read Table 3 with the results of the CSM study in mind, comparing completion rates for current selection procedure with those that would be expected if a compensatory screening model were introduced. To make this comparison, one needs to know the validity of current selection procedures. Although the actual value is not precisely known, other data suggest it is between .10 and .15. For purposes of discussion, we will assume that (a) Dod-wide, the selection ratio is 60 percent, and (b) current selection procedures have a validity between .10 and .15 in the applicant population. For these assumed values, Table 3 shows expected completion rates of 77.8 percent to 79.3 percent; these rates are close to the development sample completion rate of 78.8 percent.

What completion rates should we expect if one of the compensatory screening models is implemented? To answer this question, we will assume (a) the same 60 percent selection ratio, and (b) that the validity of a selection procedure based on a CSM model is between .20 and .30. For a validity of .20, the expected completion rate is 80.7 percent; for .25, it is 82.2 percent; for .30, it is 83.7 percent.

Table 3
Effect of Using the Models for Screening:
Expected Completion Rate as a Function of Validity
and Selection Ratio (Assumed base rate: 75% completion).

	Selection Ratio			
	50%	60%	70%	80%
Validity				
.00	75%	75%	75%	75%
.10	78.4%	77.8%	77.2%	76.6%
.15	80.2%	79.3%	78.4%	77.4%
.20	81.9%	80.7%	79.6%	78.3%
.25	83.7%	82.2%	80.8%	79.2%
.30	85.4%	83.7%	82.1%	80.2%

Basis: Taylor Russell tables, modified for dichotomous criterion (cf. Abrahams, Alf & Wolfe, 1971)

In summary, rate of service completion expected if one of the CSM models is used for selection is approximately 80 to 84 percent. This would be an increase of 1 to 5 percent at the DoD level. The expected increase is larger than this for the Army (2 to 6 percent) and the Navy (3 to 7 percent), the two Services for which the CSM models showed the highest validity. A smaller increase would be expected for the Marine Corps, the Service in which the CSM equation had the lowest validity. Only a small increase would be expected for the Air Force, which already had a completion rate of 82 percent in the 1985 sample.

Conclusions

The lion's share of the improvement potential demonstrated in this research was attributable to the contribution made by scores on the ASAP questionnaire. ASAP by itself showed a higher predictive relationship to 24-month completion than any other predictor variable; in fact, ASAP had a higher predictive validity than all the other variables in combination. Use of a compensatory screening procedure that includes ASAP could be expected to improve service completion rates appreciably, if the results here generalize to the present. Use of a compensatory procedure without ASAP would be expected to have less value, but still improve slightly over current screening practices.

Most of the improvement would be expected to occur in the Army and in the Navy; the compensatory procedure had the highest predictive validity for these two Services. It had its lowest validity for the Marine Corps, and consequently the expected improvement there would be small. The expected improvement was also small in the case of the Air Force, but for a different reason: The rate of service completion in the Air Force sample was considerably higher than that of the other Services, and little gain would be expected from the use of a compensatory screening procedure. In other words, the Air Force already appears to have highly effective selection procedures in effect, and the procedures evaluated here would probably make only a small improvement in Air Force personnel attrition.

The generally positive interpretation must be tempered with several important qualifications. First, the compensatory model equations developed in this research were based largely on analysis of data from a sample of 1985 accessions. There are two problems attendant on this fact: (a) Changes in applicant characteristics and Service environments since 1985 may have altered the relationships among the variables studied here. (b) Because of the highly selected nature of the members of the sample, the equations based on their data may contain substantial statistical bias (Defense Advisory Committee, 1991).

A second qualification has to do with the role of the ASAP questionnaire in the compensatory screening procedures. The use of the ASAP questionnaire is controversial because of the content of many of the questionnaire items. Furthermore, there is concern about susceptibility of the questionnaire to faking, coaching, and other forms of response distortion that would threaten its validity in continued operational use.

At this writing, it appears doubtful that a compensatory model containing ASAP scores will be acceptable to all the Services; if a compensatory model is adopted at all, it probably will not include ASAP scores. In that event, the models developed in the research reported here will not be useful, and it will be necessary to develop a new model or models, without ASAP scores.

REFERENCES

Abrahams, N.A., Alf, E., & Wolfe, J.H. (1971). Taylor-Russell tables for dichotomous criterion variables. *Journal of Applied Psychology, 55,* 449-457.

Defense Advisory Committee on Military Personnel Testing (1991). Chairman's letter summary of the May, 1991 meeting.

Sympson, J.B., & Candell, G. (1983) MVCOR. Unpublished computer program. San Diego, CA: Navy Personnel R&D Center.

Trent, T., & Laurence, J.H. (Eds.) (1992). *Adaptability Screening for the Armed Services*. Office of the Assistant Secretary of Defense (Force Management and Personnel).

Trent, T., Quenette, M.A., Ward, D.G., & Laabs, G.J. (1990). *Armed Services Applicant Profile (ASAP): Development and validation* (NPRDC Draft Technical Report). San Diego, CA: Navy Personnel Research and Development Center.

Note:
1. This paper is an abbreviated version of Chapter 5 in the volume entitled *Adaptability Screening for the Armed Forces* (Trent & Laurence, 1992).

A COMPENSATORY SCREENING MODEL FOR REDUCING FIRST-TERM ATTRITION IN THE U.S. ARMY[1]

Leonard A. White and Mark C. Young
U.S. Army Research Institute

A Tier system is used by the Services for classifying education credentials. High school diploma graduates (HSDG) are in Tier 1. Applicants with alternative education credentials (e.g., occupational programs) are in Tier 2, and those with no diploma or credential are in Tier 3. The Services place a premium on recruiting HSDG because earning a high school diploma is the best single measure of a person's potential for adapting to military life. First term attrition among male enlistees who have not completed high school is nearly twice as high as the rate for high school diploma graduates (Eitelberg, Laurence, Waters, & Perelman, 1984).

In recent years, there has been increasing Congressional interest in identifying applicants without a high school diploma who can perform effectively. Concurrently, advocates of non-standardized educational credentials and alternative diplomas have criticized the policy of using the high school diploma without considering other factors that might relate to attrition. In addition, HSDG are more costly to recruit than non-graduates, and with shrinking recruiting budgets there may be a growing need to recruit non-high school graduates.

The Services have been considering the feasibility of augmenting or replacing the Tier system with a compensatory screening model (CSM). The compensatory screening concept combines educational attainment with other indicators to estimate an applicant's likelihood of completing his or her first term of enlistment. Most of the measures (e.g., age, aptitude) being considered for CSM are collected during the Service application process. This compensatory screening system has the potential to improve the Services' capability to identify higher quality accessions among applicants in all levels of educational attainment.

The weights for the CSM variables are determined by their relationship to first-term attrition. Preliminary model development (McBride, et al., 1991) shows that CSM (relative to the Tier system) provides a more accurate estimate of applicants' adaptability for Service. At the present time, further technical work on CSM has been suspended while the Services evaluate CSM implementation options proposed by the Office of the Assistant Secretary of Defense (OASD).

In response to OASD, the Army recognized the potential benefits of CSM and proposed additional research to understand its impact on Army accessions. This paper describes preliminary research on the development of a CSM for Army enlisted personnel.

Method

Sample and Procedure

The sample was all Non-Prior Service (NPS) Army accessions during FY88 and FY89. In all, there are 201,793 observations. The FY88 sample was used to estimate a model of attrition and cross-validated in the FY89 sample. Similarly, a model was estimated in the FY89 sample and cross-validated in the FY88 sample.

[1] Presented at the meeting of the Military Testing Association, October, 1992. All statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Variables in the Attrition Model

The variables used in the attrition estimation were education credential, years of education, age, participation in youth military training, and the Math Knowledge (MK) subtest from the Armed Services Vocational Aptitude Battery (ASVAB). These variables were selected because they have been used successfully in past CSM research (McBride et al., 1991; Trent, Folchi, & Sunderman, 1991). All of these variables are continuous except youth military training and education credential. Dummy variable coding was used to estimate the effects of education credentials on the probability of attrition.

Tables 1 and 2 present relationships between these variables and the 24-month attrition criterion. Scoring for the attrition criterion followed closely that used by Trent et al., (1991). Records of soldiers who died, had medical disability discharges, or separated from the enlisted force to become officers were excluded from the analysis. The attrition variable was coded as 1 when a soldier was voluntarily or involuntarily discharged from the Army before 24 months of service, and as 0 otherwise.

Table 1
Army NPS Accession 24-Month Attrition Rates By Education Credential (FY 1988-1989)

Tier/Education Credential	*N*	% Attrition	*SE*
Tier 1			
High School Diploma Graduate	176,231	17.8	.1
Non-HSDG with 1 Semester College	1,653	31.3	1.1
College: 2 Years or More	5,146	15.5	.5
Adult Education	1,430	29.2	1.2
Tier 2: Alternative Credential			
High School Equivalency (GED)	11,998	36.8	.4
Occupational Program Certificate	19	31.6	11.0
High School Certificate of Attendance	95	32.6	4.8
Correspondence School Diploma	16	31.2	12.0
Home Study Diploma	14	42.9	13.7
Tier 3			
No High School Diploma or Credential	5,191	38.9	.7
Total Sample	201,793	19.6	--

Attrition Estimation

Logistic regression analyses were used to relate the explanatory variables to 24-month attrition. The general form of the logit model is $P(a_i) = 1/(1 + \exp^{-BX_i})$, where x_i is a vector of individual characteristics, $P(a_i)$ is the probability of attrition during a specified time period, and the B's are the parameters to be estimated relating the independent variables to the likelihood of attrition. The logit specification is preferred to linear regression because (a) unlike linear regression, it will always generate predicted probabilities of attrition that range from 0 to 1, and (b) it will provide unbiased estimates of the standard errors of the coefficients.

Table 2

24-Month Attrition Rates for Independent Variables in the Model

Variable	*N*	Attrition Rate	*SE*	Variable	*N*	Attrition Rate	*SE*	Variable	*N*	Attrition Rate	*SE*
Age (Tier 1)				Age (Tier 2)				Age (Tier 3)			
17	10,060	15.1	.4	17	1,478	45.3	1.3	17	1,185	43.0	1.4
18-20	128,168	17.3	.1	18-20	7,576	36.8	.6	18-20	3,270	38.2	.8
21-25	36,972	19.3	.2	21-25	2,660	32.4	.9	21-25	646	35.0	1.9
26+	9,225	24.6	.4	26+	427	32.6	2.3	26+	90	37.8	5.1
Years of Education				ASVAB-Math Knowledge				Youth Military Training Program			
8-9	1,630	44.5	1.2	29-39	7,532	26.6	.5	Yes	7,818	15.1	.4
10	3,973	40.2	.8	40-49	65,731	22.8	.2	No	193,976	19.8	.1
11	22,508	25.3	.3	50-59	79,694	18.9	.1				
12	157,958	18.5	.1	60+	43,911	14.2	.2				
13+	15,490	14.9	.3								

Results

Examination of Table 1 reveals that the attrition rates were similar for the five credential groups in Tier 2, and all but one had large standard errors. One-way analyses of variance showed no significant differences in the mean attrition rate among these five groups ($F < 1$) in the FY88 or FY89 samples. Thus, these credentials were combined to form a single category called alternative credential.

Table 3 shows the results of the logistic regression analyses of the attrition models. The difference in chi-square between the models in Table 3, and nested models which included only the three educational tiers, was highly significant (9 df, $p < .001$). This indicates that CSM significantly increases the prediction of 24-month attrition over the 3-Tier system. In the model, lower attrition is associated with more years of education, higher Math Knowledge scores, and participation in youth military programs. Soldiers with alternative credentials or no credential are considerably more likely to attrit than those with a conventional high school diploma. The positive coefficient for age and its interaction with education is due to the fact that attrition is positively related to age for high school graduates and negatively related to age for those with alternative credentials or no credential. These interactive relationships are shown in Table 2.

Table 3
Logistic Regression Coefficients for 24-Month Attrition

Variable	FY88 Accessions	FY89 Accessions	Combined 88/89[a]
Intercept	.908 (.215)	-.404 (.168)	*.075* (.172)
College (2-Years or More)	.206 (.078)	*-.042* (.073)	*-.027* (.060)
One Semester College	.543 (.086)	.593 (.073)	.501 (.057)
Adult Education	1.035 (.229)	.499 (.063)	.483 (.061)
Alternative Credential (AC)	2.705 (.241)	3.281 (.239)	3.039 (.177)
No Credential or HSDG (Non-Grad)	2.113 (.567)	2.957 (.320)	2.710 (.282)
Years of Education	-.171 (.018)	-.051 (.014)	-.109 (.014)
Age	.046 (.003)	.045 (.003)	.051 (.004)
ASVAB-Math Knowledge	-.025 (.001)	-.027 (.001)	-.024 (.001)
Youth Military Training	-.240 (.047)	-.246 (.046)	-.329 (.046)
Age X AC	-.101 (.012)	-.123 (.012)	-.117 (.009)
Age X Non-Grad	-.071 (.030)	-.105 (.017)	-.099 (.015)
Model Cross-Validation	.16 (With 1989 Sample)	.17 (With 1988 Sample)	—

Note. Standard errors are shown in parentheses. $p < .05$ for all coefficients except those in italics.
[a] Used a 38% random sample high school graduates and a 100% sample of non-graduates.

The overall fit of the models showed little shrinkage upon cross validation, with r = .16 and r = .17. These point-biserial correlations provide an index of fit, however, in practice we are more interested in how well the model predicts on average at each CSM percentile. The weights for computing a CSM percentile score for each person were derived from an OLS regression of the predicted probability of attrition onto the 11 variables in the CSM (R^2 = .99, for this model). Predicted attrition was computed using the logistic model estimated in the combined FY88/89 sample. The CSM score was then transformed to a percentile such that a low score was associated with higher predicted and actual attrition rates.

To evaluate the fit of the model within the education subgroups, we correlated the predicted and actual attrition grouped on CSM percentile scores. The best "fit" was observed for high school diploma graduates (r = .92) and the poorest for the adult education credential (r = .10, p > .05). The average across the six education groups was, *r* = .61.

At the lowest CSM percentiles there was a close fit between the actual and predicted attrition rates in all six education groups. A hypothetical CSM cut score at the 10th percentile would have excluded 46% of the non-high school graduate accessions. The attrition rate among the rejectees was 40%. This same cut score would have excluded less than .001% of the HSDG accessions. A 20th percentile CSM screen would exclude approximately 1% of the HSDG. Soldiers in this 1% HSDG sample were older (*M* = 30 years), had relatively lower scores on tne Armed Forces Qualification Test (*M* = 43rd percentile), and a 30% attrition rate, comparable to that of non-high school graduates.

One objective of CSM is to expand the recruiting market by identifying higher quality non-high school diploma graduates. Less than 0.5% of the non-high school graduates scored above the 50th CSM percentile. The predicted attrition rate for this group was approximately 15%, but the actual rate was 36%, indicating poor fit for the model in this region. In an effort to improve the fit, we estimated the attrition model with HSDG removed from the sample. In this specification, the actual and predicted attrition rates for non-HSDG in the upper 20th percentile (20% of the sample) were nearly identical, both slightly below 30%. As compared to the entire sample of non-HSDG, this elite group of non-graduates was older (22 vs. 19 years of age), averaged one more year of education, and scored higher on the AFQT (63rd vs 56th percentile). The attrition rate for this group was comparable to HSDG scoring in the lowest 20th percentile on CSM. The potential benefits of selecting higher quality non-HSDG must be weighed against costs of attracting and processing applicants in order to find sufficient numbers who qualify for enlistment under the CSM.

Discussion

The CSM can improve the prediction of Army attrition over that achieved by the Tier system. The preliminary analysis reported here indicates that a CSM can be used to identify high attrition risks in all educational categories. A CSM model targeted to non-HSDG was developed to identify higher quality non-graduate accessions. The non-HSDG scoring in the highest CSM percentiles (as compared to average non-HSDG) were more likely to complete their first two years of enlistment and had higher AFQT scores. These findings are similar to results from the Navy's CSM for non-HSDG (Trent et al., 1991). Improvements in the utility of the CSM could be achieved by incorporating additional biodata and temperament indicators of adaptability (Mael, Schwartz, and McLellan, 1992; White, Nord, Mael & Young, in press).

As compared to FY88/89 Navy and Air Force data, attrition rates showed little variation by alternative credential category in the Army sample. This limited the validity gains in the Army CSM from weighting each credential by it relationship to first-term attrition. Further research is needed to understand these Service differences in attrition rates.

References

Eitelberg, M. J., Laurence, J. H., Waters, B. K., & Perelman. L. S. (1984) *Screening for Service: Aptitude and Education Criteria for Military Entry*. Washington, DC: Manpower, Installations and Logistics, Office of Assistant Secretary of Defense.

McBride, J. R., Dempsey, J., Laurence, J., Waters, B., Belden, B., Trent, T., Folchi., & Sunderman, S. (1991, June). *Adaptability Screening Program/Compensatory Screening Model Development and Evaluation: Revised Preliminary Models.* Briefing presented at the meeting of the Military Accession Policy Working Group Technical Committee, San Diego, CA

Mael, F. A., Schwartz, A. C., & McLellan, J. A. (1992, August). Antidotes to dustbowl empiricism with objective biodata. In Rumsey, M. G. (Chair) *Biodata advances: Bridging the rational and empirical perspectives*. Symposium presented at the meeting of the American Psychological Association, Washington, D.C.

Trent, T., Folchi, J., & Sunderman, S. (1991). *Compensatory enlistment screening: A nontraditional approach*. Paper presented at the 33rd Annual Conference of the Military Testing Association, San Antonio, TX.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (in press). The Assessment of Background and Life Experiences (ABLE). In T. Trent and J. H. Laurence (Eds.), *Adaptability screening for the armed forces*. Washington, D.C.: Office of Assistant Secretary of Defense (Force Management and Personnel).

Navy Non-High School Diploma Graduate Compensatory Screening Model[1]

John S. Folchi
Thomas Trent
Steven E. Devlin

Personnel Systems Department
Navy Personnel Research and Development Center[2]
San Diego, California 92152-6800

During the 1980s, the Joint Services developed the Armed Services Applicant Profile (ASAP; Trent & Quenette, 1992) in an attempt to use applicant biographical information to improve validity when predicting attrition from military service. Although ASAP biodata was demonstrated to significantly increase validity over that generated by the traditional mental aptitude and educational attainment predictors, the ASAP has not been implemented operationally due to concerns about validity degradation. However, incremental validity gained from biodata is not expected to deteriorate over time if applicants are required to provide verification of such information.

The ASAP research led to the consideration of using a compensatory screening model (CSM) for selection purposes (Dempsey, Laurence, Waters, & McBride, 1991; Trent, Folchi, & Sunderman, 1991). Specifically, the effort described in this paper focused on using verifiable biodata to select non-high school diploma graduate (NHSDG) applicants for Navy recruitment. Research has shown that Navy NHSDG recruits fail to complete their first-term enlistments at twice the rate as high school diploma graduate (HSDG) recruits. However, the highest quality NHSDG recruits complete obligated service at a rate similar to HSDG recruits.

In recent years, Congress has become increasingly critical of selection procedures based on the current three-tiered educational classification of the applicant population. In response, research has focused on incorporating both the applicant's educational credential status and other verifiable biographical data into the selection process in a manner that captures the differences in attrition performance of individuals within each type of credential group.

The objective of this research was to develop a screening model and administrative procedures for the selection of Navy NHSDG applicants. The model and administrative procedures were required to satisfy the following criteria:

(1) The model should be sensitive to Congressional concerns about the treatment of alternative credential holders.

(2) The model should incorporate the traditional mental aptitude, educational attainment, and age predictors. In addition, it should incorporate specific biographical data predictors requested by the Bureau of Naval Personnel (BUPERS).

(3) In addition to demonstrating incremental predictive validity, the model should demonstrate sufficient face validity to be accepted and used by recruiting personnel in an operational environment.

[1] Paper presented at the 34th Annual Conference of the Military Testing Association at San Diego, California.

[2] The opinions expressed in this paper are those of the authors, are not official, and do not necessarily reflect the views of the Navy Department.

METHODOLOGY

Samples

A fiscal year (FY) 1989 Navy NHSDG sample was used for model development purposes because it was the most current and most representative sample that afforded two-year service tracking data. This sample consisted of 24,171 non-prior service Navy NHSDG applicants who were processed at Military Entrance Processing Stations (MEPS) during FY 1989. Of these applicants, 9,360 subsequently enlisted. The accession subsample included employment data from the "Worker-B" study (Cooke, 1991), which demonstrated that NHSDG accessions with eight or more months of continuous employment demonstrated lower attrition rates than accessions not satisfying this criterion.

An FY 1988 Navy NHSDG sample was used for model evaluation because it was the most current and most representative sample independent of the FY 1989 sample. This sample consisted of 22,784 non-prior service Navy NHSDG applicants who were processed at MEPS during FY 1988. Of these applicants, 7,470 subsequently enlisted.

A sample consisting of 951 FY 1985 accession records was used to simulate missing employment data in the FY 1988 and 1989 samples. Complete, although unverified, employment status data from the ASAP were available for all cases in this sample.

Model and Criterion

A logistic regression model was developed to predict service completion as a function of the predictor variable set. A maximum likelihood technique (Hosmer & Lemeshow, 1989) was applied to estimate model parameters from the FY 1989 accession sample.

The criterion, Comp24, indicates whether or not the applicant successfully completed the first 2 years of his initial tour of duty. Comp24 was coded as 1 if the applicant completed 24 months and zero if he pejoratively attrited before completing 24 months. Comp24 is computed from the applicant's Inter-service Separation Code (ISC) and the number of days served during his initial tour. Applicants with non-pejorative ISCs were excluded from development of the model (Folchi, Devlin, & Trent, 1992).

Predictor Variable Scales

The seven predictor variables included number of years of education completed, type of education credential attained, age at application, Armed Forces Qualification Test (AFQT) category, employment status, military youth program participation, and moral waiver status. Predictor scale scores were computed for each category within the seven variables, as shown in Table 1. The number of categories within each predictor was intentionally kept small so that recruiting personnel could calculate a CSM score using only a one-page application form. The scales were constructed so that all predictors were positively correlated with Comp24.

In general, the scale values represent the proportion of accessions in the corresponding category who achieved success on the criterion. Except for the Educational Credential scale scores, all 2-year service completion rates were computed from the FY 1989 accession sample.

The Education Credential scale values were determined from combined FY 1988-1989 data. Due to small sample sizes for the correspondence, home study, and occupational credentials, their scale values were estimated by the corresponding 2-year DoD completion rates, adjusted for differences between the services. The remaining scale values were the 2-year completion rates of the corresponding credentials, computed from Navy data only.

Table 1

Predictor Variable Scale Values

Category	Scale Value	Sample Size	Category	Scale Value	Sample Size
Years of Education Completed			**Education Credential**		
< 10	.552	2,208	No secondary credential	.611	8,563
10	.611	3,386	GED certificate	.628	6,610
11	.642	4,340	Correspondence school diploma	.653	115[a]
≥ 12	.675	1,909	High school certificate of attendance	.707	205
			Home study diploma	.776	98[a]
AFQT Category			Occupational program certificate	.779	147[a]
III	.607	7,957	**Age at Application**		
II	.648	3,709			
I	.746	177	17	.577	2,903
			18	.609	3,191
Employment Status			19	.647	2,163
			20	.668	1,146
< 1 month or not employed	.608	6,056	≥ 21	.646	2,440
Employed 1 to 7 months	.636[b]	2,868			
Employed ≥ 8 months	.665	436	**Military Youth Program Participation**		
			No	.621	11,624
Moral Waiver Status			Yes	.671	219
Moral waiver required	.605	3,881			
No moral waiver required	.630	7,962			

[a] Combined FY 1988-1989 DoD sample sizes.
[b] Replaced empirical value of .623, per BUPERS policy requirement.

On the basis of the Worker-B study, the applicant's employment status was included as a predictor in the model, despite missing employment data. BUPERS policy requires that applicants continuously employed for at least one month receive preference over applicants who are unemployed or have been employed less than one month. The Worker-B study did not differentiate accessions who were employed at least 1 month but less than 8 months from those who were unemployed or were employed less than 1 month. Therefore, a linear regression equation was developed to simulate the assignment of all accessions not satisfying the 8-month employment criterion into either a "Employed less than 1 month or not Employed" category or a "Employed 1 to 7 Months" category. Employment Status scale values for the two categories were then calculated as the 2-year completion rates of individuals assigned on the basis of simulation.

Operational Test and Evaluation

The NHSDG-CSM is being evaluated during an Operational Test and Evaluation (OT&E), which commenced 1 July 1992. During the OT&E, the U.S. Navy Recruiting Command is using the model to screen NHSDG applicants. Only those applicants whose CSM scores equal or exceed the operational cut score of 98 are eligible for further processing. Data collected during the OT&E will be used to achieve the following objectives:

(1) Determine the predictive validity of CSM in terms of identifying high aptitude NHSDGs who successfully complete their enlistment contracts at rates comparable to HSDGs.

(2) Determine the predictive validity of the CSM with respect to alternative credential groups that are currently classified as tier I (Adult Education and One Semester of College).

(3) Re-estimate the CSM parameters based on employment data collected in an operational environment.

(4) Evaluate new predictors for possible inclusion in future NHSDG CSMs.

(5) Monitor the applicant population serviced during the OT&E to adjust the model for preselection.

(6) Determine the extent to which Navy recruiters can improve their access to high quality NHSDG applicants without increasing their work load.

RESULTS AND DISCUSSION

Descriptive Statistics and Validity Results

Table 2 presents descriptive statistics and validity results by subgroup, as obtained from the FY 1988 NHSDG accession sample (N = 7,216). Among the education credentials, the two largest subgroups had significant validities ($r = .11$, $p < .01$ for both) while the smaller subgroups had higher two-year completion rates and mean CSM scores. The small sample of females had a higher completion rate, mean CSM score, and cross-validity than males. Within racial groups, the CSM showed significant validity coefficients for whites, blacks, and Hispanics.

Table 2

Navy Subgroup Completion Rates, NHSDG-CSM Scores, and Validity Coefficients

			CSM Score		Cross-Validity	
Tiers II and III (Cell B)	Sample Size	Completion Rate	Mean	Std Dev	r_{pbis}	p
Education						
No secondary credential	3,926	.59	56.1	24.3	.11	.001
GED certificate	3,199	.62	68.3	24.8	.11	.001
Occupational certificate	47	.89	150.0	18.1	-.13	.199
HS certificate of attendance	16	.81	118.1	19.3	-.01	.480
Home study diploma	5	.80	112.8	21.3	--	--
Correspondence diploma	3	1.00	75.7	6.5	--	--
Gender						
Males	7,189	.61	62.3	26.3	.12	.001
Females	27	.81	96.7	37.2	.45	.009
Race/Ethnicity (Males)						
White	5,424	.60	62.5	26.9	.13	.001
Hispanic	951	.65	60.9	24.7	.08	.007
Black	674	.58	62.4	23.5	.11	.002
Asian/Pacific Islander	53	.75	66.5	25.8	.16	.130
American Indian/Aleutian	35	.34	59.5	20.7	.04	.404
Total Sample	7,216	.61	62.4	26.4	.12	.001

The CSM cross-validity of .12 (point-biserial; p < .01) represents a significant increment of .04 (p < .01) over AFQT percentile score. It also represents a significant increase of .08 (p < .01) over the Success Chances for Recruits Entering the Navy (SCREEN) formula, which was used to determine the enlistment eligibility of NHSDG applicants prior to 1 July 1992.

Expectancy Comparison

Correct acceptance (true-positive) and erroneous rejection (false-negative) rates were examined using the FY 1988 accession sample. As the hypothetical CSM cut score was increased, the percentage of correct acceptances increased from a base rate of 60.7% to 73.7% at the operational cut score. This 13 percentage point improvement makes the completion rate of CSM-eligibles comparable to the 77.4% completion rate of FY 1988 HSDGs in AFQT category IIIB (N = 21,513). The percentage of erroneous rejections also increased from a base rate of 50.5% to 59.6% at the operational cut score. Correct acceptances exceeded erroneous rejections by at least 9 percentage points across all cutting scores.

An expectancy comparison of the CSM and SCREEN was conducted using six SCREEN cut scores of increasing magnitude. The percentage of correct acceptances increased from 61.2% to 73.4% for the CSM and from 60.8% to 66.1% for the SCREEN. The percentage of erroneous rejections increased from 46.3% to 59.6% for the CSM and remained fairly constant at about 60% for the SCREEN. At each cut score, the CSM achieved both a higher percentage of correct acceptances and a lower percentage of erroneous rejections than the SCREEN. As the percentage of excluded accessions increased, the percentage difference in correct acceptances for the two models generally increased (Folchi, Devlin, & Trent, 1992).

Differential Prediction and Adverse Impact

Using the FY 1988 accession sample, differential prediction and adverse impact analyses were conducted to assess the fairness of applying the operational CSM formula and cut score to select minority applicants. Regression analyses revealed that predicted completion rates were 1 to 4 percentage points lower for blacks than whites over the entire CSM score range. At the operational cut score, the difference was three percentage points, which was not significant at the .05 level. Hispanics had higher predicted completion rates than whites at the lower CSM scores, but lower predicted completion rates at the higher CSM scores. At the operational cut score, the difference of two percentage points between whites and Hispanics was not significant at the .05 level.

The proportions of accessions excluded at the cut score were compared between each group of minority males and white males. The exclusion rates were higher for blacks (94.2%) and Hispanics (94.4%) than whites (92.2%), but only the latter difference was statistically significant (p < .05). However, the white-Hispanic difference did not satisfy Cohen's (1988) small effect size criterion.

The impact of the NHSDG-CSM model and operational cut score was further examined by comparing correct acceptance and erroneous rejection rates between the three groups. The correct acceptance rate was significantly higher (p < .05) for white males (74.1%) than for black males (56.4%) and reflected a small effect size difference. The correct acceptance rate was higher for Hispanic males (81.1%) than for white males (74.1%), but the difference was not significant at the .05 level. Note that the findings concerning correct acceptance rates may be an artifact of the small numbers of black and Hispanic males (39 and 53, respectively) who satisfied the operational cut score.

The difference in erroneous rejection rates between blacks and whites (58.0% and 59.1%, respectively) was not significant at the .05 level. The difference in erroneous rejection rates between Hispanics (64.1%) and whites was significant (p < .01), but did not satisfy the small effect size criterion. These results are more reliable than the correct acceptance results because the samples were larger (4999, 635, and 898 for whites, blacks, and Hispanics, respectively).

CONCLUSIONS

The operational CSM formula described in this report can be used to improve the quality of Navy NHSDG recruits. Use of the operational CSM formula results in a significant validity increase over that generated by traditional predictors, such as mental aptitude, educational status, and age. Recruiting personnel can quickly and simply administer the seven variable CSM on a one-page application form to determine enlistment eligibility. CSM eligibles can be expected to complete their enlistment contracts at a rate comparable to HSDGs in AFQT category IIIB.

REFERENCES

Cooke, T. (1991, November). Memorandum for the Assistant Chief of Naval Personnel Policy and Career Development. Subject: *Attrition Rates for Worker-B Accessions of FY 1989* (CNA 91-2314), Alexandria, VA: Center for Naval Analyses.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dempsey, J. R., Laurence, J. H., Waters, B. K., & McBride, J. R. (1991, November). *Proposed methodology for the development of a compensatory screening model for attrition* (FR-PRD-91-17). Alexandria, VA: Human Resources Research Organization.

Folchi, J. S., Devlin, S. E., & Trent, T. (1992). *Development and evaluation of a compensatory screening model for Navy non high school diploma graduate applicants* (in review) (NPRDC-TN-92-). San Diego, CA: Navy Personnel Research and Development Center.

Hosmer, D. W., & Lemeshow, S. L. (1989). *Applied logistic regression.* New York: John Wiley & Sons.

Trent, T., Folchi J., & Sunderman, S. (1991). Compensatory enlistment screening: A nontraditional approach. *Proceedings of the 33rd Annual Conference of the Military Testing Association*, 565-570.

Trent, T. & Quenette, M. A. (1992, February). *Armed Services Applicant Profile (ASAP): Development and validation of operational forms* (NPRDC-TR-92-9). San Diego, CA: Navy Personnel Research and Development Center.

Policy Implications of Compensatory Screening[1]

by

CAPT James C. Kinney, USN
Director, Recruiting and Retention Programs Division[2]
Chief of Naval Personnel

Background

In the hiring process, the Department of Defense (DOD) is interested in predicting two critical characteristics, aptitude and adaptability. Historically aptitude has been measured primarily by the Armed Services Vocational Aptitude Battery (ASVAB). Adaptability has normally been associated with the probability of completing a term of service. This is particularly important within the Department of Defense when much of the first six to nine months of employment is spent in training. In order to recover this early investment, military recruits must have a high probability of completing the initial term of service. Empirical evidence supports the DOD use of educational credentials as an indicator of adaptability (Table 1). As a result of the attrition statistics which are associated with the attainment (or lack of) certain educational credentials, the credentials were divided into three tiers according to historical attrition rates (Table 1). Tier one (read High School Diploma Graduates HSDG's) having the lowest attrition, thus being the preferred credential. While the educational credential or tier system is a reliable adaptability screening tool, there is a two fold problem with the use of the three tier system as a qualifier in the hiring process. First, there is a high false negative in the exclusion of applicants without the preferred credentials. That is, more than fifty percent of those recruits holding Tier 2 or Tier 3 credentials have proven to be good risks. Secondly, by using the Tier system exclusively, individual's are being excluded by class, rather than being considered on the basis of individual merit.

Table 1

Tier 1: High School Graduate

Educational Credential	24-Month Service Completion Percentage
High School Diploma	82.0
Adult Education Diplcma	67.7
One Semester College(NonHSDG)	72.8
Associate Degree	85.2

[1] Paper presented at the 34th Annual Conference of the Military Testing Association at San Diego, California

[2] The opinions expressed in this paper are those of the authors, are not official and do not necessarily reflect the views of the Navy Department.

Professional Nursing Diploma	81.0
Baccalaureate Degree	85.9
First Professional Degree	80.0
Master's Degree	82.2
Post Master's Degree	87.5
Doctorate Degree	80.0
(High School Senior)	80.5

Tier 2: Alternative Credential Holder

Educational Credential	24-Month Service Completion Percentage
Test-Based Equivalency Diploma	62.7
Occupational Program Certificate	81.9
Correspondence School Diploma	67.9
Home Study Diploma	83.0
High School Certificate of Attendance	73.5
(Credential Near Completion)	67.2

Tier 3: Non-High School Graduate

Educational Credential	24-Month Service Completion Percentage
Less than High School Diploma	60.5

An additional problem manifest recently in the three tier system is the movement of educational credentials from one tier to another for political purposes rather than on the basis of empirical attrition evidence. For example, in 1989, Adult Education diploma holders were moved from Tier Two to Tier One even though their attrition rates clearly justified Tier Two status.

Historically, the services have recruited primarily from Tier One credential holders in order to minimize the costs associated with first term attrition.

Discussion

In an effort to overcome the identified shortcoming of the current Three Tier System, Navy has investigated the development of an additional screening tool for Tier Two and Tier Three credential holders. The Compensatory Screening Mode (CSM) is designed to compensate Tier Two and Tier Three applicants for lack of preferred educational credentials by identifying compensating personal characteristics which would lower the hiring risk. (The CSM research is documented in another paper presented to this conference). When the research indicated that there was a reasonable chance to successfully identify low risk applicants from among the available Tier Two and Tier Three applicant pool, Navy implemented a test to recruit and track these applicants through two years of active service to observe attrition behavior and prove the hypothesis. Navy began the test in July of this year. The resulting demographics of recruits contracted through the first three months of the test are provided at Table 2. We do not yet have enough tracking data to document the success of the CSM in reducing attrition among the Tier Two and Tier THREE applicants.

Table 2

Navy CSM Contracts (JUL-SEP 92) - Demographic Data

Subgroup	N	Percent
Race/Ethnicity		
White	805	88.0
Black	93	10.2
Other	17	1.8
Hispanic[a]	166	(18.1)
Education		
GED Equivalency	486	53.1
No Secondary Credential	265	29.0
H.S. Cert. of Attendance	148	16.2
Occupational	9	1.0
Home Study	6	0.7
Correspondence School	1	0.1
Years of Education		
≤9	3	0.3
10	161	17.6
11	466	51.0
≥12	283	31.0
Moral Waiver		
Not Required	781	85.4
Required	134	14.6
AFQT Mental Group		
CAT I	50	5.5
CAT II	557	60.9
CAT IIIA	307	33.6
Employment		
Employed < 1 month of not employed	187	20.4
Employed 1 to 7 months	169	18.5
Employed 8 mos. or more	559	61.1
Youth Program		
Participated	69	7.5
Did not participate	846	92.5

Age in Years		
17	33	3.6
18	169	18.5
19	309	33.8
20	189	20.7

[a] Persons of Hispanic origin may be of any race.

Note: All CSM contracts were males (N=915).

The advantage to the Navy is that the individuals who qualify under the CSM concept are primarily high aptitude people. Both our demographic research and our early results indicate that 2 out of every 3 applicants hired are Cat I or Cat II upper mental groups.

It has become apparent that the use of CSM has several important policy implications. In the following paragraphs we will address each issue separately.

Attrition Standards

While DOD hiring policies have historically been concerned with attrition, we have not established a goal or standard to apply to the hiring criterion. The unofficial policy, evident in all four services recruiting practices is to minimize attrition by using the traditional High School diploma credential as the primary adaptability screening criterion. This is certainly supportable empirically. Our goal in our CSM effort was to identify from the Tier Two and Tier Three population those individuals whose personal characteristics indicate a potential to adapt to military life to same degree a High School Diploma predicts adaptability. This in effect codifies the attrition standard in the hiring process to be that demonstrated by High School Diploma graduates. It would seem prudent, if the CSM approach to predicting attrition proves valid, that DOD officially establish High School Diploma Graduate attrition behavior as the hiring criterion for adaptability screening.

High School Completion

From a policy perspective, it would seem important that DOD, as the largest employer in the nation of 17-21 year old youth, consistently uphold the requirement to complete High School. If DOD were perceived to undervalue the traditional High School Diploma, it could undermine high school completion. Establishing the High School Diploma as the hiring standard and requiring other credentials to complete additional screening to be eligible for employment will send the right message while allowing the military to identify and hire qualified individuals without a diploma.

Quality Scorecard

Currently the Services compete for quality applicants in the marketplace and their recruiting success is to a large degree measured against their attainment of high quality standards relative to the other services' performance. The primary quality standards routinely reported are percent of High School Diploma Graduates (actually Tier I) and the percent of Upper Mental Groups hired. If CSM works, and those individuals hired adapt like HSDG's, then the DOD recruiting scorecard should be modified to remove the HSDG reporting standard and reflect other criterion which are critical success indices such as percentages of Upper Mental Groups and/or minority attainment relative to Service EEO goals.

Artificial Ceiling

Within the Navy, when we have hired from the Tier Two/Tier Three pool, we have constrained the number of Non High School Diploma Graduates (NHSDG) to a limited percentage of the total. This was done for two reasons. First, was a concern about the significantly higher attrition associated with this market, Second, and probably of more primary importance to policy makers, was a concern for the DOD recruiting "scorecard". The number of NHSDG's hired was in no way reflective of the number who could quality under the old standards. If the CSM approach works, there should be no artificial ceiling placed on the number of NHSDG's hired. If, in fact, these recruits adapt to the same degree as HSDG's, we should let the marketplace dictate the number hired.

Three Tier System

The current three Tier system subdivides educational credentials into three categories based on historical attrition rates associated with the credential. In reality the services have essentially looked at the marketplace as containing HSDG's and NHSDG's and primarily avoided the second group. The CSM approach recognizes this distinction and acknowledges the traditional HSDG as the criterion for passing the adaptability portion of the employment screening process. Those lacking a High School diploma must be screened in more detail to determine potential for employment. Since the model weights each individual type of educational credential on its own merits, there will no longer be any reason for the three tier system. The ultimate object ought to be to classify those educational credentials which reflect attrition at or below the DOD standards (HSDG attrition) as qualifying and all other credentials would require CSM screening.

Summary

It would appear that the successful validation of the Compensatory Screening Model will place the Department of Defense in a "win/win" situation. Without compromising the validity of the traditional High School Diploma, we will have addressed the issue of class exclusion and expanded the recruiting market providing a tool to enable us to identify and employ the highest aptitude recruits available.

NAVAL RESERVISTS AND OPERATIONS DESERT SHIELD/STORM*

Herbert George Baker, PhD
Navy Personnel Research and Development Center
San Diego, CA 92152-6800

ABSTRACT

This paper presents information drawn from responses to the 1991 Naval Reserve Survey. In this project we sought reservists' insights into events preceding their deployment for Desert Storm, perceptions and experiences during active duty, and consequences of their deployment upon return to civilian life. For the first of these, the critical elements appear to be length of notification prior to reporting for active duty and the time which elapsed between activation and deployment. The timing of these events undoubtedly affects employment status and family burdens. During the active duty tour, areas of particular concern focused on inprocessing and outprocessing as well as reception by active duty commands and the role reservists were "allowed" to play. Finally, in examining the outcomes, or consequences, of their deployment, we will present findings on reservists' overall assessment of their call-up experience, perceived effects on overall well-being, and the impact of the call-up on intention to remain in the reserves. The paper concludes with implications for future mobilization events.

Introduction

The 1991 Naval Reserve Survey was administered from November through December 1991, to a sample of 31,763 reservists (10% of reservists not recalled for Operations Desert Shield and Desert Storm; 25% of recalled reservists in medical occupations; and 100% of recalled reservists in non-medical occupations). Survey topic areas included: overall reserve experience, in- and outprocessing, active duty experiences, and return to civilian life. The primary focus was on elements of policy and practice whose modification could increase reservist job satisfaction and Naval Reserve mobilization readiness.

The survey requested the respondent's Social Security Number. This allowed acquisition of personal data from the reserve master tape without elicitation in the questionnaire itself. In addition, respondents were asked to indicate the dates of significant events in their mobilization. From these responses were calculated certain information such as time between notification and entry on active duty. Respondents were also invited to provide write-in responses.

Only those surveys received on or before 31 March 1992 were included in the data analyses. After subtracting for surveys which were undeliverable due to faulty addresses, the adjusted response rate for the survey was 44 percent. There was an 89 percent successful match between returned questionnaires and the reserve master tape (i.e., SSN matchup).

*The opinions expressed are the author's, and do not necessarily reflect official Department of the Navy policy.

Results

Sample Characteristics

More than two-thirds of the respondents were enlisted personnel, and slightly more than one-fourth were in medical occupations. Women comprised 21 percent of the sample.

The sample was fairly evenly split among those not recalled to active duty, those recalled and assigned to a base in continental U.S. (CONUS), and those who were assigned to a forward area, with 35, 38, and 27 percent, respectively.

Seventy-seven porcent were over 30 years, 40 percent over 40. The highest racial presence was that of Black, which, at three percent for officers and eight percent for enlisted, was well below their representation in society at large.

Of those with spouses, only 18 percent reported their spouses to be unemployed either by choice or involuntarily. If the 29 percent of the respondents having no spouse are not considered, the proportion of working spouses is even higher.

Readiness and Satisfaction

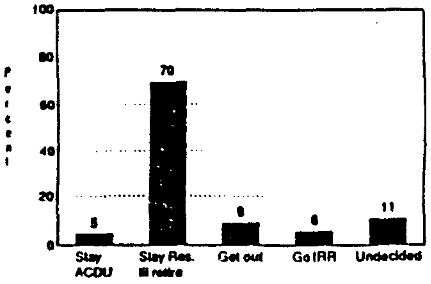
Pre-Mobilization

The majority of respondents indicated their intention to remain in the reserves until retirement (70%), while 15 percent intended either to get out of the resrves or transfer to the Individual Ready Reserve (Figure 1). Eleven percent were undecided.

Reasons for intending to leave were fairly evenly split between 12 alternatives provided. However, conflict between reserve obligations and civilian jobs and family responsibilities edged higher at seven percent each. Such conflicts may cause reservists to sacrifice civilian income or domestic stability, or, on the other hand, to miss reserve meetings, which may result in a "bad year" for reserve retirement purposes. Highest of all reasons for getting out was lack of meaningful work (10%).

Thirty-five percent never expected to be called to active service.

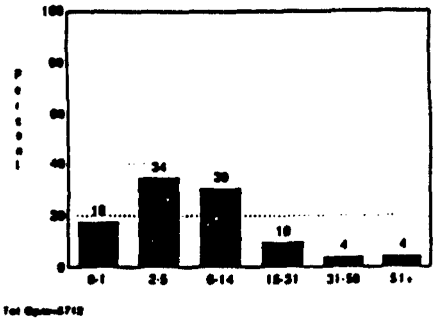
CAREER INTENT IN RESERVES (Q2)



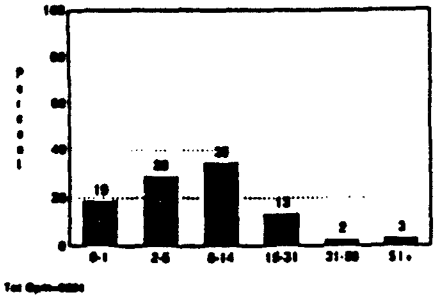
During Mobilization

Figure 2 shows that 82 percent of recalled respondents were on active duty within two weeks of initial notification! Of those, 83 percent were in a forward area within two weeks of activation (Figure 3).

DAYS FROM NOTIFICATION TO ACDU (Q24-25)



DAYS FROM START ACDU TO DEPART CONUS (Q25-26)

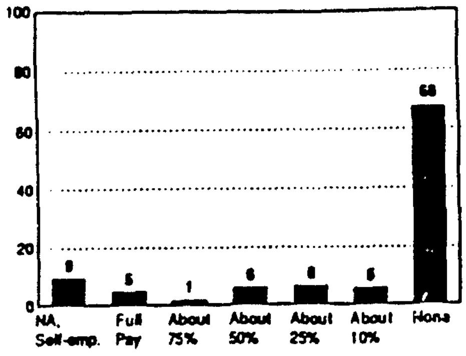


Naval Reservists reported excellent support from their civilian employers, with 70 percent indicating that their employers' personnel policies were supportive. This value would be even higher if the 12 percent "not applicable" were excluded (those being self employed or non-working individuals). Only five percent reported their employers having non-supportive policies.

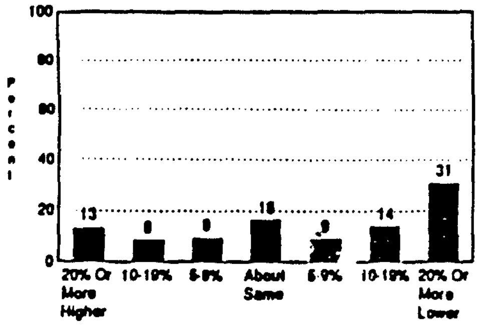
Seventy-two percent indicated that their business or profession was not threatened by the call-up. However, several specific occupations (e.g., doctors, dentists) were threatened.

Most recalled reservists (68%), did not have their civilian pay continued (Figure 4). Only about five percent drew full civilian pay while on active status. The "bottom line" reflected in Figure 5 is that for about 16 percent of the respondents, activation made no impact on their income. Thirty percent gained in income by coming on active duty, while 54 percent lost money. Obviously, these percentages hide the fact that, for some (e.g., doctors), there was an enormous income differential between civilian and military income.

CONTINUING PAY FROM EMPLOYER (Q34)



WAS MILITARY ACDU INCOME HIGHER THAN CIVILIAN (Q35)



Tot Gp/n=7008

Sixty-four percent of the respondents were given three days or more to report for active duty. Interestingly, 87 percent said they could realistically get their affairs in order within a week.

The condensed responses to eight survey items are shown in Figure 7, the first five having to do with the process of entering on active duty, and the last three items dealing with circumstances at the assigned duty station. In-processing was generally timely and professionally done for the majority of respondents; however, 15 percent thought inprocessing was not timely.

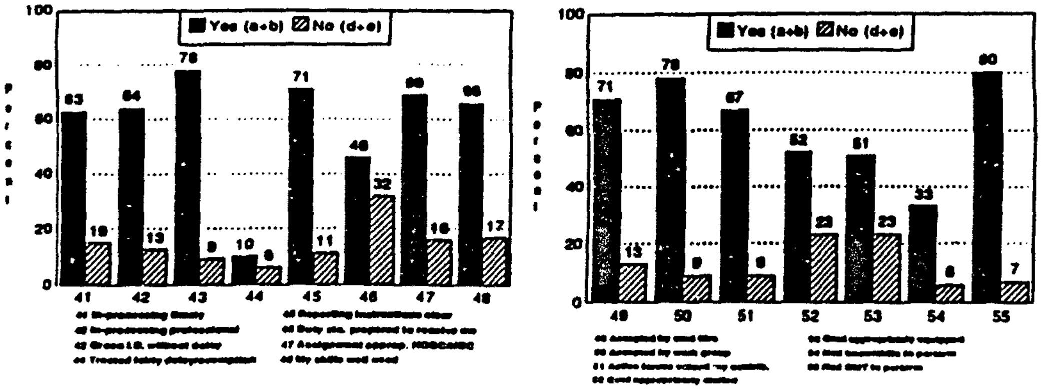
Almost a third of the respondents (32%) felt that their duty station was not prepared to receive them. And, although two-thirds of the respondents agreed that their assignments were appropriate to their NOBC/NEC and that their skills were well utilized, these items also had levels of disagreement in excess of 15 percent.

Three areas of interest are covered by Figure 7: unit acceptance, base support, and skills and training. As can be seen in the chart, very few respondents felt that they were not accepted by their leaders or their co-workers, or that their contribution was not valued. However, nearly a fourth of the respondents indicated that their receiving commands were not appropriately staffed or not appropriately equipped.

Sixty-one percent felt they were well prepared with respect to general military training, but a fifth of the respondents felt their occupational and operational training had not prepared them well for their assignments.

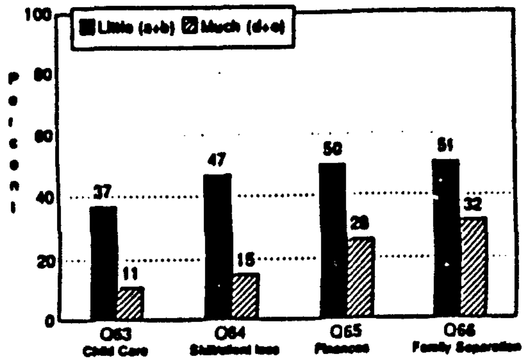
AGREE THAT: (Q41-48)

AGREE THAT: (Q49-55)



The four survey items addressed by Figure 8 asked how much problem was had with certain things which are incidental to activation. Although only 11 percent of the respondents said child care was much of a problem, when those who marked "not applicable" are excluded, 24 percent of those who have children had child care problems. Fifteen percent felt that their call-up threatened them with loss of skills or loss of clients. Many of these were, of course, the self-employed individuals. Those same persons make up a good share of the 26 percent who had a problem with finances.

HOW MUCH PROBLEM:
(Q63-66)

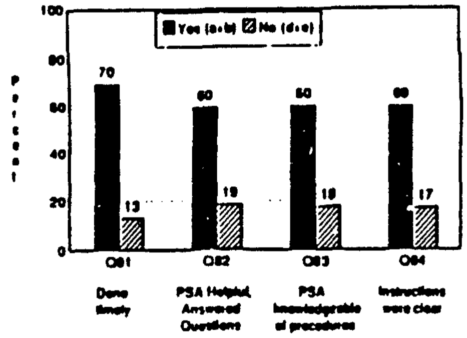


Post-Mobilization

For those who had been released from active duty, 70 percent felt their outprocessing was accomplished in a timely fashion (see Figure 9). Sixty percent agreed that Personnel Support Activity (PSA) staff were helpful and knowledgeable, and instructions were clear. However, the same items had disagreement percentages above 15 percent.

The survey asked, of those who had been released from active duty, whether they had returned to their civilian job. Eighty-seven percent had, to the same or a similar job. Of those who did not return to their pre-activation job, The data show that 27 percent had either lost the job or were pressured to leave.

OUT-PROCESSING (Q81-84)

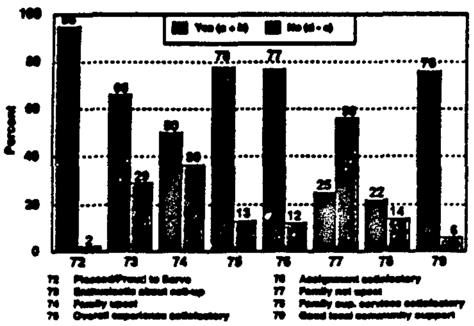


Overall Experience

Forty-two percent indicated that their civilian job was not related to the job they were called up to perform. As expected, the medical occupations were more likely to be working in an occupation similar to their active duty assignment than were non-medical reservists.

Eight survey items are dealt with by Figure 10. Ninety-five percent were pleased and proud to serve. Agreement with being enthusiastic about being called-up was at sixty-six percent. More than three-fourths agreed that their overall recall experience and their duty assignments while on active duty were satisfactory.

Agree That: (Q72-79)



More than 70 percent of the respondents said they were generally satisfied with the reserves prior to DS/S. Nearly three-fourths had planned to stay in the reserves until retirement prior to DS/S. Seventy-eight percent had made no changes to their career plans as a result of DS/S, and only 20 percent reported that their current level of satisfaction or dissatisfaction was due to DS/S. Of those who were recalled to active duty, 78 percent said that, overall, their recall experience was satisfactory. Finally, there was overwhelming agreement that they were proud to serve thair country during Operations Desert Shield and Desert Storm (95%).

Implications

1. Age and family data indicate that reservists have developed relatively stable career patterns, and more complex family situations, both of which interact with mobilization concerns.

2. Because of dissimilarity between civilian job and active duty assignments, a concern with skills loss and skills retention by reservists is justified.

3. Reservists being pressured to leave civilian jobs or being terminated flies in the face of established law.

4. Indications are that the families of most Naval Reservists, like those of the majority of Americans, depend on two wage earners. This situation should be considered in reserve mobilization.

5. Meaningful work is important, and its importance can only increase as the levels of ability and the technical competence of the Navy's people increase; the Navy is the most tecnnical of all the services.

6. In- and outprocessing procedures need improvement.

7. The rapidity with which the great majority reported for active duty and moved to forward areas is gratifying and speaks well of the readiness and responsiveness of Naval Reservists.

8. Most reservists are satisfied with the overall reserve situation.

9. For the overwhelming majority, Desert Shield/Storm was a positive experience.

A Multipurpose Occupational Approach to Understanding the Federal Manager

Donna J. Gregory, Randolph K. Park, and Michele A. Armitage
U.S. Office of Personnel Management

Introduction

The Office of Personnel Rese-:rch and Development (PRD) of the U.S. Office of Personnel Management (OPM) conducted a Governmentwide occupational study of Federal executives, managers, and supervisors. The primary objective of the study was to establish an empirically-based continuum of executive, managerial, and supervisory behaviors and competencies to guide curriculum design and evaluation efforts and to update and revise the Management Excellence Framework (MEF). The MEF has been used by Federal agencies as the basis for the selection and development of managers for the past ten years. The revised MEF will be issued by OPM's Office of Executive and Management Policy and will provide a description of managerial functions and the competencies necessary for managerial effectiveness.

An additional objective of the study was to establish a single source of occupational information for the development of consistent and job-related products (e.g., selection criteria, performance standards, training curriculum) to support human resource management (HRM) programs and policies for executives, managers, and supervisors.

Method

The study employed the Multipurpose Occupational Systems Analysis Inventory--Closed-ended (MOSAIC), a process that gathers data for many HRM purposes using survey methodology and computer statistical analyses to make decisions about the data. As a first step in survey construction, the PRD researchers developed a conceptual model of the effective executive, manager, and supervisor.

Survey Design

Tasks and Competencies. A review of the management and psychological literature was undertaken to examine and document managerial work dimensions and managerial competencies (skills and abilities) from previous management studies. The review also includad sources from the current literature on social, demographic, and economic environments and their impact on the role of managers today and in the future. The result of the literature review was a report entitled *Dimensions of Effective Behavior: Executives, Managers, and Supervisors* (Corts & Gowing, 1992). This research report includes, in a single source, an integration of the behaviors (tasks) and competencies with descriptions as they appear in the literature. The report's foundation was Howard & Bray's 30-year research effort on 26 behavioral dimensions (1988) that had Leen found to correlate significantly with ratings of management potential and with subsequent achievement for American Telephone & Telegraph Company executives. The Corts and Gowing report provided the empirical basis for the managerial tasks and competencies that were used to design two major components of the Leadership Effectiveness Survey (LES).

An OPM research team with expertise in occupational analysis procedures examined the management studies and job information cited in the Corts and Gowing report. The researchers conducted a comparison of the various sources to ensure complete and comprehensive coverage of managerial dimensions and to eliminate much of the confusion in terminology. The Howard and Bray dimensions were linked to other leading models of managerial behavior, including the MEF (1985) and the Federal Executive Institute's Executive Competency Study (1989). This comparison, or "crosswalk," identified common elements in the various models. This review and linkage ensured that all major competencies were identified. The "crosswalk" also resulted in the creation of two categories of competencies, personal and organizationally oriented. It is important to note that each competency has been linked to the work behaviors or task definitions in the literature. These tasks were rationally sorted and categorized under three managerial functional areas: Program Management and Direction, Resources Management, and Interpersonal Skills and Relationships. This provided an organizing framework for the LES.

Workforce Quality. The initial conceptualization of the variables contributing to managerial effectiveness was based on a model for assessing workforce quality. The model was the outgrowth of a Governmentwide research program on the quality of the Federal government workforce initiated by OPM in 1988. OPM and the Merit Systems Protection Board convened a national Advisory Committee on Federal Workforce Quality Assessment. Through discussions with this committee, PRD developed the **Workforce Quality Assessment and Improvement Model** (Dye, 1990; Gowing & Payne, 1991).

The model shows workforce quality to be multidimensional with individual attributes interacting with situational factors (including organizational) and the combination of these variables leading to individual, group/team, and organizational outcomes. Individual attributes are those fixed and dynamic qualities and competencies that individuals bring to the organization or develop while within it. Examples are: knowledges, skills, abilities, motivation, self-esteem, attitudes, values, beliefs, and interests. Situational factors are the circumstances surrounding the existing work situation that can affect the quality of work or service provided. These factors may be internal to the organization or external environmental factors impacting the organization. Some examples include: organizational policies, practices, and conditions, such as culture and climate; objectives; and availability of resources.

The second major step in the design of the LES was thorough reviews by various independent groups, including a focus group comprised of Federal executives on assignment to the National Academy of Public Administration and nine agency representatives serving on an interagency focus group for the MEF project. The LES also was pilot tested with approximately 400 participants at the three OPM Executive Seminar Centers and the Federal Executive Institute in August of 1991. Revisions were made to the LES based on the comments from the focus groups and the pilot test. The final version of the LES was designed to collect background and demographic information, task and competency ratings on a number of scales, classification information, and organizational style and culture data.

Description of the Leadership Effectiveness Survey

The **Leadership Effectiveness Survey** consisted of five parts which are described below. Each respondent, however, did not have to complete all five parts. The sampling plan and survey design permitted OPM to collect the information needed to develop many HRM products without unduly burdening individual respondents. This design resulted in five different survey booklets. All participants were asked to complete Parts I, II, III; and either Part IV or V. The survey forms also differed in the rating scales that participants were asked to use in Parts II and III.

> Part I - Background: This part contained demographic and employment history questions including some used in the workforce quality assessment program studies and **The Federal Executive Survey**. This survey was recently constructed to explore differences between Presidential Rank Award winners and nonwinners (Corts, Anderson, Baker, & Gowing, 1992). Part I also contained additional items reflecting special concerns germane to the sample (e.g., number of employees supervised; size of budget administered). These items provided a profile of the Federal supervisors, managers, and executives who participated in the survey.

Part II - Managerial Tasks: This section asked respondents to check tasks performed from a list of 151 tasks and to rate those tasks on one of three different scales: time spent, importance, and learning difficulty.

Part III - Managerial Competencies: Survey respondents were asked to rate the 22 competencies derived from the OPM competency study on two rating scales -- importance and one of the following: proficiency needed at entry, development of competency, and distinguishes superior from barely acceptable workers.

Parts II and III provided information that identified patterns of task requirements and the competencies needed for effective performance by managers at the three levels. These data are particularly important for succession planning and for defining the training needs continuum from supervisor to executive.

Part IV - Occupational Description: In this section, respondents were asked to complete a number of questions about the duties and responsibilities of their current jobs. The questions were based on factors that occupational analysts have found to be common to Federal managerial and supervisory jobs. This can be useful in identifying factors or behaviors by grade level to help structure positions, develop model position descriptions, and create job evaluation procedures.

Part V - Personal and Organizational Style: This section asked respondents to indicate their own preferred way of dealing with a number of managerial situations, as well as to express opinions about how they evaluate their organization's style. The Quality Orientation portion addresses the eight criteria of the President's Award for Quality and Productivity Improvement.

Sampling Design

The sample was drawn from the population of all Federal executives, managers, and GS-11 to GM-15 supervisors. The sampling plan was designed to obtain a sample representative of the Federal managerial population. To insure adequate representation of smaller sized groupings, personnel from small and medium sized agencies, minorities, and females were over-sampled. Among the Senior Executive Service, the sample consisted of approximately 32 percent of the white males and all of the minorities and females. Among the small agencies, the whole management population was sampled. Among the medium agencies, the whole management population was sampled, except for GM-13 to GM-15 supervisors of whom 92 percent were sampled. Among large agencies, a random sample of about 6.7 percent of the managers and supervisors was drawn.

Four forms of the LES were administered to a total sample of 20,664 executives, managers, and supervisors. Equal numbers of the four different survey forms were distributed across the sample. In addition, a fifth survey form which was used as part of a separate study of Administrative Law Judges. The surveys were mailed to about 1,400 Federal personnel offices for distribution in October 1991.

Overall, 10,061 completed surveys were returned (a response rate of about 49 percent). Table 1 shows the numbers of executives, managers, and supervisors in: the Federal population, the survey sample, and the identifiable returned surveys. A total of 7,938 incumbents who returned surveys categorized themselves in one of the three managerial levels. The remaining 2,123 survey respondents identified themselves as project team leaders, senior scientists, special assistants, other, or gave no response.

Table 1. Sampling Design for the Leadership Effectiveness Survey

	Executives	Managers	Supervisors	Total
Federal Population	8,038	27,842	144,299	180,179
Survey Sample	4,198	4,888	11,578	20,664
Survey Returns	1,763	2,659	3,516	7,938

The original sampling design was weighted to adjust for over-sampling on small and medium agencies, minorities, and females. The survey respondents were representative of the Federal managerial population.

Results and Discussion

Analyses of the LES identified important competencies across the three management levels. Survey respondents rated competencies for their importance for effective job performance and whether proficiency in the competencies was needed upon entry into the position. Ratings on competencies were examined by grouping respondents into three subsamples: executives, managers, and supervisors. These groupings were identified based on self-report.

Important competencies needed upon entry to the job were identified by using respondent ratings on competency importance and need at entry. Table 2 shows the percentage of executives, managers, and supervisors who rated each competency as "very important" or "crucial." Competencies were identified as important when the percentage of respondents rating the competencies as very important or crucial was high for a given sample. For example, at least 95.9 percent of the executives rated Oral Communication as a highly important competency. Table 2 also shows the percentage of respondents who rated the competencies as being needed immediately upon entry (i.e., first day) on the job. For example, 86.3 percent of the executive sample rated Oral Communication as a competency needed upon entry.

For each competency, the percentage levels were combined over the two scales to yield a composite rating score reflecting the value of the competency's importance and need at entry. Inspection of the composite rating scores was made across the three managerial levels to identify basic competencies, that is, those competencies that received high ratings of importance and need at entry across all three levels. These Basic competencies are shown as the first nine competencies in Table 2. Oral Communication, for example, was considered a basic competency since it was rated highly on importance and need at entry for executives, managers, and supervisors.

The competencies differentiated the job content of executives, managers, and supervisors. First-level competencies were those that an incumbent must master at the supervisory level before advancing to the managerial level. These competencies had relatively high composite ratings (need at entry and importance) at the supervisory level that increased at the managerial level. For example, Team Building was high at the supervisory level and even higher at the managerial level, indicating that mastery of this competency was increasingly necessary for effective performance at the higher managerial levels. Mid-level competencies were those with ratings that were either highest at the managerial level (e.g., Internal Controls/Integrity) or high at this level and increased at the executive level (e.g., Creative Thinking). Higher-level competencies had composite scores that were highest at the executive level. Vision and External Awareness were rated highest by executives, but were rated much lower by managers and supervisors. This comparative analysis resulted in a continuum of competencies across the three management levels that can be used for succession and development planning.

Conclusions and Recommendations

The research-based information obtained from the analysis of the LES will support the revision of a new MEF, ensuring its continued validity, as well as consideration of the impact of new areas such as office technology, total quality management, and workforce diversity. The initial result of the LES data analysis is a continuum of competencies that reflects the progression of needed capabilities across the three management levels. These findings led to a model for managerial succession and development planning.

The competency continuum derived from the LES ratings yielded different patterns of competencies for the three leadership levels, and identified competencies that are needed by all management jobs. The competency continuum also underscores the emergence of competencies that reflect changes in management work. Added emphasis is given to technology management, diversity, and Total Quality Management.

References

Corts, D. B., Anderson, C. H., Baker, D. R., & Gowing, M. K. (1992). *Characteristic behaviors, attitudes, skills, and work styles of Distinguished Presidential Rank Award Federal executives.* Washington, DC: U. S. Office of Personnel Management, Personnel Research and Development. (In draft)

Corts, D. B., & Gowing, M. K. (1992). *Dimensions of effective behavior: executives, managers, and supervisors.* Washington, DC: U. S. Office of Personnel Management, Personnel Research and Development. (In draft)

Dye, D. A. (1991). *An explication of a model for assessing the quality of the Federal workforce* (WQR-91-03). Washington, DC: U. S. Office of Personnel Management, Personnel Research and Development.

Gandy, J. A., Mann, W. G., Jr., & Outerbridge, A. N. (April, 1990). *Job performance criteria and biodata validity: Comparisons and considerations.* Paper presented at the Annual Convention of the Society for Industrial and Organizational Psychology, Miami Beach, Florida.

Gowing, M. K., & Payne, S. S. (1992). Assessing the quality of the Federal workforce: A program to meet diverse needs. In S. E. Jackson (Ed.), *Working through diversity: Human resource initiatives.* New York: The Guilford Press.

Howard, A., & Bray, D. W. (1988). *Managerial lives in transition: Advancing age and changing times.* New York: Guilford Press.

U.S. Office of Personnel Management. Office of Training and Development. (1985). *The Management Excellence Framework: A competency-based model of effective performance for Federal managers.* Washington, DC: U.S. Government Printing Office.

Table 2. Percentage of Respondents Indicating Competency Importance or Needed at Entry.

Competency	Percentage of Respondents Indicating Competency is Very Important or Crucial			Percent of Respondents Indicating Competency is Needed at Entry		
	Executives	Managers	Supervisors	Executives	Managers	Supervisors
BASIC COMPETENCIES						
Oral Communication	95.9	90.6	87.2	86.3	74.9	72.6
Written Communication	90.7	85.9	81.4	88.8	80.0	76.2
Problem Solving	92.1	87.0	84.1	72.9	60.0	59.0
Leadership	89.2	89.1	79.6	66.8	59.5	49.5
Interpersonal Skill	79.3	77.8	72.0	67.3	58.4	55.1
Self-Direction	76.6	72.5	68.5	68.3	63.0	60.3
Flexibility	80.7	80.3	74.2	65.7	59.7	54.6
Decisiveness	80.8	80.0	74.0	58.2	51.8	44.0
Technical Competence	63.8	63.7	70.9	54.9	44.6	48.1
FIRST-LEVEL COMPETENCIES						
Human Resource Management	75.8	75.9	62.8	39.6	30.3	23.3
Influencing/Negotiating	74.3	66.9	59.8	37.1	31.1	24.6
Team Building	71.1	70.6	62.5	37.3	29.5	27.0
Conflict Management	67.2	68.5	62.8	36.9	32.9	27.8
Managing Diverse Workforce	50.4	56.4	49.1	48.2	45.1	40.5
MID-LEVEL COMPETENCIES						
Creative Thinking	72.8	57.9	50.2	55.3	37.5	33.8
Planning & Evaluation	64.2	61.6	51.3	31.3	23.9	19.0
Client Orientation	68.2	65.8	62.3	28.9	25.2	22.0
Internal Controls	42.2	50.9	42.9	18.9	21.2	11.7
Financial Management	44.5	46.7	35.8	20.6	20.3	11.4
Technology Management	42.5	45.5	43.0	18.3	14.0	13.0
HIGHER-LEVEL COMPETENCIES						
Vision	68.1	51.6	37.3	39.1	22.4	15.8
External Awareness	58.8	45.8	39.8	22.9	13.9	11.0

Sample sizes are 1,780 executives, 3,442 managers, and 3,732 supervisors on the Importance scale, and 575 executives, 1,184 managers, and 1,303 supervisors on the Needed at Entry scale.

Comparison of the Governmentwide and Department of Defense Managerial Competencies

Randolph K. Park, Michele A. Armitage, and Daniel B. Corts
U.S. Office of Personnel Management

Human resource management in today's changing organizational environment calls for innovative and flexible systems. One way of building such a system is by identifying the tasks and competencies essential for effective management. In doing so, a foundation is formed with which to build a human resource development system for selecting and developing managers.

For several years, the Management Excellence Framework (MEF; 1985) has been the official description of executive, managerial, and supervisory competencies needed for successful job performance. The MEF is a competency-based model which provides the conceptual basis for managerial selection, and training and development programs. The model is used by individual agencies and by OPM for these purposes.

Currently, OPM is investigating whether the MEF will hold for the Federal manager of the 1990's. First, a comprehensive review of the literature collected descriptions of all managerial tasks and competencies within a single source (Corts & Gowing, 1992). An OPM research team with expertise in occupational analysis procedures then examined the information from the Corts and Gowing report with other leading models of managerial behavior to ensure that all major competencies were identified. In addition, the competencies and tasks were organized into functional areas.

In 1991, OPM conducted a Governmentwide survey of white collar executives, managers, and supervisors. One purpose of the survey was to collect occupational analysis information on Federal managers (Gregory, Park, & Armitage, 1992). Four different types of competencies were identified from the results of the survey. Basic competencies were those rated important and needed at entry at all three levels of leadership. First-level competencies were those that an incumbent must master prior to entering the supervisory level. Similarly, Mid-level and Higher-level competencies were those that must be mastered prior to entry into the managerial and executive levels, respectively. These competencies are summarized in the first column of Table 1 and the competency definitions are shown in Figure 1.

Implicit with the differentiation of competencies is the accummulation of competency mastery for higher leadership levels. Supervisors need to have mastered both basic and first-level competencies, and managers need basic, first-level, and mid-level competencies. By the executive level, all competencies would have been mastered.

The accummulation of competency mastery for higher leadership levels is the basis of the MEF and is intended to aid agencies in managerial succession and development planning. Of interest in this study is the extent to which the competency continuum will be applicable to specific agency use. This study compares the competency results of civilian managers from several agencies: the Department of Defense, the Department of the Navy, the Department of the Army, and the Department of the Air Force.

Method

Description of the Survey. The **Leadership Effectiveness Survey** consisted of five parts which are described below. Each respondent, however, did not have to complete all five parts. The sampling plan and survey design permitted OPM to collect the information needed to develop many HRM products without unduly burdening individual respondents. This design resulted in five different survey booklets. All participants were asked to complete Parts I, II, III; and either Part IV or V. The survey forms aiso differed in the rating scales that participants were asked to use in Parts II and III. Part I, Background, contained demographic and employment history questions. Part II, Managerial Tasks, contained questions asking respondents to check tasks performed from a list of 151 tasks and to rate those tasks on one of three different scales: time spent, importance, and learning difficulty. Part III, Managerial Competencies, contained questions asking respondents to rate the 22 competencies derived from the OPM competency study on two rating scales -- importance and one of the following: proficiency needed at entry, development of competency, or distinguishes superior from barely acceptable workers. Information from Parts II and III can be used to identify patterns of task requirements and the competencies needed for effective performance by managers at the three levels. These data are particularly important for succession planning and for defining the training needs continuum from supervisor to executive. In Part IV, Occupational Description, respondents were asked to complete a number of questions about the duties and responsibilities of their current jobs. In Part V, Personal and Organizational Style, respondents were asked to indicate their own preferred way of dealing with a number of managerial situations, as well as to express opinions about how they evaluate their organization's style.

Subjects. The sample was drawn from the population of all Federal executives, managers, and GS-11 to GM-15 supervisors. The sampling plan was designed to obtain a sample representative of the Federal managerial population. To insure adequate representation of smaller sized groupings, personnel from small and medium sized agencies, minorities, and females were over-sampled. The whole management population was sampled from small and medium agencies, except for GM-13 to GM-15 supervisors of whom 92 percent were sampled. Among large agencies, a random sample of about 6.7 percent of the managers and supervisors was drawn.

Four forms of the Governmentwide survey were administered to a total sample of 20,664 executives, managers, and supervisors. Equal numbers of the different survey forms were distributed across the sample. The surveys were mailed to about 1,400 Federal personnel offices for distribution in October 1991.

Overall, 10,061 completed surveys were returned (a response rate of about 49 percent). A total of 7,938 incumbents who returned surveys categorized themselves in one of the three managerial levels. The remaining 2,123 survey respondents identified themselves as project team leaders, senior scientists, special assistants, other, or gave no response.

Results

The Governmentwide percentages of respondents who indicated a competency was important or crucial are shown in Table 1. The percentages are shown for the three leadership levels: executives, managers, and first-line supervisors. Table 1 also shows the raw percentage differences for the Departments of Defense, the Army, the Air Force, and the Navy.

For each agency sample, the differences in percentages are reported between the agency sample and the Governmentwide sample. Since each agency sample is subsumed by the Governmentwide sample, the Governmentwide percentages were adjusted when agency members were extracted from the Governmentwide sample. The difterence between percentages was tested for statistical significance on independent samples. These results are also reported in Table 1.

Inspection of Table 1 suggests some general patterns in competency importance for military related agencies that are different from the Governmentwide norms. For example, all four agencies, although not uniformly, exceeded the Governmentwide importance ratings on Technical Competence, Client Orientation, and Financial Management.

The differences, for the most part, are agency-specific. In the Department of Defense, executives rated Leadership, Interpersonal Skills, Human Resource Management, and Managing Diverse Workforce lower in importance than the Governmentwide norms, and Technical Competence higher. Managers rated Team Building, and Vision higher, and managers and supervisors rated Client Orientation higher. In the Department of the Army, executives rated higher Leadership, Technical Competence, and Creative Thinking, and Internal Controls was rated lower. Department of the Army managers rated higher Technical Competence, and supervisors rated higher Creative Thinking and Technology Management, and lower Problem Solving. Department of the Army managers and supervisors rated Client Orientation and Financial Management higher and Interpersonal Skills lower. In the Department of the Air Force, managers rated lower Oral and Written Communications, Interpersonal Skills, Influencing/Negotiating, Team Building, Conflict Management, Planning and Evaluation, and External Awareness. Department of the Air Force supervisors rated higher Client Orientation, Financial Management, Vision, and External Awareness, and rated lower Interpersonal Skills. In the Department of the Navy, Technology Management and Vision were rated higher in importance at all leadership levels, and Technical Competence, Creative Thinking, Client Orientation, and Financial Management at the manager and supervisor levels. Department of Navy managers also rated higher Planning and Evaluation, and supervisors rated lower Managing Diverse Workforce.

Discussion

The governmentwide continuum tends to hold for the four agencies, although the results show possible exceptions in the placement of some competencies. For example, Financial Management is higher in importance than in the governmentwide continuum and could be identified as a basic competency for these agencies.

The governmentwide managerial competency continuum can be used as a general guide in developing the agency-specific continuums. For example, the U.S. OPM recently completed an occupational analysis of Federal Deposit Insurance Corporation supervisors, managers, and executives that provided agency-specific task, competency, and task-competency linkage information needed for training curriculum development (Park, Armitage, Gregory, and Polak, 1992). The unique characteristics of defense agency populations suggest that they may also benefit from agency-specific needs analyses.

References

Corts, D.B., & Gowing, M.K. (1992). Dimensions of effective behavior: executives, managers, and supervisors. (Report No. PRD-92-05). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development.

Gregory, D.J., Park, R.K., & Armitage, M.A. (1992). A Multipurpose Occupational Approach to Understanding the Federal Manager. Paper presented at the 1992 annual meeting of the Military Testing Association in Del Mar, CA.

Park, R.K., Armitage, M.A., Gregory, D.J., & Polak, R. (1992). Occupational Study of Federal Deposit Insurance Corporation executives, managers, supervisors, and team leaders. (Report No. PRD-92-26). Washington, DC: U.S. Office of Personnel Management, Personnel Research and Development.

U.S. Office of Personnel Management. Office of Training and Development. (1985). The Management Excellence Framework: A competency-based model of effective performance for Federal managers. Washington, DC: U.S. Government Printing Office.

Table 1. Differences in Percentages of Respondents Indicating Competency is Important or Very Crucial.

	Governmentwide			Dept of Defense			Dept of the Army			Dept of the Air Force			Dept of the Navy		
	E	M	S	E	M	S	E	M	S	E	M	S	E	M	S
Sample N	1780	3442	3732	102	152	174	77	308	389	33	254	194	75	250	262
BASIC COMPETENCIES															
Oral Communication	95.6	90.6	87.5	1.6	2.0	-1.4	2.9	-3.3	-1.3	-0.4	-4.6■	0.8	-0.1	3.7	0.7
Written Communication	90.8	86.7	84.0	1.9	2.6	2.3	2.7	-1.4	-2.0	-2.8	-5.4■	-1.6	1.9	4.1	2.6
Problem Solving	92.2	87.6	85.7	-3.7	-3.6	0.8	-0.2	-2.5	-4.6■	3.0	-3.4	-3.8	-0.4	-1.1	2.4
Leadership	88.0	88.3	79.0	-6.8■	0.5	-0.1	9.0■	1.0	-1.4	12.0	1.2	0.5	6.0	0.8	1.1
Interpersonal Skills	79.3	78.1	73.1	-11.5◆	1.1	-3.0	0.2	-5.8■	-8.2◆	-7.1	-12.1◆	-10.7◆	5.9	-1.8	-4.4
Self-Direction	75.3	73.5	69.3	-1.4	-0.7	2.7	7.7	1.6	-3.4	-6.4	-5.1	0.1	3.9	0.5	-2.8
Flexibility	80.1	79.6	72.9	-0.9	-4.7	C.1	5.2	0.2	-1.8	6.7	-0.3	-1.9	5.8	1.5	0.8
Decisiveness	80.7	80.4	73.7	-2.3	0.6	-0.5	-2.3	-0.2	-0.9	0.2	-5.0	0.9	1.5	-1.2	-4.8
Technical Competence	63.0	63.3	69.4	12.1◆	2.1	0.9	17.2◆	11.4◆	4.7	-6.7	1.0	-0.8	10.3	13.9◆	6.3■
LOWER-LEVEL COMPETENCIES															
Human Resource Management	74.3	73.8	58.5	-10.6■	3.6	6.3	9.3	3.4	4.6	-7.2	-3.4	5.5	-1.4	4.4	1.4
Influencing/Negotiating	73.0	66.0	58.5	3.0	0.4	3.8	1.4	-0.9	-3.7	-1.8	-8.7◆	1.2	5.9	5.0	3.8
Team Building	69.9	69.3	60.7	-3.1	8.9■	-0.8	3.6	-1.2	1.1	6.7	-6.1■	-2.3	4.4	2.0	2.3
Conflict Management	66.6	69.4	62.7	-6.2	3.8	3.9	-2.4	-4.4	-6.1■	-11.2	-8.6◆	1.2	7.1	0.5	-2.5
Managing Diverse Workforce	50.2	57.0	50.4	-11.4■	-4.4	-0.1	6.7	0.7	-1.2	-11.5	-3.3	-2.4	-1.6	-2.9	-6.1■
MID-LEVEL COMPETENCIES															
Creative Thinking	70.0	56.9	46.9	2.1	1.0	5.7	10.9■	1.6	6.7■	14.3	-1.2	1.7	8.3	9.9◆	10.8◆
Planning & Evaluating	63.5	61.3	49.8	4.4	5.8	-0.5	2.2	2.9	2.6	7.1	-11.2◆	4	4.3	7.7■	2.8
Client Orientation	62.9	60.8	53.1	2.6	12.5◆	16.4◆	4.4	8.8◆	8.0◆	4.0	1.4	9.4■	8.0	13.4◆	9.9◆
Internal Controls	41.1	50.2	40.0	-5.3	-0.7	2.0	-9.8	4.4	0.3	4.9	-3.2	4.5	-2.3	0.8	2.6
Financial Management	40.8	42.1	24.6	-4.9	4.2	6.4	4.5	13.5◆	16.0◆	16.5	-1.9	15.6◆	4.4	13.8◆	13.3◆
Technology Management	39.8	43.8	37.4	4.5	7.1	2.5	8.3	5.1	8.6◆	12.5	-2.8	5.3	13.5◆	9.5◆	8.1◆
HIGHER-LEVEL COMPETENCIES															
Vision	65.3	49.4	32.8	2.3	12.1◆	5.0	6.2	5.4	3.8	6.9	-0.2	7.2■	16.0◆	8.0■	10.3◆
External Awareness	58.3	44.4	38.1	-5.5	2.7	8.6	-7.3	5.0	0.0	-8.0	-6.1■	9.1■	5.5	0.9	-3.1

Notation: E = Executive, M = Manager, S = Supervisor;
■ $a < .05$, ◆ $a < .01$.

Figure 1. MANAGERIAL COMPETENCIES

1. **Written Communication**--Expresses facts and ideas in writing in a succinct and organized manner.

2. **Oral Communication**--Expresses ideas and facts to individuals or groups effectively; makes clear and convincing oral presentations; listens to others; facilitates an open exchange of ideas.

3. **Problem Solving**--Identifies and analyzes problems; uses sound reasoning to arrive at conclusions; finds alternative solutions to complex problems; distinguishes between relevant and irrelevant information to make logical judgments.

4. **Interpersonal Skills**--Considers and responds appropriately to the needs, feelings, and capabilities of others; adjusts approaches to suit different people and situations.

5. **Managing Diverse Workforce**--Is sensitive to cultural diversity, race, gender, and other individual differences in the workforce; manages workforce diversity.

6. **Vision**--Takes a long-term view and initiates organizational change for the future; builds the vision with others; spots opportunities to move the organization toward the vision.

7. **Creative Thinking**--Develops new insights into situations and applies innovative solutions to make organizational improvements; designs and implements new or cutting-edge programs/processes.

8. **Flexibility**--Is open to change and new information; adapts behavior and work methods in response to new information, changing conditions, or unexpected obstacles; effectively deals with pressure and ambiguity.

9. **Decisiveness**--Makes sound and well-informed decisions; perceives the impact and implications of decisions; commits to action, even in uncertain situations, in order to accomplish organizational goals; causes change.

10. **Leadership**--Inspires, motivates and guides others toward goal accomplishment; coaches, mentors, and challenges subordinates; adapts leadership styles to a variety of situations; models high standards of honesty, integrity, trust, openness, and respect for the individual by applying these values to daily behaviors.

11. **Conflict Management**--Manages and resolves conflicts, confrontations, and disagreements in a positive and constructive manner to minimize negative personal impact.

12. **Self-Direction**--Demonstrates belief in own abilities and ideas; is self-motivated and results-oriented; recognizes own strengths and weaknesses; seeks feedback from others and opportunities for self-learning and development.

13. **Influencing/Negotiating**--Persuades others; develops networks and coalitions; gains cooperation from others to obtain information and accomplish goals; negotiates to find mutually acceptable solutions; builds consensus through give and take.

14. **Planning and Evaluating**--Determines objectives and strategies; coordinates with other parts of the organization to accomplish goals; monitors and evaluates the progress and outcomes of operational plans; anticipates potential threats or opportunities.

15. **Financial Management**--Prepares, justifies, and/or administers the budget for program area; plans, administers and monitors expenditures to ensure cost-effective support of programs and policies.

16. **Human Resources Management**--Empowers people by sharing power and authority; develops lower levels of leadership by pushing authority downward and outward throughout the organization; shares rewards for achievement with employees; ensures that staff are appropriately selected, utilized, appraised, and developed, and that they are treated in a fair and equitable manner.

17. **Client Orientation**--Anticipates and meets the needs of clients; achieves quality end-products; is committed to improving services.

18. **External Awareness**--Identifies and keeps up-to-date on key agency policies/priorities and economic, political, and social trends which affect the organization; understands where the organization is headed and how to make a contribution.

19. **Team Building**--Manages group processes; encourages and facilitates cooperation, pride, trust, and group identity; fosters commitment and team spirit; works with others to achieve goals.

20. **Technology Management**--Integrates technology into the workplace; develops strategies using new technology to manage and improve program effectiveness; understands the impact of technological changes on the organization.

21. **Internal Controls/Integrity**--Assures that effective internal controls are developed and maintained to ensure the integrity of the organization.

22. **Technical Competence**--Understands and appropriately applies procedures, requirements, regulations and policies related to specialized expertise, e.g., engineering, physical science, law, or accounting; maintains credibility with others on technical matters.

DEVELOPING SELECTION PROCEDURES FROM A MULTIPURPOSE JOB ANALYSIS

Donald E. McCauley, Jr.
U.S. Office of Personnel Management

Introduction

The U.S. Office of Personnel Management (OPM) recently conducted a large-scale project using MOSAIC (Multipurpose Occupational Systems Analysis Inventory--Closed-Ended), a task/competency-based approach to occupational analysis. The MOSAIC framework was applied to study the Revenue Agent occupation of the Internal Revenue Service for the purpose of developing selection procedures for that occupational series.

Revenue Agents typically work in one of three major functional areas: Examination (Exam), Employee Plans (EP), and Exempt Organizations (EO). Revenue Agents in Examination conduct examinations of individual, business, and corporate taxpayers to determine substantially correct tax liability. EP Revenue Agents ensure compliance of pension plans with Federal tax laws and determine substantially correct tax liability. EO Revenue Agents ensure compliance of tax-exempt organizations with Federal tax law. The type of work done by Agents in all three functional areas requires knowledge of accounting, auditing, and various areas of tax law. Traditionally, knowledge of accounting has been a prerequisite for entrance to the occupation. Knowledge of tax law is acquired through training. For approximately the first two years after being hired, Revenue Agents are involved in phases of formal classroom training, structured on-the-job training, and limited independent duties of increasing complexity.

In 1991, the IRS entered into an interagency agreement with the Office of Personnel Research and Development (PRD) in OPM to conduct the occupational analysis using its MOSAIC methodology. MOSAIC is an inventory-based occupational analysis methodology that collects information for many different human resource management purposes at one time. MOSAIC eliminates the costly duplication of effort that results from performing separate occupational analyses for separate purposes (e.g., selection, performance management, training). IRS recognized that a MOSAIC analysis would provide the information they needed to develop new selection procedures and would also give them information that would be useful for other human resource management applications, such as training.

In this analysis, Revenue Agents in Grades 5 through 12 in three functional areas (Exam--General Program, EP, and EO) were studied.

Method

Development of the Inventory. In June, 1991, a panel of supervisory Revenue Agents was convened to review the tasks and competencies identified by an IRS study team that had gathered much information on the occupation through interviews with, and observation of, Revenue Agents. From the work of this panel, an inventory was developed and tried out on a small sample of Revenue Agents. Revisions were made to the inventory based on the try-out. The final inventory consisted of background questions, 111 tasks, 58 competencies, and a comment section.

As was mentioned above, the MOSAIC methodology collects information for many human resource management purposes at one time. This is done principally by asking respondents to rate the tasks and competencies on a variety of different rating scales. The tasks were to be rated on three rating scales: relative time spent, relative importance, and relative difficulty to learn. The competencies were to be rated on four rating scales: importance, the degree to which proficiency in the competency distinguishes between barely acceptable and superior Revenue Agents, the degree to which proficiency in the competency is needed at entry to the job, and the best means of becoming proficient in the competency. In order to minimize the burden on any one respondent, two versions of the inventory were created. These two versions differed only in the rating scales that respondents were asked to apply to the tasks and competencies. In other words, no one respondent was asked to rate the tasks and competencies on all of the rating scales used.

Form 1 of the inventory required respondents to rate the tasks on two scales:

> relative time spent, and
>
> relative importance;

and to rate the competencies on two scales:

> importance, and
>
> the degree to which proficiency in the competency distinguishes between barely acceptable and superior Revenue Agents.

Form 2 also required respondents to rate the tasks on two scales:

> relative time spent, and
>
> relative difficulty to learn;

and to rate the competencies on two scales:

> the degree to which proficiency in the competency is needed at entry to the job, and
>
> the best means of becoming proficient in the competency.

Sampling Plan. Each of the five grades within each of the three functional areas was of interest in this project. Therefore, the sampling plan considered each of the 15 grade-within-functional-area combinations separately.

Ten of the grade-within-functional-area combinations (Exam-5, EP-5, EP-7, EP-9, EP-12, EO-5, EO-7, EO-9, EO-11, and EO-12) had relatively few incumbents in each combination; thus, the entire populations for these ten combinations were included in the sample.

The remaining five combinations (Exam-7, Exam-9, Exam-11, Exam-12, and EP-11) had large enough numbers of incumbents so that surveying the entire population was not cost-effective. For these five combinations, random samples were drawn from the IRS personnel databases.

For 3 of these combinations (Exam-7, Exam-9, and EP-11), samples were drawn so as to have at least 100 respondents in each cell, assuming a 40% return rate. It was calculated that a sample size of 100 respondents would ensure that the standard error of the mean time spent ratings would be within plus or minus two-tenths.

The populations for the Exam-11 and Exam-12 groups were large enough so that stable data could be obtained with samples that equalled relatively small percentages of the populations. The sizes of the random samples drawn for the these two grade-within-functional-area combinations were 20% for Exam-11 and 30% for Exam-12. Drawing larger samples would not have been cost-effective.

For the five combinations for which random samples were drawn (Exam-7, Exam-9, Exam-11, Exam-12, and EP-11), care was taken to ensure representativeness in terms of race, sex, national origin, and geographic region.

Administration of the Inventory. The two forms of the inventory were reproduced and distributed in November, 1991. Care was taken that each of the 15 subgroups of interest received equal numbers of each form of the inventory. The inventories were mailed to the home addresses of the Revenue Agents in the sample along with a cover letter that explained the purpose of the project and asked for the agents' participation.

Results

Return Rate. Of the 2,673 inventories that were sent out, 2,008 (75%) were returned. Of the 2,008 returned surveys, 217 were considered unusable. The elimination of these 217 returns resulted in 1,791 usable inventories. The usable returns constituted comparable percentages in all of the grade-within-functional-area combinations, and all of these percentages exceeded the estimated 40% return upon which the sample sizes were calculated.

The research sample (i.e., the usable returns) was found to be representative of the population in terms of grade, functional area, geographic location, sex, race, and national origin.

Task Results. The percentages of respondents performing each task and the percentages of respondents who considered each task to be important were calculated for each grade-within-functional area combination.

There were 22 tasks performed by 50% or more of the agents in all three functional areas. At the 11 and 12 grade levels, the Revenue Agent series became more homogeneous with 62-63 (56%-57%) of the 111 tasks performed across all 3 functional areas. The case for homogeneity becomes stronger when one considers that there were 13 tasks for which no subgroup had 50% of its members performing. Thus, 63% to 64% of the 98 performed tasks were performed by 50% of the Grade 11 and 12 Revenue Agents in all functional areas. Finally, when the whole sample was subjected to a CODAP analysis that compares the tasks performed by different subgroups, 1,675 incumbents (93.5% of the research sample) were placed in a single group based on a common set of core tasks. On the basis of these results, one can conclude that, despite the differences in functional area, most Revenue Agents at the journey level perform similar work activities.

There were 16 tasks considered important by all agents in the research sample. Again, the Revenue Agent series became more homogeneous at the 11 and 12 grade levels, with 52 and 46 (47% and 41%) of the 111 tasks performed across all 3 functional areas. There were 15 tasks for which no subgroup had 50% of its members giving an importance rating of 3 or higher. Thus, 48% to 54% of the 96 important tasks were considered important by 50% of the Grade 11 and 12 Revenue Agents in all functional areas.

The analyses of these two task rating scales revealed that, at least at the Grade 11 and 12 levels, there was considerable overlap in the work activities performed and in those considered to be important across the three functional areas. It was decided that the 52 tasks that were performed by 50% or more of the Grade 11 agents in the three functional areas and considered to be important by those agents would be considered the description of the typical Revenue Agent job. This determination was made for two reasons: (1) these 52 tasks constituted the largest degree of homogeneity in the research sample, and (2) it can be reasonably expected that all lower-graded agents will reach Grade 11. These tasks are presented in Table 9.

Competency Results. The 58 competencies were rated on three scales that are directly pertinent to selection. These three scales were (1) importance, (2) the degree to which proficiency in the competency distinguishes between barely acceptable and superior agents (distinguishing ability), and (3) the degree to which proficiency in the competency is needed at entry to the job (need at entry). As with the tasks, the results for the competencies were analyzed for each grade within each program area. Inadvertently, all of the EP-5 and EO-5 agents who responded had received Form 1 of the inventory. Thus, no EP-5 or EO-5 agents rated the competencies on the need at entry scale.

There were 28 competencies considered to be important by all agents in the research sample. Grade 5 respondents considered one additional competency to be important. Grade 7 respondents considered six additional competencies to be important. Grade 9 respondents thought seven additional competencies to be important. Grade 11 and Grade 12 respondents considered five additional competencies to be important. The number of competencies considered important was very consistent across almost all grade levels (i.e., Grades 7 through 12), ranging from 33 to 35.

There were 19 competencies considered to be distinguishing by all agents in the research sample. Grade 5 respondents considered two additional competencies to be distinguishing. Grade 7 respondents considered 13 additional competencies to be distinguishing. Grade 9 respondents thought 12 additional competencies were distinguishing. Grade 11 respondents considered ten additional competencies to be distinguishing. Grade 12 respondents considered 11 additional competencies to be distinguishing. The number of competencies considered to be distinguishing was very consistent across almost all grade levels (i.e., Grades 7 through 12), ranging from 29 to 32.

There were 13 competencies considered to be needed at entry by all agents in the research sample. There were no Grade 5 respondents in two of the functional areas. Grade 7 respondents considered five additional competencies to be needed at entry. Grade 9 respondents considered 11 additional competencies to be needed at entry. Grade 11 respondents considered ten additional competencies to be needed at entry. Grade 12 respondents considered six additional competencies to be needed at entry. The number of competencies considered to be needed at entry was very consistent across almost all grade levels (i.e., Grades 7 through

12), ranging from 18 to 24, with the most homogeneity at the Grade 9 and 11 levels.

Linkage of Tasks and Abilities. Since it was at the Grade 11 level that the greatest degree of homogeneity of work behaviors occurred across the three functional areas and since it can be reasonably expected that all lower-graded Revenue Agents will progress to the Grade 11 level, the competencies that had been identified as important, needed at entry, and distinguishing for the Grade 11 level were linked to the tasks that were performed and considered important by at least 50% of the Grade 11 respondents. Each competency was found to be important to the performance of some (12 to 52) of the tasks. The measurement of these competencies formed the basis of the selection process for the Revenue Agent series.

Conclusions

It is clear from the above that the MOSAIC methodology is eminently suited to provide information for the development of selection procedures. The scales used provided a wealth of information both about the job under study and about the relative importance, distinguishing ability, and need at entry of the competencies necessary to perform that job. MOSAIC furnishes firm support for any selection procedure developed from it.

MOSAIC is also effective in the study of multiple grade levels within a job series. In the study described above, MOSAIC was effective in determining that one selection procedure could be used to select applicants for five grade levels within three functional areas. In addition, the information that MOSAIC provides for selection, training, performance management, and other human resources management uses can be broken out by grade and by grade-within-functional area, providing a wealth of information for personnelists and managers alike.

Finally, the study described above should make very clear the importance of sampling design in effective MOSAIC studies. Since the purpose of this study was to gain information about five grade levels within three functional areas, it was critical to the success of the project that these 15 cells be sampled effectively.

Personality and Biodata: Is There A Relationship?

Stephanie Booth-Kewley and Marie D. Thomas

Navy Personnel Research & Development Center
San Diego, CA

Biographical information, or biodata, is being used with increasing frequency in personnel selection. Past research indicates that biodata measures are capable of predicting a variety of important criteria, such as occupational choice, academic performance, job performance, training success, creativity, and tenure and adjustment to the military (Owens, 1976; Hough, 1987). In fact, the results obtained for biodata measures have been so impressive that researchers have come to regard biodata measures as one of the few legitimate alternatives to skill and ability measures for personnel selection (Ghiselli, 1966; Reilly & Chao, 1982).

Despite the demonstrated utility of biodata measures, very little research has been directed towards increasing our theoretical understanding of biodata. The need for some theoretical framework for biodata has been pointed out by numerous researchers (e.g., Hough, 1987; Russell, Mattson, Devlin, & Atwater, 1990). Specification of the constructs that underlie valid biodata items and scales should increase our understanding of the psychological prerequisites necessary for success in organizational settings.

A number of biodata researchers (e.g., Ghiselli, 1966; Mumford, & Owens, 1987) have drawn a sharp distinction between biodata measures and personality or temperament measures. However, as was pointed out by Mael (1991), "Many items termed 'biodata' are indistinguishable from the types of self-report items found in temperament and attitude measures" (p. 764). Moreover, many of the biodata factors that researchers have extracted from biodata instruments seem to have clear linkages to personality. For example, factor analysis of biodata has revealed factors that measure social participation (e.g., Childs & Klimoski, 1986; Owens & Schoenfeldt, 1979); these seem to overlap conceptually with the widely studied personality construct of Extraversion. Similarly, in some biodata instruments, factors measuring drive, ambition, or achievement motivation have been identified (Baehr & Williams, 1967; Owens & Schoenfeldt, 1979); these seem conceptually similar to the personality construct of Conscientiousness. Also, in some biodata measures, factors that assess personal adjustment or emotional stability have been identified (Owens & Schoenfeldt, 1979); these biodata factors seem conceptually similar to the widely studied personality construct of Neuroticism.

Given recent research in organizational psychology suggesting that broad personality variables translate into important organizational outcomes (George, 1989; Staw, Bell, & Clausen, 1986; Barrick & Mount, 1991), and given that the relationship between biodata and personality has not received much research attention, research is needed to determine the degree to which personality and biodata measures are, in fact, related.

Recent advances in personality measurement research indicate the existence of five major, robust dimensions of personality (Costa & McCrae, 1985; Digman, 1990). These dimensions are Extraversion (E), Neuroticism (N), Agreeableness (A), Conscientiousness (C), and a fifth dimension variously identified as Intellect, Culture, or Openness to Experience (O). These five dimensions have been consistently identified using different methodologies (e.g. peer ratings vs. self-report), different measures, different populations, and in different cultures. Two additional personality variables--self-esteem and self-efficacy--were also included in the present study because we thought that they might be related to biodata.

The objective of the present study was to examine the relationship of broad personality constructs to validated biodata measures in a sample of Navy personnel. It was hypothesized that substantial associations between the biodata factors and the personality dimensions would be found.

Method

Subjects

The subjects were 484 male U.S. Navy recruits completing basic training in San Diego. Because only male recruits receive basic training in San Diego, females were not included in our sample. The sample was made up of 68 percent nonHispanic whites, 13 percent blacks, 12 percent Hispanics, and 4 percent Asians, with other race/ethnic groups making up the remaining 3 percent. All subjects had high school diplomas or high school equivalency diplomas (GED); 32% had also completed some college. The subjects ranged in age from 17 to 33, with a mean age of 20.4. Ninety percent of the subjects were single (never married); the remaining 10 percent were either married (8%) or divorced (2%).

Measures

Owens Biographical Questionnaire. The Owens Biographical Questionnaire (BQ) is a widely used, 118-item biodata instrument. It assess diverse areas of an individual's background, such as academic achievement, social activities, family background, sports participation, hobbies and interests, and temperamental attributes such as sensitivity to criticism and social confidence. Most of the BQ items refer to the high school years. The BQ was developed using a rational approach assisted by factor analysis. A set of 659 items was successively administered to samples of subjects, factor analyzed, and pared down, until the item set was reduced to 118 (Owens and Schoenfeldt, 1979). Factor analysis of the final 118-item BQ indicates the existence of 13 male (and 15 female) factors. The male factors are shown below (Stokes, Mumford, & Owens, 1989, p. 515).

1. Warmth of parental relationship--Existence of close, warm, interested, supportive parental relationship.
2. Intellectualism--Interest and participation in cultural, literary, and/or scientific pursuits.
3. Academic achievement--History of striving for high levels of academic achievement.
4. Social introversion--Extremely introverted and ineffective in social situations.
5. Scientific interest--Great interest in, and enjoyment of, scientific courses.
6. Socioeconomic status--High parental educational, occupational, and income level.
7. Aggressiveness/independence (verbal)--High verbal and persuasive skills and interests.
8. Parental control vs. freedom--Great amount of parental control and direction over activities.
9. Positive academic attitude--Exhibited positive attitude toward high school academic activities.
10. Sibling friction--Large degree of friction or competition with siblings.
11. Religious activity--Strong religious beliefs and high activity in religious or charitable organizations.
12. Athletic interest--Very active in athletic activities and performed very well.
13. Social desirability--Little concern or desire to behave in socially desirable ways.

The stability of the factor structure, and the reliability and validity of the individual factors, has been established in numerous studies (Eberhardt & Muchinsky, 1982; Mumford & Owens, 1984; Shaffer, Saunders, & Owens, 1986).

The internal consistencies (coefficient alphas) of the BQ factors for the present data ranged from .54 to .84, with a median of .73.

Armed Services Applicant Profile. The 50-item Armed Services Applicant Profile (ASAP) was designed to tap background dimensions that might predict an individual's propensity to adapt well to the military. Like the BQ, the ASAP measures diverse areas of a person's background, including academic attitude, delinquency, athletic interest, and work history. Unlike the BQ, the ASAP does not focus primarily on the high school years. The ASAP has adequate criterion validity (i.e., it predicts Navy outcomes such as retention), and is reasonably resistant to faking (Trent, 1992).

Item scoring of the ASAP uses an empirically-developed key (horizontal percent method), which was validated and cross-validated on very large samples (Trent & Quenette, 1992). Although factor analysis was not used to develop and refine the ASAP, principal components analysis of this measure indicates six factors (Trent, 1992), listed below.

1. School achievement--Academic achievement and positive attitude towards school courses, activities, and teachers.
2. Delinquency--Past substance use, truancy, and getting into trouble in high school.
3. Work ethic--History of stable employment and superior job performance.
4. Independence--History of continuous employment and/or schooling, maturity, stable friends, social conformity.
5. Social adaptation--Confidence, sociability, persuasiveness, persistence, and autonomy from parents.
6. Physical involvement--Involvement in athletic activities and superior athletic performance.

The internal consistencies (coefficient alphas) of the ASAP factors for the present data ranged from .35 to .64.

NEO Personality Inventory. The 118-item NEO Personality Inventory (NEO-PI; Costa & McCrae, 1985) is a widely used personality inventory developed to measure the "Big Five" personality dimensions: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. The NEO-PI scales have adequate internal consistency, test-retest reliability, and validity (Costa & McCrae, 1985). The coefficient alphas of the five NEO scales ranged from .72 to .90 for the present sample.

Rosenberg Self-Esteem Scale. The Rosenberg Self-Esteem Scale (Rosenberg, 1965) is a widely used, 10-item measure of self-esteem. This scale has adequate internal consistency, test-retest reliability, and validity (Rosenberg, 1965). For the present sample, the scale had a coefficient alpha of .85.

Self-Efficacy Scale. The 17-item General Self-Efficacy scale developed by Sherer et al. (1982) was used to measure self-efficacy. This scale was designed to assess an individual's general expectation that he or she is capable of successfully executing behaviors necessary for producing the outcomes that he or she desires. The General Self-Efficacy scale has adequate internal consistency and construct validity (Sherer et al., 1982). The coefficient alpha for the present sample was .90.

Procedure

Besides completing the biodata and personality measures, respondents also provided their age, race/ethnicity, education, and marital status. Subjects were administered the questionnaires in groups of 10 to 60, as a scheduled part of Navy basic training. Participation was voluntary, but only 11 recruits out of 495 (2%) refused to participate.

Results

Correlation coefficients were computed between the seven personality variables and the 13 BQ factors. These results are presented in Table 1. As the table shows, most of the personality-BQ correlations were of modest size, with a mean correlation of only .16. However, because of the fairly large sample, most of the correlations (66%) were statistically significant ($p < .05$).

The largest personality-BQ association was the correlation of .63 between Neuroticism and the BQ Social Desirability factor. Although Owens and Schoenfeldt (1979) labeled this factor Social Desirability, its strong inverse association with

Table 1
Correlations Between Personality Variables and BQ Factors

	N	E	C	A	O	Self-Efficacy	Self-Esteem
F1	-.12	.21	.16	.21	-.07	.21	.25
F2	-.20	.07	.05	.05	.13	.09	.06
F3	-.27	.23	.33	.09	.09	.33	.25
F4	.12	-.39	-.04	.09	.02	-.12	-.20
F5	-.19	.14	.19	.15	.22	.25	.17
F6	-.15	.17	.02	-.08	.10	.14	.21
F7	-.04	.32	.17	.09	.46	.24	.19
F8	.06	.04	-.05	-.09	.02	-.02	-.04
F9	-.21	.15	.25	.31	.08	.26	.13
F10	.14	.04	.02	-.03	-.02	-.03	-.08
F11	.06	-.10	-.04	.12	-.08	-.05	-.13
F12	-.17	.29	.21	.04	-.03	.26	.22
F13	-.63	.08	.27	.22	.03	.47	.48

Note. Correlations of .09 and above are significant at the .05 level; correlations of .12 and above are significant at the .01 level. F1 = Warmth of Parental Relationship. F2 = Intellectualism. F3 = Academic Achievement. F4 = Social Introversion. F5 = Scientific Interest. F6 = Socioeconomic Status. F7 = Aggressiveness/Independence. F8 = Parental Control vs. Freedom. F9 = Positive Academic Attitude. F10 = Sibling Friction. F11 = Religious Activity. F12 = Athletic Interest. F13 = Social Desirability.

Table 2
Correlations Between Personality Variables and ASAP Factors

	N	E	C	A	O	Self-Efficacy	Self-Esteem
School Achievement	-.23	.20	.24	.20	.08	.29	.24
Delinquency	-.27	.07	.29	.29	-.06	.21	.15
Work Ethic	-.11	.16	.11	.05	.07	.18	.13
Independence	.00	.08	.03	.09	-.06	-.01	-.03
Social Adaptation	-.10	-.21	-.07	.07	-.05	-.09	-.04
Physical Involvement	-.28	.25	.26	.16	.01	.33	.29

Note. Correlations of .09 and above are significant at the .05 level; correlations of .12 and above are significant at the .01 level.

Neuroticism, coupled with its item content suggest that "Positive Adjustment" would be a more accurate label. This same BQ factor (Social Desirability or Positive Adjustment) also correlated substantially with the personality dimensions Self-Efficacy (r = .47) and

Self-Esteem (r = .48). Aside from these three correlations, there was only one other personality-BQ correlation that exceeded .40: this was the correlation between Openness to Experience and the BQ factor Aggressiveness/Independence (r = .46). Individuals high on Openness to Experience tended to be verbally aggressive, independent, and unconventional.

Correlation coefficients were computed between the personality variables and the ASAP factors; these are shown in Table 2. As was found for the BQ, the ASAP factors had only modest correlations with the personality dimensions, with a mean correlation of .15. Again, due to the large sample size, most of the correlations (62%) were statistically significant. The personality-ASAP correlations were of a magnitude similar to the personality-BQ correlation^s, with mean correlations of .15 and .16, respectively.

Because the ASAP factors had somewhat lower coefficient alphas than the BQ factors, all personality-biodata correlations were corrected for unreliability of the biodata factors to permit a more accurate comparison. The coefficient alphas of the biodata factors were used in making the corrections. Although the correction for unreliability increased the personality-biodata correlations, the overall effect was minimal (the corrected correlations are not shown). Correcting for unreliability raised the mean correlation for the BQ from .16 to .19 and the mean correlation for the ASAP from .15 to .22. The overall pattern of results remained essentially unchanged: neither biodata instrument related strongly to personality.

Discussion

Our prediction that there would be substantial associations between biodata and personality was not confirmed. For both biodata instruments, overlap with the personality dimensions was minimal. These findings were surprising, given that the BQ and, to a lesser degree, the ASAP, contain a number of items resembling those typically found on personality scales, and given that both measures contain factors (e.g., Social Introversion on the BQ and Work Ethic on the ASAP) that seem related to personality.

The reasons for the lack of substantial associations between the biodata and personality measures are not clear. It may be that the particular personality constructs that we chose to measure in this study are not the ones that relate the most strongly to biodata. It is also possible that biodata and personality are truly distinct content domains. Research in which a large number of diverse personality constructs are measured in conjunction with biodata is needed before we can conclude that biodata and personality are truly separate domains.

The fact remains that many items that are referred to as biodata are, as Mael (1991) pointed out, indistinguishable from items that appear in personality and temperament measures. Yet it is often claimed that because they assess factual, objective information, biodata instruments are less subject to distortion and social desirability effects and have other related advantages compared to "softer" self-report measures, such as personality scales (e.g., Owens, 1976; Stokes, Mumford, & Owens, 1989). To the degree that biodata measures are made up of nonfactual, subjective items, they are similar to personality measures, and the claims of their greater objectivity and associated advantages seem unfounded.

We suggest that biodata researchers more clearly establish the boundaries of the biodata domain. This could be done either by limiting the domain to content that is truly biographical (i.e., items that assess discrete, objective past behaviors or objective characteristics of the person) or by explicitly broadening the biodata domain so that "softer" items assessing personality, feelings, attitudes, likes and preferences are included. A clearer definition of what does and does not constitute biodata would eliminate a lot of the confusion that seems to currently characterize the biodata literature.

References

Baehr, M. E., & Williams, G. B. (1967). Underlying dimensions of personal background data and their relationship to occupational classification. *Journal of Applied Psychology, 51*, 481-490.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Childs, A., & Klimoski, R. J. (1986). Successfully predicting career success: An application of the biographical inventory. *Journal of Applied Psychology, 71*, 3-8.

Costa, P.T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory Manual.* Odessa, FL: Psychological Assessment Resources.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Eberhardt, B. J., & Muchinsky, P. M. (1982). An empirical investigation of the factor stabiiity of Owens' Biographical Questionnaire. *Journal of Applied Psychology, 67*, 138-145.

George, J. M. (1989). Mood and absence. *Journal of Applied Psychology, 74*, 317-324.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests.* New York: John Wiley and Sons, 1966.

Hough, L. M. (1987). *Literature review: Utility of temperament, biodata, and interest assessment for predicting job performance.* Minneapolis: Personnel Decisions Research Institute.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44*, 763-792.

Mumford, M. D., & Owens, W. A. (1984). Individuality in a developmental context: Some empirical and theoretical considerations. *Human Development, 27*, 84-108.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 1-31.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.) *Handbook of Industrial and Organizational Psychology.* Chicago: Rand McNally.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. *Journal of Applied Psychology, 64*, 569-607.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Russell, C. J., Mattson, J., Devlin, S. E., & Atwater, D. (1990). Predictive validity of biodata items generated from retrospective life experience essays. *Journal of Applied Psychology, 75*, 569-580.

Shaffer, G. S., Saunders, V., & Owens, W. A. (1986). Additional evidence for the accuracy of biographical data: Long-term retest and observer ratings. *Personnel Psychology, 39*, 781-809.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B., & Rogers, R. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports, 51*, 663-671.

Staw, B. M., Bell, N. E. & Clausen, J. A. (1986). The dispositional approach to job attitudes: A lifetime longitudinal test. *Administrative Science Quarterly, 31*, 56-77.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (1989). Life history prototypes in the study of human individuality. *Journal of Personality, 57*, 509-545.

Trent, T., & Quenette, M. A. (1992). *Armed Services Applicant Profile (ASAP): Development and validation of operational forms* (NPRDC Technical Report 92-9). San Diego, CA: Navy Personnel Research and Development Center.

Trent, T. (1992). The adaptability screening profile (ASAP). In T. Trent & J. H. Lawrence (Eds.), *Adaptability screening for the armed forces.* Washington, D. C.: Office of Assistant Secretary of Defense (Force Management and Personnel).

Effects of Coaching on Validity of a Self-Report Temperament Measure[1]

Mark C. Young and Leonard A. White
U.S. Army Research Institute

Scott H. Oppler
American Institutes For Research

Introduction

The U.S Army is considering implementing a new personnel selection and classification instrument called the Assessment of Background and Life Experiences (ABLE). ABLE measures temperament constructs for predicting the "will-do", motivational components of performance in Army occupations. Results from several investigations show that ABLE provides incremental validity over the Armed Services Vocational Aptitude Battery (ASVAB) and educational attainment as a predictor of first-term attrition, leadership potential, and indiscipline (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; White, Nord, & Mael, 1990, White, Nord, Mael, & Young, in press).

A key problem with self-report instruments like ABLE is the potential for persons to raise their scores by selecting the most socially desirable responses. This social desirability response style, or faking good, distorts the self-report information and may invalidate test scores. If the Army implements ABLE, self-help books might be written to coach applicants on how to do well on the test. Moreover, in order to meet quotas, recruiters might encourage or even train applicants to respond in a particular manner.

Dunnette, McCartney, Carlson, & Kirchner (1962) reported that score variances on the Adjective Checklist decreased when salesmen and applicants were instructed to fake. Moreover, there was a complete loss of validity when salesman completed the test under faking versus honest instructions. These researchers noted that "even the relatively small amount of faking occurring among a subgroup of applicants seems sufficient to attenuate the validities of the various Checklist scores" (pp. 23-24).

Hough and her associates (Hough et al., 1990) have suggested that faking good may not be a serious threat to the validity of ABLE. Using a research sample of Army soldiers, they compared ABLE validities for two groups which differed in socially desirable responding. Validities for the "overly desirable" group were only slightly lower than for those labeled as "accurate" responders. The soldiers in the Hough et al. concurrent validity research completed ABLE for research purposes. Thus, there was little incentive for faking good, and the amount of faking was low.

Research we presented at the Military Testing Association convention last year raises new concerns about the effects of faking and coaching on ABLE's validity. Coaching raised ABLE scores .5 standard deviation higher than when persons were simply asked to fake good (Young, White, & Oppler, 1991). This effect is particularly striking given how simple and short the coaching instructions were (i.e., instructions and 3 practice items with feedback were provided within a 5-minute period). Using the same dataset, we now examine the effect of coaching on ABLE's validity.

[1] Presented at the meeting of the Military Testing Association, October, 1992. All statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Method

Subjects

Coaching Experiment. A total of 973 Regular Army accessions were administered the 199-item ABLE at U.S. Army Reception Battalions at Ft. Knox, KY (n=606) and Ft. Jackson, SC (n=367). In this sample, 80% percent were male, 27% minorities, and 95% high school graduates. The recruits were tested in groups ranging in size from 7-182.

Longitudinal Validation (LV) Sample. A total of 48,731 Regular Army soldiers from the Project A Longitudinal Validation sample completed the 199-item ABLE. The Longitudinal Validation sample consists of 50,255 soldiers who were administered a battery of predictor measures upon entering the Army during FY86/87. These recruits were told that their test scores would be used for research purposes only. (See Campbell & Zook, 1991, for a description of the data collection procedures).

Instruments

Assessment of Background and Life Experiences (ABLE). The ABLE consists of 7 content scales to measure temperament constructs (i.e. Work Orientation, Dominance, Dependability, Adjustment, Cooperativeness, Internal Control, and Physical Condition) and 2 validity scales. The Social Desirability scale was developed to identify individuals who raise their scores by faking good. The Nonrandom Response scale was developed to detect careless (or random) responding. See Hough et al., 1990, and White et al. (in press) for a detailed description of ABLE.

Procedure

Participants were told the research purpose was to learn how different test-taking strategies affect ABLE scores. All were asked to imagine themselves as civilians applying to join the Army. Subjects were randomly assigned to one of three instructional conditions. The important components of each condition are as follows:

Honest. Subjects were asked to select the response which "best describes your background, opinions, or feelings."

Ad-lib Faking. Subjects in this condition were asked to choose the answers "that you think will impress the Army the most ... and make sure that the Army selects you."

Coached. Respondents were given the Ad-lib Faking instructions followed by three practice items. The coach identified the "correct" response to each item, and explained why it was the best answer for impressing the Army. The practice items measured attributes of dependability, cooperativeness, and speed of learning. The "correct" responses were those indicating that the respondent possessed the highest levels of these attributes.

Attrition Criterion

Eighteen-month attrition was the criterion used for the validity analyses. This criterion was coded as 1 when a soldier separated before 18 months of service, and as 0 otherwise. Attrition status was obtained from the Army's Enlisted Master File (EMF). The data for 8 soldiers who died or separated from the enlisted force to become officers were excluded from these analyses.

Results

Screening for Missing Data and Random Responding

Coaching Experiment. Respondents who omitted more than 10% of ABLE items (20 items) or who were classified as responding randomly were excluded from the data analyses. An ABLE scale score was set to missing if more than 10% of the items on that scale were omitted. If fewer than 10% were omitted, scores on the missing items were imputed as the average of the person's scores on the remaining items for that scale. Using these criteria, 56 subjects (less than 6% of the sample) were dropped from subsequent data analyses.

Project A Longitudinal Validation Sample. Using the same criteria (described above) for screening data, 5.595 subjects (less than 12%) were dropped from the analyses.

Manipulation Checks

Coaching Experiment. The three practice items were included within ABLE. These items were analyzed to assess coaching effectiveness at the individual level. Eighty-seven percent chose the coached response on two of the three items, while 74% chose the coached response on all three items. In sum, a majority of examinees followed the coach's instructions, although over 10% did not.

Effect of Instructional Condition

Table 1 presents the means, standard deviations, and effect size estimates for ABLE scales by instructional condition. Eight single-factor ANOVAs were used to evaluate the effect of instructional condition on the ABLE scales. This main effect was highly significant ($p<.0001$) for all scales. The highest F value was obtained for the Social Desirability scale, $F(2,914) = 216$, $p<.0001$). Individual comparisons of scale means by condition were made using the Scheffe test. All group means differed significantly from one another on all scales.

Table 1
Effects of Coaching on Mean ABLE Scores

		Honest (*n*=255-256)		Ad-Lib Faking (*n*=318-320)		Coaching (*n*=340-341)		Effect Size[a]	
Scale	Number of Items	*M*	*SD*	*M*	*SD*	*M*	*SD*	Coaching vs. Honest	Coaching vs. Faking
Work Orientation	28	63.8	10.1	71.1	10.4	78.2	7.5	1.7 (2.2)	.8 (1.2)
Dominance	19	42.3	7.1	46.2	7.6	51.8	6.6	1.4 (2.1)	.8 (1.3)
Dependability	21	50.2	6.0	53.8	5.8	56.3	5.7	1.0 (1.5)	.4 (.8)
Adjustment	15	34.4	5.7	37.3	5.7	40.6	5.2	1.1 (1.7)	.6 (1.1)
Cooperativeness	10	24.6	3.4	26.2	3.1	28.0	2.7	1.1 (1.6)	.6 (1.0)
Internal Control	13	34.8	3.6	36.3	3.1	37.6	2.3	.9 (1.3)	.5 (.8)
Physical Condition	8	17.5	4.0	19.2	3.9	21.7	3.2	1.2 (1.6)	.7 (1.1)
Social Desirability	11	15.7	3.6	18.6	4.8	23.8	5.7	1.7 (2.3)	1.0 (1.4)

[a] Effect size is the difference in group means divided by the pooled standard deviation. Positive effect sizes indicate that coached respondents score higher than the comparison groups. Effect sizes for respondents (*n*=251) who gave the coached response to all 3 manipulation check items are presented in parentheses.

As compared to those responding honestly, ABLE content scale scores of coached respondents were raised by an average of 1.2 standard deviations. More importantly, coaching raised ABLE scores significantly higher (.6 standard deviation) than those of uncoached respondents who were simply asked to distort their answers in a positive direction.

Since 26% of respondents did not fully comply with the coaching instructions, we also compared the scores of those who answered all three manipulation check items correctly (74%) with those in the Honest and Ad-lib Faking conditions. The effect sizes for these comparison are presented in parentheses within Table 1. The effects of coaching on ABLE scores were much greater for those who followed the coach's guidance. Relative to the Honest and Ad-lib Faking conditions, effectively coached respondents raised their scores an average of 1.7 and 1.0 standard deviations, respectively.

Validities of ABLE Scales By Condition

As shown in Table 2, all content scales were negatively correlated with attrition (all $p < .05$) in the Project A Longitudinal Validation (LV) sample. Since the response set for the Honest condition of our coaching experiment is equivalent to that in the Longitudinal Validation sample (in which soldiers were asked to complete ABLE for research purposes only), we hypothesized that the same predictor-criterion relationships would be found in the Honest condition. This was generally supported by the results shown in Table 2. In the Honest condition, all but one content scale (Dominance) had the predicted negative relationship with attrition. Two of these (Adjustment and Physical Condition) were significant ($p < .05$, one-tailed). The lack of statistical significance for most correlations may be due to the small sample sizes.

Table 2
Effects of Coaching on ABLE's Validity for Predicting 18-Month Attrition

ABLE Scale	LV Sample[a] (*n*=40,192-43,136)	Honest (*n*=250-251)	Ad-Lib Faking (*n*=316-318)	Coaching (*n*=339-340)	Combined (*n*=906-909)
Work Orientation	-.05*	-.01	-.05	.07	-.01
Dominance	-.04*	.05	-.11*	.08	-.01
Dependability	-.09*	-.09	-.06	.08	-.02
Adjustment	-.10*	-.18*	-.07	.11	-.05
Cooperativeness	-.07*	-.09	-.06	.03	-.05
Internal Control	-.05*	-.07	-.08	.10	-.03
Physical Condition	-.09*	-.13*	-.11*	.05	-.07*
Social Desirability	.00	-.03	.00	.13*	.03
Attrition Rate	.16	.15	.17	.14	.15

[a] These new recruits were told their test scores would be used for research purposes only and were asked to respond honestly.

*$p < .05$. With the exception of Social Desirability, all tests were one-tailed.

ABLE scale validities for the Ad-Lib Faking group were highly similar to those obtained in the Honest condition. None of the correlations across these two conditions was significantly different (all $p > .05$, two-tailed).

ABLE content scales were not negatively correlated with attrition in the Coaching condition (p>.05, one-tailed). Unexpectedly however, the magnitudes of their correlations were typically as large as those in the other conditions, albeit opposite in sign. The reason for this complete reversal in scale/criterion relationships is unclear. The majority cf content scale validities for coached subjects (i.e., Dependability, Adjustment, Internal Control, and Physical Condition) differed significantly (p<.05, two-tailed) from those obtained in the Honest condition.

Finally, validities for the combined group (i.e., Honest, Ad-Lib Faking, and Coaching) are presented in Table 2. This group provides a wide range of response sets, as might be expected in an operational context. Although the majority of individuals were asked to distort their scores, all validities for this group were in the predicted direction; with the Physical Condition scale being statistically significant (p<.05, one-tailed)

Validities by Social Desirability Responding

In our final analyses, we used the Social Desirability scale to detect high levels of faking among persons in the combined group. Persons scoring at or above 24 on the Social Desirability scale were identified as highly effective fakers. This cut point (24) corresponds to the mean score obtained in the Coaching condition. For persons scoring lower on Social Desirability (n=662), all ABLE content scales were negatively correlated with attrition, and all but two of these correlations were significant (p<.05, one-tailed). Thus, the observed validities for this group are what we would expect to find when subjects are responding honestly. In contrast, validities for highly effective fakers (n=247) were generally of opposite sign, as was observed among the Coaching group.

Discussion

In this experiment, coaching caused a complete reversal of the negative relationships expected between ABLE content scales and attrition. Unlike persons instructed to respond honestly or fake good, coached respondents who scored high on ABLE scales were actually more likely to attrit. This suggests that coaching could be a problem for the operational use of ABLE, unless effective countermeasures (i.e., the detection of faking, warning statements during testing, and the adjustment of scores for faking; see White, Nord, Mael, & Young, in press) were used.

This research also demonstrated the utility of ABLE's Social Desirability scale to detect faking. In an operational context, it is assumed that a wide range of response sets would be used by persons completing ABLE. While some would respond honestly, others might be coached to dramatically inflate their scores. Our results show that ABLE's validity can be preserved by excluding respondents identified as effective fakers.

Future research should address why faking on a temperament measure may alter the direction of the predictor criterion relationship. This type of reversal was also reported by Dunnette et al. (1962). An understanding of this phenomenon might help in developing strategies to control faking.

In future research, we plan to follow up with this and other samples to determine if the loss of validity that results from coaching can be recovered through the adjustment of scores for faking. We will also continue to explore the use of objective biodata as a way of measuring ABLE constructs without some of the inherent problems of traditional temperament measures (Mael & Schwartz, 1992; Mael, Schwartz, & McLellan, 1992; White & Kilcullen, 1992).

References

Campbell, J. P., & Zook, L. M. (Eds.). (1991). Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A (ARI Research Report 1597). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Dunnette, M. D., McCartney, J, Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. Personnel Psychology, 15, 13-24.

Hough, L.M., Eaton, N.K., Dunnette, M.D., Kamp, J.D., & McCloy, R.A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities, Journal of Applied Psychology, 75, 581-595.

Mael, F. A., & Schwartz, A. C. (1992, April). Capturing adaptability constructs with objective biodata. Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology, Montreal, Quebec.

Mael, F. A., Schwartz, A. C., & McLellan, J. A. (1992, August). Antidotes to dustbowl empiricism with objective biodata. In Rumsey, M. G. (Chair) Biodata advances: Bridging the rational and empirical perspectives. Symposium presented at the meeting of the American Psychological Association, Washington, D.C.

White, L. A., & Kilcullen, R. N. (1992, August). The validity of rational biodata scales. Paper presented at the Annual Meeting of the American Psychological Association, Washington, D.C.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (in press). The Assessment of Background and Life Experiences (ABLE). In T. Trent and J. H. Laurence (Eds.), Adaptability screening for the armed forces. Washington, D.C.: Office of Assistant Secretary of Defense (Force Management and Personnel).

Young, M. C., White, L. A., & Oppler, S. H. (1991, October). Coaching effects on the Assessment of Background and Life Experiences (ABLE). Paper presented at the 33rd Annual Conference of the Military Testing Association, San Antonio, TX.

The Validity and Adverse Impact of Various Biodata Keying Procedures [1]

Robert N. Kilcullen
U.S. Army Research Institute

Kirk K. Thor and Elizabeth R. Carroll
Consortium of Universities of
the Washington Metropolitan Area

The U.S. Army Research Institute is currently evaluating new selection instruments for first-line civilian supervisory positions in the Department of the Army. Included in this battery is a biodata instrument measuring prior behavior and reactions to life events thought to be relevant to supervisory performance. Previous research suggests that biodata is predictive of a wide variety of job behaviors including leadership, turnover. trainability, and performance ratings (Mumford & Stokes, 1991). Reilly and Chao (1982) reviewed 44 studies using biographical information as a predictor and reported an average cross-validity of 0.35 across occupations and criteria.

Empirical keying strategies are traditionally used to score biodata. The empirical method involves weighting item responses based on their relationship to the criterion of interest. This method has been criticized because empirically-keyed instruments often show high validity initially, but substantial shrinkage across samples and over time (Dunnette, Kirchner, Erikson, & Banas, 1960; Schwab & Oliver, 1974; Walker, 1985; White & Kilcullen, 1992). In addition, item selection and scoring is atheoretical, which makes it difficult to understand what is being measured or why different criterion groups respond differently to the biodata items. Awareness of these problems has led to increasing interest in rational keying strategies. This typically involves identifying constructs (e.g., Need for Achievement) likely to predict the criterion of interest and writing items that tap these constructs. Response weights are determined a priori based upon the presumed relationship between the response and the construct measured. The scored item responses are then summed to form scale scores having substantive meaning. The potential advantages of rational keying include a greater theoretical understanding of the phenomenon under study as well as stable validities (Mumford & Stokes, 1991; Mumford, Uhlman & Kilcullen, 1992).

Contradictory results come from studies comparing the relative validity of rational and empirical keys. Mitchell & Klimoski (1982) reported higher validity for an empirical key compared to a rational key. On the other hand, Uhlman, Reiter-Palmon & Connelly (1990) used a biodata questionnaire to predict various measures of academic performance and adjustment for 1,969 college freshmen. Compared to the empirical keys, the rational keys were found to have higher cross-validities for all but one of the performance and adjustment criteria, with rational coefficients as high as .68. Another study by Schoenfeldt (1989) compared the validities obtained using empirical, rational, and other keys for predicting supervisory ratings of overall performance and service orientation as well as absence/tardiness criteria for 867 service employees in a large corporation. Both a priori and post hoc rational scales proved to be better predictors than the empirical scale. Finally, Clifton, Kilcullen, Reiter-Palmon & Mumford (1992) found that the validities of rational keys for predicting achievement were consistently more stable compared to empirical keys across a variety of sample sizes.

The Army's biodata instrument for first-line civilian supervisors is designed to accommodate both rational and empirical scoring techniques, as well as a combination of these methods. This paper describes the validity of rational, empirical and rational/empirical keys for predicting supervisory performance. Since the biodata instrument is intended for operational use black/white and male/female differences are also assessed. Comparisons are made with respect to differential validity, slope differences and the potential for adverse impact at various cutoff scores.

[1] Presented at the meeting of the Military Testing Association, October, 1992. All statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

Method

Subjects and Procedure

A total of 2,044 first-line civilian supervisors in the Department of the Army served as subjects. A wide variety of occupations and grade levels were represented in the sample. The demographic composition of the subjects were as follows: 30 percent were female, 74 percent were White, 17 percent were Blacks, and the remainder were Asians and Hispanics. Nearly all of the subjects completed high school, 43 percent had some college experience, 23 percent graduated from college, and 10 percent had graduate or professional degrees. Half of the subjects were randomly assigned to a key-construction group and half to a cross-validation group.

Data collection took place at 54 locations throughout the Continental United States. Subjects were briefed about the purpose of this research and were then administered the selection battery consisting of three instruments, one of which was the biodata inventory. The order of administration was randomized. In a separate session, ratings of the subjects' job performance were collected.

Biodata Instrument

To accommodate rational scoring, biodata scales were developed to measure individual characteristics identified as having potential for predicting the performance of supervisors based on a job analysis and a literature review. A panel of psychologists reviewed the construct definitions and each generated 10 to 15 items referring to past behaviors and life events thought to be indicative of the construct in situations most individuals would be exposed to by adulthood. Candidate items were reviewed by the panel for construct relevance, response variability, relevance to the Army Civilian population, readability, non-intrusiveness and neutrality with respect to social desirability. A consensus decision was then reached concerning the 20 to 40 best items for each construct, and response options were weighted on a continuum to reflect the presumed relationship between the item responses and the predictor construct. The surviving items were reviewed by a second panel of psychologists and were then pilot-tested and revised based on item analysis.

Criteria

Job performance ratings from the subjects' immediate supervisors and their supervisors-once-removed served as one criterion. Subjects were evaluated using 19 separate scales. Six scales tapped generic job performance dimensions such as Work Quantity, Accuracy, and Job Knowledge. Thirteen scales were derived from subject-matter-expert workshops where first-line supervisory performance dimensions and their behavioral anchors were identified. Included were scales pertaining to maintaining employee morale, resolving conflicts, providing personal/career counseling, coordinating with supervisors in other units, and maintaining standards and discipline. The alpha reliability of the 19-scale instrument was .963.

A second criterion was a self-report measure of administrative records of job performance over the past four years. Included were items relating to the frequency of awards, letters of commendation, performance-based pay raises, disciplinary actions as well as several other verifiable indicators of performance. Responses were summed to form a composite score. Previous research in the U.S. Army (Campbell, 1987) indicated that self-reports were more accurate than official records due to errors in processing personnel actions and delays in updating personnel files. The alpha reliability of this measure was .65.

Keying Methodologies

In the rational approach response weights were preassigned by the panels of psychologists. Alpha reliabilities were calculated for each scale in the key-construction sample and revisions were made in some cases to improve internal consistency. The scales were regressed onto the criteria in the key-construction sample and the resulting regression weights were applied to the cross-validation sample.

The rational/empirical (R/E) approach mimicked the rational approach except that items were retained only if they correlated above .10 with the criteria in the key-construction sample. Scales were obtained by factor analyzing the surviving items. The regression weights of these scales were derived and applied in the same way as in the rational approach.

In the empirical approach unit weights of 0, 1 and 2 were applied to item response options based on the average criterion rating of individuals selecting those responses in the key-construction sample. Items that correlated higher than .10 with the criteria in the key-construction sample were retained and summed in the cross-validation sample to obtain predictor scores.

Results

The internal consistency of the rational scales in the cross-validation sample are presented in Table 1. Coefficient alphas range from .65 for the nine item Communication Skill scale to .85 for the 27 item Stress Tolerance scale. All but five of the scales have reliabilities at or above .70.

Table 1.
Rational Biodata Scale Statistics (N=962)

Scale	No. of Items	M	SD	Alpha
Cognitive Ability	37	3.31	.38	.82
Social Maturity	14	4.08	.34	.69
Self-Esteem	18	3.52	.39	.69
Dominance	24	3.16	.36	.76
Harm Avoidance	19	2.73	.38	.70
Consideration	17	3.35	.42	.78
Object Belief	33	2.97	.36	.83
Defensiveness	17	2.92	.41	.74
Social Alienation	14	2.64	.47	.79
Achievement	22	3.45	.48	.77
Interpersonal Skills	29	3.03	.40	.82
Dependability	23	3.90	.37	.79
Planning/Organizing	18	2.88	.42	.73
Supervising Skills	10	2.93	.56	.69
Communication Skills	9	3.11	.58	.65
Work Motivation	15	3.56	.46	.69
Practical Intelligence	28	3.27	.36	.81
Stress Tolerance	27	3.21	.42	.85
Energy Level	11	3.40	.49	.73
Self-Monitoring	25	2.84	.36	.79
Need for Security	25	2.72	.40	.83
Need for Approval	15	3.07	.46	.76

The cross-validities of the three keys are presented in Table 2. Each key yields validities higher than .30 for both criteria. Ratings are best predicted by the R/E (r=.38) and Empirical (r=.37) keys. Performance records are best predicted by the R/E (r=.48) and Rational (r=.43) keys. No correction for range restriction is applied to the validities. Validity shrinkage from the key-construction sample to the cross-validation sample averaged .15 for the Empirical key and .05 for the Rational and R/E keys.

Table 2.
Cross-Validation Results

	Criterion	
Scoring Method	Ratings	Performance Records
Rational	.32 (N=962)	.43 (N=1022)
Rat/Empir.	.38 (N=805)	.48 (N=820)
Empirical	.37 (N=937)	.39 (N=1002)

Table 3 presents subgroup means for each key. In general, effect sizes for race compare favorably to those of other selection procedures. Only when the Rational key predicted the rating criterion did a substantial effect size occur. Effect sizes in gender comparisons are generally larger but favor females.

Table 3.
Predictor Statistics by Subgroup

		Subgroup					
				Effect			Effect
Criterion	Key	Whites	Blacks	Size	Males	Females	Size
Ratings	Rat.	.04	-.19	.59	-.03	.06	-.23
		(.38)	(.38)		(.39)	(.41)	
	R/E.	.03	.01	.05	-.02	.15	-.45
		(.38)	(.38)		(.38)	(.37)	
	Emp.	1.22	1.21	.06	1.20	1.27	-.41
		(.17)	(.16)		(.17)	(.17)	
Ratings	Rat.	2.68	2.68	.00	2.68	2.69	-.07
		(.14)	(.15)		(.14)	(.16)	
	R/E.	2.69	2.69	.00	2.67	2.73	-.32
		(.19)	(.21)		(.19)	(.19)	
	Emp.	.99	1.04	-.26	.98	1.04	-.32
		(.19)	(.19)		(.19)	(.19)	

Table 4 presents validities for predicting ratings in White, Black, male and female subgroups. All tests for differential validity among Whites and Blacks were not significant. Similarly, no evidence of differential validity was found between males and females. The rational key had the largest subgroup differences, with minority validities seven to eight percentage points below majority validities.

Table 4.
Differential Validity for Ratings Criterion

Scoring Key	Whites	Blacks	Significance
Rational	.32 (N=716)	.25 (N=172)	n.s.
Rat/Empir	.38 (613)	.35 (123)	n.s.
Empirical	.38 (706)	.37 (153)	n.s.
	Males	Females	Significance
Rational	.34 (667)	.26 (295)	n.s.
Rat/Empir	.38 (563)	.37 (242)	n.s.
Empirical	.37 (656)	.35 (281)	n.s.

Subgroup validities for predicting performance records are found in Table 5. Again, tests of differential validity between subgroups were not statistically significant. In this case the largest subgroup differences were obtained with the Empirical key.

Table 5.
Differential Validity for Performance Records Criterion

Scoring Key	Whites	Blacks	Significance
Rational	.43 (N=730)	.39 (N=189)	n.s.
Rat/Empir	.50 (638)	.46 (118)	n.s.
Empirical	.40 (754)	.32 (162)	n.s.
	Males	Females	Significance
Rational	.42 (689)	.41 (319)	n.s.
Rat/Empir	.48 (559)	.47 (256)	n.s.
Empirical	.35 (689)	.43 (305)	n.s.

Slope differences between subgroups were also examined. The pattern of results was the same across keys and criteria. In every case no Black/White or Male/Female slope differences were detected.

Finally, Table 6 and Table 7 present selection ratios at three cutoff scores, the 20th, 40th and 60th percentiles, to assess possible adverse impact. When predicting ratings the rational key demonstrated adverse impact against Blacks at each cutoff score. No other evidence of adverse impact was found. In fact, female selection ratios were consistently higher than male selection ratios across all keys.

Table 6.
Adverse Impact for Predicting Ratings

		Selection Ratio				
Scoring Method	Predictor Cut-Score	Whites	Blacks	Males	Females	Adverse Impact
Rational	20	.84	.64	.78	.84	Blacks
	40	.65	.41	.59	.63	Blacks
	60	.43	.26	.37	.46	Blacks
Rat/Empir	20	.80	.77	.76	.90	n.a.
	40	.60	.58	.55	.71	n.a.
	60	.40	.37	.35	.51	n.a.
Empirical	20	.79	.81	.77	.94	n.a.
	40	.59	.60	.55	.76	n.a.
	60	.40	.39	.33	.60	n.a.

Table 7.
Adverse Impact for Predicting Performance Records

		Selection Ratio				
Scoring Method	Predictor Cut-Score	Whites	Blacks	Males	Females	Adverse Impact
Rational	20	.80	.79	.80	.81	n.a.
	40	.59	.58	.60	.61	n.a.
	60	.38	.41	.39	.41	n.a.
Rat/Empir	20	.80	.79	.78	.85	n.a.
	40	.59	.64	.57	.68	n.a.
	60	.38	.44	.35	.50	n.a.
Empirical	20	.79	.85	.78	.85	n.a.
	40	.58	.67	.57	.66	n.a.
	60	.38	.52	.37	.49	n.a.

Discussion

Each of the three keys were able to predict the performance of first-line civilian supervisors, with uncorrected validities surpassing .30 in every case. It remains unclear whether the rational or empirical approach yields the best validities. The empirical key was better at predicting ratings given by supervisors (r=.37) but the rational key was better at predicting performance records (r=.43). Given the validity shrinkage commonly associated with empirical keys (e.g., White & Kilcullen, 1992), the rational key may be the preferred approach if consideration is given only to validity.

However, subgroup analyses reveal that the rational key may adversely affect Blacks. When predicting ratings the rational key had the highest effect size for race, the highest White/Black differences in validity, and also demonstrated adverse impact at each cutoff score. Still, the rational key may be salvageable. Previous research with empirically keyed instruments indicates that White/Black differences can be eliminated with modest losses in test validity by deleting items having poor validity for Blacks (Gandy, Outerbridge, Sharf, & Dye, 1989). Since rational items are not criterion-based it may be possible to delete items contributing to subgroup differences without a loss in validity. Future research will explore this issue.

The results also suggest that the Rational/Empirical key may be a viable alternative to the strictly rational or empirical approach. In this research the R/E key yielded the best validities for both criteria, it's White/Black effect size was small, differences in validities between Whites and Blacks were also small, no subgroup slope differences were found, and no evidence of adverse impact was observed. The stability of R/E validities are unknown, but research by White & Kilcullen (1992) found that biodata items with stable validities over a four year period were similar to the R/E items in that they mimicked rational response weighting but were also initially correlated with the criterion. Future research will examine the stability of R/E keys.

References

Campbell, J.P. (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year. (ARI Technical Report 746). (AD A133 343)

Clifton, T.C., Kilcullen, R.N., Reiter-Palmon, R., & Mumford, M.D. (1992). Comparing different background data scaling procedures using triple cross validation. Paper presented at the 100th annual meeting of the American Psychological Association, Washington DC.

Dunnette, M.D., Kirchner, W.K., Erikson, J., & Banas, P. (1960). Predicting turnover among female office workers. Personnel Administration, 23, 45-50.

Gandy, J.A., Outerbridge, A.N., Sharf, J.C., & Dye, D.A. (1989). Development and initial validation of the Individual Achievement Record (IAR). U.S. Office of Personnel Management, Washington DC

Mitchell, T.W., & Klimoski, R.J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. Journal of Applied Psychology, 67, 411-418.

Mumford, M.D., & Stokes, G.S. (1991). Developmental determinants of individual action: Theory and practice in the application of background data. In M.D. Dunnette (ed.) The handbook of industrial and organizational psychology. (2nd edition). Orlando, FL: Consulting Psychologists Press.

Mumford, M.D., Uhlman, C.E., & Kilcullen, R.N. (1992). The structure of life history: Implications for the construct validity of background data scales. Human Performance, 5, 109-137.

Reilly, R.R., & Chao, G.T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

Schoenfeldt, L.F. (1989). Biographical data as the new frontier in employee selection research. Address presented at the annual meeting of the Division of Evaluation, Measurement and Statistics of the American Psychological Association, New Orleans.

Schwab, D.P., & Oliver, R.L. (1974). Predicting tenure with biographical data: Exhuming buried evidence. Personnel Psychology, 27, 125-128.

Uhlman, C.E., Reiter-Palmon, R., & Connelly, M.S. (1990). A comparison and integration of empirical keying and rational scaling of biographical data items. Paper presented at the Southeastern Psychological Association

Walker, C.B. (1985). The fakability of the Army's Military Applicant Profile (MAP). Paper presented at the Association of Human Resources Management and Organizational Behavior proceedings, Denver CO.

White, L.A., & Kilcullen, R.N. (1992). The validity of rational biodata scales. Paper presented at the annual 100th meeting of the American Psychological Association, Washington DC.

TRAINING REQUIREMENTS FOR IMPLEMENTATION OF A COMPUTERIZED EXAMINATION DEVELOPMENT SYSTEM (MADMAX)

Suzanne G. Fischer
GSCS David A. Power

NAVAL EDUCATION AND TRAINING PROGRAM MANAGEMENT SUPPORT ACTIVITY (NETPMSA)

A mission of the Navy Advancement Center Department at the Naval Education and Training Program Management Support Activity (NETPMSA) is to develop advancement-in-rating examinations for all Navy enlisted occupations.

BACKGROUND

In many ways, NETPMSA is the heart of the Navy's enlisted advancement system. At NETPMSA, advancement exams for the Navy's enlisted occupations (ratings) are developed, produced, distributed, controlled, and statistically analyzed. Over 87 advancement exam development teams support over 96 separate enlisted ratings. An advancement exam development team for a rating consists primarily of a Supervisory Instructional Systems Specialist (SISS) or an Instructional Systems Specialist (ISS) as team leader and an exam writer who is an active-duty senior enlisted member in the rating and serves as the resident subject matter expert for that rating. To support a single rating, the advancement exam development team must develop and produce up to seven different 150-item, norm-referenced advancement exams per year. In 1991, NETPMSA's exam development teams developed over 630 separate advancement exams enabling NETPMSA to provide more than 343,272 advancement exams to active-duty U.S. Navy personnel and 39,368 exams to U.S. Naval Reserve personnel.

In the Navy's enlisted advancement system, the advancement-in-rating exams represent an important tool. First, they allow the individual Navy member a means to enhance his or her own opportunities for advancement by studying. Second, they allow each rating community a fair way to rank order its own candidates for promotion. Most importantly, they serve as an equitable tool the Navy can use to carry out its manning priorities by selecting only the best from a multitude of highly qualified people. By both serving as an important management tool and allowing the individual a level of control over his or her career, the advancement-in rating exams form an integral part of the integrity long attributed to the Navy's enlisted advancement system.

The Navy's enlisted advancement system is unique because it must represent and affirm the important roles of Navy enlisted personnel as they carry out their daily jobs in the isolated world of floating platforms and remote duty stations. Navy ratings are occupations that provide the multitude of crafts, trades, skills, expertise, services, and specialties required to keep our ships afloat, manned, armed, serviced, and ready to carry out the Navy's role in the Department of Defense. Each Navy rating is divided into promotional levels called paygrades. Each paygrade encompasses a group of job-related requirements, skills, knowledges, training, performance elements, duties, responsibilities, and professional commitments that comprise the full-performance, on-the-job reality of an enlisted person in that paygrade within the larger technical arena of his or her rating. This is why each Navy advancement-in-rating examination is designed to test the whole person within the scope of the entire paygrade and rating and not just a specific task or level of expertise concerning a particular system or piece of equipment. Although diversity and isolation may be intrinsic to naval service, the advancement exam guarantees to each member of a Navy rating the opportunity to compete for promotion against his or her peers on the same day, at the same time, by means of the same testing instrument, and through the same advancement-in-rating process--regardless of duty station or command.

To develop a reliable, norm-referenced advancement exam that will test the whole person takes considerable planning and work. In selecting the subject matter sections for an advancement exam, the development team must make certain that the technical arena of the paygrade and rating are equitably represented. To do this, our exam writers must rely upon years of experience and knowledge. This is why only experienced senior enlisted personnel are assigned to NETPMSA as exam writers. Although the team leader is responsible for three or more different ratings, the exam writer is always an experienced, active-duty E-7, E-8, or E-9 who writes only the exams for his or her own rating.

IMPLEMENTATION OF MADMAX

For two decades, NETPMSA'S teams developed Navy advancement-in-rating exams by use of a 5-by-7-inch item card format to develop, store, accumulate, and retrieve individual test items. The system was simple, reliable, presented very few problems, and resulted in norm-referenced advancement exams of very high quality. What the traditional system did not offer was the storage, retrieval, development, production, and analytical potential of a computerized system. In perpetuating the traditional system, no matter how effective, NETPMSA was perceived to be falling behind the times.

Late in 1990, NETPMSA had a opportunity to acquire computer equipment and initiate a computerized exam development process. In addition to their continuing exam development responsibilities, four NETPMSA exam writers developed the Microcomputer-based Automated Development and Maintenance of

Best Copy

Advancement Examination: (MADMAX) system. MADMAX is a dBASE IV runtime program used with an IBM-compatible microcomputer with a removable hard disk for security. MADMAX promised exam development teams a computerized method to define the exam sections, store and maintain the item bank, build the exam, review the exam, and print both rough and smooth versions of the exam. In other words, MADMAX would make it possible for the Navy's advancement-in-rating exam development teams to maintain much of the reliability of the traditional system and yet capitalize on the speed and storage potential of the computer.

MADMAX was a new system with great potential, but it was not perfect. Once NETPMSA had acquired the computer equipment, exam development personnel were feeling the pressure to get the system started and the writers trained to use it. First, MADMAX had to be developed and tested. Next, a training program and operations manual had to be developed. Finally, the exam writers had to be trained. During all these stages, the bugs in MADMAX had to be worked out.

The exam writers accepted MADMAX with mixed emotions. The hackers and computer-literate writers welcomed MADMAX, or at least tolerated the opportunity to learn how it worked. Other exam writers were apprehensive, afraid, or even hostile. They did not understand why a traditional, effective system had to be replaced by automation. Although they were subject matter experts in the technical world of their ratings, many of the writers had never used a computer system and were not enthusiastic about having to learn to use MADMAX.

Regardless of their acceptance or computer training levels, several exam writers were scheduled to attend MADMAX training. Each work station in the training room consisted of a hard disk drive mounted on a CPU. The monitor was placed on top of the hard disk drive. After the writers took their places, the instructors told them to use to figure program to draw the frame for a figure. The figure program failed to work. The instructors quickly moved to the next program and told the writers to type the statistical data from a control item card and enter it into MADMAX by "hitting the RETURN button." Most of the writers managed to complete this task, but one reluctant E-7 said, "There isn't a thing on my keyboard that says RETURN!" He was right. The button he needed to push was labeled ENTER. When he did push ENTER, he was unable to proof what he had typed because we had placed his monitor too high for him to be able to see through the bifocal portion of his glasses. All this writer knew was that we had forced him to learn a system he did not want, told him to draw a figure with a program that had failed to work, told him to push a button that was not there, and asked him to enter and proof data he could not see--all within five minutes. He summarized his experience in one word that echoed along the back row of writers as MADMAX crashed at three more work stations.

It was a severe moment of truth for everyone. Now, all the exam writers were apprehensive. They were worried about how they were going to use a system that did not work to develop and produce their required seven advancement exams per year on time. The instructors and programmers were also discouraged because after months of intense work, MADMAX had failed to work at a time when the exam writers were being trained to use it. The team leaders were also devastated. After meticulously developing and producing advancement-in-rating exams that tested thousands of people they never saw, they forgot to pretest their own exam writers to establish each writer's level of computer literacy. In developing and presenting a sophisticated computerized advancement exam development system, we all seemed to forget some fundamental realities.

The same spirit of mutual self-support that keeps the Navy's ships afloat and moving emerged in the middle of this crisis. First, the computer equipment was placed back into the daily work stations and arranged so each of the writers could use MADMAX more effectively and comfortably. Next, the computer-literate writers formed help-desk teams to work with the writers with limited or no computer experience. All writers, regardless of training, ability, or acceptance levels, contributed valuable documentation that allowed the programmers to implement MADMAX. Within only nine months, all the exam writers were using MADMAX. Even the E-7 who could not find the RETURN button successfully used MADMAX to develop and produce all his exams. Working together, NETPMSA's advancement exam development teams had performed a miracle.

The implementation process continues. Currently, the operations manual for MADMAX includes sections for documentation or suggestions for improvements. By use of these comments sheets, suggestions for improvements or documentation of potential areas for improvement are submitted regularly to the programmers The help-desk teams continue to provide on-the-job training to the exam writers and act as a resource to the programmers as improvements to MADMAX continue. MADMAX is not yet perfect, but it is steadily getting better.

ADVANTAGES AND DISADVANTAGES

No new computer system is without its advantages and disadvantages. Currently, the exam writers are realizing the advantages and regarding the disadvantages as a source of motivation to improve MADMAX.

One of the most important advantages of MADMAX is that it alleviates the use of cards. The removable hard disk drive and a system of passwords provides for the required security of individual test items. In the traditional system, the exam writer had to cut and paste the statistical results of each exam on the 5- by 7-inch card containing the statistical history of

Best Copy

the item. With MADMAX, statistical results for each item are stored in the item bank.

Another advantage of MADMAX is its validation capability. Not only will MADMAX create and produce traditional documentation, such as a test plan and outline, references to OCCSTDs, authorized subject matter publications for each item, and P-value worksheet (control item inventory sheet), it will also perform validation checks for response (alt) distribution, number of items included in each section and in the exam, weights of the subject matter sections, duplicate items, number of consecutive items with the same response (alt) as the correct answer, and spelling errors.

MADMAX also alleviates the continual passing back and forth of item cards among members of the exam development team while the exam is in its rough (working) stages. MADMAX offers the exam writer continuing control over the development of the exam from the test plan and outline through the production of the camera-ready smooth exam.

We are implementing MADMAX because even a good system will have some disadvantages. An initial disadvantage is that the exam writer is required to have some typing skills or computer experience, or must receive training to acquire these skills. Another disadvantage is that the writer must enter and exit several programs to accomplish a function. In the case of special symbols and mathematical formulas and equations, MADMAX is not as user friendly as it could be. The exam writer must use a variety of codes to gain accessibility to special symbols. The symbols are small, and often the typing staff must create formulas and equations on the typewriter and cut them into the smooth exam.

Another disadvantage of MADMAX is that the writers must print out several copies of the rough exam before rendering the exam in its camera-ready format. This task results in an excessive use of computer paper. Even with numerous roughs, the exam writers have lost some control over the layout of the exam items, especially in the use of figures and instructions (blurbs) to the candidates who will take the exam.

The disadvantages provide motivation for NETPMSA's exam development teams to work together to improve MADMAX. Once the exam item banks expand as more control items are entered into MADMAX and some of the remaining bugs are worked out, our automated exam development process should become much more user friendly and efficient.

AUTHORING SYSTEM CAPABILITY DATABASE

by

Gerry Costello, Director
Applied Courseware Technology, Inc.
P.O. Box 95, Newcastle, NB, Canada, E1V 3M2

Captain Dan Hansen, Staff Officer C.A.L.
&
Lt(N) Roger St-Pierre, Staff Officer Sim & Trainers
Deputy Commandant (Development)
Canadian Forces Fleet School, FMO Halifax, NS, Canada

INTRODUCTION

The task of selecting an appropriate Authoring System for a wide variety of Computer Aided Learning (CAL) applications can be challenging. No single system has the capability of satisfying every user requirement and situation. Each Authoring System has its own strengths and weaknesses which makes it appropriate for certain applications and less advantageous for others. It is, therefore, imperative that users determine which system best meets the needs identified during the analysis process.

When developing courseware design specifications and statements of work, end-users can either stipulate the Authoring System to be employed, or leave the choice to the contractor. In either case, it is possible for the selected Authoring System to have limitations which may restrict its ease of use and/or effectiveness. Additional costs may be necessary to maintain/ upgrade the coursewares once delivered. It is therefore necessary to select an Authoring System based both on the requirements of an application, and the experience/training of the maintenance staff.

The evaluation of Authoring Systems presents two major problems. The first involves accurately describing system capabilities using terminology that can be applied fairly to all systems, quantifying these capabilities, and then assessing the accuracy of the specifications. The second problem area comprises the assessment of the relative effectiveness of each system; that is, to objectively measure and scientifically quantify the learning time, user friendliness, maintenance costs, and vendor support associated with each product.

OBJECTIVE

The objective of this study was to quantify the capabilities and relative ease of use of six Authoring Systems specified by the Canadian Forces Fleet School (Halifax) (CFFSH). This list consists of the Authoring Systems either being considered for or are being used to develop computer-based training for the Fleet.

OVERVIEW OF THE METHODOLOGY

The work associated with this project was divided into two distinct phases as follows:

Phase I - Authoring System Capability Database; and

Phase II - Assessment of System Productivity.

Phase I - Authoring System Capability Database

This phase of the study consisted of the creation of a database which objectively quantified the capabilities of each specified Authoring System. The resulting database is capable of ranking each Authoring System according to an objective set of criteria which provides users with an indication of each system's capabilities relative to the other systems. Clients are able to list each system that meet capabilities derived from the specifications provided by the vendor and assessed by Applied Courseware Technology.

The database is complemented by a search facility which permits users to select critical specifications for a project from lists of capabilities/features provided by the database interface and weight each according to its importance to the users. The search facility then lists the products which meet the critical specifications and ranks products according tc the capabilities, weightings and effort model evaluation data.

The database should therefore help to reduce the work involved in reviewing proposals, and make vendor selection more reliable.

Phase II - Assessment of System Productivity

This phase involved the design of an overall Effort Model and the administration of an objective test of the areas defined in the model.

The development of a courseware authoring effort model was based on the objective measurement of the total effort required for a typical user to develop, modify, and debug a sequence of courseware screens. In essence, the Effort Model was designed to measure all of the features of an Authoring System which contribute to its effectiveness. These include:

Instructional design aids;

Availability of models, templates, sample code, and prompted screens;

Developer interface and intuitiveness;

Ease of maintenance;

Adaptation to author skills levels;

Editing and debugging capabilities;

Help routines available on-line; and

Effective manuals/documentation.

Finally, the Effort Model is comprised of three distinct parts:

System Speeds	
- Response times	- 8 items
- Task timings	- 4 items
Product Support	
- Vendor response rates	- 3 items
- Documentation and support	- 3 items
Development Effort	
- Actions to complete task	- 6 items

RESEARCH AND ANALYSIS OF CAPABILITIES

Prior to proceeding with the analysis, Applied Courseware Technology and the Deputy Commandant (Development) (DComdt(D)) staff of CFFSH agreed to the following definition:

> An Authoring System is a production tool which includes preprogrammed instructional sequences and permits users to input and structure content to create computer-based instruction.

The specifications defining the Authoring System capabilities were then developed from a review of research materials and the personal experience of the study team. A list of generic capabilities and features was subsequently developed and approved by the DComdt(D) staff prior to the evaluation of the most recent versions of the specified Authoring Systems. Care was taken to define those attributes which are necessary in an authoring tool from the view of a developer or instructional designer and which are founded in the well established principles of learning. This approach was based on the principle that courseware should be developed based on pedagogy, rather than on the capabilities of the Authoring System.

A second principle guiding this evaluation was the understanding that products should be given credit for those features which are integral parts of the Authoring System. Conversely, products should not be given credit for features which are not part of the Authoring System, but, rather, are features of an operating system or some other product. If, however, an Authoring System contained a feature which took advantage of the capabilities of the operating system or environment in which it ran, the product was given credit. Finally, if a feature required the user to exit the Authoring System or write additional code in order to develop it, the Authoring System was not given credit for that feature.

Based on these criteria, the following seven categories of capabilities were specified:

Hardware and Devices Supported - describes the ability of each Authoring System to support the 30 devices (grouped into six sub-categories categories of hardware and other devices) identified by the contractor and the staff of DComdt(D).

User Interface - describes the attibutes of an Authoring System interface which increase productivity in the design, development, and production of courseware. This category is comprised of 32 features grouped into seven sub-categories.

Lesson Authoring - documents the 132 possible features of an Authoring System which define limitations on the development, presentation, answer analysis, branching, etc.

CMI Capabilities - defines the 30+ possible features of an Authoring System which control/facilitate training administration, the production of student performance reports, the design reports, and the security of the system.

Lesson Presentation - quantifies those features of an Authoring System which permit the use of student record keeping, lesson summaries, review facilities, glossaries, help routines, book marks, and comments to/from Instructors or Authors.

Documentation and Support - itemizes the number of methods of documentation and support (of the 24 identified by the contractor) possessed by each Authoring System. Further, this category identifies the support and services provided by each vendor.

Associated Costs - specifies the costs associated with the purchase and ongoing use of each Authoring System. The costs associated with the purchase of single and multiple station copies in both the development and delivery modes are documented as are distribution royalties and run-time fees.

Further, evaluation statistics for system speeds, product support, and development effort were also identified and examined.

SUMMARY

This study provided both valuable information on the capabilities of six Authoring Systems and the means to simplify the decision-making process with respect to courseware development. The resulting objectively-derived database lists Authoring System capabilities along with a rating of the effort required to produce courseware using each Authoring System. This database, and its associated Effort Model, can be of value in:

- reducing the time and costs associated with reviewing proposals and tenders;
- improving the specifications for Requests For Proposals;
- providing data to vendors who wish to improve their product;
- reducing project delays and cost overruns; and
- offering information for use in the development of programs to train design, programming, and maintenance staff.

REFERENCES

Albin, Marilyn. (1991). CBT authoring system selection: Features and benefits. CBT Directions. June, p. 20-26.

Becker, Robert S. (1988). How to evaluate CBT authoring systems. CBT Guide 1988. Boston: Weingarten Publications, Inc.

Booz-Allen, Hamilton. (1988). Computer-based training trade study report. Washington: Booz-Allen, Hamilton, Inc.

Collins, M.A.J. (1989). Problems associated with the selection and use of authoring languages and authoring systems. The Supplementary Proceedings of the Second International Conference on Computer Assisted Learning. Dallas: Technical Report - 05, 89, 46-47, 1989.

Costello, G.S. (1992). Developing computer based instruction - A systems design approach. Training Officer. Manchester: Marylebone Press, Vol. 28, No 2. March 1992.

Department of the Army. (1989). Computer based training: Army authoring system requirements. U.S.A.

Gery, G. (1987). Making CBT Happen. Boston: Weingarten Publications, Inc.

Jones, Mark K. (1989). Human computer interaction: A design guide. New Jersey: Educational Technology Publications.

Kearsley, G. (1983). Computer-based training: A guide to selection and implementation. Ontario: Addison Wesley Publishing Company, Inc.

Matthews, T.E. (1988). A model for the instructional design of computer assisted instruction lessonware for adult learners. (unpublished thesis, St. Francis Xavier University, 1988).

Meredith, J.C. (1981). The CAI author/instructor. New Jersey: Educational Technology Publications.

Savage, D.A. et al. (1991). From storyboard to keyboard: A design guide for CBT. Santa Clara: CEIT Systems, Inc.

U.S. Army Research Institute. (1989). Selection of a computer based training authoring system: Functional requirements and evaluation criteria. U.S.A.

Interactive Hypermedia: An Affordable Methodology for Army Training

Dwight J. Goehring

U.S. Army Research Institute

The U.S. Army expends tremendous effort and resources to assure that its personnel are adequately trained. Training and skill sustainment are a complex and continuous endeavor because of many factors including the implementation of new systems, personnel turbulence and skill decay, and an enormous variety of specific jobs to be trained. Not only must soldiers and leaders be proficient in their individual specialties but it is essential that they be capable of working together as crews, teams and progressively larger units. The Army refers to the training of groups as collective training. Today all of these challenges must be met under the shadow of diminishing resources.

As a part of collective training, Army units train to operate in a combined arms and services environment. Army Combat, Combat Support and Combat Service Support branches must all function together with the personnel, high technology weaponry and systems from other services in a highly integrated manner to survive on the battlefield and accomplish the mission. To prepare for such integrated operations requires extensive training in a simulated combat environment.

The Army conducts collective training operations at its Combat Training Centers. The National Training Center (NTC) is the prime facility for large-scale armor and mechanized infantry combat training. It is the most developed of the Combat Training Centers, located in the Mojave desert at Fort Irwin near Barstow, California. The NTC consists of over 1,000 square miles of terrain, accommodating large maneuver and live fire exercises plus nap-of-the earth flying, firing of air defense weapons and practice in the use of electronic warfare.

The Army Research Institute has established an archive of data generated by NTC training exercises conducted at the Combat Training Centers. It develops methodologies for utilization and supports a wide range of analyses conducted with sponsorship of interested agencies throughout not only the Army but other military services and non-defense agencies.

One of the challenges of Army training is how to exploit modern technological advances to improve the effectiveness and efficiency of training, especially collective training. There have been significant advances in both learning technologies and computer sciences which together offer important contributions not only to Army training but to training in other institutional settings. One such advance serves as the foundation for the project reported here.

Terminology

The origins of the concept of connecting ideas electronically in an other than sequential way are traceable back nearly fifty years. An automated information access scheme was envisioned which included the capability of fast accessing and linking of information, of storing the trail of links, and of annotating the retrieved information (Bush, 1945). Englebart (1963) discussed the functionalities necessary for computers to augment human abilities, including links between texts, document libraries, and separate private space for computer users' personal files, computer screens with multiple display windows, and the facilitation of work done by multiple persons working in collaboration. Later, he introduced the mouse computer pointing device, outliner and idea processor, and on-line help systems.

The term ***hypertext*** was coined by T.H. Nelson in 1965. The simplest definition is that hypertext is "nonsequential writing or reading." Other writers (Slatin, 1991) have emphasized that a true hypertext can only be realized in a computer-based system, that it is a system the existence of which is completely dependent upon a computer. When the concept of hypertext was first implemented in operational programs they contained textual material exclusively simply because textual displays were the first non-numerical output of early computers.

In a broader context, hypertext can be thought of as a mode of information organization which is like an n-dimensional web. Items of information are associated without bound. A hypertext user or reader can proceed through the information in a highly individual way consistent with his or her information requirements. In fact, a personal system of information storage based on just this principle of information organization has been recently implemented (Phillips, 1991).

As a generalization of hypertext, ***hypermedia*** is a technology for organizing discrete quantities of information, termed information nodes, for utilization in an arbitrary, nonsequential manner by traversing links between information nodes. Conceptually, an information node is unrestricted as to type of information or amount of information it may contain. It may consist of textual material, still or moving pictures, video and/or sound recordings, independent computer programs, or other entire hyperdocuments.

A link is an association between nodes enabling the user to move from the current node to another. A hyperdocument is a structure of nodes and links which is self-contained, dealing with a particular subject domain for some specific purpose. These concepts are well developed and extensive discussions are available (Nielsen, 1990, Berk & Devlin, 1991)

What is novel about the current time is that the technologies have now matured sufficiently that these techniques are readily available for application in training systems and direct application to the training environment can be made using currently available hardware and software at a realistic cost (Hannafin and Peck, 1988).

Technological Leverage

Constrained resources is emerging as the keynote of the decade. In the arena of Army training as elsewhere the message is to make do with less. One of the ways of moving toward this goal is to take advantage of technological advances when they produce higher quality output at the same or reduced cost, or same output at reduced cost. Using technology for this purpose has been termed technological leverage.

The application of computers for training has held such promise for decades. The effectiveness of interactive videodisk technology in certain Department of Defense training systems has been established in an extensive meta-analysis (Fletcher, 1990). Therefore, the effectiveness of interactive hypermedia, of which interactive videodisk is merely one form, can reasonably be assumed. Thus, the time for fulfillment of the promise of using computers for more effective training is now at hand. Furthermore, because delivery costs continue to decrease, quality training using computers and related technologies is becoming more and more affordable.

Project Goals

The purpose of the research reported here was to explore the integration of existing training exercise data from the NTC into a functioning computer-based system employing interactive hypermedia technology. It is also intended to provide a preliminary evaluation of the principle of coupling this technology and NTC data to provide effective training in fundamental ground-warfare military tactics. The desired outcome was to develop a proof-of-principle system (POPS) for evaluation of this training methodology.

METHOD

One of the greatest strengths of interactive hypermedia for training is that the learner is an active participant in the training process. Beyond making responses and receiving feedback the learner actively structures the sequence in which the material is encountered. The metaphor guiding the project from its beginning is the idea of placing the learner into the highly stimulating environment of an intellectual playpen. The envisioned environment has boundaries, but the concept is to have the world filled with wonderful and stimulating information entities so that the biggest problem of the learner at any point is what to explore, experience and learn about next. The assessment and feedback component assures that all material will be mastered to standard prior to completion by the learner, while leaving the ordering largely to the choices of the learner.

POPS was designed as a single module requiring the learner approximately two hours to complete. The target learner was envisioned as a junior Army leader familiar with basic military terminology. Emphasis was placed on exploring the principles and the range of possibilities of the technologies involved rather than on developing an operational training system.

The NTC training exercise used in the development of POPS provided several different types of information which, together constitute the training exercise data set. Each type of data will now be described. For a more thorough discussion of data generated at the Combat Training Centers, including at the NTC see Hamza (in press).

Instrumented data at NTC are collected through a telemetry system. They include digital records of vehicle positions, firing and engagement information of instrumented players, and several other types of information pertinent to the specific training event. The development of POPS used these data collected at one- and five-minute intervals. Data for air players based on 10-second intervals were also explored.

Take-home Packages are textual descriptions and evaluations of the training unit's performance during the exercise for various echelons and subunits comprising the battle task force. Tables present quantified battle damage assessment data.

Video tape recordings were available of After Action Reviews following completion of the training exercise. These are conducted at various echelons within hours of the exercise conclusion and recorded without editing. An After Action Review is essentially a debrief of the exercise participants where intentions, actions, and outcomes are reviewed.

Audio communications are recorded on 40 channels with time tagging. Additional information about the training exercise was collected through interviews with personnel who participated in the training exercise.

Finally, several additional data sources were available and used in developing POPS. These included acetate overlays to maps and a variety of paper-based information such as orders and task organization listings

Although the training exercise data set contains extensive audio and video material, the decision was made at the design phase of the project to exclude audio and video capabilities from POPS for reasons of economy. However, the information contained in these data sources can still be readily used in a textual medium. Vocal audio recordings can be transcribed to text and video material can be represented in various graphical forms or described in textual.

No hardware acquisition was necessary in support of the project. MS-DOS computers, scanners and a variety of software were available and used in the project.

In order to implement POPS, selection of suitable computer-based training software was necessary. More than 70 MS-DOS based authoring systems are on the market and nearly as many development tools exist for the Microsoft Windows environment (Wilkens, 1991).

While capabilities of such systems change at a rapid pace, none meets every purpose. Any decision involves tradeoffs among alternatives. Gery (1987) identifies three dimensions of tradeoffs--productivity versus creativity, structure versus freedom, and power versus simplicity--and discusses these considerations in detail. She identifies a number of additional factors along which software systems vary: training needs and the training audience; system features, functions and requirements; authoring features, support and training; costs; vendor capabilities; existing courseware; and consultation services.

The authoring system software chosen for this implementation was ***HyperWriter*** (Ntergaid, 1990). It satisfied the paramount criteria for the project: authoring and system capabilities and requirements, and immediate availability.

RESULTS

The functioning POPS is comprised of 129 nodes with nearly 300 links among them. POPS incorporates two stand-alone software entities into the commercial authoring system[1]. Approximately two and a half megabytes of disk storage are used for the system and its various supporting data files. Each of the distinct types of display is now described to provide a detailed understanding of POPS.

Textual data were taken directly from various data sources with minor editing. Misspellings were corrected and dates and other identifications were expunged.

Five types of static graphics can be distinguished in POPS. They are terrain maps, data graphs, system linkage maps, adorning graphics and responsive graphics.

[1] The assistance of Richard Crenshaw, ARI, and Jack Briscoe, BDM Inc., with software development is appreciated.

Maps are an integral part of any military planning and of the understanding of a military operation. Figure 1 shows a large-scale map of the NTC displaying the operational sectors, axes of advances and phase lines for the training event. Color coding provides a representation of the gross geographical features. Color has been shown to contribute substantially to data presentation (Hudson, 1985). Five kilometer grid lines and map reference coordinates are included.

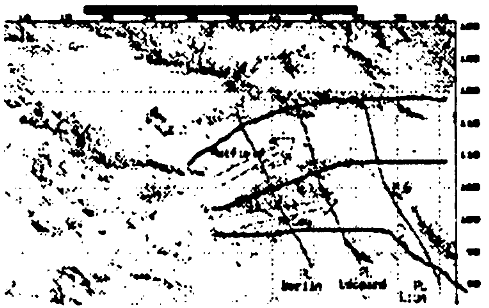


Figure 1. Screen showing annotated NTC map.

A second type of static graphic included in POPS is data graphs generated by other software. The input is various statistical data from the training event data set and the output is a screen display.

A third type of static graphic display available in POPS is a system linkage map, in some hypertext systems termed a graphical browser. Such a display shows the structure of links between all nodes. The user can actuate any link displayed on the map and consequently move directly to that node. Thus, the link map provides a random access means for navigation of the hyperdocument.

The fourth type of static graphics is adorning graphics. These are intended to give the screen visual appeal and to maintain the interest and motivation of the learner.

The fifth type of static graphics used in POPS is responsive graphics. Responsive graphics are characterized by producing an action in the system when activated. Typically they are an area of the screen where, when the cursor is placed within the boundary and the mouse button or key is depressed, some activation of the software occurs. The sensitive area of the screen is typically graphically visible to the learner, as in form of an icon such as a three-dimensional-appearing button. Most often this action is traversing a link to another node but it may initiate other functions, such as terminating the hypermedia program. Figure 2 shows a screen from POPS which contains both adorning and interactive graphics.

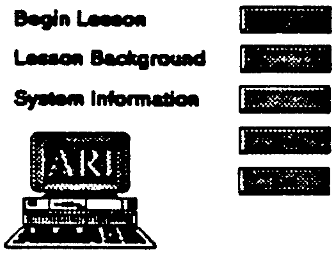


Figure 2. POPS screen showing adorning and interactive graphics.

Two types of dynamic graphics, based on the instrumented data of the training exercise data set, have been incorporated into POPS. One is a software package adapted from TBAT/MPART (Nichols & Shillcock, 1990) and the other is an animation effect created by displaying an automatically changing sequence of screen displays. The other dynamic graphic system uses a sequence of graphic files, generated by existing external software that scales and replays NTC instrumented player location data. These files are then displayed to the learner as a sequence by the authoring system to show player movement. Both display systems are used in POPS to support visualization of the vehicle positions over time.

The software developed for the assessment and feedback component of POPS has functionalities which include: (1) multiple entry points depending upon from where it is called in the lesson, (2) maintenance of learner records using a learner-generated password, (3) presentation of fixed response items and context-dependent responses, (4) capability of being exited for return to lesson at any point with restart at the same position, and (5) presentation of the learner performance record and current progress through the assessment and feedback submodule including identification of sections of lesson requiring completion. For additional information see Goehring (in press).

DISCUSSION

This project lead to several conclusions about using hypermedia technology for training military tactics by employing extant NTC data. We found it is feasible to combine the data and technology for training purposes; the concept of the project is sound.

While it seems clear that most of the data is usable, some filling of gaps in the training exercise story through the generation of simulated data is necessary to provide the learner with a complete picture of the training event. The need to minimize the risk of erroneous inferences and interpretations in producing such simulated data underscores the necessity of including staff with expertise in NTC and Army training on the project development team.

The work here has been intended as a feasibility demonstration of what is possible and a determination of what is required for full system development rather than an evaluation of effectiveness. Business and industry are recognizing the need to change the methods of evaluation for interactive systems and multimedia because traditional methods based on verbal methods are simply not appropriate for evaluation of visual processes (Marlino, 1990). The development of POPS demonstrates that interactive hypermedia technology exists today to produce quality instructional materials of moderate complexity for training fundamental tactics

The full development of the training system will be challenging, requiring a range of talents and sufficient resources and appropriate institutional support. In the paragraphs which follow, attention will of the resources which are necessary for full development of the hypermedia-based training system.

What is needed for a full development of the concept demonstrated by POPS? As with most projects the key to a successful development of the interactive hypermedia tactical trainer will depend upon the planning phase. Solid sponsorship and elucidation of project goals and objectives will be critical. Quality is expensive but is often a savings in the long run.

The complexity of computer-based training development should not be underestimated. Fourteen distinct roles involved in the development process of the typical computer-based trainer have been identified (Gery, 1987). These include Project Manager, Program Sponsor, Instructional Designer, Subject Matter Expert, Writer, Editor, Data Entry Specialist, Authoring System Host Computer Programmer, Media Expert, Graphics Designer, Authoring System Specialist, Learner Evaluator, Production Administrator, and Trainer Administrator. Assembling the appropriate development team and resourcing it adequately are critical for project success.

Computer-based instruction development ratios reported in the literature range from 1 to 4000 hours of development time for every hour of instruction with modal dominance in the 200 to 350 range. Reasons for high costs vary, most are situationally dependent (Orlansky & String, 1979; Jay, 1989). Gery (1987) presents the basis for a more refined model of cost estimation for computer-based trainer development. Fourteen courseware variables, five technological factors, twelve human variables and five additional factors impact the cost and development time. Collectively these factors yield a range of from 85 to 300+ development hours per computer-based training hour.

Judging the complexity of POPS on the relevant dimensions produces an estimate near the low end of the range. Substantial enhancement of the assessment and feedback component discussed above can be expected to move the cost toward the center of the range. Adding video and audio components would likely have costs toward the higher end of the range. Further, a shorter lesson, 4-8 hours of instruction, could be expected to have higher per hour development costs than a longer lesson of 60-80 hours of instruction.

CONCLUSION

Indiscriminate application of technology is not the answer to the challenges of Army training. However, judicious use of information technology holds exhilarating potential. This has been expressed succinctly, "because information technology moves so very much more quickly than other kinds of technology--halving in price and doubling in power every two years on the average" (Feigenbaum, McCorduck & Nii, 1988).

Finally, interactive hypermedia, and hypertext more generally, may be viewed as one stop along the technological roadway toward training worlds based on virtual reality (Krueger, 1991). Such training milieus, in which the learner is immersed in a simulated sensory environment to maximize learning and transfer, have been identified as particularly appropriate where (1) the environment is hostile, (2) there is a high benefit from rehearsal and practice, and (3) there is high potential for damage to equipment or danger to individuals (Middleton, 1991). These characteristics accurately describe combat on the modern battlefield.

Our experience on this project has taught us that though the effort is challenging, the capabilities are at hand to create complex, quality, interactive hypermedia-based training programs for the Army. Following moderate development costs, we believe production costs will be nominal. The software and hardware technology has come a long way and the time to fully utilize them is now.

REFERENCES

Berk, E. & Devlin, J. (Eds.) (1991). ***Hypermedia/Hypertext Handbook***. New York: McGraw-Hill.

Bush, V. (1945). As we may think. ***The Atlantic Monthly***. 176.1, 101-108.

Englebart, D.C. (1963). A conceptual framework for the augmentation of man's intellect. In P.D. Howerton & D.C. Weeks, (Eds) ***Vistas in Information Handling***. Washington, DC: Spartan Books, 1-29.

Feigenbaum, E.A., McCorduck, P. & Nii, P.H. (1988). ***The Rise of the Expert Company***. New York: McGraw-Hill.

Fletcher, J.D. (1990). Effectiveness and cost of interactive videodisc instruction in defense training and education. Paper P-2372, Alexandria, VA: Institute for Defense Analyses.

Gery, G. (1987). ***Making CBT Happen***. Boston, MA: Weingarten Publications.

Goehring, D.J. (in press). Interactive hypermedia for tactical training. Technical Report. Alexandria, VA: Army Research Institute.

Hamza, A.N. (in press). Combat Training Center Archive Catalog. ARI Research Product, Alexandria, VA.

Hannafin, M.J. & Peck, K.L. (1988). ***The design, development, and evaluation of instructional software***. New York: MacMillan Publishing Company.

Hudson, P.T.W. (1985). What does color add to a display that can't be done in black and white? ***Proceedings of Color in Information Technology and Visual Displays***. IRE Publication No. 61, University of Surrey, London, 33-37.

Jay, J. (July, 1989). CBT Development Costs (Part 3). ***CBT Directions***.

Krueger, M.W. (1991). Artificial reality: Past and future. In S.K. Helsel & J.P. Roth (Eds) ***Virtual Reality: Theory, Practice and Promise***. Westport, London: Meckler.

Marlino, M.R. Evaluating multimedia: Lessons learned from the past. ***Multimedia Review***. Fall 1990, 14-17.

Middleton, T. (1991). The potential of virtual reality technology for training. ***Proceedings of Interactive Multimedia '91 Conference***. Warrenton, VA: Society for Applied Learning Technology, 129-131.

Nichols, J.J. & Shillcock, D.M. (1990). User's Guide to the CTC Analyst's Workstation. Working paper. , Monterey, CA: BDM Inc.

Nielsen, J. (1990). ***Hypertext and Hypermedia***. New York: Academic Press.

Ntergaid, Inc. (1990). ***HyperWriter!*** Fairfield, CT.

Orlansky, J. & String, J. (April 1979). Cost effectiveness of computer-based instruction in military training. Institute for Defense Analyses Science and Technology Division, Contract DAHC15 73 C 0200, Task T134.

Phillips, R.L. (1991). An interpersonal multimedia visualization system. ***IEEE Computer Graphics and Applications***, ***11***(3), 20-27.

Slatin, J.M. (1991). Composing hypertexts: A discussion for writing teachers. In E. Berk & J. Devlin, (Eds.) ***Hypermedia/Hypertext Handbook***. New York: McGraw-Hill.

Wilkens, K. (1991). The transition to multimedia: Issues in the development of training materials. In ***Interactive Multimedia '91***. Warrenton, VA: Society for Applied Learning Technology.

AUTOMATING THE ARMY TRAINING DEVELOPMENT SYSTEM

Scott E. Graham and Ray S. Perez

U.S. Army Research Institute
Automated Instructional Systems Technical Area

For nearly a decade the Systems Approach to Training (SAT) has been the mainstay of Army training development and execution. The goal of SAT is to integrate evaluation, analysis, design, development, and implementation in determining the who, what, where, when, why, and how of training. SAT has, however, a predominantly paper-based system, and as with any paper-based system that generates and shares large quantities of information, it is often cumbersome. In response, the Training and Doctrine Command (TRADOC) Deputy Chief of Staff for Training (DCST) is working to build an Automated SAT (ASAT) that will increase efficiency by taking advantage of recent advances in computer technologies. In support of the ASAT development process, this paper takes a quick look at the training development process at the U.S. Army Armor School (USAARMS) and at the process that is being automated.

Under sponsorship of TRADOC's Army Training Board (ATB), the Army Research Institute (ARI) conducted research to design a comprehensive, computer-based training development system in 1984. The initial focus was on automating the development of Army Training and Evaluation Program (ARTEP) documents, which are now called Mission Training Plans (MTPs). In 1985, ARI conducted a front end analysis to specify how emerging computer and data base management technologies could be used to automate the ARTEP production system. The aim of the Computerized ARTEP Production System (CAPS) was to provide improved proficiency, responsiveness, and less costly production of ARTEPs. Based on the CAPS recommendations, ATB, with the technical assistance of ARI, funded the development of a prototype ASAT system. The ASAT prototype was built and installed by SAIC Inc. at the US Army Logistics Center and Quartermaster School in 1989 with a formative evaluation of the prototype system conducted in 1990.

Based on the ASAT prototype's successful evaluation at a combat service support school, TRADOC initiated a follow-up study to derive a functional description based on the needs of all TRADOC schools. In 1991, a contract was awarded to CAE-Link to develop the functional description (FD) for an operational ASAT system. The ASAT FD lays out a set of modular functions which cover the entire SAT process. In addition, CAE-Link conducted a comparative analysis of SAT in the TRADOC schools which described the SAT process as it was being done in the schools. i.e., primarily without the aid of automation.

In an attempt to further integrate advanced computer technologies into the ASAT process, the TRADOC DCST expressed interest in the potential application of knowledge-based expert systems and artificial intelligence tools as a means to facilitate the training development process. The future ASAT system with automated intelligent tools is referred to in this paper as ISAT, the Intelligent Systems Approach to Training.

Purpose The purpose of this research is to examine the automated training development needs of a combat arms school, and in particular, the U.S. Army Armor School. The Armor School, test site for the second ASAT installation, has an earned reputation for being one of the best schools for organization, productivity, and training development. The research findings may therefore apply well only to other well-organized training development organizations. The current effort: (1) examines the training analysis and development process at the Armor School, (2) identifies and discusses the adequacy of automation procedures being used by the Armor School and their perceived needs for improved automated tools, and (3) identifies areas in which automated expert tools are potentially most beneficial in the ASAT system as the basis for future ARI ISAT research.

Method

The method consisted of interviewing key training development personnel from the U.S. Army Armor School (USAARMS) Directorate of Training Development (DOTD). Much of the information came from the branch chiefs of the Analysis Branch and Unit Training Branch of the DOTD Training Division. These branches have the primary responsibilities for task analysis and MTP development, respectively. We also interviewed personnel from the Futures Branch, Simulation, Automation, and Technology Division, and ARI Fort Knox Field Unit personnel who have worked with DOTD over the years. Altogether 12 individuals were interviewed, including civilians, officers, and NCOs. The primary goal was to elicit the best ideas from each individual as to how they thought automation could be used to improve the task analysis and training development process. While we are confident that the vast majority of the information presented is accurate, the information is fundamentally opinions.

Results

The results of the interviews have been summarized and are presented in topical categories. The categories include: the Armor School training development process, task analysis for new equipment, standard setting, and automated MTP procedures.

The Armor School Training Development Process The Armor School has completed full task analyses for all of enlisted tasks and, as specified by SAT, these are contained in task analyses work sheets (TAWS). The Armor School is in a flux as to who is responsible for the collective analyses and the writing of the MTPs. In principal, it should be the training departments, e.g., Command and Staff, in that they contain the subject matter expertise. Most of the completed analyses have, however, been done within DOTD. Perhaps the biggest problem in the USAARMS training development process is getting the task analysis information to the training developers. Presently this requires getting a paper copy of the task analysis onto the training developer's desk.

Fortunately, this is one of the main problems ASAT seeks to remedy in that it would provide an electronic copy of the file on the training developer's PC.

Historically, training developers have either not been aware that analyses are completed or have believed that using them were more trouble than they were worth. ASAT will remedy this situation to the extent that using the analysis makes the developer's job easier. An alternative and/or complementary approach would be to hold the developer responsible for showing that his MTP or other training follows the analysis. An automated intelligent tool might identify where and how developed training significantly differs from that prescribed by the analysis.

One of the goals of the Analysis Branch is to minimize unnecessary changes to tasks, and in particular, the renaming of tasks at different skill levels. Last year there were over 250 tasks changed. Their ability to reduce duplication has increased considerably since the TAWS are now stored and organized in computerized databases. Likewise, ASAT should support the reduction of task redundancy. To do so, the system must be intelligent enough to recognize tasks that are fundamentally the same.

Tank Gunnery Task Analysis One of the challenges of Armor collective task analysis is the integration of gunnery with command, control, and maneuver training and standards. Historically, the training of tank gunnery has largely been separate from the training of command, control, and maneuver tasks. There are MTP tasks that address gunnery, but these tasks were not constructed with the level of specificity required in all situations. Specifically, the MTP tasks involving gunnery do not sufficiently represent the essence of combat gunnery, namely speed and accuracy. In the past it may have been acceptable to separate gunnery from maneuver in training development, but with the development of advanced weapon system technologies this approach may no longer be appropriate. Fortunately, improved simulation capabilities will likely help integrate gunnery and maneuver training.

Armor School Databases The foundation of the ASAT system design is a database with tables which will contain individual and collective task analysis information and the resultant task lists. The Armor School has been able on their own to computerize some of the functions proposed in ASAT including several task analysis databases. From these they produce some very useful products including a master task list which specifies: task number, task title, MOS or specialty code, skill level, training status, the department responsible, products from the task, and the need for common task manuals. The information can be sorted in a number of ways to include functional category, e.g., tactics. Among other things the task dictionary and database is used for is to count tasks. Prior to the database the analysts spent many hours counting tasks, e.g., how many Scout tasks does the school teach. Updating and maintaining the databases are, however, very time consuming tasks.

Task Analysis for New Equipment One of the areas in which intelligent ASAT tools may be most useful is in the area of requirements for new systems. The Army has recently, for example, completed testing of prototype M1A2 tanks. While the M1A2 task analysis was satisfactorily completed in time, the analysis might have been different. Task analyses do not address the bigger picture as to how new equipment can best be employed to achieve battlefield success. In the case of the M1A2, this would include use of the Position Navigation (POSNAV) system, the Commander's Independent Thermal Viewer (CITV), and the Intervehicular Information System (IVIS). Consider the tank platoon task, "Conduct a deliberate attack." The task steps includes: maneuver the platoon toward the objective; occupy the assault position; and assault the objective. As the task steps are defined, the addition of POSNAV, CITV, and IVIS has little effect on the successful performance of these steps. If, however, a forward-thinking task analysis required the optimum utilization of the advanced capabilities, the task would likely be very different. For example, how did the platoon leader use the additional information that was available as the result of the new equipment?

In the M1A2 case, the expert model would be structured on how commanders and strategists who fully understand the potential of the technological enhancements would use the new systems to fight. Currently with new systems the analysis question becomes," Which tasks change because we now have the CITV?" By contrast, the question might also be, "Given that Armor and the Army now has this capability, how can we restructure all that we do to fully exploit this new capability." Clearly, the latter question exceeds the perspective of those personnel typically required to do task analysis. Instead there might be an expert system built from the ideas of forward-thinking strategists who fully understand the potential and complexities of technological enhancements.

MTP Development One of key individuals interviewed is the head of the Unit Training Branch. For him, the most useful automation tool would be a template that would walk the MTP developer through the MTP process. The template would contain the major fields of the MTPs and have general information already included. The developer would largely just have to fill in the blanks. He also sees the need to link the MTP development system to the analysis databases. If there was a direct link to the collective analysis database, MTP development would largely require only a tweaking of the analysis information to ensure that format was correct. One important point needs to be underscored. The success of the ASAT system will depend on the degree to which quality subject matter expertise is built into the analysis. If the analysis is strong, then the training generated from the analysis will likewise be strong.

One of the likely major changes to the MTPs in the next five to ten years will be a greater reliance on the specification of simulation use in the ARTEPs. The MTPs being developed include an appendix which lists tasks that can be trained with simulation along with a list of those simulators that are available to train particular tasks. In the future, the MTP standards may include specified levels of performance on the simulators themselves, e.g., the Close Combat Tactical Trainer (CCTT).

Task Crosswalks One persistent problem is that it is tempting for training developers and MTP writers to crosswalk between tasks based only on task titles. The MTPs contain instances where it is appears that tasks have crosswalked from too high a level to the individual tasks, for example, to crosswalk from the platoon task to the individual task. That is not always an obvious relationship. The proper procedure requires the platoon tasks to be crosswalked to both the platoon leader and platoon sergeant tasks and then to the crew collective tasks. You would then crosswalk from the crew collective task to the individual tasks. An automated system should include procedures that would automatically cross-walk tasks and keep the cross-walks updated.

Shared Task Management The Army Training Support Center (ATSC) is by doctrine the shared task manager whereby they track all shared tasks. A shared task is such that one proponent is using another proponent's task. As shared task manager, ATSC puts out a catalog of all officer and enlisted shared tasks. The current paper-based system makes is awkward given the number and frequency of task changes. ASAT should greatly facilitate the shared task manager function. The FD describes a system which would be on-line to all of the schools. If one wanted to get a task summary or look up a task, there would be a central TRADOC-wide database that would tell what is available about a task, be it a task summary or a full TAWS. That could then be printed out. Furthermore, if major changes were made to a task, each of the schools' users would receive some notice that a change had been made. Communication between the schools would not be a daily activity even if a direct link was in place. If there was a common database between the proponent schools, it probably would be sufficient for updated information to be sent up in batch files on a weekly basis. By contrast, information posted in a local data base between DOTD, the training departments, and the training developers should be updated on a regularly scheduled basis.

Standard Setting Regarding standard setting for individual tasks, some are done in DOTD and some are done in the training departments. A reasonable concern with having the training department personnel do the work is that the same individual is responsible for being the SME, instructor, doctrine writer, as well as for doing the task analysis. When the instructor/SME sits down to do the task analysis, it is generally difficult for the SME to analyze the task in a way other than the way he teaches it. The task analysis therefore tends to look like an old lesson plan that he has been using. This is sometimes called "Reverse SAT." The SME is going back and putting on paper as task analysis that which he is already teaching.

Armor School analysts prefer write process standards for command, control, and maneuver tasks. The rationale is that is hard to come up with tactical product standards in that the tasks can be performed under a variety of METT-T conditions. By contrast, the process standards identify which steps are critical to the completion of the task. Once a unit is in the field for training and evaluation, it should be possible for the unit to develop its own product standards. From the school's point of view, the MTPs are designed to be guidelines for the battalion commander. With the MTP, he gets process standards. It is the prerogative and the responsibility of the battalion commander to add product standards based on the METT-T conditions set in the exercise. Currently the Analysis Branch does not think that it can or should include product standards by assuming what the METT-T will be.

There has been some discussion lately about developing standard training exercises, e.g., on SIMNET, which have "Frozen METT-T." In this case it would be possible to develop more product standards for command, control and maneuver tasks in that the METT-T conditions would be nearly identical for all units training with the exercise. A similar approach is currently being tried out as part of the revamping of Army Reserve and National Guard training under the rubric of Bold Shift. In this case, small units train on a series of well-defined situational training exercises, which are referred to as "Lanes," until they meet some standard.

Automated MTP Development Currently the MTP provides general guidelines and STXs that must be tailored by the units to include the specifics of their mission essential task list (METL). The MTP standards, likewise, are general in that they cannot anticipate METT-T changes which are made to reflect the unit's METL. As the unit tailors the MTP STXs, the standards should also be tailored to provide a rigorous assessment of performance in the tailored situation.

A future ASAT automated tool might "understand" how the basic STXs are modified and automatically modify the performance standards to reflect the conditions in the tailored STX. This tool would allow the trainer to sit down at his computer and quickly generate an STX and standards based on a particular wartime mission, or specific contingencies in particular areas of operation (AO), e.g., South West Asia or Panama. Based on the unit METL, the mission, and AO, the automated expert tool could generate the conditions, tasks, and standards to which the must train.

Conversion to a New System When any new system is introduced, there is always a problem of converting information from the old system. The Analysis Branch is concerned about this problem, but based on their experience said that they would populate the new database themselves. When they previously converted to their database using an interim system called TaskMaster, the best they could get was 90% conversion. They thought that even if the conversion was 95% complete that finding the 5% error would be more trouble than re-typing the information. Even though most of the information was converted accurately, it took considerable time to find the remaining errors. When DOTD used TaskMaster to develop the second soldier's manual, they basically skipped the automated conversion process. Instead they just block copied the old files into the template which took a temporary employee only about a week to complete. ASAT is considering the development of a process that will convert everything into the new program. This conversion requirement may be unnecessary and prohibitively expensive. The key would be to give the schools enough time to gradually enter their old data.

In Closing While it is obvious from our recent success is South West Asia that we have a well trained force, it is also true that there are still some significant areas for improvement in the TRADOC training development process. It is interesting to consider the potential applications of automated expert tools, but there is far more basic automation work that needs to be completed first. Automating the SAT process at the training developer level is a first step. In the end, ASAT should lead to a marked improvement in the Army training development process and to the overall readiness of the Army.

Alleviation of Chemical Protective Mask Effects on Stinger Team Performance

Joan Dietrich Silver and John M. Lockhart
Army Research Institute for the Behavioral and Social Sciences
Fort Bliss Field Unit, Texas

During Operations Desert Shield and Desert Storm (1990-1991), soldiers were under continuous threat of chemical attack. Survival during chemical attack requires that the combat soldier be encapsulated in a flexible system of protection known as Mission Oriented Protective Posture (MOPP). The system consists of an overgarment (jacket and trousers), boots, gloves with liners, and a mask. These articles of clothing, while performing a life-preserving function, can act individually or in concert to adversely affect tasks of psychomotor coordination, manual dexterity, and body mobility (Bensel, Teixeira, & Kaplan, 1987; Johnson & Sleeper, 1986).

The negative effects of the MOPP gear on performance were seen when Stinger teams engaged subscale aircraft in an engagement simulation facility (Johnson & Silver, 1992; in press); the performance of both the team chief and the gunner were significantly impaired when they were encapsulated in protective clothing.

The chemical protective mask is the component of the MOPP gear which is the most likely source of the Stinger team performance decrement observed during visual detection, identification, and tracking of aircraft. The mask is believed to impair performance because it reduces the FOV of the team chief's binoculars and of the gunner's missile sight by over 50 percent.

The hypothesis that the chemical protective mask is the primary source of the Stinger performance decrement finds support in the work of Bensel et al, (1987), Harrah (1985), and Kobrick and Sleeper (1986). They have documented the reduction in FOV caused by the mask and also its subsequent degrading effect on performance.

Silver and Lockhart (1992) were able to restore most of the FOV not only of the military binoculars but also of the Stinger missile sight when these devices were used in combination with the chemical protective mask. They modified the binoculars by replacing the eyepieces with a device used by underwater photographers to restore the FOV reduced by the combination of a diving mask and a camera encased in an underwater housing. The Stinger sight was modified simply by enlarging the rear peepsight from 1/8 inch to 5/8 inch. The modified devices were tested in a laboratory setting using Stinger personnel. Silver and Lockhart found that the FOV were significantly increased for both the binoculars and the Stinger sight. About 90 percent of the binocular FOV and about 70 percent of the Stinger sight FOV was restored.

If indeed the reduced FOV is the cause of the performance decrement seen with the Stinger teams (Johnson & Silver, 1992; in press), it follows that restoration of the FOV should result in improved engagement performance. The purpose of this study was to investigate that hypothesis.

Method

Participants

Participants were twelve Stinger teams (team chief and gunner). Ages ranged from 18 to 29 years, with a mean of 21.04 years.

Apparatus

Testing took place at the Army Research Institute's Range Target System (RTS) engagement simulation facility located at White Sands Missile Range, New Mexico. Participants employed the Stinger Tracking Head Trainer (THT) in simulated engagement of subscale fixed-wing and rotary-wing models of US and Soviet aircraft. The THT used in this research is a Stinger training device which develops and maintains gunner proficiency in tracking aircraft and firing the Stinger weapon. Gunner actions such as "acquisition" and "fire" were automatically recorded by data acquisition stations (DAS) and team chief actions such as "detection" and "identification" were entered by data collectors on DAS computer keyboards located at each of four weapon positions.

Procedure

After giving informed consent, participants received eight engagement trials in each of 3 FOV conditions: (1) binoculars or Stinger sight alone (no mask), (2) chemical protective mask with binoculars or Stinger sight (mask), and (3) chemical protective mask with modified binoculars and modified Stinger sight (mask/modifications). The 3 FOV conditions were counterbalanced over weapon stations.

Results and Discussion

Hypotheses established a priori stated that Stinger teams would perform better both in the no mask and mask/modifications FOV conditions than in the mask FOV condition. Data were collected on task performance measures (TPM) for fixed-wing and rotary-wing aircraft. TPMs are expressed as ranges for fixed-wing aircraft and elapsed time for rotary-wing aircraft. Within-subjects planned comparisons (Keppel, 1973) were performed on the appropriate means. The rotary-wing and fixed-wing engagement means are represented graphically in Figures 1 and 2, respectively.

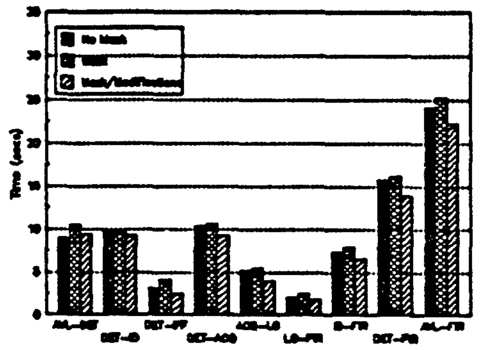


Figure 1. Rotary-wing engagement events by FOV condition.

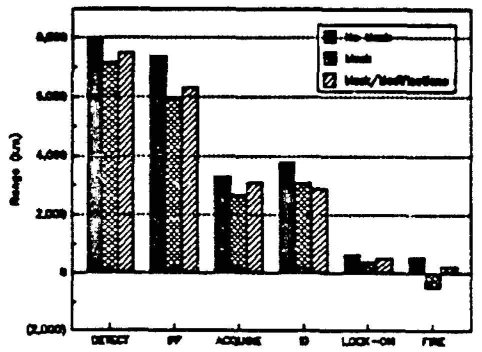


Figure 2. Fixed-wing engagement events by FOV condition.

Our hypotheses, that performance in both the no mask and in the mask/modification FOV conditions would be superior to that in the mask FOV condition were not supported by statistical analysis. Only one of 30 comparisons was significantly different, LO-FIR. However, 29 of 30 comparisons were in the predicted direction. Performance was better both in the no mask and mask/modifications FOV conditions than in the mask FOV condition.

The absence of significant differences between the mask and mask/modifications FOV conditions seemingly indicates that restoration of the FOV to the binoculars and Stinger sight when used with the chemical protective mask has no beneficial effect on Stinger engagement performance. Additionally, the lack of significant differences between the no mask and mask FOV conditions suggests that the mask does not impair Stinger engagement performance. Nonetheless, the presence of a body of research (Johnson & Silver, 1992; in press) which showed clearly that MOPP gear significantly impaired the engagement performance of three different groups of Stinger teams tested on three separate occasions warrants caution before accepting such conclusions. In each Stinger test, the performance of both the team chief and the gunner was significantly degraded by the MOPP gear.

Furthermore, before accepting these conclusions, one must take into consideration that 29 of 30 comparisons of Stinger engagement performance were in the predicted direction. The regularity and orderliness of the data strongly suggested that the modifications to the binoculars and to the Stinger sight were improving performance.

Given the evidence from prior research documenting the degrading effects of the MOPP gear on Stinger performance and the trends present in our data, it became incumbent on us to search for possible alternative explanations for our results. The ensuing search uncovered three problems--small sample size, substantial variability, and the possibility of learning effects inherent in the test environment--which, acting together, held the potential for reducing the effect of our independent variable.

The effects of sample size and variability must stand, but the effects of learning the test environment were isolated by examining the data at a point at which they had not had an opportunity to confound performance--during the first sequence of trials given each test day. We define "learning the test environment" as assembling pieces of information about the RTS engagement simulation facility which can subsequently reduce the effects of the independent variable over the course of the test day (i.e., reduce the disadvantges of wearing a mask).

If indeed acquisition of information about the test environment was reducing the effects of our independent variable, then evidence to support this hypothesis should be found in the data from the first sequence of trials administered to each team on each test day. If this hypothesis can be supported, then the data from these trials should reveal that the differences between the no mask and mask FOV conditions, and the mask and mask/modifications FOV conditions are substantially larger during the eight sequence one trials than the differences obtained when the data were combined over all three sequences of engagement trials. That is exactly what was found.

The no mask engagement means were compared to the mask means for the fixed-wing and rotary-wing events. The mask means were also compared to the mask/modifications means. The rotary-wing engagement events are presented first. Because the sample size was now reduced to four, statistical analyses were not performed on the data; only descriptive statistics are provided.

Rotary-wing engagement events: Sequence one. The sequence one rotary-wing engagement event means for each of the FOV conditions are represented graphically in Figure 3.

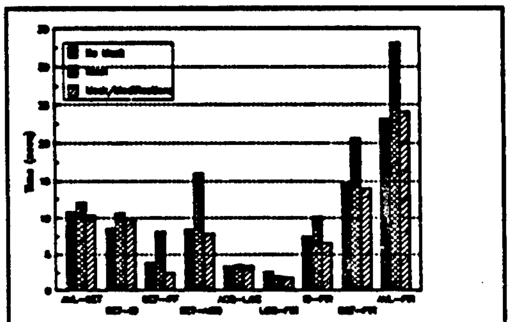


Figure 3. Sequence one rotary-wing engagement events by FOV condition.

Comparing Figures 1 and 3, it can be seen in the former that the rotary-wing engagement event means from the 3 FOV conditions are very similar to each other. As previously noted, we have hypothesized that this is so, at least in part, because of acquiring information about the test environment. On the other hand, the sequence one data displayed in Figure 3 reveal that the mask FOV condition means are generally considerably larger than those from the no mask and mask/modifications FOV conditions. Performance in the no mask and mask/modifications FOV conditions is virtually identical—results which were originally predicted. Restoring the FOV should produce performance similar to that when no mask is worn.

The rotary-wing sequence one data are offered as evidence in support of the hypothesis that acquiring information about the test environment repressed the effect of the independent variable. We acknowledge, of course, that this conclusion is tenuous until supported by further research.

Fixed-wing engagement events: Sequence one. The sequence one fixed-wing engagement event means are represented graphically in Figure 4. Once more, it is helpful to compare the data from all three sequence of engagement trials (Figure 2) to those from only the first sequence of trials in which the effects of learning the test environment are believed not to be present (Figure 4).

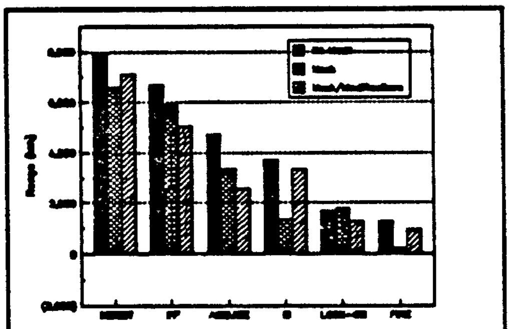


Figure 4. Sequence one fixed-wing engagement events by FOV condition.

The results from the fixed-wing events do not present as strong a case for the consequences of learning the test environment as do those from the rotary-wing events. The no mask FOV condition means are better than the mask FOV condition means for five of six engagement events, but, the mask/modifications FOV condition means are superior to the mask FOV condition only for half of the events. Interestingly, however, the means for the critical team chief event--identification--are as predicted; the no mask and mask/modifications FOV conditions are considerably better than the mask FOV condition. The beneficial effects of the modified binoculars are apparent for this engagement event.

The modification to the Stinger sight did not produce a clear-cut benefit for the gunner. He actually performed poorer on three of four engagement events with the modified sight, unlike his rotary-wing performance. Nonetheless, for the critical gunner event--fire--the modified sight evidently aided performance.

Summary

Although this research offers some support for the contention that restoration of FOV improves Stinger engagement performance, the lack of significant results tempers this conclusion. Nevertheless, these findings should serve as the impetus for further research investigating the effects of restoration of FOV, not only on Stinger engagement performance, but on any combat task affected by the chemical protective mask.

References

Bensel, C. K., Teixeira, R. A., & Kaplan, D. B. (1987). The effects of US Army chemical protective clothing on speech intelligibility, visual field, body mobility and psychomotor coordination of men (Technical Report Natick TR-87/037). Natick, MA: U.S. Natick Research Development and Engineering Center.

Harrah, D. M. (1985). Binocular scanning performance for soldiers wearing protective masks (Technical Memorandum 14-85). Aberdeen Proving Ground, MD: U. S. Army Human Engineering Laboratory.

Johnson, D. M., & Silver, J. D. (1992). Stinger team performance during engagement operations in a chemical environment: The effect of cuing (ARI Technical Report 947). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, D. M., & Silver, J. D. (in press). Stinger team performance during engagement operations in a chemical environment: The effect of experience (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, R. F., & Sleeper, L. A. (1986). Effects of chemical protective handware and headgear on manual dexterity. Proceedings of the Human Factors Society 30th Annual Meeting, Dayton, OH, 2, 994-997.

Keppel, G. (1973). Design and Analysis: A Researcher's Handbook. New Jersey: Prentice-Hall, Inc.

Kobrick, J. L., & Sleeper, L. A. (1986). Effect of wearing protective clothing in the heat on signal detection over the visual field. Aviation, Space, and Environmental Medicine, 57, 144-148.

Silver, J. D., & Lockhart, J. M. (in press). The effect of restoration of field of view on Stinger team performance in a chemical environment (ARI Technical Report). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Restoration of Field of View
Impaired by the Protective Mask

Joan Dietrich Silver and John M. Lockhart
Army Research Institute for the Behavioral and Social Sciences
Fort Bliss Field Unit, Texas

Soldiers in combat in a chemical environment are required to wear a flexible system of protection known as Mission Oriented Protective Posture (MOPP). This system provides five levels of protection, MOPP0 through MOPP4. A soldier in level 5, or MOPP4, is fully encapsulated in the protective garb which consists of an overgarment, boots, gloves, and a mask.

Each component of the MOPP gear can act alone or in concert to impede body mobility, psychomotor coordination, or manual dexterity (Bensel, Teixeira, & Kaplan, 1987; Johnson & Sleeper, 1986). Adverse effects of the MOPP gear on combat tasks are not uniform, but depend on the nature of the task to be performed.

One combat task affected by the MOPP gear is the engagement performance of Stinger teams. A significant performance decrement was seen for both the team chief and the gunner when they engaged subscale aircraft in an engagement simulation facility (Johnson & Silver, 1992; in press). The chemical protective mask is the component of the MOPP gear which has been identified as the primary source of the performance decrement seen with the Stinger teams. The reduced field of view (FOV) caused by the chemical mask when it is used in combination with the binoculars and the Stinger sight is believed to be the reason for the impaired performance. Substantial support for this hypothesis is found in the research of Bensel et al. (1987), Harrah (1985), and Kobrick and Sleeper (1986), who documented both the reduction in FOV caused by the mask and its subsequent detrimental effects on various performance measures.

If indeed the reduced FOV is the source of the performance decrement seen with the Stinger teams, then substantially increasing the FOV of the military binoculars and the Stinger sight when they are used with the chemical mask should result in improved engagement performance.

The purpose of the present study was to investigate the effect of modifications to the M19 military binoculars and to the Stinger sight on increasing the FOV of these devices when they are used in combination with the chemical protective mask.

Method

The FOV was determined for each of six experimental conditions: (1) Stinger sight, (2) M40 chemical protective mask and Stinger sight, (3) M40 chemical protective mask and modified Stinger sight, (4) M19 binoculars alone, (5) M40 chemical protective mask and M19 binoculars, and (6) M40 chemical protective mask and modified M19 binoculars.

Participants

Twelve Stinger soldiers (16S Military Occupational Specialty [MOS]) participated in this research. Ages of the participants ranged from 19 to 39 years, with the mean age being 27.25 years.

Apparatus

The apparatus for FOV measurement consisted of a white, free-standing, 8 feet (horizontal plane) by 6 feet 3 inches (vertical plane) board upon which 6 feet long by 3 3/4 inches wide strips of white paper were centered in the vertical and horizontal plane, forming a cross. A small white light (less than 1/4 inch diameter) was mounted at the center of the cross, providing a fixation point for the participants. The strips of white paper which formed the four arms of the cross were each marked in 1/2 inch increments from the center of the board to the end of each arm (3 feet each in length or 72 1/2 inch increments). A black pointer could be moved along each of the four arms to and from the center of the apparatus via a system of pulleys mounted on the rear of the apparatus. Four flood lights were directed at the stimulus to provide a constant source of illumination in the otherwise darkened room.

Two M19 standard military issue 7 x 50 binoculars were used; one with modifications and one without. One pair of binoculars was modified by replacing each of the eyepieces with a device used by underwater photographers to overcome the reduced FOV which results from the combination of the diving mask and the camera's underwater housing. The purpose of the modification was to restore the maximum amount of the FOV of the binoculars when they are used with the chemical protective mask.

Two Stinger sight assemblies were used; one with modifications and one without. The sight of one assembly was modified by increasing the size of the eyepiece from 1/8 inch in diameter to 5/8 inch. The purpose of the modification was to restore as much as possible of the FOV of the sight when it was used with the chemical protective mask.

Both pairs of binoculars and both Stinger sight assemblies were individually attached as required by experimental conditions via wing-nut and screw to a tripod mounted on a section of plywood board. The tripod was adjustable in the horizontal and vertical planes.

Procedure

After giving informed consent, each participant received a period of instruction and demonstration. Participants were assigned to no mask or mask conditions in counterbalanced order. Within the no mask and mask conditions, use of the binoculars (with and without modifications) and Stinger sight (with and without modifications) occurred in randomized order. The order of direction of trials (top, bottom, left, right) was randomized within 6 FOV conditions: (1) no mask and binoculars, (2) no mask and Stinger sight, (3) mask and binoculars, (4) mask and Stinger sight, (5) mask and modified binoculars, and (6) mask and modified Stinger sight.

Twelve trials were conducted in each of the 6 FOV conditions. There were six vertical FOV trials and six horizontal FOV trials (three in each direction—top to center, bottom to center, left to center, and right to center).

Participants were seated on an adjustable chair which was placed on the plywood platform holding the tripod. Each participant wore a black eyepatch over the left eye during Stinger sight trials only.

Prior to a sequence of trials in each FOV condition, participants were instructed to center the binoculars and the Stinger sight assembly on the white fixation light in the middle of the stimulus. Participants were cautioned not to shift their gaze in any direction, but to remain fixated on the white light throughout each sequence of trials. They were told that a black pointer would be moved inward toward the fixation light in the center of the stimulus from one of the four directions and that they should say "stop" as soon as they detected the pointer with their peripheral vision. They were then to name the direction from which the pointer appeared (top, bottom, left, right). A data collector recorded the position at which the participant detected the pointer. One set of practice trials was administered before each of the six FOV conditions. A set of practice trials consisted of moving a pointer from the extremity of an arm toward the center of the stimulus in each of the four directions until the participants detected the pointer and said "stop."

Testing occurred over a period of 3 days, with four participants being tested each day. Participants were seated 20 feet from the stimulus for all binocular testing and 11 feet from the stimulus for Stinger sight testing. The platform with the tripod and chair was moved into position as required by experimental condition. The M40 protective mask was used by all participants.

Results and Discussion

Hypotheses established a priori stated that the FOV measured when no mask was worn and when the mask was used with the modified binoculars and Stinger sight would be significantly greater than those measured when the mask was used with the unmodified binoculars and Stinger sight. The results of the within-subjects planned comparisons (Keppel, 1973) performed on the FOV means are listed in Table 1. Modifications restored about 90 percent of the FOV to the binoculars when used with the mask and about 70 percent of the Stinger sight FOV when used with the mask.

The experimental hypotheses were fully confirmed; the FOV for the mask condition were significantly smaller than those for the other two FOV conditions.

If indeed a reduced FOV is the cause of the performance decrement seen with the Stinger teams during the engagement sequence, then restoration of a substantial portion of that FOV should result in improved performance.

Table 1

Planned Comparisons of Mean Fields of View

Means (degrees)		Results	
Vertical Field of View			
M19 Binoculars			
No Mask	Mask		
6.40	1.89	*F*(1,11) = 1356	*
Mask	Mask/Mods		
1.89	5.39	*F*(1,11) = 2458	*
Stinger Sight			
No Mask	Mask		
13.50	6.37	*F*(1,11) = 204.48	*
Mask	Mask/Mods		
6.37	9.25	*F*(1,11) = 138.25	*
Horizontal Field of View			
M19 Binoculars			
No Mask	Mask		
6.72	2.20	*F*(1,11) = 422.69	*
Mask	Mask/Mods		
2.20	5.81	*F*(1,11) = 340.13	*
Stinger Sight			
No Mask	Mask		
15.67	6.78	*F*(1,11) = 389.48	*
Mask	Mask/Mods		
6.78	10.76	*F*(1,11) = 93.49	*

* *p*<.001

References

Bensel, C. K., Teixeira, R. A., & Kaplan, D. B. (1987). The effects of US Army chemical protective clothing on speech intelligibility, visual field, body mobility and psychomotor coordination of men (Technical Report Natick TR-87/037). Natick, MA: U.S. Natick Research Development and Engineering Center.

Harrah, D. M. (1985). Binocular scanning performance for soldiers wearing protective masks (Technical Memorandum 14-85). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Johnson, D. M., & Silver, J. D. (1992). Stinger team performance during engagement operations in a chemical environment: The effect of cuing (ARI Technical Report 947). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, D. M., & Silver, J. D. (in press). Stinger team performance during engagement operations in a chemical environment: The effect of experience (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, R. F., & Sleeper, L. A. (1986). Effects of chemical protective hardware and headgear on manual dexterity. Proceedings of the Human Factors Society 30th Annual Meeting, Dayton, OH, 2, 994-997.

Keppel, G. (1973). Design and Analysis: A Researcher's Handbook. New Jersey: Prentice-Hall, Inc.

Kobrick, J. L., & Sleeper, L. A. (1986). Effect of wearing protective clothing in the heat on signal detection over the visual field. Aviation, Space, and Environmental Medicine, 57, 144-148.

The Impact of Military Clothing on Human Engineering Measurement Techniques

Steven P. Paquette
U.S. Army Natick Research,
Development and Engineering Center
Natick, MA

Introduction

This paper reports the results of a pilot study initiated to address existing gaps in knowledge about the clothed body size and the nude and clothed joint limits of U.S. Army soldiers. The investigation was primarily designed to evaluate the applicability of available methods and measurement devices for collecting accurate and repeatable clothed anthropometric and range of motion data. This was accomplished by conducting formal comparisons among competing data acquisition strategies with the ultimate goal of identifying an optimal methodology for obtaining these data in larger scale clothed military surveys.

The U.S. Army currently has extensive information on the nude body size of its active duty personnel as a result of a major anthropometric survey completed in 1988 (Gordon et al., 1989). Data from this survey includes over 200 dimensions measured on approximately 9000 individuals. However, no comparable database exists that addresses changes in body size and range of motion as personal protective clothing and equipment items are worn. These data are needed so that accurate design information about the overall size, reach, and clearance requirements of clothed operators can be integrated into military crewstations. Unfortunately, reference documents such as MIL-STD-1472D (1989), Human Engineering Design Criteria for Military Systems; and MIL-HDBK-759B (1992), Human Factors Engineering Design for Army Materiel, provide limited information to designers of military hardware about the clothed (functional) body size and mobility of soldiers.

Materials and Methods

A market search and literature review were first conducted in order to identify previously applied methods and existing measurement devices that had potential to yield accurate and reliable clothed anthropometric and range of motion data. Based on the findings, two approaches were selected for acquiring clothed anthropometric data: traditional anthropometry and electromechanical digitization using the FARO Metrecom. Three devices were chosen for quantifying nude and clothed joint ranges of motion. These included the Polhemus 3SPACE ISOTRAK digitizer, the MEDmetric Penny and Giles Electrogoniometer and the Lafayette Plurimeter-V inclinometer.

The anthropometric and range of motion components of this study were conducted separately since each required a completely different set of measurements and procedures. For each part of the study, a separate group of six subjects (3 males and 3 females) participated. A wide range of body size variability was sought to determine whether the methods chosen for evaluation would perform equally well on individuals at both ends of the anthropometric distribution. Male and Female subjects were matched separately for both height and weight for the 5th, 50th, and 95th percentile values based on U.S. Army anthropometric data (Gordon et al., 1989). Additional matching criteria were also employed between the two test subject groups in order to maximize the comparability of the data in terms of individual body proportions.

Clothing Conditions

Identical clothing and equipment items were used in both parts of the study. They consisted of an eight layer U.S. Army combat vehicle crewman (CVC) ensemble and a six layer U.S. Aviation ensemble that included cold weather, ballistic, and chemical protective clothing. The specific clothing items associated with each layer of the ensembles are presented in Table 1. Since the Army does not have females assigned to combat vehicles, only males were tested in the CVC condition. Both clothing ensembles were chosen to represent worst case situations, (i.e., multiple layers of bulky clothing) in order to submit measurement techniques to rigorous test conditions.

Anthropometric Measurement Procedures

Subjects participating in the anthropometric phase of this study were measured for over one hundred dimensions in both clothing ensembles using conventional anthropometric techniques and the FARO Metrecom. The dimensions were chosen for their utility in workspace design and computer modeling of clothed human figures. Measuring equipment used in conventional anthropometry included sliding and spreading calipers, anthropometers, a headboard, and a steel tape. These items were used in accordance with standard procedures for collecting nude anthropometric data, however, it was necessary to generate a unique set of measurement definitions for many dimensions since nude measurement terminology was not directly applicable to the clothed body.

The Metrecom was also used to measure clothed body dimensions. The Metrecom is essentially a computerized 3-D digitizer developed as a skeletal diagnostic tool. The system consists of a six-degree-of-freedom arm which is maneuvered so that the tip of a touch probe makes contact with a given landmark. The arm apparatus is linked to a microcomputer which automatically records the 3-D coordinates at the tip of the probe. Point-to-point distances were recorded using the Metrecom, however, given the difficulty in accurately tracing the probe around the circumference of a given body segment. it was not used in this study to measure circumferential dimensions.

Skeletal landmarks under the clothing were identified by using the heights of nude landmark reference points on the body to approximate their location. The height for a measurement was established as the distance from the standing surface to a specific landmark. The heights used to landmark the depths, breadths, and circumferences were then systematically transferred to outer clothing layers using chalk. Depths and breadths were measured without compressing the clothing and recorded at the point of contact with the calipers. Circumferences were measured using a spring tension scale attached to a steel measuring tape to maintain consistent compression. Eighty grams of tension was used as the standard for most clothed measurements since this value approximates that used in most nude measurements that specify no visible tissue compression. Dimensions of the hips and shoulders were measured in a more fully compressed state (200 gm of tension for light clothing layers and 600 gm of tension for heavy clothing layers).

Table 1. A Description of Clothing Ensembles used in the Study.

Layer	Combat Vehicle	Aviation
0.	Nude Measurments	Nude Measurements
1.	Long Johns, Socks	T-shirt, Briefs, Socks
2.	Quilted Liner (Coat & Trouser)	Quilted Liner (Coat & Trouser)
3.	Ballistic Vest	Coverall, Boots, Helmet & Gloves
4.	Coverall, Boots, Helmet & Gloves	Cold Weather Overalls
5.	Cold Weather Overalls	Cold Weather Jacket
6.	Cold Weather Jacket	Survival-Recovery Vest w/ Insert
7.	SRU-21/P Survival Vest	
8.	Chemical Protective Overgarment	

Initial and repeat measurements of all dimensions were made in order to assess measurement reliability. These measurements were first made using conventional anthropometry and then the Metrecom during the same measurement session. Subjects were measured for all dimensions while nude and then remeasured for each dimension as layers of clothing were added to the body until all layers were donned. Some clothing layers were defined as a single garment, and only those dimensions impacted by it were remeasured for each layer.

Range of Motion Measurement Procedures

As previously noted, three devices were chosen for quantifying nude and clothed joint ranges of motion. The first of these was the Polhemus 3SPACE ISOTRAK. The ISOTRAK is an electromagnetic device that provides position and orientation, (i.e., six degrees of freedom, of a moving sensor relative to a fixed source). Using low frequency magnetic fields, the system records the three translational (x,y, and z) and the three rotational (azimuth, elevation, and roll) coordinates of a body upon which the sensor is attached. Signals which uniquely define the position and orientation of the sensor are directly output in a microcomputer. A major advantage of the ISOTRAK for the application at hand is that the sensors, which are attached directly to a subject's skin, need not be removed as additional clothing is added.

The MEDmetric Penny and Giles electrogoniometer (elgon) utilizes flexible strain gauge transducers to record measurements of angular displacement. Data are transmitted to a microcomputer and up to four planar motions can be recorded simultaneously. A single axis sensor is used for one degree of freedom joints, twin axis sensors are used for joints with two degrees of freedom, and triple axis sensor permits measurement of longitudinal rotation. The triple axis sensor was not yet available for use in this study. Data from measurement trials using the elgon are recorded manually from the LCD display. As with the ISOTRAK, the elgon sensors are applied directly to the skin thus facilitating a single outfitting per measurement session.

The Lafayette Plurimeter-V is essentially an inclinometer that measures the angle that a body segment forms with the vertical as movement occurs. The Plurimeter is a liquid pendulum inclinometer with a dial mounted on a base. Measurements are made with this device by placing it externally over clothed body segments and manually recording the angle at the desired point in the motion.

Subjects participating in this phase of the analysis were measured for forty-two joint-specific planar motions. In addition, adduction and abduction limits for every 15 degrees of flexion were measured at the shoulder and hip joints. A total of four trials for each measurement were conducted. Subjects were initially measured for all motions while nude and then each dimension was remeasured as clothing layers were added. Initial subject position was defined as the standard anatomical standing position with the shoulder rotated slightly so that the thumb pointed forward. All joint movements were defined in terms of the anatomical plane it which they occurred. The majority of motions were defined so that only the segment being recorded moved during the measurement. In situations where subjects were not able to stabilize themselves during execution of a motion, a portable support stand was used to maintain balance, but not as a weight bearing device to increase the range of motion. External markers such as eye level targets for visual fixation and foot prints were used to standardize subject position. Through careful pre-test planning, a strategy was devised to instrument and measure subjects for all motions with all three devices simultaneously. Therefore, the reported values for each measurement are directly comparable. Upper and lower body measurements were made on separate days since the extensive battery of measurements could not be completed in a single measuring session.

Table 2. A Comparison of the Mean Deltas for all Subjects Between Traditional Anthropometry and the Metrecom. (mm) Selected Dimensions from Layer Zero and Layer Six of the Aviation Ensemble are presented.

Anthropometric Dimension	Layer Zero T.A.	Metr.	T.A.-Metr.	Layer Six T.A	Metr.	T.A.-Metr.
Midshoulder Height	3	3	0	3	5	-2
Popliteal Height	3	3	0	3	2	1
Sitting Height	1	2	-1	2	3	-1
Stature	3	2	1	2	3	-1
Bideltoid Breadth	4	8	-4	4	7	-3
Forearm Breadth	1	2	-1	5	3	2
Hand Breadth	1	2	-1	1	5	-4
Midthigh Breadth	2	3	-1	4	6	-2
Calf Depth	1	4	-3	3	3	0
Chest Depth	2.	7	-5	2	9	-7
Neck Depth	2	7	-5	4	9	-5
Wrist Depth	1	2	-1	3	4	-1

Table 3. A Comparison of the Mean Values for all Subjects Between Traditional Anthropometry and the Metrecom. (mm) Selected Dimensions from Layer Zero and Layer Six of the Aviation Ensemble are Presented.

Anthropometric Dimension	Layer Zero T.A.	Metr.	T.A.-Metr.	Layer Six T.A	Metr.	T.A.-Metr.
Midshoulder Height	1414	1408	6	1453	1448	5
Popliteal Height	408	416	-8	381	394	-13
Sitting Height	909	898	11	958	955	3
Stature	1702	1700	2	1774	1768	6
Bideltoid Breadth	458	455	3	511	515	-4
Forearm Breadth	87	92	-5	134	138	-4
Hand Breadth	151	154	-3	238	239	-1
Midthigh Breadth	164	165	-1	224	221	3
Calf Depth	116	121	-5	192	201	-9
Chest Depth	245	256	-11	359	361	-2
Neck Depth	108	110	-2	165	155	10
Wrist Depth	41	42	-1	68	66	2

Results

Anthropometric Analysis

Measurement reliability was examined by comparing the initial and repeat values for each measurement obtained by using each method. The mean absolute difference (MAD) between the two measurement trials was computed for all subjects for each layer of each ensemble. The replications are close for both methods (see Table 2), however traditional anthropometry consistently demonstrated better repeatability than the Metrecom for most anthropometric dimensions. The third column in Table 2 presents the difference between the MAD values of each technique. A negative value denotes those measurements where traditional anthropometry demonstrated superior repeatability than the METRECOM. The effects of multiple clothing layers on measurement repeatability are also demonstrated in Table 2 as a comparison of the Layer 0 and Layer 6 deltas reveal. In general, the Metrecom was better at measuring heights and lengths than depths and breadths, since these measurements are made between well defined landmarks and placement of the probe tip on depths and breadths is subject to greater variability. It is also noteworthy that many dimensions of the arms and legs tend to show proportionally greater variation than larger dimensions of the chest and hips. This is primarily attributed to differences in the fit of clothing on the body.

Additional comparisons were also made between traditional anthropometry and the Metrecom by examining differences between the actual measurement values. Mean values were computed for all measurements for each layer of each ensemble, and then a delta was calculated by subtracting the Metrecom mean from the traditionally measured mean. Table 3 presents the results of a comparison for selected measurements. A positive value indicates that traditional anthropometry values are larger than those of the Metrecom. Traditional measurement tended to produce height values that were larger, and depths and breadths that were smaller than the Metrecom.

Range of Motion Analysis

As with the anthropometric phase of this study, analysis of the range of motion data focused primarily on determining the most reliable and accurate data collection method. Initially, reliability estimates were generated for each measurement technique by computing an average deviation for each dimension. The average deviation measures consistency among the trials and reflects the repeatability of a given technique. The average deviation is computed by first calculating the average over all four measurement trials, then finding the absolute difference between this average and each trial, and finally computing the average of these differences. As Table 4 illustrates, the overall results indicate good repeatability among the trials for all techniques with electricgoniometry demonstrating slightly more consistency than either the ISOTRAK or plurimeter. The addition of clothing layers did not appear to effect the repeatability of either technique.

Table 4. Average and Average Deviation of Four Trials For Selected Joint Motions In Layer Six of the CVC Ensemble (Degrees): 95th Percentile Male

	Average			Average Deviation		
Joint Motion	Isotrak	Plurim.	Elgon	Isotrak	Plurim.	Elgon
Neck Flexion	55	53	50	3.0	0.8	1.6
Elbow Flexion	113	116	112	2.5	4.0	2.4
Wrist Flexion	47	49	50	2.6	3.5	2.8
Hip Flexion	85	85	50	3.3	4.8	2.5
Knee Flexion	105	102	104	2.0	1.9	1.8
Ank. Plan-Flex.	33	31	22	1.3	1.5	1.3

Table 5. Difference and Absolute Difference Between Techniques for Selected Joint Motions in Layer Three of the Aviation Ensemble (Degrees): 5th Percentile Female

	Difference			Absolute Difference		
Joint Motion	I-P	I-E	P-E	I-P	I-E	P-E
Neck Flexion	-1.0	-11.8	-10.8	1.0	11.8	10.8
Neck Lat. Left	-0.3	16.5	16.8	1.3	16.5	16.8
Elbow Flexion	-2.5	37.3	39.8	3.0	37.3	39.8
Wrist Extension	-4.8	21.8	26.5	4.8	21.8	26.5
Wrist Flexion	-3.3	9.0	12.3	3.3	9.0	12.3
Hip Flexion	1.5	28.5	27.0	1.5	28.5	27.0
Hip Abduction	0.0	26.5	26.5	1.0	26.5	26.5
Knee Flexion	-2.3	-2.0	-0.3	2.3	3.0	2.3
Ank. Plan-Flex.	3.5	1.3	-2.3	3.5	3.3	2.8

Further comparisons were made by examining differences among the three measurement techniques. This was done by constructing pair-wise comparisons between techniques which yielded three sets of comparisons for each: ISOTRAK-Plurimeter, ISOTRAK-Elgon, and Elgon-Purimeter. For each comparison, the difference and absolute difference were calculated. The absolute difference for each pair was derived by taking the [absolute] difference between the measurement values for each trial and then averaging these absolute differences across all measurements. In general, the data demonstrate strong agreement between the ISOTRAK and the plurimeter, particularly at Layer 0, with the Elgon consistently different from either technique. These differences are primarily attributed to problems with the Elgon related to extreme sensitivity to placement of the sensors as well as the tendency for the fragile sensors to break and report erroneous values. Table 5 presents differences and absolute differences for a representative set of joint motions in layer three of the aviation ensemble.

Discussion and Conclusion

Based on the results of the measurement comparisons between conventional anthropometry and the Metrecom, it is apparent that the former approach is best suited for collection of clothed anthropometric data. Traditional anthropometry on the whole generated more consistent and repeatable data using measuring tools that were proven to be more easily adapted to the clothed body than the Metrecom. Difficulties with the Metrecom were primarily related to positioning and placement of the probe tip on the clothed body surface so that measurements were recorded along a single plane. Subject movement between the time the measurement endpoints are located and digitized also contributes to this problem. The relatively poor performance of the Metrecom in replicating depth and breadth measurements exemplifies this point.

It should be noted that the Metrecom holds certain advantages over traditional methods in that the amount of time required to complete the anthropometric measurement is less and the resulting values are transmitted directly to a computer. While the Metrecom has demonstrated utility for accurately recording anthropometric dimensions, it is concluded that traditional anthropometric techniques combined with slightly modified definitions to accommodate the clothed body, are best suited for quantifying changes in body size of clothed individuals.

The electrogoniometer produced consistent results for most range of motion measurements, however its values were systematically different than those of either the ISOTRAK or the plurimeter. The ISOTRAK has never been used to measure joint motion (nude or clothed) prior to this study, while the plurimeter belongs to a class of instruments that constitute the accepted standard in range of motion analyses. Since the ISOTRAK compared so well to this standard and since the elgon compared so poorly, it is argued that the ISOTRAK is superior to the elgon in terms of accurately measuring nude and clothed joint ranges of motion. Moreover, from a practical standpoint, the elgon proved to be excessively fragile and routinely prone to failure when used under these test conditions.

The close agreement between the ISOTRAK and plurimenter revealed in this study suggest that both devices provide reasonably accurate and repeatable results under identical test conditions. In addition, based on criteria such as the practical application and use of each device, no major limitations for either technique were uncovered in this study. However, there exist a number of factors that strongly indicate the ISOTRAK is preferable to the plurimeter for use in clothed studies. First, since the ISOTRAK sensors are applied only once to a subject's body, the motions are recorded from precisely the same location for each clothing layer. Exact placement cannot be assured with the plurimeter and this becomes even more difficult to control as additional clothing layers are added. Secondly, the ISOTRAK provides real-time data recovery directly to a computer which eliminates the possibility of data recording and transcription errors. The ISOTRAK is also more versatile than the plurimeter in that it can recorded information about the entire trajectory of a motion, and it can be used to monitor the status of adjacent segments as movement occurs. Other advantages over the plurimeter include a significantly shorter test period and the need for fewer technicians to collect data. Thus, the plurimeter does have demonstrated utility as a portable, low-cost device for collecting accurate clothed range of motion data. Nevertheless, the ISOTRAK is by far the best technique examined for measuring nude and clothed ranges of joint motion. Based on the above findings, conventional anthropometry and the Polhemus 3Space ISOTRAK will be utilized in future data collection efforts aimed at understanding the functional changes in body size and mobility that personal protective clothing and equipment impart on soldiers.

References

Gordon, C.C., Churchill, T, Clauser, C.E., Bradtmiller, B., McConville, J.T., Tebbetts, I., and Walker, R.A. (1989). 1988 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics (Technical Report TR-89/044). U.S. Army Natick Research, Development and Engineering Center. Natick, MA.

MIL-STD-1472D, (1989). Human Engineering Design Criteria for Military Systems, Equipment, and Facilities. U.S. Government Printing Office: 14 March, 1989.

MIL-HDBK-759B, (1992). Human Factors Engineering Design for Army Materiel. U.S. Government Printing Office. In Press.

Acknowledgments

I would like to thank Mr. H. Case and Mr. J. Annis from Anthropology Research Project, Yellow Springs, OH, and Dr. R. Maulucci from MOCO Inc., Scituate, MA for their outstanding contributions to this study.

PROBLEMS IN USING SURVEY DATA TO GENERATE PSYCHIATRIC DIAGNOSES

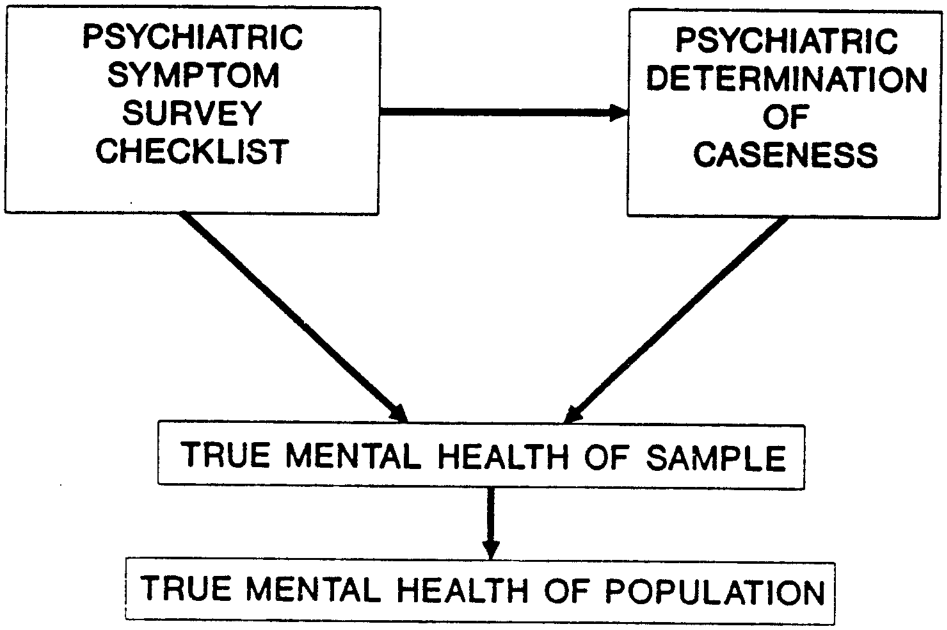
Captain Mark A. Vaitkus and Colonel James A. Martin
U.S. Army Medical Research Unit-Europe

The field of psychiatric epidemiology has grown to hold considerable importance for military researchers. Largely as a result of studies designed to estimate levels of post-traumatic stress disorder among veterans of the Vietnam War (e.g., Kulka, et al., 1990), policymakers have come to at least partially depend on such epidemiological data to help assess mental health care needs and project resource allocations.

One of the methods used to measure psychiatric symptomatology in epidemiological samples is the self-report symptom checklist. In particular, we are speaking about symptom checklists that are filled out as part of completing a more general survey questionnaire and do not involve the presence of a clinician or trained interviewer. While some take care to distinguish "symptomatology" from "diagnosis" with respect to such instruments, there is a tendency to establish "cut-off scores" that dichotomize respondents into "psychiatric case" versus "psychiatric noncase." This procedure implies that those in the "psychiatric case" category would be diagnosed as having a psychiatric problem were they to undergo a clinical examination.

There are numerous pitfalls in using self-administered surveys to determine psychiatric caseness, several of which we outline below. From a purely logical standpoint, the basic problem can be described in terms of measurement error associated with each of the elements shown in Figure 1.

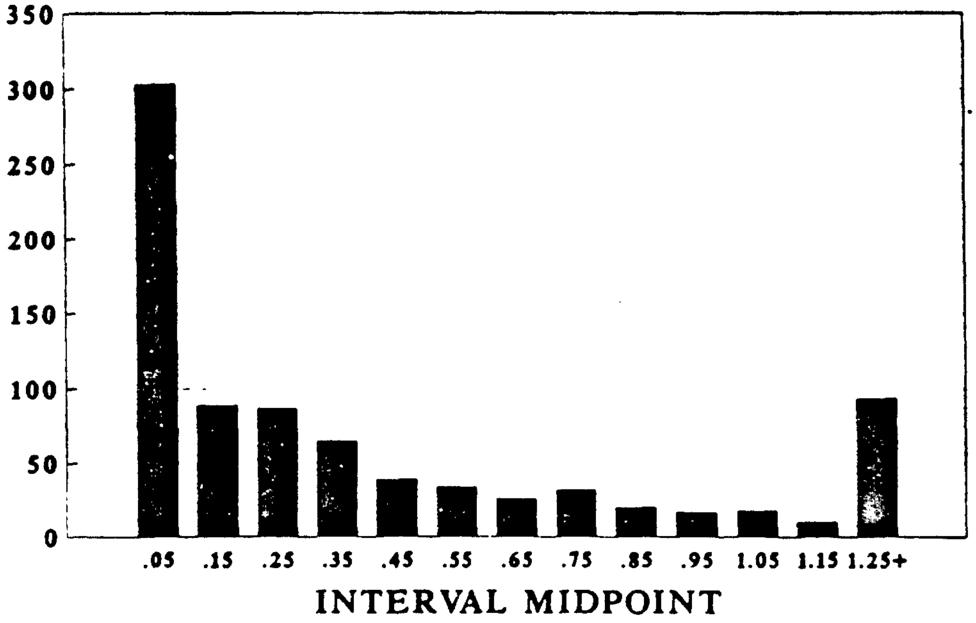
Figure 1.


PSYCHIATRIC SYMPTOM SURVEY CHECKLIST
PSYCHIATRIC DETERMINATION OF CASENESS
TRUE MENTAL HEALTH OF SAMPLE
TRUE MENTAL HEALTH OF POPULATION

The PSYCHIATRIC DETERMINATION OF CASENESS (PDC) is the result of a complex social interaction between clinician and interviewee that has its own source of error, as witnessed by the fact that clinicians may disagree on whether a "psychiatric case" label is appropriate for a given individual following independent interviews. A PSYCHIATRIC SYMPTOM SURVEY CHECKLIST has a source of error as well in terms of measuring an individual's true mental health condition. However, when the CHECKLIST becomes a tool used only to approximate the PDC, a third source of error is introduced, which may be described as the CHECKLIST's lack of sensitivity and/or specificity vis-a-vis the PDC. The TRUE MENTAL HEALTH OF A GIVEN SAMPLE is not directly addressed in such "validations," and estimates of the TRUE MENTAL HEALTH OF THE POPULATION are even more remote.

The fact that any proposed cut-off score is to some extent arbitrary is illustrated in Figure 2. The data come from a survey of Operation Desert Storm veterans conducted in June 1991 that utilized the Brief Symptom Inventory (BSI), a standard psychiatric listing of symptoms which is a shortened form of the SCL-90-R (Derogatis & Spencer, 1982). For each of 53 symptoms, respondents were asked whether they had experienced none, a little bit, moderate, quite a bit, or extreme discomfort resulting from it over the past week. These response categories were then coded from 0 to 4 respectively, and the mean response

FIGURE 2. BSI GENERAL SEVERITY INDEX RAW FREQUENCY HISTOGRAM



ODS VETERANS-JUNE '91 N=846 (E1-E6)

value computed. A mean value of .58 across all 53 symptoms indicates "caseness" on the General Severity Index of the BSI, according to the instrument's documentation. That the General Severity Index is measuring at least a somewhat continuous rather than dichotomous phenomenon is clear from Figure 2, as is the observation that the interval containing the cut-off score (.55) possesses no special power in splitting the distribution into two meaningful parts.

Table 1 presents data that demonstrate two additional problems with estimating caseness from survey data, especially for military samples. The first relates to whether adult civilian norms for caseness are appropriate for military populations, especially since such norms routinely turn up much higher levels of caseness than found in nonpatient civilian samples. Table 1 shows that estimates of caseness drop precipitously when adolescent norms are used, and in fact this may be because the average age of the Desert Storm veteran sample (24) more closely resembles that of the adolescent group. This table also suggests that unless subjects are tracked longitudinally for more than a few weeks or months, exaggerated caseness estimates may result that are more the effect of a transitory environmental state than a chronic condition.

TABLE 1.

ESTIMATIONS OF CASENESS VARYING BY NORM USED AND TIME OF SURVEY

BSI SCALE/SUBSCALE	CASE ESTIMATE ADULT NORM	CASE ESTIMATE ADOLESCENT NORM
Depression Subscale (FEB '91-Pre ODS)	45%	10%
Depression Subscale (JUN '91-Post ODS)	29%	7%
General Severity Index (JUN '91-Post ODS)	27%	5%

MATCHED SUBJECTS PRE AND POST ODS (N=846)

From a clinical point of view, one of the disturbing qualities of symptom scale scores is how possibly critical information is lost during score computation. Three hypothetical cases are offered in Table 2 to show how rather different BSI symptom patterns that involve more versus less serious symptom reporting would in fact be given exactly the same mean severity score. Such mean severity scores, the ones that are used to generate caseness cutoff scores, clearly neglect symptom extensiveness (number of reported symptoms) in addition to the relative seriousness of the reported symptoms themselves. Measuring average frequency and/or severity of individual symptoms does not begin to address the issue of how symptoms should be weighted with respect to one another or how symptom breadth/extensiveness should be factored into a total mental health score.

Based on the foregoing, it should not be surprising that even among those who are not labeled cases, it is possible to find significant percentages of those suffering from at least mild psychiatric symptomatology (see Figure 3). They may not warrant the kind of resources dedicated to the "caseness group," but to ignore their mental health needs entirely would seem foolish in light of any policy that includes a preventive component.

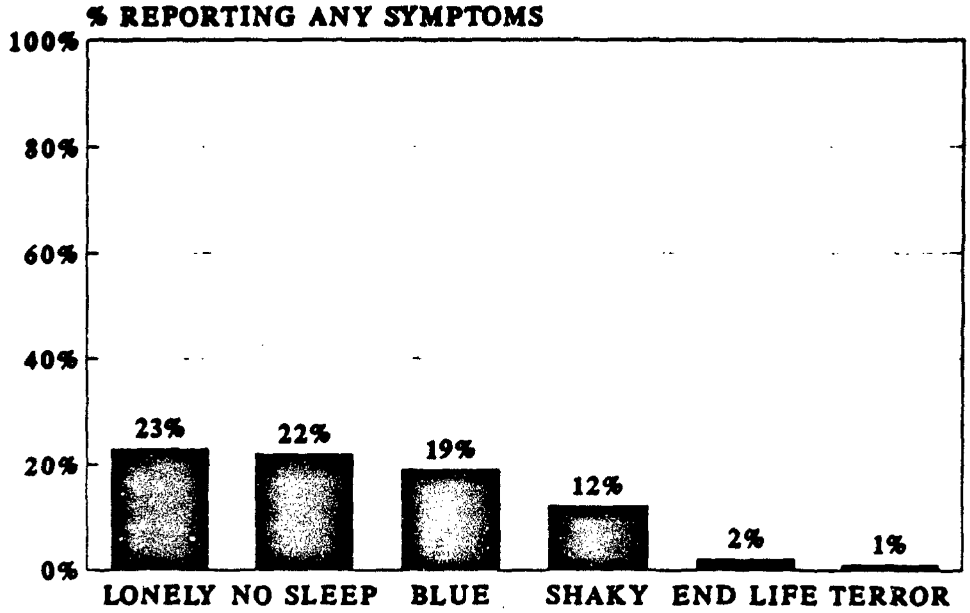
Table 2.

INFORMATION LOSS AS A RESULT OF UNWEIGHTED SCORING

	RESPONSE PATTERN		
BSI SYMPTOM	CASE1	CASE2	CASE3
Feeling Lonely	MOD	LIT	NONE
Feeling Blue	MOD	LIT	NONE
Spells of Terror or Panic	NONE	LIT	MOD
Thoughts of Ending Your Life	NONE	LIT	MOD
MEAN SCORE	1	1	1
EXTENSIVENESS SCORE	2	4	2

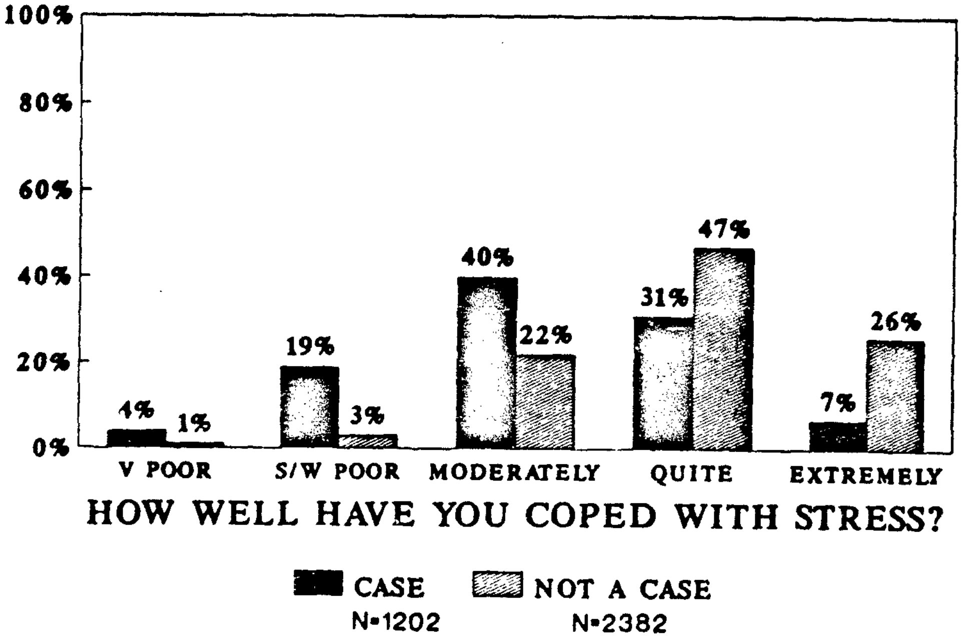
LIT=A LITTLE BIT MOD=MODERATE

Figure 3. SYMPTOMATOLOGY OF BSI NONCASES
Selected Symptom Examples



N=2410 ODS VETERANS FALL '91

Figure 4. RELATIONSHIP BETWEEN BSI CASENESS AND COPING ABILITY



ODS VETERANS IN USAREUR FALL '91 (E1-E6)

Although Figure 4 points to a significant relationship between BSI caseness and reported ability to cope with stress, the association is far from a perfect one. It may ultimately be a more productive exercise to look at reports of coping or successfully meeting work and family obligations before attempting to predict the numbers of those who are truly likely to appear for psychiatric counseling or referral.

Reported psychiatric symptomatology is in fact an important predictor of coping or ability to handle stress on the job and in one's personal life. However, as Table 3 reveals, having to rely on a dichotomous caseness variable instead of a continuous BSI General Severity score unduly restricts one's ability to account for variation in five-point scale measures of such coping or functional abilities. Even worse, as a dependent variable, caseness limits multivariate models to logistic or other maximum likelihood techniques.

Future research should (1) attempt to link symptomatology with help-seeking or dysfunctional behavior, and (2) relate the effects of diagnosis, intervention, and treatment on subsequent symptomatology.

REFERENCES

Derogatis, L. R., & Spencer, P. M. (1982). The Brief Symptom Inventory (BSI): Administration, scoring & procedures manual-I. Baltimore: Johns Hopkins University School of Medicine.

Kulka, R. A., Schlenger, W. E., Fairbank, J. A., Hough, R. L., Jordan, B. K., Marmar, C. R., & Weiss, D. S. (1990). Trauma and the Vietnam War generation. New York: Brunner/Mazel.

Table 3. EFFECT OF DICHOTOMOUS VS. CONTINUOUS MEASURE ON EXPLAINED VARIANCE

DEPENDENT VARIABLE	CASENESS	BSI SCORE
Stress Effect on Military Performance	.18	.25
Stress Effect on Personal Life	.24	.30
Ability to Cope With Stress	.15	.19

REGRESSION METHOD--R-SQUARED STATISTICS N=3544

Factors Mediating the Relationship of the CAGE to Hazardous Drinking in Army Personnel

John P. Allen, Ph.D. and Joanne B. Fertig, Ph.D., Treatment Research Branch, Division of Clinical and Prevention Research, National Institute on Alcohol Abuse and Alcoholism, Room 14C-20, 5600 Fishers Lane, Rockville, Maryland 20857

The CAGE (Mayfield, McLeod and Hall, 1974) is a commonly employed alcoholism screening measure. Items are dichotomously scored and a total scale score of two or higher is generally interpreted as presumptive of alcohol dependence. While most research on the CAGE has been conducted in medical care settings, the CAGE has been employed in at least three general community studies (Saunders & Kershaw, 1980; Smart, Adlaf & Knoke, 1991; Tejera, Santolaria, Gonzalez-Reimers, Batista, Jorge & Hernandez-Nieto, 1991). Investigations to assess the relationship of the CAGE to levels and frequency of drinking by non-alcoholics have also been performed (Moore & Malitz, 1986; Spencer, Bartu & Harrison-Stewart, 1987; Lee & DeFrank, 1988; Smart, Adlaf & Knoke, 1991). Only two studies have explored whether demographic variables mediate its association of the CAGE with alcohol consumption (Cutler, Wallace & Haines, 1988; Lee & DeFrank, 1988). Both considered the possible role of gender. Interestingly, their results were contradictory.

The current project investigates the role of gender, age, marital status, ethnicity, and rank group in affecting the relationship of the CAGE to self-reported weekly alcohol intake.

Method

Subjects were active duty Army personnel completing the Health Risk Appraisal questionnaire in the context of in-processing, a periodic physical exam, or a pre-physical fitness test. Demographic characteristics of the sample are included in Table 1.

Table 1

Subjects' Demographic Characteristics and Odds Ratios of Positive CAGE Score Being Associated with Hazardous Alcohol Consumption

Characteristic	Percent of Sample	Odds Ratio
Gender		
Male	91%	5.4
Female	9%	9.4
Ethnicity		

Caucasian	60%	6.1
Black	28%	4.4
Hispanic	8%	5.9
Other	4%	9.6
Marital Status		
Never Married	33%	5.0
Married	58%	5.9
Previously Married	9%	4.7
Rank Group		
Enlisted	84%	5.2
Warrant Officer	2%	8.1
Commissioned Officer	14%	10.7
Age Group		
17 to 20	14%	5.7
21 to 25	31%	5.2
26 to 30	21%	5.2
31 to 35	14%	4.5
Over 35	20%	7.0

CAGE items and a question on number of drinks in a typical week are embedded in the Health Risk Appraisal.

Drinks per week responses were dichotomized. For males 21 or more was scored as hazardous drinking. For females the lower end cut off score was 15. (While admittedly somewhat arbitrary, these values suggest increased risk for a variety of medical problems (Babor, Kranzler & Lauerman, 1987).) The CAGE was scored as positive or negative.

Results

Table 2 indicates the overall specificity and sensitivity of the CAGE.

Table 2

	Hazardous Drinking Present	 Absent
CAGE Positive	*1,339* *(41%)*	*7,249*
CAGE Negative	*1,920*	*58,838* *(89%)*
Total	*3,259*	*66,087*

Demographic variables, their interactions and CAGE results were entered into a logistic regression analysis (BMDP LR). Hazardous drinking served as the criterion. Starting with the fully saturated model, several models were evaluated. Selection of the final model was based on three goodness of fit measures (chi-square, Hosmer-Lemeshow and C.C. Brown). Chi-square of the model was 53.391 (df=22, p=.000) with sensitivity and specificity of .41 and .89 respectively. Age and marital status mediated the relationship.

Relationships between hazardous drinking and CAGE score were also evaluated by level for each demographic variable. Odds ratios are provided in Table 1. Differences between odds ratios (ratio of major to minor diagonal in the 2 x 2 contingency table) for the various strata of the demographic variables were tested with the Maentel-Haenzel test. The CAGE score was more discriminating for females than for males. Race also mediated the relationship, with "other" (Native Americans and Pacific Islanders) displaying the highest association. The relationship between hazardous drinking and CAGE scores was also related to rank, with commissioned officers having the highest odds ratio. Finally, marital status and age also influenced the relationship. The CAGE performed best with older subjects and with married soldiers.

Discussion

While being highly specific in its relationship with hazardous drinking, the CAGE is low in sensitivity and positive predictive value. Three possible reasons may account for this. The base rate for self-reported hazardous consumption is quite low with only 4.7 percent of respondents acknowledging drinking in this range. Second, three of the four CAGE items deal specifically with emotional reactions to drinking. These responses may be more associated with past or current alcohol dependence than with current hazardous consumption. Finally, it is possible that some soldiers who score positive on the CAGE are "binge" drinkers who occasionally drink excessively but whose typical weekly consumption is not in the hazardous range.

Although the CAGE was designed as a screening measure for alcoholism, it evidences some value also as a screen for hazardous drinking. Its brevity, ease of scoring, and health implications are such that it (or a more accurate measure) warrant continued inclusion in the Health Risk Appraisal. Efforts to more precisely gauge frequency, quantity, intensity, and patterning of alcohol use in the Health Risk Appraisal might also prove helpful.

References

Babor, T.F., Kranzler, H.R. & Lauerman, R.J. (1987). Social drinking as a health and psychosocial risk factor: Anstie's limit revisited. In M. Galanter (ed.), Recent developments in alcoholism, Vol 5. New York: Plenum Press, 373-402.

Mayfield, D., McLeod, G. & Hall, P. (1974). The CAGE questionnaire: Validation of a new alcoholism screening instrument. American Journal of Psychiatry, 131(10), 1121-1123.

Cutler, S.F., Wallace, P.G. & Haines, A.P. (1988). Assessing alcohol in general practice patients--A comparison between questionnaire and interview (findings of the Medical Research Council's general practice research framework study on lifestyle and health). Alcohol and Alcoholism, 23, 441-450.

Lee, D.J. & DeFrank, R.S. (1988). Interrelationships among self-reported alcohol intake, physiological indices and alcoholism screening measures. Journal of Studies on Alcohol 49(6), 532-537.

Smart, R.G., Adlaf, E.M. & Knoke, D. (1991). Use of the CAGE scale in a population survey of drinking. Journal of Studies on Alcohol, 52(6), 593-91.

Moore, R.D. & Malitz, F.E. (1986). Underdiagnosis of alcoholism by residents in an ambulatory medical practice. Journal of Medical Education, 61, 46-52.

Spencer, J., Bartu, A. & Harrison-Stewart, A. (1987). Observations on community screening for alcohol problems: A pilot project to assess the feasibility of identifying heavy drinkers in a community setting. Alcohol and Alcoholism, 22(1), 65-69.

Saunders, W.M. & Kershaw, P.W. (1980). Screening tests for alcoholism--Findings from a community study. British Journal of Addiction, 75, 37-41.

Tejera, F., Santolaria, F., Gonzalez-Reimers, E., Batista, N., Jorge, J.A. & Hernandez-Nieto, L. (1991). Alcoholic intake in a small rural village. Alcohol and Alcoholism, 26(3), 361-366.

Personality Tests to Predict Success in Navy Pilot Training

D. R. Street, K. T. Helton, and D. L. Dolgin

Naval Aerospace Medical Research Laboratory
Pensacola, Florida 32508-5700

The increasing cost of training aircrew to operate modern naval aircraft and the simultaneous decline in retention rate for these same trained aircrew increase the importance of utilizing the best selection methods available. This importance is underscored by the fact that every aircrew selectee who fails to complete training contributes to a potential operational personnel shortage if expected replacements necessary to maintain military readiness do not materialize as planned. Pilot selection research to date has generally focused on the testing of various psychomotor and cognitive abilities (Carretta, 1986; Davis, 1989; Dolgin & Gibb, 1989; Hilton & Dolgin, 1990). While these abilities would seem logically necessary for successful performance in flight training, some failures may be due, at least in part, to personality and/or motivational factors (Helmreich, 1982).

Historically, researchers have tried to find the ideal aviator personality profile among numerous personality measures. This ideal aviator personality profile has often been anecdotally called "the right stuff." Promising results have been found in identifying characteristics that improve the likelihood of later success in aviation such as persistence, motivation, coolness under pressure (clear thinking), and novel problem solving (e.g., Retzlaff & Gibertini, 1987). Other researchers have considered personality factors with varying degrees of success (Dolgin & Gibb, 1989; Hunter & Burke, 1991).

Certain personality characteristics or traits may correlate highly with success in initial/primary flight training. For example, interpersonal orientation, self-assertiveness, and achievement motivation are associated with pilot attitude and performance (Helmreich, Sawin, & Carsrud, 1986). Important developments in personality assessment have included attempts to avoid response bias by masking the personality dimension of interest and to screen for positive attributes, in contrast to a past emphasis on psychopathology (Picano, 1991).

Personality testing has improved with tools that assess specific attributes as opposed to the general approach of most personality measures, which are composed of numerous questions and whose responses are then analyzed in search of trends (Hollenbeck & Whitener, 1988). One recent study (Picano, 1991) utilized a measure designed to assess 31 behavioral traits commonly found in working environments. That study focused on experienced Army aviators and found significant differences between nonaviators and aviators on 22 of 31 administered subtests of the Occupational Personality Questionnaire. The emergence of increasingly effective personality measures has prompted the Air Force to reconsider personality testing (Davis, 1989). Ongoing research at the Naval Aerospace Medical Research Laboratory (NAMRL) has generated data on a variety of personality measures including validation of a "risk test" with recommendations for naval aviation implementation (Blower & Dolgin, 1991; Dolgin, Shull, & Gibb, 1987). The Navy does not currently have an operational personality measure for use in pilot selection.

The present paper deals with an automated personality assessment, the Pilot Personality Questionnaire (PPQ). The PPQ is an attribute self-report inventory. It was designed to take advantage of those useful assessment elements found in various paper-and-pencil tests that have historically shown promise in tapping specific aviation-linked personality characteristics. The PPQ was compared to a pass/fail criterion by Shull and Dolgin (1989). In that study, subjects' PPQ scores were compared to primary flight training outcome. They found a low, but significant, relationship between various PPQ scores and the pass/fail criterion in primary flight training. We believe that certain personality traits as measured by the PPQ could increase the accuracy of training success predictions.

METHOD

Subjects. The subjects who participated in this study took the Aviation Qualification Test/Flight Aptitude Rating (AQT/FAR) prior to selection for aviation training. The AQT/FAR is the primary selection test battery (paper-and-pencil) currently used by the U. S. Navy and Marine Corps for entrance into flight training. Each of the subjects also took the PPQ at NAMRL between 1989 and 1991 as part of a continuing selection research project. The subjects participated in the study on a voluntary basis. Before administering the test, all subjects were informed that the test results obtained would not affect their status in the flight program and would not be entered into their service record.

The initial data pool consisted of 245 subject PPQ and AQT/FAR cases collected while the subjects were waiting to enter primary flight training. Only subjects who later passed through advanced flight training or failed due to academic- or flight-related failures in any flight training phase were included in the analysis. Subjects who had failed flight training due to nonflight- or nonacademic-related failure were not included in the analysis. The final analysis sample ($N = 211$) consisted of 201 males and 10 females ranging in age from 21 to 29 years ($M = 22.77$, $SD = 1.36$). The sample was further divided into two groups for analysis: those who had passed ($N = 168$) and those who had failed ($N = 43$) during any phase of flight training.

Apparatus. The PPQ was administered as part of a 3-4 h assessment battery. The first 91 subjects were given the test on an Apple IIe microcomputer system with an Amdek Color Plus I monitor. The remaining 120 subjects were administered the test on a Zenith 248 with a Zenith monochrome monitor. Response entry on both systems was via a numeric keypad.

Materials. The PPQ is a self-administered, untimed, personality inventory containing 112 multiple-choice questions answered to via a computer keyboard. The test is a combination of four different personality tests: 1) Locus of Control (LOC), 2) Work and Family Orientation (WOFO), 3) Personality Attributes Questionnaire (PAQ), and 4) Social Desirability Scale (SDS). These four tests were included because of their prior use as pilot personality measures.

The LOC (Rotter, 1966) was designed to measure an individual's attribution or cause and control of life events. The scale separates causal attribution as being either self-controlled (internal) or controlled by others (external). The WOFO (Helmreich & Spence, 1978) is a measure of achievement motivation and attitudes toward family and career. The PAQ (Spence, Helmreich, & Holahan, 1979) measures socially undesirable behaviors such as hostility and aggressiveness. The SDS (Crowne & Marlowe, 1960) was included as a measure of motivation and as a way of reducing response bias by measuring self-report distortion.

Subjects' responses were partitioned into 12 scales that were designed to measure (1) self-assertiveness, (2) interpersonal orientation, (3) aggressiveness, (4) hostility, (5) verbal aggressiveness, (6) submissiveness, (7) high-mastery motivation, (8) high-work motivation, (9) competitiveness, (10) self-control, (11) fatalism, and (12) high-social desirability. (See Dolgin and Gibb (1989) for a discussion).

The AQT/FAR, which contains four multiple-choice tests, is the primary nonmedical instrument that the U. S. Navy/Marine Corps uses to screen officer flight training applicants. The Academic Qualification Test (AQT) is a single test that measures such attributes as general intelligence, verbal and quantitative abilities, clerical skills, and situational judgement. The FAR is made up of three different tests. The Mechanical Comprehension Test (MCT) assesses mechanical aptitude and the ability to perceive physical relationships. The Spatial Apperception Test (SAT) is a measure of spatial orientation that involves determining the angle of bank at which various aircraft are configured. The Biographical Inventory (BI) samples personality history, interests, and attitudes while assessing acquired aviation knowledge; it is the only untimed test of the group. In this research, only the raw AQT/FAR scores, not the stanine scores, were used for analysis.

Research Design. We separated the subjects into a pass or fail group based on their performance during all stages of flight training. Failure may have occurred at any phase of flight training due to academic- or flight-related difficulties. Next, we compared the PPQ scale score and AQT/FAR subtest score means for the pass and fail groups. Student's *t* tests were conducted for the pass and fail groups to explore simple group mean differences. We then conducted a series of multivariate analyses to assess the predictive value of group subtest differences. Discriminant analysis was used to further describe the multivariate relationships in the data. The PPQ and AQT/FAR scores were first entered into a forward stepwise discriminant function analysis based on the pass/fail criterion to reduce the set of variables to the smallest number of predictor variables with maximal prediction of the criterion. In this procedure, variables with the highest relationship with the criterion were added to a regression equation. The partial correlations were used to indicate the degree of relationship. As variables were added, the multiple correlation was recomputed. When the changes in R at each step were no longer significant, variables were no longer added. A priori, we also decided to retain at least the AQT and FAR as well as any other variable that significantly added to a prediction equation. At this point, the prediction equation included only those variables that were predictive of the pass/fail criterion. The variables remaining in the equation after stepwise discriminant analysis were then entered into a standard discriminant analysis to determine a classification model.

RESULTS

Means and standard deviations for the pass and fail groups are presented in Table 1. We analyzed the group me_ns to determine possible differences.

Table 1. Means and Standard Deviations (SD) for Pass and Fail.

Variable	Pass (N = 168) Mean	 *SD*	Fail (N = 43) Mean	 *SD*
	Pilot Personality Questionnaire			
Self-assertion	26.88	3.01	25.95	3.15
Interpersonal orientation	21.95	4.09	22.44	4.80
Aggression	19.74	3.39	19.37	3.81
Hostility	13.76	4.46	12.98	4.74
Verbal aggression	4.30	2.66	4.79	2.77
Submissiveness	5.42	2.66	5.28	2.29
Mastery motivation	22.96	4.03	22.40	3.72
Work motivation	22.21	2.04	22.16	3.50
Competitiveness	16.05	2.62	14.63	4.04 ***
Self-control	21.67	4.16	22.42	4.26
Fatalism	15.51	8.09	15.21	6.26
Social desirability	75.96	14.91	78.01	13.33
	Aviation Selection Test Battery			
AQT	5.79	1.19	5.37	1.16 *
FAR	7.13	1.70	6.35	1.78 **
SAT	23.45	4.29	21.86	5.38 *
MCT	50.73	7.93	47.58	8.30 *
BI	41.79	11.52	38.65	9.10

$^{*}p < .05$
$^{**}p < .01$
$^{***}p < .005$ (two-tailed)

The results of t tests for independent samples are also presented in Table 1 for the pass and fail groups during overall flight training (i.e., primary, intermediate, and advanced). The group means for the AQT, MCT, and SAT scores were significantly different at the $p < .05$ level, while the group means for the FAR (a linear composite of the SAT, MCT, and BI) and the Competitiveness scale were significantly different at the $p < .01$ level. This procedure was employed by Picano (1991) to describe differences between experienced Army pilots and a nonaviation standardization sample of the Occupational Personality Questionnaire. As mentioned, the simultaneous application of 17 separate t tests resulted in an increased probability of significance through chance. The probabilities were not adjusted to account for this because we employed multivariate techniques to assess the value of differences in the prediction of flight training success.

To assess the contribution of the various PPQ and AQT/FAR variables to a linear prediction equation, we conducted a stepwise discriminant function. Five variables met the tolerance (.01) requirements for independence and remained in the equation. These were the AQT/FAR, and the Verbal Aggression, Competitiveness, and Self-Control scales of the PPQ. The variance accounted for by the 12 remaining AQT/FAR and PPQ variables not included in the equation was accounted for by those retained in the equation. Although five variables met the tolerance test for independence, only the PPQ Competitiveness scale met our a priori requirement and added significant variance to the prediction equation. This was also the only PPQ variable to be significantly different in the comparison of the pass/fail group means. The MCT, SAT, and BI did not meet the redundancy test and were dropped from the analysis. This was not surprising, since the FAR is a composite of these subtests.

The three significant variables remaining in the equation after stepwise discriminant analysis were next entered into a standard discriminant analysis. For precision purposes, Table 2 presents the *F*'s to Remove and Wilk's lambda values produced in the standard discriminant analysis with five predictors retained. The *F*'s to Remove show the relative weights of the scales in the equation. Summary statistics for the standard discriminant function include Wilk's lambda equal to .92194 and an approximated *F*(3, 207) of 5.84 ($p <$ .0007).

Table 2. Summary Statistics for Standard Discriminant Function Analysis.

Variable	Wilks' lambda	Partial lambda	*F* to remove	*p*-level
Competitiveness	.952380	.968045	6.833145	.0096
FAR	.944001	.976637	4.951822	.0271
AQT	.936931	.984007	3.364370	.0681

A classification equation was developed for use with the unstandardized raw scores remaining in the equation after stepwise analysis. Table 3 presents the discriminant function classification matrix with the five AQT/FAR and PPQ variables remaining in the equation. Pass and fail means were significantly different for the distribution of discriminant function scores calculated for the two groups (χ^2 (3) = 16.86, $p < .001$). A Pearson correlation coefficient of .28 was obtained. The discriminant function explained 7.7% of the total variance. The discriminant function was able to accurately classify 70.1% of the cases. To reduce attrition by 50%, the prior probabilities were adjusted to 57% and 43% for the pass and fail groups, respectively. This level of attrition reduction was obtained at a cost of 41 out of the 168 (24%) student naval aviators who would have otherwise passed through advanced flight training.

Table 3. Classification Matrix. *

	Predicted Group Membership		Cases
Actual Group	Pass	Fail	
Pass	127 75.6%	41 24.4%	168
Fail	22 48.8%	21 51.2%	43

* Percent of grouped cases correctly classified: 70.1%

DISCUSSION

Using the PPQ, we found that the competitiveness personality trait in successful student naval aviators was significantly different from students who fail. This difference coincided with differences found on the naval aviation selection test battery for the same groups. Furthermore, student naval aviators who passed through advanced flight training were also more likely to score higher on the MCT and SAT subtests of the AQT/FAR. In other words, increases in MCT and SAT scores appear to be related to an increased probability of success in flight training. However, there was no difference in their BI scores of the FAR.

We also found that pass and fail students were statistically different on competitiveness as measured by the PPQ. This difference was greater than that found on any AQT/FAR variable. These results are consistent with those of previous researchers (Picano, 1991; Retzlaff & Gibertini, 1987) and indicate that those successful aviators in our study were different on some personality characteristics from their unsuccessful peers.

The practical value of the differences obtained in our study was suggested through discriminant analysis and reveals that the PPQ may increase the accuracy of decisions regarding likelihood of succeeding through advanced flight training. The contribution of the PPQ to the existing AQT/FAR predictors in our prediction equation is statistically significant. In fact, the PPQ competitiveness scale explained the greatest amount of variance in the final prediction equation. Taken as a whole, the results describe a picture of the successful naval aviator based on high general cognitive ability (AQT), high spatial reasoning (SAT), high mechanical reasoning (MCT), and high competitiveness (PPQ). There is a related cost in terms of false rejections who would have otherwise passed through advanced flight training. A decision to implement a system including the PPQ should weigh the cost of lost aviators against the savings gained through reduced attritions. Although preliminary, our findings demonstrate the value of the PPQ Competitiveness scale as a predictor of aviation training success. Future cross-validation studies with the discriminant model described in this report are necessary to establish the value of the PPQ Competitiveness scale in predicting flight training success.

ACKNOWLEDGMENTS

We thank Ms. K. Mayer for her editorial support in the preparation of this report. We also thank Ms. S. Starling for assistance in data collection and Mr. P. Collyer for computer-programming support.

The research reported in this report was completed under the Naval Medical Research and Development Command work unit No. 63706N M0096.002-M00960.01. The views expressed in this report are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. Trade names of materials and/or products of commercial or nongovernment organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such materials and/or products. Volunteer subjects were recruited, evaluated and employed in accordance with the procedures specified in the Department of Defense Directive 3216.2 and the Secretary of the Navy Instruction 3900.39 series. These instructions are based upon voluntary informed consent and meet or exceed the provisions of prevailing national and international guidelines. Reproduction in whole or in part is permitted for any purpose of the United States Government.

REFERENCES

Blower, D. J. & Dolgin, D. L. (1991). *An evaluation of performance-based tests designed to improve naval aviation selection* (Report No. 1363). Pensacola, FL: NAMRL-Naval Aerospace Medical Research Laboratory.

Carretta, T. R. (1986). The basic attributes test (BAT) system: A preliminary evaluation of three cognitive subtasks. *Proceedings of the 30th Annual Meeting of the Human Factors Society* (pp. 1321-1325). Santa Monica, CA: Human Factors Society.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting and Clinical Psychology*, *24*, 349-354.

Davis, R. D. (1989). *Personality: Its use in selecting candidates for US Air Force undergraduate pilot training* (Report No. AU-ARI-88-8). Maxwell Air Force Base, Alabama: USAF-Airpower Research Institute.

Dolgin, D. L. & Gibb, G. D. (1989). Personality assessment in aviation selection: Past, present and future. In R. Jensen (Ed.), *Aviation Psychology* (pp. 285-319). London: Gower Publishing Group.

Dolgin, D. L., Shull, R. N., & Gibb, G. D. (1987). Risk assessment and the prediction of student pilot performance. *Proceedings of the Fourth International Symposium on Aviation Psychology* (pp. 480-485). Columbus, OH: Ohio State University Aviation Psychology Laboratory.

Helmreich, R. L. (1982). *Pilot selection and training.* Paper presented at the American Psychological Association Annual Meeting, Washington, DC.

Helmreich, R. L., Sawin, L. L., & Carsrud, A. L. (1986). The honeymoon effect in job performance: Temporal increases in the predictive power of achievement motivation. *Journal of Applied Psychology*, *71*, 185-188.

Helmreich, R. L., & Spence, J. T. (1978). The work and family orientation questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. *JSAS Catalog of Selected Documents in Psychology*, *8*, 35.

Hilton, T. R., & Dolgin, D. L. (1991). Pilot selection in the military of the free world. In R. Gal & A. D. Mangelsdorff (Eds.), *Handbook of Military Psychology* (pp.88-101). Sussex, England: John Wiley and Sons.

Hollenbeck, J. R., & Whitener, E. M. (1988). Reclaiming personality traits for personnel selection: Self-esteem as illustrative case. *Journal of Management*, *14*, 81-91.

Hunter, D. R., & Burke, E. F. (1991). *Meta analysis of aircraft pilot selection measures* (Report No. ARI-AVSCOM-TR). Fort Rucker, AL: ARI-Army Research Institute.

Picano, J. J. (1991). Personality types among experienced military pilots. *Aviation, Space, and Environmental Medicine*, *62*, 517-520.

Retzlaff, P. D., & Gibertini, M. (1987). Air Force pilot personality: Hard data on the "Right Stuff." *Multivariate Behavioral Research*, *22*, 383-399.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*, 1-28.

Shull, R. N., & Dolgin, D. L. (1989). Personality and flight training performance. *Proceedings of the 33rd Annual Meeting of the Human Factors Society* (pp. 891-895). Denver, CO: Human Factors Society.

Spence, J. T., Helmreich, R. L., & Holahan, C. K. (1979). Negative and positive components of psychological masculinity and femininity and their relationships to self-reports of neurotic and acting-out behaviors. *Journal of Personality and Social Psychology*, *37*, 1673-1682.

Personality hardiness predicts soldier adjustment to combat stress[1]

Paul T. Bartone, Ph.D., David H. Marlowe, Ph.D.,
Robert K. Gifford, Ph.D. & Kathleen M. Wright, Ph.D.

Walter Reed Army Institute of Research
Department of Military Psychiatry
Washington, DC 20307-5100

ABSTRACT

U. S. Army soldiers (N = 8,632) who participated in the Gulf War were surveyed to determine psychological effects of combat exposure, and the possible role of personality hardiness as a stress moderator. Personality hardiness is a trait or cognitive style that theoretically influences how stressful experiences get processed and integrated by individuals. Survey measures included a 31-item combat exposure inventory; soldier cohesion scales; a short personality hardiness scale (30-items); and several measures of psychiatric outcome such as the Brief Symptom Inventory (BSI), the Impact of Events Scale (IES), and an indicator of risk for Post-Traumatic Stress Disorder based on DSM-III(R) diagnostic criteria. Multiple regression analyses show that personality hardiness (-) is a significant independent predictor of psychiatric distress, along with combat exposure (+) and unit cohesion (-). These regression results were replicated in a smaller sample of soldiers (N=777) for which both pre- and post-combat measures were available. Here, hardiness and combat exposure showed an interaction effect, supporting a stress-buffering hypothesis. The findings of this study suggest that renewed attention be paid to psychological variables in research on soldier adjustment and performance. Individual psychological characteristics (which may interact with and be shaped by social/contextual factors) can influence how soldiers process and make cognitive sense of extremely stressful events.

[1] Presented at the 34th Annual Conference of the Military Testing Association, 1992. This report draws on a large scale study of stress and adaptation in U.S. Army soldiers who participated in Operation Desert Storm, directed by Dr. David H. Marlowe, Chief, Department of Military Psychiatry. Thanks are due to Joseph M. Rothberg, Ph.D., Charles Hoover, and William Deerfield for their assistance with various aspects of this work. The views of the authors do not necessarily reflect those of the Department of the Army, or of the Department of Defense (para. 4-3, AR 360-5).

Personality hardiness predicts soldier adjustment to combat stress

Some of the most extreme human stressors are those encountered by soldiers in combat (Marlowe, 1986). Traumatic wartime experiences can have both short and long-term damaging impact on the mental health of military personnel (e.g., Solomon & Flum, 1988; Kulka et. al., 1990). The recent Gulf War presented an opportunity to study the psychological impact of combat stress on U.S. soldiers, and the possible role of psychological factors as stress moderators. While social processes such as unit cohesion, morale and leadership have received continued attention as factors in military performance and breakdown (Manning, 1991), research examining individual psychological processes in military samples (coping skills, cognitive appraisals, personality etc.) is less prominent. The present study draws on a larger program to examine a specific question: does the personality style described as "hardiness" contribute to healthy adjustment of soldiers exposed to combat stress?

As part of a larger research effort, post-combat debriefing surveys were collected on 8,632 U.S. Army soldiers who deployed to Saudi Arabia during the Gulf War. Surveys were also obtained on a control group of 465 soldiers who did not participate in the wartime deployment (Bartone et. al., 1992). All post-combat surveys were administered 3-12 months following the end of the war. Measures included a 31-item combat exposure inventory; soldier cohesion scales, both "horizontal" (soldier-soldier) and "vertical" (soldier-leader); personality hardiness (a short, 30-item form); several measures of psychiatric/psychological outcome such as the Brief Symptom Inventory (BSI; Derogatis, 1975), the Impact of Events Scale (IES; Horowitz, Wilner & Alvarez, 1979); and a specially constructed indicator of risk for Post-Traumatic Stress Disorder based on DSM-III(R) diagnostic criteria. A short pre-combat survey was administered to a smaller group of soldiers (N=833) just prior to the ground war offensive, permitting some prospective analyses to be done.

Personality Hardiness

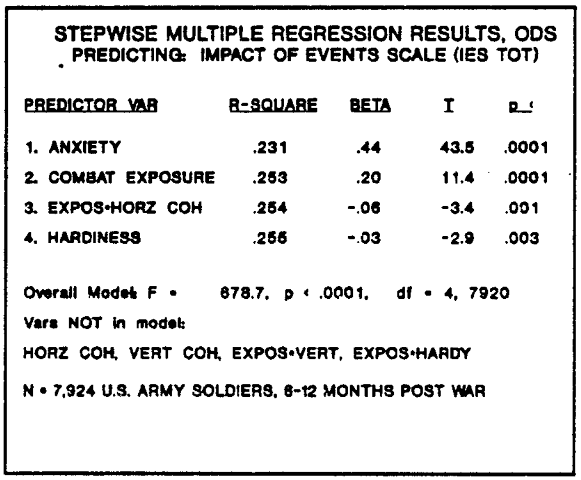
Originally developed by Salvatore Maddi and Suzanne Kobasa Ouellette (Maddi, 1967, 1970; Kobasa, 1979; Maddi & Kobasa, 1984), the concept of personality hardiness is grounded in existential psychology and personality theory (Kobasa & Maddi, 1977; Kierkagaard, 1954; Keen, 1970). Theoretically, as a function of their own psychosocial developmental history, hardy ("authentic" in an existential sense) persons are more open to experience on a variety of levels, and are more solidly grounded and confident in their sense of self and place in the social world. The critical implication for stress research is that hardy types are not as easily threatened or psychologically disrupted by ordinarily painful aspects of the human condition. This theoretical underpinning sets hardiness apart from such apparently related constructs as "optimism" (Scheier & Carver, 1985) or "hope" (Snyder et. al. 1991), which generally posit a much simpler process whereby stressful or painful experiences are disregarded or ignored. Of particular relevance to the domain of combat stress is Maddi's suggestion (1976) that the hardy person is not as vulnerable to the threat of imminent death. Empirical studies have confirmed personality hardiness is a promising individual differences variable that seems to influence the relation between psychosocial stress and health outcomes (e.g., Bartone, 1989a; Contrada, 1989; Kobasa, Maddi & Kahn, 1982; Roth et. al, 1989; Wiebe, 1991).

Perhaps partly as a function of its theoretical depth and complexity, the construct of hardiness has proved difficult to measure (Funk & Houston, 1992). The present work utilizes a refined 30-item hardiness scale ("Dispositional Resilience Scale") that corrects many of the problems of earlier hardiness measures (Bartone et. al., 1989; Bartone, 1991). It includes ten items each to measure the three general characteristics of commitment, control, and challenge that Kobasa (1979) suggested hardy persons possess. The Dispositional Resilience Scale is fully balanced for positive and negative items, with an equal number (15) of each. The correlation between the 30-item form and 27 non-overlapping items from the original hardiness scale (6 alienation from self, 2 alienation from work, 7 powerlessness, 10 security, 2 cognitive structure) is -.74 (bus driver sample, N=753; Bartone, 1991). Scores on the short form have demonstrated appropriate correlations with theoretically related (convergent) and unrelated (discriminant) variables, and are generally predictive of continued mental and physical health under a variety of environmental stressors (e.g., Bartone et. al., 1990; Bartone, 1989b). For example, scores on this measure were found to discriminate Army disaster assistance workers who remain healthy from those reporting stress-related symptoms over time (Bartone et. al., 1989).

The six-month stability coefficient is .57 (N=80 Army officers), and three-month is .58 (N=21 Army personnel workers). Cronbach's alpha for the total scale ranges from .70 to .85 depending on the sample. While reliability and factor analyses with various samples generally confirm the presence of the 3 sub-dimensions of commitment, control and challenge, internal consistency for the challenge scale is low (.62). It is important to remember that the three subscales were never intended to fully describe "hardiness" from a theoretical point of view, but rather were suggested by Kobasa as "three general characteristics" that "hardy persons are considered to possess" (1979, p. 3). As such, when taken together these characteristics can provide a useful operational (if theoretically incomplete) indicator of the personality style or type described by Maddi (1976) as authentic, individualistic, or "hardy".

Results/Discussion

Analyses for this report focus on the "Impact of Events Scale" (IES) as the primary outcome measure of psychiatric distress. The IES is thought to be an especially sensitive indicator of reactions to extreme or traumatic stressors (Horowitz, Wilner & Alvarez, 1979), and yields scores on subscales of Avoidance and Intrusion, as well as a Total score. Figure 1 presents results from a multiple regression analysis, predicting IES Total scores for the post-combat (N=7924) sample only. Scores on a generalized anxiety measure were entered as a covariate in an attempt to control for the possible confounding effects of neuroticism (Costa & McCrae, 1985; Funk, 1992). This probably represents an overly conservative approach in that anxiety scores in this sample very likely indicate realistic responses to actual stressful circumstances, to some degree. Thus, in partialling out the effects of anxiety on IES scores, we are probably removing some of the very effect we are trying to identify. Still, the problem of neurotic contamination of stress and illness reports is so important that this was deemed an appropriate compromise. Not surprisingly, anxiety was a strong independent predictor of IES total scores. The model also showed significant independent effects for Combat Exposure and Hardiness, as well as a Combat Exposure X Horizontal Cohesion interaction.



STEPWISE MULTIPLE REGRESSION RESULTS, ODS
PREDICTING: IMPACT OF EVENTS SCALE (IES TOT)

PREDICTOR VAR	R-SQUARE	BETA	T	p ‹
1. ANXIETY	.231	.44	43.5	.0001
2. COMBAT EXPOSURE	.253	.20	11.4	.0001
3. EXPOS•HORZ COH	.254	-.06	-3.4	.001
4. HARDINESS	.256	-.03	-2.9	.003

Overall Model: F • 678.7, p ‹ .0001, df • 4, 7920

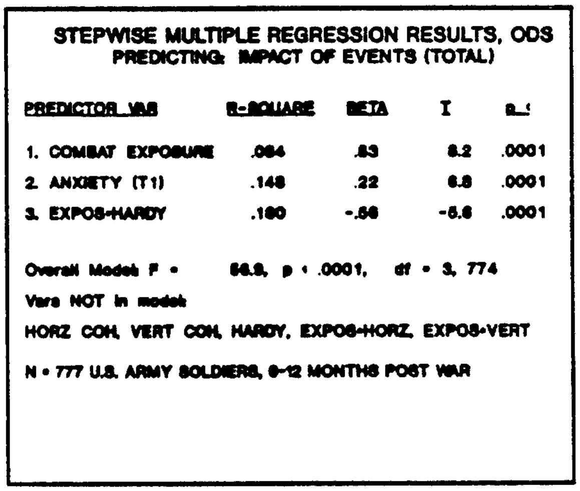
Vars NOT in model:

HORZ COH, VERT COH, EXPOS•VERT, EXPOS•HARDY

N • 7,924 U.S. ARMY SOLDIERS, 6-12 MONTHS POST WAR

Figure 1

Prospective data were available for a smaller sample (N=777) of soldiers who had completed both pre- and post-combat surveys. For this group, the pre-combat survey was administered 1-4 weeks before the launch of offensive ground operations. Using these data, it was possible to control for actual pre-combat generalized anxiety levels, rather than post-combat levels. This provides a better control for generalized anxiety or neuroticism, since scores could not be influenced by actual combat events that came later in time. Still, it is probably a conservative approach in that pre-combat anxiety scores may reflect a realistic appraisal of events likely to follow, rather than generalized worry or neuroticism. As Figure 2 shows, pre-combat anxiety is a significant predictor of post-combat IES scores. With the effects of pre-combat anxiety removed, the remaining significant effects in the model are for Combat Exposure and a Combat Exposure X Hardiness interaction term. Thus, the independent main effect of hardiness on soldier adjustment in the cross-sectional data becomes an interaction effect (with Combat Exposure) when pre-combat levels of anxiety are controlled for. These findings are supportive of a "stress-buffering" hypothesis, wherein persons who possess the personality style of hardiness are less vulnerable to the disruptive effects of severe stress.



STEPWISE MULTIPLE REGRESSION RESULTS, ODS
PREDICTING: IMPACT OF EVENTS (TOTAL)

PREDICTOR VAR	R-SQUARE	BETA	T	P <
1. COMBAT EXPOSURE	.084	.53	8.2	.0001
2. ANXIETY (T1)	.148	.22	6.8	.0001
3. EXPOS-HARDY	.180	-.56	-6.6	.0001

Overall Model F = 66.9, p < .0001, df = 3, 774

Vars NOT in model

HORZ COH, VERT COH, HARDY, EXPOS-HORZ, EXPOS-VERT

N = 777 U.S. ARMY SOLDIERS, 6-12 MONTHS POST WAR

Figure 2

Similar regression analyses were performed with the IES subscales of Avoidance and Intrusion as dependent variables. In the larger sample (N=7924), both hardiness and horizontal cohesion enter as independent significant (negative) predictors of Avoidance scores, along with Combat exposure and the control variable of Anxiety. Cohesion did not predict Intrusion scores, though it earlier showed an interaction effect with Combat Exposure on IES Total scores. This suggests that the healthy effects of cohesion are felt primarily with regard to decreased tendencies to avoid or deny disturbing responses to trauma. It would seem that soldiers in highly cohesive units, like high hardy soldiers, are more open to and accepting of stressful events they've experienced.

These results demonstrate that the personality style of "hardiness" is one of several factors that has a significant effect on post-combat adjustment for U.S. soldiers. This effect pertains even when the possible confounding influence of generalized anxiety is controlled for, and is independent of social variables such as cohesion. Perhaps more importantly, hardiness was found to interact with combat exposure, lending support to the theorized function of hardiness as a stress-resistance resource (Maddi, 1976). With these empirical effects demonstrated in samples of U.S. soldiers exposed to actual combat, it now remains for additional research to identify the manner and form in which this personality style develops, how it is influenced by group processes such as leadership and cohesion, and the degree to which it can be trained or modeled. It is also the task of future research to identify the underlying mechanisms, cognitive, emotional, and physiological, wherein the generalized orientation to the world summarized as "hardiness" serves to buffer the ill-effects of environmental stressors.

REFERENCES

Bartone, P.T. (1989a). Predictors of stress-related illness in city bus drivers. J. of Occupational Medicine, 31, 657-663.

Bartone, P.T. (April, 1989b). Hardiness, optimism, and health: A construct validity study. 60th Annual Meeting of the Eastern Psychological Association, Boston, MA.

Bartone, P.T. (June, 1991). Development and validation of a short hardiness measure. Presented at the 3rd Annual Convention of the American Psychological Society, Washington, DC.

Bartone, P.T., Gifford, R.K., Wright, K.M., Marlowe, D.H. & Martin, J.A. (June, 1992). U.S. soldiers remain healthy under Gulf War stress. Presented at the 4th Annual Convention of the American Psychological Society, San Diego, CA.

Bartone, P.T., Ursano, R.J., Wright, K.W. & Ingraham, L.H. (1989). The impact of a military air disaster on the health of assistance workers: A prospective study. J. of Nervous and Mental Disease, 177, 317-328.

Bartone, P.T., McCarroll, J.E., Wright, K.M., Ursano, R.J. & Fullerton, C.S. (August, 1990). Personality hardiness and resiliency in high-stressed military populations. Presented at the 98th Annual Convention of the American Psychological Association, Boston, MA.

Contrada, R.J. (1989). Type A behavior, personality hardiness, and cardiovascular responses to stress. J. of Personality and Social Psychology, 57, 895-903.

Costa, P.T. & McCrae, R.R. (1985). Hypochondriasis, neuroticism, and aging: when are somatic complaints unfounded? American Psychologist, 40, 19-28.

Derogatis, L.R. (1975). Brief Symptom Inventory. Baltimore: Clinical Psychometric Research.

Funk, S.C. (1992). Hardiness: A review of theory and research. Health Psychology, 11 (5), 335-345.

Funk, S.C. & Houston, B.K. (1987). A critical analysis of the hardiness scale's validity and utility. J. of Personality and Social Psychology, 53, 572-578.

Horowitz, M., Wilner, N. & Alvarez, W. (1979). Impact of Event Scale: A measure of subjective stress. Psychosomatic Medicine, 41 (3), 209-218.

Keen, E. (1970). Three Faces of Being: Toward an Existential Clinical Psychology. New York: Appleton-Century-Crofts.

Kierkagaard, S. (1954). The sickness unto death. New York: Doubleday.

Kobasa, S.C. (1979). Stressful life events, personality, and health: An inquiry into hardiness. J. of Personality and Social Psychology, 37, 1-11.

Kobasa, S.C. & Maddi, S.R. (1977). Existential personality theory. In R. Corsini (Ed.), Existential Personality Theories. Itasca, IL: Peacock.

Kobasa, S.C., Maddi, S.R., & Kahn, S. (1982) Hardiness and health: A prospective study. J. of Personality and Social Psychology, 42, 168-177.

Kulka, R.A., Schlenger, W.E., Fairbank, J.A., Hough, R.L., Jordan, B.K., Marmar, C.R., & Weiss, D.S. (1990). Trauma and the Vietnam War Generation. New York: Brunner/Mazel.

Maddi, S.R. (1967). The existential neurosis. J. of Abnormal Psychology, 72, 311-325.

Maddi, S.R. (1970). The search for meaning. In M. Page (Ed.), Nebraska Symposium on Motivation. Lincoln, Nebraska: University of Nebraska Press.

Maddi, S.R. (1976). Personality Theories: A Comparative Analysis (3rd Edition). Homewood, IL: Dorsey Press.

Maddi, S.R. & Kobasa, S.C. (1984). The Hardy Executive. Homewood, IL: Dow Jones-Irwin.

Manning, F.R. (1991). Morale, cohesion and esprit de corps. In R. Gal & A.D. Mangelsdorff, Handbook of Military Psychology. Chichester, UK: Wiley.

Marlowe, D.H. (1986). The human dimension of battle and combat breakdown. In R.A. Gabriel (Ed.), Military Psychiatry: A Comparative Perspective. Westport, CN: Greenwood.

Roth, D.L., Wiebe, D.J., Fillingim, R.B. & Shay, K.A. (1989). Life events, fitness, hardiness, and health: A simultaneous analysis of proposed stress-resistance effects. J. of Personality and Social Psychology, 57, 136-142.

Scheier, M.F. & Carver C.S. (1985) Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. Health Psychology, 4, 219-247.

Solomon, Z. & Flum, H. (1988). Live events, combat stress reaction and post-traumatic stress disorder. Social Science & Medicine, 26 (3), 319-325.

Snyder, C.R., Harris, C., Anderson, J.R., Holleran, S.A., Irving, L.M., Sigmon, S.T., Yoshinobu, L., Gibb, J., Langelle, C., & Harney, P. (1991). The will and the ways: Development and validation of an individual differences measure of hope. J. of Personality & Social Psychology, 60, 570-585.

Wiebe, D.J. (1991). Hardiness and stress moderation: A test of proposed mechanisms. J. of Personality and Social Psychology, 60, 89-99.

Mood Changes Associated With Litter Carrying

William Tharion[1], Valerie Rice, and Marilyn Sharp

United States Army Research Institute of Environmental Medicine
and GEO-CENTERS INC.[1]
Natick, Massachusetts 01760

INTRODUCTION

The carrying and lifting of patient litters readily induces muscular fatigue due to sustained muscular contractions (Lind & McNicol, 1968). Psychological perception has been shown to affect physical performance and contributes greatly to the ability to perform at physiological limits (Morgan, 1985). Assessments of athletes' mood using the Profile of Mood States (POMS) has shown the best performances are by those who exhibit the iceberg profile consisting of higher vigor and lower negative moods than college norms (Morgan, 1985). Attention to mood changes is particularly important because mood changes often occur prior to changes in physical performance (Opstad, Ekanger, Nummested, & Raabe, 1978).

The purpose of this study was to examine mood changes when using a shoulder harness during litter carrying. A secondary purpose was to ascertain mood changes resulting from carrying in 2 vs. 4-person teams, between female vs. male litter bearers, and also for a 15 min bout of repeated, rapid, short, litter carries vs. a single, moderate speed prolonged carry.

METHOD

Physically active military volunteers, 12 male and 11 female, carried a military litter with a patient manikin while walking on a treadmill. Heart rate monitors recorded and stored heart rate data. Subject characteristics are shown in Table 1.

Immediately before and within 10 min after litter carrying, the POMS (McNair, Lorr, & Droppleman, 1971) was completed. The POMS is a 65-item questionnaire with feelings rated for each item on a five-point scale from "not at all" to "extremely". Responses are how one feels "right now". The items assess six different mood scales: tension, depression, anger, vigor, fatigue, and confusion.

A brief description of the various litter carrying parameters is provided below. These parameters have been described in more detail previously (Tharion, Rice, Sharp, and Marlowe, 1992).

Carry Type. Two different types of litter carrying were examined. The first carry task simulated carrying and loading as many patients as possible into an ambulance in 15 mins. Subjects carried a litter with an 81.6 kg manikin 50 m on a treadmill. To simulate loading the litter into an ambulance subjects dismounted the treadmill and walked/ran to a weight stack machine where they lifted the weight they carried (40.9 kg if part of a 2-person team and 20.5 kg if part of a 4-person team) to a height of 135 cm. Subjects then returned to the treadmill and ran 50 m to get the next patient. Treadmill speed was self-paced and ranged from 4.0 km/hr to 20.9 km/hr. The second task, the prolonged carry, simulated carrying a litter without rest breaks for as long as possible up to 30 min at a constant rate, 4.0 km/hr.

Table 1. Male and female physical characteristics (Mean ± Standard Deviation).

CHARACTERISTIC	MALE	FEMALE
Height (cm)	178.4 ± 7.5	162.6 ± 7.1
Weight (kg)	79.2 ± 13.1	58.1 ± 6.2
Body Fat (%)	15.4 ± 4.0	24.9 ± 6.5
Bench Press Max (kg)	85.8 ± 19.5	38.8 ± 6.4
Dead Lift Max (kg)	135.1 ± 23.2	82.1 ± 11.5
Age (yrs)	20.8 ± 2.6	23.6 ± 4.0

Team Size. Subjects carried the head-end of the litter, in either 2 or 4-person simulated teams. Therefore, during a simulated 2-person carry one subject supported the front end of the litter by holding both head-end litter handles. For the simulated 4-person carry the front was supported by 2 subjects, each holding one of the head-end handles. The rear end was supported by ceiling straps.

Harness Use. Subjects carried the litter both with and without a harness. The harness was designed to shift support of the carried weight from the hands to the back and shoulders.

Mood changes were evaluated using a repeated measures ANOVA by these variables: gender (male/female), exercise (pre/post), type of carry (15 min/prolonged), team size (2-person/4-person) and harness condition (harness/no harness). Duncan's post hoc test determined differences between means. Team size and harness condition were randomized. Prolonged carries followed 15 min carries.

RESULTS

Raw scores for the individual mood scales by the independent variables are summarized in Table 2. Higher levels of tension ($p \leq 0.05$) and fatigue ($p \leq 0.001$) were associated with the 15 min carry than with the prolonged carry. Heart rate was also greater during the 15 min carries compared to the prolonged carries ($p \leq 0.001$). Greater subjective fatigue ($p \leq 0.01$) was felt after compared to before carrying the litter for both

Table 2. POMS raw scores (Mean ± S.D.) by sex, carry type, harness, team size, and exercise condition.

	SEX		CARRY TYPE	
	Female	Male	15 Min	Prolonged
TENSION	4.2 ± 2.9	5.4 ± 4.1	5.4 ± 4.2	* 4.3 ± 2.8
DEPRESSION	2.5 ± 4.9	2.8 ± 6.2	3.2 ± 6.6	2.2 ± 4.1
ANGER	12.4 ± 2.3	13.3 ± 3.3	3.3 ± 3.3	2.5 ± 2.3
VIGOR	15.4 ± 9.1	11.0 ± 7.1	12.6 ± 6.4	13.6 ± 4.0
FATIGUE	2.4 ± 3.1	3.0 ± 4.1	3.5 ± 4.1	* 1.8 ± 2.8
CONFUSION	2.8 ± 2.5	3.4 ± 2.9	3.4 ± 3.2	2.8 ± 2.3

	HARNESS		TEAM SIZE	
	Yes	No	2-Person	4-Person
TENSION	5.5 ± 3.6	4.8 ± 3.6	4.9 ± 3.8	4.8 ± 3.4
DEPRESSION	2.4 ± 4.9	2.9 ± 6.1	3.2 ± 6.8	2.1 ± 3.9
ANGER	2.6 ± 2.6	3.4 ± 3.0	2.9 ± 3.0	2.8 ± 2.7
VIGOR	13.1 ± 8.3	14.7 ± 8.5	12.5 ± 8.3	13.7 ± 8.5
FATIGUE	2.7 ± 3.7	2.6 ± 3.6	3.0 ± 4.7	2.4 ± 3.1
CONFUSION	3.1 ± 2.5	3.1 ± 3.0	3.4 ± 3.2	2.8 ± 2.2

	EXERCISE STATE	
	Pre	Post
TENSION	4.7 ± 3.7	5.0 ± 3.6
DEPRESSION	2.9 ± 5.7	2.5 ± 5.3
ANGER	2.8 ± 2.8	2.9 ± 2.8
VIGOR	13.3 ± 8.1	12.9 ± 8.7
FATIGUE	2.0 ± 3.4	* 3.8 ± 3.6
CONFUSION	3.3 ± 3.0	2.9 ± 2.5

* Main effect differences via ANOVA at ($p<0.05$)

types of carries. Significant interaction effects between carry type and the effect of the exercise were observed for depression ($p \leq 0.04$) and fatigue ($p \leq 0.01$). After the 15 min carry depression decreased and fatigue increased.

Subjects felt more vigor after the exercise when wearing a harness compared to not wearing a harness ($p \leq 0.02$). Lower heart rates were exhibited during exercise when wearing the harness ($p \leq 0.001$). Vigor was reduced significantly after exercise for male 2-person teams compared to male 4-person or female 2 or 4-person teams as shown in Figure 1, ($p \leq 0.03$).

Figure 1. Exercise state by sex by team size interaction for POMS mood of vigor.

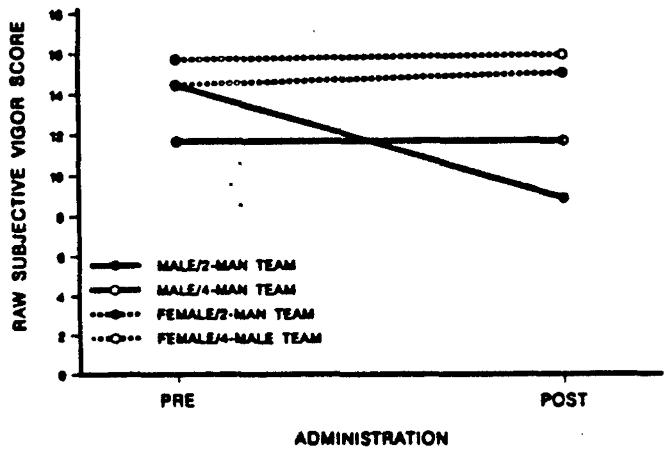
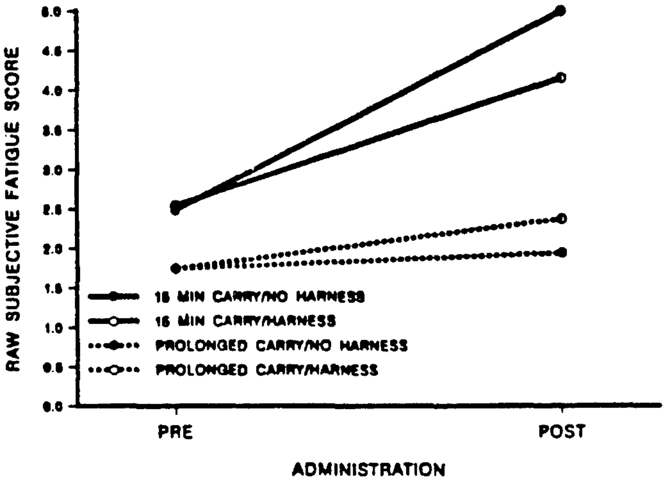


Figure 2. Exercise state by carry type by harness condition interaction for POMS mood scale of fatigue.



The greatest fatigue was felt after the 15 min carry when a harness was not used. Figure 2 shows this significant 3-way interaction. Similarily heart rate during the exercise was greatest in the 15 min carry when no harness was used ($p \leq 0.05$).

DISCUSSION

Higher fatigue and tension scores were associated with the 15 min carry resulting from faster self-paced repeated carries combined with a heavy lift. Using a harness reduces subjective feelings of fatigue in both the 15 min and prolonged carries. Heart rate was also shown to be higher during the 15 min carry but was reduced when a harness was worn. Those tasks that are more physically rigorous are likely to have greater mood changes. Bugge, Opstad and Magnus (1979) reported a high correlation between negative mood induced by the physical rigors of a military ranger training course and impaired performance on a number of cognitive tests, e.g. logical reasoning. Although, cognitive measures were not assessed in the present study, there are a number of cognitive tasks the medic must perform, e.g. assessing medical condition, triage, etc. Therefore, the importance of minimizing the negative psychological impact of litter carrying is important for the medic to successfully complete their mission. As Opstad, Ekanger, Nummested, and Raabe (1978) remark, mood changes often occur before changes in performance, serving as a warning signal to potential performance decrements.

The characteristics of higher fatigue and tension scores associated with the 15 min carry suggest this carry was psychologically more stressful. From a human factors standpoint the task was more complex. The combination of carrying, lifting and running to retrieve the next patient is much more difficult physically and mentally than carrying at a steady state.

It is likely the high aerobic demand of the 15 min repetitive carry/lift was responsible for the negative mood changes observed after that type of carry. Previous research (Tharion, Harman, Kraemer and Rauch, 1990) evaluating mood states in various strength training routines found the more negative moods were associated with workouts with a large aerobic component. Those workouts with more repetitions using a lower weight, less rest and more total work were found to be psychologically more distressing than workouts where the rest was greater, less work was performed and fewer repetitions were done with more weight.

Depression, although significantly reduced post-carry, was below college norms (McNair, et al., 1971) even prior to litter carrying. This change perhaps should be viewed as subjects becoming happier after completing the litter carry. Subjects may have experienced a sense of accomplishment from completing a strenuous, challenging task. Although the POMS scale doesn't specifically have a happiness dimension, the lessening of the depression score, which was well below college norms to begin with, suggests a happiness dimension as opposed to depression.

Since past research has shown that positive mood patterns are associated with better physical performance, strategies which reduce the negative psychological effects of the exercise should be employed. The use of a shoulder harness and carrying in 4-person teams are associated with a more positive mood which in turn is likely to produce better physical and cognitive performances in litter bearers.

REFERENCES

Bugge, J.F., Opstad, P.K., & Magnus, P.M. (1979). Changes in the circadian rhythm of performance and mood in healthy young men exposed to prolonged, heavy physical work, sleep deprivation, and caloric deficit. *Aviation, Space and Environmental Medicine, 50*: 663-668.

Lind, A.R. & McNicol, G.W. (1968). Cardiovascular responses to holding and carrying weights by hand and by shoulder harness. *Journal of Applied Physiology, 25*: 261-267.

McNair, D.M., Lorr, M. & Droppleman, L.E. (1971). *EDITS manual for the Profile of Mood States.* San Diego, CA: Educational and Industrial Testing Service.

Morgan, W.P. (1985). Selected psychological factors limiting performance: a mental health model. In D.H. Clarke & H.M. Eckert (Eds.), *Limits of human performance* (pp. 70-80). Academy Papers, No. 18. Champaign, IL: Human Kinetics.

Opstad, P.K., Ekanger, R., Nummestad, M., & Raabe, N. (1978). Performance, mood, and clinical symptoms in men exposed to prolonged, severe physical work and sleep deprivation. *Aviation, Space, and Environmental Medicine, 49*: 1065-1073.

Tharion, W.J., Harman, E. A., Kraemer, W.J. & Rauch, T.M. (1991). Effects of different weight training routines on mood states. *Journal of Applied Sport Science Research, 5*: 60-65.

Tharion, W.J., Rice, V., Sharp, M.A. & Marlowe, B.E. (1992). *The effects of litter carrying on rifle shooting.* Manuscript submitted for publication.

Do Dispositional Variables Play a Role in Job Satisfaction?

Stephanie Booth-Kewley

Navy Personnel Research and Development Center
San Diego, CA

Most past research on job attitudes and job satisfaction has emphasized the importance of situational factors. The contribution of dispositional variables to job satisfaction has been largely ignored. Recently, however, dispositional constructs as predictors of job attitudes have begun to receive research attention, and evidence suggesting that disposition may be relevant to job attitudes is beginning to accumulate (Staw, Bell, & Clausen, 1986; Levin & Stokes, 1989).

Just as job satisfaction is influenced by objective characteristics of the job and the organization, it may also be influenced by ongoing emotional states and dispositions of the individual. One study, for example (Staw, Bell, & Clausen, 1986), found that affective disposition in early adolescence was a significant predictor of job satisfaction in adulthood. Another investigation (Levin & Stokes, 1989) found a significant negative association between negative affect--the tendency to experience negative emotions--and job satisfaction, even after the effects of job characteristics were controlled. Brief and his colleagues (Brief, Burke, George, Robinson, & Webster, 1988) found that negative affect was significantly associated with both job satisfaction and job stress. Similarly, George (1989) found that negative affect was related to turnover intentions and job tenure, and that positive affect was related to job tenure.

The past decade has witnessed a resurgence of interest in personality variables as predictors of organizational attitudes and behaviors. Recent research suggests that broad personality variables may be associated with a variety of important organizational outcomes such as absenteeism (George, 1989), training performance (Hough, 1987), and job performance (Barrick & Mount, 1991). Given evidence that personality variables such as Neuroticism and Extraversion have substantial associations with life satisfaction, it seems likely that personality variables will be related to job attitudes and job satisfaction. More research is needed to determine the effects of personality on job and organizational attitudes.

The purpose of this study was to examine the association of positive and negative affect and six personality variables in relation to job and organizational attitudes in a Navy sample. It was hypothesized that positive affect would be positively related to job satisfaction and qsatisfaction with the Navy, and negative affect would be negatively related to these variables. It was hypothesized that Extraversion, Agreeableness, Conscientiousness, Optimism, and Self-Esteem would be positively related to job satisfaction and satisfaction with the Navy, and that Neuroticism would be negatively related to these variables.

Method

The study was conducted as part of a larger project on quality of life in the Navy (Booth-Kewley & Thomas, 1992).

Sample

One hundred and thirty two enlisted Navy personnel served as respondents. Respondents were from four duty locations, two in San Diego and two in Norfolk, Virginia.

There were 68 males and 64 females in the sample. The sample was 52 percent nonHispanic White, 20 percent Elack, 17 percent Hispanic, seven percent Asian; the remaining four percent were from "other" race/ethnic groups. Respondents ranged in age from 18 to 46, with a mean age of 28 years (SD = 6.9 years). All but two respondents had high school diplomas; the other two had high school equivalency degrees. The pay grades of the respondents ranged from E-1 to E-9, with a mean pay grade of 4.6 (SD = 1.6).

Fifty percent of the respondents were in white-collar clerical or administrative job ratings (e.g., Disbursing Clerk) and 46 percent were in blue-collar aviation ratings (e.g., Aviation Electronics Technician). The remaining four percent were in ratings that were neither aviation-related nor clerical.

Measures

Affect Balance Scale. Positive and negative affect were measured using the Bradburn Affect Balance Scale (Bradburn, 1969). The Affect Balance scale is a ten-item measure, made up of two subscales: Positive Affect (5 items), and Negative Affect (5 items). The Affect Balance Scale has been widely used, and has good psychometric properties (Bradburn, 1969; George & Bearon, 1980). In the present sample, internal consistencies (coefficient alphas) of .67 and .70 were found for the Positive Affect and Negative Affect scales, respectively.

NEO Five-Factor Inventory. Costa and McCrae's (1989a) NEO Five-Factor Inventory (NEO-FFI) was used to assess four personality dimensions: Neuroticism, Extraversion, Agreeableness, and Conscientiousness. (The fifth scale of the NEO, Openness to Experience, was not used because of time constraints and because it was not thought to be relevant to quality of life.) The NEO-FFI is the short, 60-item version of the NEO Personality Inventory (Costa & McCrae, 1985), a widely used personality instrument. Each of the NEO-FFI scales has adequate psychometric properties (Costa & McCrae, 1989a; 1989b). For the present sample, the internal consistencies of the NEO-FFI scales were .85 for Neuroticism, .72 for Extraversion, .68 for Agreeableness, and .86 for Conscientiousness.

Optimism. Optimism was measured using a four-item scale developed by Scheier et al. (1989). The four-item measure is a short version of Scheier and Carver's (1985) Life Orientation Test (LOT), a widely used measure of Optimism. In the present sample, the four-item LOT had an internal consistency of .58.

Rosenberg Self-Esteem Scale. Self-Esteem was measured using the Rosenberg Self-Esteem Scale (Rosenberg, 1965), a widely used, 10-item questionnaire. This measure has adequate psychometric properties (Rosenberg, 1965). In the present sample, this measure had an internal consistency of .85.

Job Satisfaction. Job satisfaction was an index consisting of responses to two questions: (1) the interview question, "How do you feel about the job you have right now?", and (2) the questionnaire item, "How do you feel about your job?" In both instances, respondents were asked to indicate their feelings about their jobs on a seven-point response scale. Responses to the two job satisfaction questions were averaged to form a Job Satisfaction index. The coefficient alpha of this index was .85.

Satisfaction with the Navy. Satisfaction with the Navy was an index consisting of responses to two questions: (1) the interview question, "How much do you like being in the Navy?" and (2) the questionnaire item, "How do you feel about the Navy?". Both questions were answered using a seven-point response scale. Responses to these two questions assessing feelings towards the Navy were averaged to form Satisfaction with the Navy index. The coefficient alpha of this index was .86.

Navy Stress. The measure of Navy stress was the single interview question, "How stressful do you find being in the Navy?" Responses were given on a seven-point response scale.

Intention to Reenlist. The measure of intention to reenlist was the interview question, "When you complete your current enlistment, do you plan to re-enlist?". The possible responses of "Yes", "Unsure", or "No" were assigned codes of 3, 2, and 1, respectively.

Demographics. Respondents completed a demographic questionnaire which asked for their sex, age, race/ethnicity, education level, job rating, marital status, and pay grade.

Procedure

Respondents were interviewed individually, and then administered a set of paper-and-pencil questionnaires, which included the measures of affect, personality, and demographics described above.

Results

The affect and personality variables were correlated with the measures of job and Navy satisfaction; these results are presented in Table 1.

Table 1

Correlations of Personality and Affect Variables
with Job and Navy Satisfaction

	Job Satisfaction	Satisfaction with the Navy	Navy Stress	Intention to Reenlist
Positive Affect	.14	.01	-.12	.01
Negative Affect	-.31**	-.33**	.30**	-.19*
Neuroticism	-.14	-.11	.27**	-.06
Extraversion	.11	.16	-.02	.15
Agreeableness	.26**	.30**	-.13	.15
Conscientiousness	.17	.10	-.02	-.01
Optimism	.16	.22**	-.19	.03
Self-Esteem	.06	-.04	-.14	-.10

** $p < .05$

* $p < .01$

Affect

Negative affect had a significant inverse association with all four of the job and Navy satisfaction measures: job satisfaction ($r = -.31$, $p < .01$), satisfaction with the Navy ($r = -.33$, $p < .01$), Navy stress ($r = -.30$, $p < .01$), and intention to reenlist ($r = -.19$, $p < .05$). Thus, respondents high in negative affect were less satisfied with their jobs, less satisfied with the Navy, found the Navy to be more stressful, and were less likely to plan to reenlist than respondents low in negative affect. Contrary to expectation, positive affect was not related to any of the job or Navy satisfaction variables.

Personality

Three of the six personality variables had significant associations with the job and Navy satisfaction variables. Agreeableness was positively associated with both job satisfaction ($r = .26$, $p < .01$) and with satisfaction with the Navy ($r = .30$, $p < .01$). Neuroticism was positively associated with Navy stress ($r = .27$, $p < .01$). Optimism was positively associated with satisfaction with the Navy ($r = .22$, $p < .01$).

Contrary to expectation, the other three personality variables--Extraversion, Conscientiousness, and Self-Esteem--were not associated with any of the job or Navy satisfaction variables.

Discussion

This study found that negative affect was associated with job satisfaction, satisfaction with the Navy, perceived stressfulness of the Navy, and intention to reenlist. Positive affect, however, was not associated with any of these variables. It appears that negative affect is closely tied to people's evaluations of their jobs and attitude towards the Navy. However, positive affect does not seem to be related to these evaluations.

The present results are consistent with past research linking negative affect with job satisfaction (Brief et al., 1988; Levin & Stokes, 1989). Given evidence that people high in negative affect perceive their jobs as containing lower amounts of desirable characteristics (Levin & Stokes, 1989), it may be that negative affect influences the way in which job-related information is processed. Relative to their low negative affect counterparts, individuals high in negative affect may focus on, and give greater cognitive emphasis to, the negative aspects of their jobs.

It is somewhat surprising that positive affect was not associated with any of the job or Navy attitude variables. Intuitively, one would expect both positive and negative affect to have roughly equivalent but opposite effects on job and organizational attitudes. However, the present results are consistent with a study by George (1989), which found negative affect but not positive affect to be linked with turnover intentions. It may be that positive affect simply does not affect the way in which job-related information is processed, whereas negative affect does.

Three of the personality variables--Agreeableness, Neuroticism, and Optimism--were associated with at least one of the job or Navy attitude variables. These findings are consistent with the subjective well-being literature, which has found that personality variables are associated with life satisfaction (Costa & McCrae, 1980; Diener, 1984). It seems intuitively plausible that agreeable and optimistic people would evaluate most aspects of their lives including their jobs and organizations more positively than their less agreeable, less optimistic counterparts. Conversely, it seems plausible that neurotic individuals would tend to evaluate most aspects of their lives in a negative manner.

It should be pointed out that the amount of variance in job satisfaction explained by the affect and personality variables in this study was fairly small. Situational variables, such as characteristics of the job, and interactive variables, such as person-environment fit, may have an equal or greater impact on job and organizational satisfaction than dispositional variables.

Further research is needed to clarify the ways in which dispositional tendencies influence on-the-job behavior, job performance, and job attitudes. For example, what mechanisms are responsible for the relationship between negative affect and lower job satisfaction? What are the effects of positive and negative affect on workplace interactions? What are the effects of Agreeableness and Neuroticism? Do dispositional variables affect job performance? Do they affect organizational commitment and job involvement? These and related questions should be addressed in future research.

In conclusion, the present study found that dispositional variables make a significant contribution to job and organizational satisfaction. It is recommended that future research on job satisfaction include measures of dispositional variables whenever possible, so that the disposition-job satisfaction relationship can be more fully understood.

References

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Booth-Kewley, S., & Thomas, M. D. (1992). *The subjective quality of life of Navy personnel.* San Diego: Navy Personnel Research and Development Center. Manuscript under preparation.

Bradburn, N. M. (1969). *The structure of psychological well-being*. Chicago: Aldine.

Brief, A. P., Burke, M. J., George, J. M., Robinson, B. S., & Webster, J. (1988). Should negative affect remain an unmeasured variable in the study of job stress? *Journal of Applied Psychology, 73*, 193-198.

Costa, P.T., Jr., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology, 38*, 668-678.

Costa, P.T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory Manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P.T., Jr., & McCrae, R. R. (1989a). *NEO Five-Factor Inventory: Form S.* Odessa, FL: Psychological Assessment Resources.

Costa, P.T., Jr., & McCrae, R. R. (1989b). *NEO PI/FFI Manual Supplement.* Odessa, FL: Psychological Assessment Resources.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*, 542-575.

Emmons, R. A., & Diener, E. (1985). Personality correlates of subjective well-being. *Personality and Social Psychology Bulletin, 11*, 89-97.

George, J. M. (1989). Mood and absence. *Journal of Applied Psychology, 74*, 317-324.

George, L. K., & Bearon, L. B. (1980). *Quality of life in older persons: Meaning and measurement*. New York: Human Sciences Press.

Hough, L. M. (1987). *Literature review: Utility of temperament, biodata, and interest assessment for predicting job performance*. Minneapolis: Personnel Decisions Research Institute.

Levin, I., & Stokes, J. P. (1989). Dispositional approach to job satisfaction: Role of negative affect. *Journal of Applied Psychology, 74*, 752-758.

Scheier, M. F., & Carver, C S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology, 4*, 219-247.

Scheier, M. F., Matthews, K. A., Owens, J. F., Magovern, G. J., Lefebvre, R. C., Abbott, R. A., & Carver, C. S. (1989). Dispositional optimism and recovery from coronary artery bypass surgery: The beneficial effects on physical and psychological well-being. *Journal of*

Personality and Social Psychology, *57*, 1024-1040.

Staw, B. M., Bell, N. E. & Clausen, J. A. (1986). The dispositional approach to job attitudes: A lifetime longitudinal test. *Administrative Science Quarterly*, *31*, 56-77.

VALIDATING TRAINING USING TASK ANALYSIS DATA

Amiel T. Sharon and James M. Reeder
U.S. Office of Personnel Management

INTRODUCTION

Background

The purpose of this paper is to describe the method used to validate a U.S. Army Corps of Engineers training program for skilled workers in the power plant industry. The Corps operates hydroelectric power plants that supply electricity to a wide area of the United States. The power plants are maintained and operated by skilled power plant electricians, mechanics, and shift operators who are trained for their duties in a formal program called the Hydropower Training Program.

The Hydropower Training Program, which consists of classroom instruction, on-the-job training (OJT), and independent study, is three years long for electricians and mechanics and four years long for shift operators. Classroom instruction consists of academic courses, such as physics and mathematics, and job-related courses, such as generator relays. It also includes specific trade skill instruction such as welding. The classroom instruction is "frontloaded" by being given primarily during the first year of the training program. The remaining time in the program is devoted primarily to OJT and to independent study. OJT is structured by providing to each trainee written guidelines that specify what the trainee is expected to learn on the job and in independent study.

A structured oral examination, which is the primary training measure of the program, is administered to trainees at the end of each training period to assess knowledge acquired in the classroom, on the job, and in independent study. The oral examination functions both as an evaluation and a didactic tool. It serves to assess trainee progress and at the same time to provide trainees with specific feedback on their knowledge, skill, and performance. The oral examination session for a single trainee is conducted by three examiners with experience in the trainee's trade. The examiners, who are always from a different power plant than the trainee, ask the questions and evaluate the trainee's answers. Although the examiners follow a standard evaluation outline that lists the topics and subtopics to be covered on the examination, they themselves determine the specific questions to ask about each topic.

At the conclusion of the oral examination session, the examiners independently evaluate the trainee's performance by assigning points to each topic by following a standard evaluation outline. The outline indicates the maximum number of points to be given to each examination topic. The examiners pool their independent judgments, agree on a single score for each topic, and add the scores to calculate the total score on the oral examination.

There is no passing point or cutoff score on the oral examination. The total oral examination score is added to other numerical scores to make up a composite score for each trainee for each training period. Depending on the training period, the composite score may include points from the oral examination, on-the-job performance appraisal by the trainee's work supervisor, and/or grades from classroom instruction.

Validation strategy

Content validity was deemed to be the most suitable approach to validate the Hydropower Training Program because the goal of the training program is to impart knowledge and skill necessary for successful job performance. Content validation makes it possible to identify the ability, skill, and knowledge requirements of the job and compare the requirements to the subject-matter taught in training. Unlike criterion-related validity, content validity does not require large samples of trainees for statistical analyses. Instead, it relies on the judgments of subject matter experts in identifying and comparing the content of training to the content of the job.

METHOD

The first step in the content validation of the training program was a task analysis. Since the Program targets three job specialties (electrician, mechanic, and shift operator), each specialty required a separate task analysis. The task analysis procedure consisted of developing task inventories with assistance of subject-matter experts, administering the inventories to job incumbents, and analyzing the task data.

The task inventories were developed by the project staff with three SME panels, one panel for each job specialty. The SMEs were hydropower plant personnel selected by Corps of Engineers headquarters to ensure that they represent the diversity of the total workforce, which was a relatively small -- a total of 74 workers. Each inventory had two major parts:

> **Tools and equipment** - A list of tools and equipment used in the occupation was presented. Respondents indicated how frequently they used each tool or piece equipment.
>
> **Job tasks** - A list or an inventory of job tasks in each occupation, organized under major job duties, was presented.

The number of tasks and tools identified by the SMEs for each occupation is indicated in Table 1.

Table 1

The Number of Tasks and Tools by Job Specialty

Occupation	Number of Tasks	Number of tools
Electrician	177	91
Mechanic	199	95
Shift operator	245	30

For each task respondents indicated:

1. Whether they performed the task (yes or no)
2. How much time they spent on it (5-point scale)
3. How important it was to their job (5-point scale)
3. How difficult it was to learn (5-point scale)
4. Whether they learned it in the classroom (yes or no)
5. Whether they learned it on the job (yes or no)

For each tool the respondents indicated on a four-point scale how frequently it was used.

Following the analysis of the survey data the OPM project staff met with the three SME panels for the second time to review the results and to:

a. identify the tasks and tools of the job that require training;

b. link the tasks and tools of the job to the subjects taught in the training program;

c. specify the relative emphasis to be given to each subject in classroom instruction and on-the-job training.

Survey administration and response rate

The task inventories were distributed to all electricians, mechanics, and operators in the District. Respondents were given prepaid business reply envelopes for the return of the inventories directly to OPM without the possibility of review by supervisors. A total of 67 subjects, or 91 percent of the total population of skilled workers, (26 operators, 21 electricians, and 20 mechanics) returned completed inventories.

Data analysis

The questionnaire responses were analyzed separately for the three occupations in the following manner:

c. The percentages of responses were calculated to evaluate the frequency of use of various tools and equipment in the occupation.

d. Six statistics were computed for each job task: percent performing the task, the mean time spent, mean relative importance, mean learning difficulty, and the percent indicating whether the task was learned in the classroom, on the job, or both.

RESULTS

Tasks and tools

The job analysis questionnaire results verified the performance and use of tasks and tools as identified by the SME panels. All of the tasks and tools listed in the questionnaires were reported to be performed or used in varying degrees by at least some of the job incumbents. As can be seen from Table 2, virtually all the tasks were reported to be performed by at least 50 percent of the workers. This result suggests that the three occupations are fairly homogeneous, with little variation in the way work is done within an occupation. Thus, from a training perspective, the issue was how to best train all tasks and how much emphasis to give them in training, rather than which tasks to train.

Table 2

The Number and Percent of Tasks Performed by at least half of the Workers in each Job Specialty

Occupation	Number of Tasks	Percent of Total
Electrician	175	99
Mechanic	187	94
Shift operator	237	97

After reviewing the statistical results of the task analysis, the SME panels concluded that training is required for virtually all tasks, even those tasks performed relatively infrequently or only at one or two power plants. They reasoned that journeyworkers must have training for all tasks because they often move from one plant to another, where they may encounter infrequently performed tasks.

Linking job requirements to the training program

The training program that was being examined in this study was documented in the form of an outline that detailed the subject matter to be learned during each three- or six-month period of the program.

Using the training program outline and the task analysis results, the project staff and the SMEs identified the tasks and tools that were being taught under each topic or subject in the training program. The panels determined which tools and tasks support each subject in training and whether training is provided for each task of the job. This linkage procedure, which constituted the validation of the training program, resulted in a listing that showed which tasks fall under each subject matter taught in the training program. Thus, with a few exceptions, all tasks were linked to one or more training topics and all training topics linked to one or more tasks.

To ensure that both the training program and the principal training measure of the program, the oral examination, reflect the relative importance of the components of the job, the SME panels reviewed the weights assigned to the various subject matter areas on the oral examination (these weights were assigned by an SME panel in a prior study) and specified the number of hours to be allocated to each subject area in classroom instruction and on-the-job training. In allocating these hours the panels considered the total number of hours available for training, the breadth and difficulty of each subject, and the extent a subject can be best learned in the classroom, on-the-job, or in independent study. The procedure for allocating the time for each subject area was based on pooled SME judgments, not a rigid formula. SMEs considered the task analysis data as well as their own knowledge of the subject in arriving at their judgments.

The SME panels noted that their recommendations for training time for each subject area as they applied to on-the-job training are approximate hours, not rigid guidelines. They pointed out that the schedule of the actual work in a power plant often governs the assignments that can be given to a trainee and, consequently, it may be difficult to follow the guidelines rigidly. Nevertheless, when both trainees and supervisors are aware of the hours specified in the training plan, they can map out the work assignments to support the training plan. Doing so will increase the proximity of time between classroom and on-the-job training on the same subject and therefore improve the effectiveness of training.

SUMMARY

Using task analysis data, this study compared the content of a training program for skilled hydroelectric power workers to the content of the job. The results indicated that the subject matter taught in the training program was related to the tasks performed and tools used on the job. The content of the oral examination, which is the principal evaluation technique used in the training program, was also found to reflect the content of the job and the training program. The number of hours that should be allocated to each subject area in the classroom and on-the-job training were also determined in this study.

REFERENCE

Equal Employment Opportunity Commission; Civil Service Commission; Department of Labor; and Department of Justice. Uniform guidelines on employee selection procedures (1978). Federal Register, 43(166), 38290-38315.

A - 815

Improving the Efficiency and Effectiveness of Training

Michele M. Morales
Mark S. Teachout
U.S. Air Force Armstrong Laboratory

J. Kevin Ford
Douglas Sego
Michigan State University

Introduction

Some issues the Air Force is now facing include fewer resources and increasing job complexity. Thus, improving the quality of Air Force training has never been more critical. The purpose of this paper is to describe an ongoing effort to improve the efficiency and effectiveness of Air Force training.

Background

Three important purposes of a training evaluation system are content validity, training efficiency and training effectiveness. Content validity ensures that the content of training is relevant to the job. This can be accomplished through a direct comparison of tasks performed on the job with the tasks taught in the training program. The greater the overlap between the job and training domains, the greater the content validity of the training program.

While a program may be training the appropriate tasks, that is, the training is content valid, the amount of emphasis placed on those tasks may not match their "need" for training as indicated by job content information. Training efficiency ensures that the appropriate amount of emphasis is placed on each task taught in training. Training efficiency can be determined through a direct comparison, or matching, of the emphasis placed on tasks during training, with the training needs identified from the job domain. The greater the match between job and training emphasis, the greater the efficiency of the training program.

Finally, training effectiveness determines whether trainees learned the material that was taught in training. This requires information about the performance of trainees at the end of the training program. The greater the amount of learning that has occurred, compared to a specified standard, or criterion for success, the greater the effectiveness of the training program.

In summary, content validity, training efficiency and training effectiveness information are important components of a training evaluation system. Each component contributes information that can be used to assess and change the content, emphasis, and conduct of a training program.

Methods, Procedure and Results

Research Context

The current effort integrates training efficiency and effectiveness information to facilitate changes in training course content and emphasis. A Training Efficiency and Effectiveness Model (TEEM) was developed to improve the quality of technical training by utilizing evaluation information at the task level. The technical training program examined was the Aerospace Ground Equipment (AGE) Airman Basic-in-Residence (ABR) course at Chanute AFB, IL. The course consisted of 18 weeks of instruction regarding equipment used to support aircraft.

Content Validity

Initially, the content validity of the program was assessed to ensure that the tasks taught in training were the same as those performed on the job. While the job content domain is specified at the task level by the Occupational Survey Report (OSR), the ABR training content domain is identified at the task cluster level by the Specialty Training Standard (STS). Therefore, it was not possible to compare the two domains at the task level of specificity. Consequently, we developed a questionnaire that asked ABR course instructors to link OSR tasks to the course Plan of Instruction to determine which OSR tasks were taught in training. These instructors were asked to identify OSR tasks taught in the training program, their corresponding instruction block, and if previously taught, the reason for their deletion from the training program. Additional instructors and course supervisors verified their responses. Fourteen blocks of instruction and 152 OSR tasks were found to compose the AGE ABR course.

Training Efficiency

Training efficiency was examined to determine if the amount of emphasis placed on tasks in the training program matched the amount of emphasis recommended from OSR data. Ford and Wroten's (1984) Matching Technique was used to link the emphasis information from the job and training content domains to determine training efficiency. Figure 1 displays conceptually the Matching Technique that compares actual training emphasis from the training domain with recommended training emphasis from the job domain. Thus, potential training excesses, deficiencies, and "hits" could be determined. These were indicative of tasks that were potentially over-trained, under-trained tasks and matched tasks, respectively.

Actual emphasis was determined by course instructors who estimated instruction time per task. Some of the tasks taught were matched with the Plan of Instruction training objectives which listed time spent on each objective. For the remaining tasks, instructors completed a questionnaire to link those tasks to the POI. They also estimated the amount of time spent training each task. Recommended training emphasis measures were ratings collected from the field for the Occupational Survey Report.

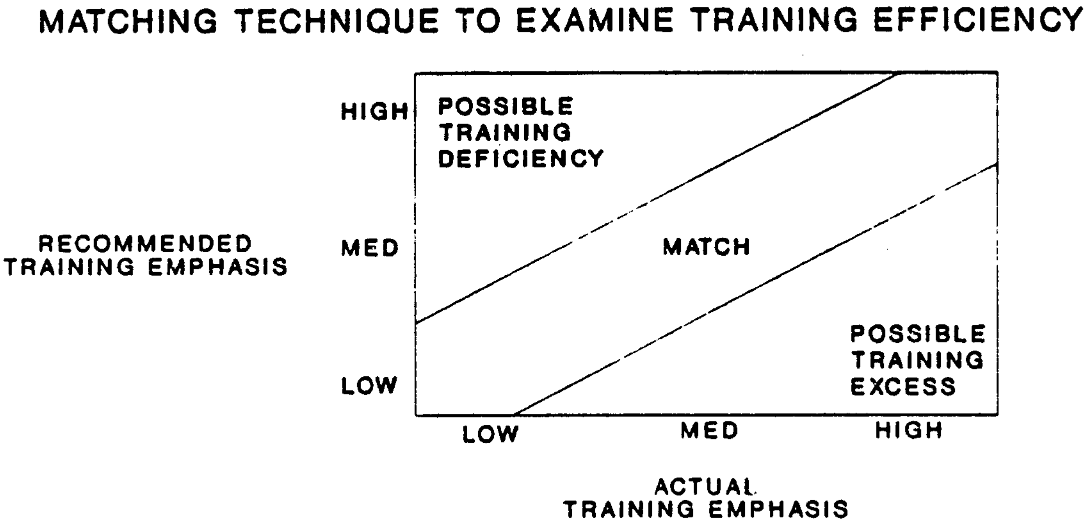

MATCHING TECHNIQUE TO EXAMINE TRAINING EFFICIENCY
RECOMMENDED TRAINING EMPHASIS
HIGH
MED
LOW
POSSIBLE TRAINING DEFICIENCY
MATCH
POSSIBLE TRAINING EXCESS
LOW
MED
HIGH
ACTUAL TRAINING EMPHASIS

Figure 1: The Matching Technique

Results were found for each of the three potential outcomes. A task given a high emphasis in the schoolhouse and a low recommended emphasis rating from the field was identified as a potentially over-trained task. A task with a low emphasis in the schoolhouse and a high recommended emphasis rating from the field was identified as a potentially under-trained task. Tasks with approximately equal schoolhouse and field ratings were considered a match.

TRAINING EFFECTIVENESS

Training effectiveness was examined by developing an end of course knowledge test to determine whether trainees learned the material that was taught in training. Multiple-choice knowledge test items were developed for 19 tasks sampled from the training course content. These tasks were chosen to be representative of tasks performed on the different types of AGE equipment used in the training course. The knowledge information was then integrated with the training efficiency results obtained from the Matching Technique, as suggested by Ford and Sego (1990). This Training Effectiveness and Efficiency Model (TEEM) is displayed in Figure 2. The performance level of both over- and under-trained tasks is depicted in this model. In this study, knowledge data was used to indicate performance level. For purposes of this example performance is dichotomized as "performed well" or "not performed well".

		JOB PERFORMANCE LEVEL	
		NOT PERFORMED WELL	PERFORMED WELL
MATCHING TECHNIQUE RESULTS	TRAINING EXCESSES	ELIMINATE FROM TRAINING FIND OTHER OPTIONS	REDUCE/MAINTAIN TRAINING EMPHASIS
	TRAINING MATCHES	INCREASE TRAINING EMPHASIS FIND OTHER OPTIONS	MAINTAIN CURRENT EMPHASIS
	TRAINING DEFICIENCIES	INCREASE TRAINING EMPHASIS	MAINTAIN CURRENT EMPHASIS

Figure 2. Training Effectiveness and Efficiency Model

These results can be utilized to facilitate training course changes. For example, training time might be reduced for over-trained tasks that were performed well, while training time might be increased for under-trained tasks that were performed poorly.

Discussion and Summary

Results obtained from integrating performance and knowledge data with the Matching Technique results showed tasks in all domains of the conceptual model. It appeared that regardless of the Matching Technique's results, little or no action was necessary when task performance was good. However, if performance was poor, training emphasis could be increased, the task could be taught on the job, or performance could be enhanced through the use of job aids.

This methodology is programmed on IBM compatible software to present this information at the task level, in meaningful formats to the user. For example, task information can be clustered by equipment or function and tasks can be ordered by training emphasis and Automated Training Indicator ratings. The TEEM methodology is a simple approach that expedites training course revisions by linking training content, training efficiency and training effectiveness information.

REFERENCES

Ford, J.K. and Wroten, S.P. (1984). Introducing new methods for conducting training evaluation and for linking training evaluation to program redesign. Personnel Psychology, 37, 651-665.

Ford, J.K. and Sego, D. (1990). Linking training evaluation to training needs assessment: A conceptual model. Air Force Human Resources Laboratory Training Systems Division. AFHRL-TP-90-69.

Results From A Comparison of Intel and Scout Activities, and Unit Performance at the National Training Center

Brian J. Bush
U.S. Army Research Institute Field Unit
Presidio of Monterey, California

Background

This paper discusses a study of unit performance at the National Training Center (NTC). The NTC is one of three major Army combat training centers (CTCs), two in the United States and one overseas. The CTCs developed to provide intensive and realistic combined arms and interservices training to increase unit readiness, develop leaders, standardize doctrine, and provide feedback to the Army and other participants.

The NTC is characterized by an area large enough for force-on-force maneuver, live fire training, air space for joint operations, deployment training, instrumentation to document actual events, a dedicated opposing force (OPFOR), observers/controllers, and detailed after action reviews.

The intent of the NTC experience is to provide training enhancement by 'stretching' the visiting unit to permit identification of strengths and weaknesses. This focusing facilitates a rotational unit's opportunity to develop it's capabilities.

The study focuses on the Intel Battlefield Operating System (BOS), which includes scout activites, and unit performance at the National Training Center (NTC). The information on the Intel BOS and unit performance is derived from Observer/Controller (O/C) comments found in Take Home Packages (THPs) provided to units at their home station. The THPs include a by mission critique of their NTC training with a lessons' learned section. The O/Cs are Army officers assigned to the BOS for which they are subject matter experts and have school training and field experience.

The THPs are part of the Combined Training Center's (CTC's) archives maintained by the ARI-POM Field Unit under the guidance of the Combined Arms Command - Training (CAC-T) located at FT. Leavenworth.

The purpose of the study is to provide feedback to units, commanders and trainers, and schools to improve field and garrison training.

Method

The author surveyed two-hundred and sixty-six TF missions conducted between Feb. 1986 and Jan. 1990. There are generally 6 or 7 task force missions per rotation. The intel and scout sections of NTC Take Home Packages THPs) were reviewed to identify the activities conducted by mission, and determine the range of comments describing the conduct of the activities.

Based upon O/C comments, the performance of each activity was rated on a scale of 1 thru 5. A (1) = an O/C reported activity as should have been done but not attempted; a (2) = attempted but done poorly; a (3) = done adequately or mentioned as done without a positive or negative comment; a (4) = done well; and a (5) = done very well.

The maneuver section was used for O/C reported comments relating to mission success of the TF. The range of scores and comments were: (1) = the TF was defeated in detail, or was unable to clear an area or reach it's objective. (2) = the O/C comments were inadequate to determine mission success or failure. (3) = the TF was able to secure the objective, or the TF caused the Opposing Force (OPFOR) to commit it's reserves.

Findings

Activity performance: The average performance rating for ninety-nine intel activities surveyed was $\underline{M} = 2.24$, and scout activities, $\underline{M} = 2.25$.

The ten highest (plus ties) and lowest rated intel activities with means ( ) and number of missions in which the activity is mentioned (out of a possible 266 missions) are as follows:

Highest			Lowest		
Oral OPORD brief	(3.00)	4	plan R & S routes	(1.79)	14
brigade update	(2.76)	66	enemy cbt multipliers	(1.78)	9
id enemy positions	(2.76)	4	conduct battle handoff	(1.75)	8
use of PWs for intel	(2.75)	4	enemy air threat	(1.67)	24
intel estimate	(2.73)	110	DST execution	(1.63)	8
intel brief overall	(2.71)	98	enemy FASCAM threat	(1.60)	10
situation template	(2.71)	189	plan retrans support	(1.57)	7
terrain analysis	(2.66)	178	control measures	(1.50)	4
friendly aves of app	(2.60)	77	DST use by TF	(1.33)	3
IPB planning	(2.56)	39	R & S checkpoints	(1.00)	2
id enemy composition	(2.56)	32			

The overall highest and lowest (plus ties) rated scout activities with means ( ) and frequencies are:

Highest			Lowest		
platoon initiative	(3.05)	21	scout training	(2.00)	12
TF FRAGO	(2.82)	22	scout fire plan	(2.00)	14
pre-combat checks	(2.71)	86	sleep plan	(2.00)	89
spot reports	(2.68)	136	medical plan	(2.00)	6
tracking the enemy	(2.62)	157	route recon	(1.96)	23
finding the enemy	(2.59)	134	target handoff	(1.94)	17
scout WARNO	(2.58)	26	effect of scouts kia	(1.88)	49
weapons check	(2.56)	16	vehicle check	(1.86)	7
plt ldr effectiveness	(2.55)	42	scout rehearsal	(1.84)	50
OP positions	(2.53)	120	reconstitution of kia	(1.71)	7
			plt ldr recon	(1.57)	45

Mission success: The following activities were significantly related to mission success. Next to the level of significance is the number of missions in which the activity was mentioned.

Intel Activities

Chi Square			Kendall' Tau		
situation template	(p<.01)	189	situation template	(p<.01)	189
spot reports	(p<.01)	136	spot reports	(p<.01)	136
battle damage asses.	(p<.01)	149	battle damage asses.	(p<.01)	149
local security	(p=.01)	31	local security	(p<.01)	31
intel integration	(p<.02)	79	intel integration	(p<.01)	79
R & S execution	(p<.04)	94	R & S briefing	(p=.01)	23
			R & S execution	(p<.02)	94
			aves of approach	(p<.04)	97
			S2 commo w/recon	(p<.05)	42
			counter-recon plan	(p=.05)	53

Scout Activities

Chi Square			Kendall's Tau		
actions on contact	(p<.01)	72	mission effect.	(p<.01)	112
mission effect.	(p<.02)	112	actions on contact	(p<.01)	72
			knew friendly sit.	(p<.04)	96
			commo link external	(p<.05)	83
			TF OPORD	(p<.05)	93

Interpretation of the performance of the activities must be done within the context of the National Training Center. Therefore, the scale values should be interpreted as relative scores and not absolute score values.

Discussion

Activity performance: The averages for intel (M = 2.24) and scout activities (M = 2.25) are below the scale midpoint of 2.50. This tendency to have more negative than positive remarks is reasonable when considering that the O/Cs try to point out deficencies in performance so that the TF can correct them to enhance success on the next mission. The trend to be somewhat more critical is equally consistent for both intel and scout O/Cs.

The primary intel activities, those with several subordinate activities (the intel estimate, intel briefing, situation template, terrain analysis, IPB planning, and DST development) generally received higher ratings than secondary or subordinate activities (plan R & S routes, identify enemy combat multipliers, identify enemy air threat, identify enemy FASCAM threat, plan retrans support, develop control measures, provide checkpoints, and execution and use of the DST). The exception is the Decision Support Template (DST).

In contrast, most of the subordinate scout activities were rated higher than the primary activities. An exception to this tendency was the TF FRAGO submitted to the scouts. Like many of the primary intel activities, it received a higher rating than it's subordinate activities.

Mission success: Ten of the intel activities were significantly correlated with mission success. Four were a direct outcome of scout performance (spot reports, local security, R & S execution, communications). Five were attributed to the S2 and S2 section (the situation template, intel integration, R & S briefing, avenues of approach, and the counter-recon plan. The tenth was attributed to the line units during contact with the OPFOR (battle damage assessments).

Five scout activities were significantly related to mission success. The rating for scout mission effectiveness is the most representative of the relationship between overall scout performance and TF mission success. Actions on contact with the enemy suggests the importance of the scouts not becoming decisively engaged and compromising their mission. Establishing and maintaining commo with the S2/TOC is imperative for the dissemination of spot reports. And knowing the friendly situation impacts on scout fratricides and their survivability on the battlefield.

The omission of any intel and scout activities from the findings in regard to mission success does not necessarily imply that other activities did not effect mission success. In several cases, the sample size of activities was too small for analysis with mission success.

THE MEASUREMENT AND EVALUATION OF MENTAL WORKLOAD: PROBLEM, PROGRESS, AND PROMISE

Richard E. Christ, Ph.D.
Army Research Institute Field Unit - Fort Bliss, TX

INTRODUCTION

The concept of operator workload (OWL). Both physical and mental work depend not only on the particular task to be accomplished, but also upon the availability of the internal resources required of the operator to perform the task. Thus, operator workload (OWL) is defined in terms of the interaction between the work imposed on an operator by a task and the operator's capacity to perform that work. (For a discussion of the conceptual foundations of workload see Lysaght et al., 1989.)

The OWL program objectives and accomplishments. There has been considerable research concerned with workload, the majority conducted in laboratory settings. Of the applied research, most has been associated with aviation systems. To fill the void, the U.S. Army Research Institute (ARI) sponsored a three-year exploratory development effort called the OWL Program. The challenge of the OWL Program was to apply and validate the most relevant of the workload measurement techniques and use the results to formulate practical guidance. To meet this challenge, five objectives were established for the OWL Program. These objectives and the major research product produced when each was achieved are given below.

1. Determine the current status of OWL in the Army. Hill, Lysaght, et al. (1987) present the results of a review of Army and Defense Department requirements documents and an analysis of interviews with prospective users of the guidance which was to be produced by the OWL Program.

2. Identify and evaluate the techniques and methodologies currently available for the assessment of OWL. Lysaght et al. (1989) document the results of a comprehensive review and evaluation of the concept of workload and methods for its assessment.

3. Select and apply the most promising OWL assessment techniques to several Army systems. Hill, Iavecchia, et al. (1992) directly compare four workload techniques in a series of eight separate studies across three Army systems. Hill, Byers, et al. (1992) present and discuss the results obtained from these studies and other studies in terms of their meaningfulness or validity for a number of different practical topic areas.

4. Use the results of achieving Objective 2 and Objective 3 to synthesize guidance as to which OWL techniques should be used for a given system at a given stage in development. Harris, Hill, Lysaght, and Christ (1992) describe the rationale, capabilities, and features of the Operator Workload Knowledge-based Expert System Tool (OWLKNEST).

5. Synthesize overall lessons learned from the OWL Program and provide the managers of Army systems what they need to know about OWL. Christ, Hill, Bulger, and Zaklad (1990) prepared a pamphlet for the managers of Army systems that describes the need and some procedures for insuring that OWL issues and concepts are incorporated into the materiel acquisition process.

The purpose of this report. In addition to describing and identifying the major accomplishments of the overall ARI OWL program, this report has two major goals: (a) to highlight some examples of experimental and analytical work conducted during the pursue of the application and validation of workload assessment techniques and (b) to identify several promising and important areas for future research. The remainder of this report is organized around these two goals.

OVERVIEW OF THE PURPOSES, METHODS, AND RESULTS OF THE OWL PROGRAM PRIMARY RESEARCH STUDIES

General purpose of the OWL studies. A major purpose of the OWL Program primary research studies was to evaluate the **applicability and validity** of workload assessment techniques for Army systems. The concept of applicability is based upon very practical issues such as how many resources are required to employ a technique and operator acceptance. The concept of validity must be examined as a multi-dimensional continuum concerned with the "degree of reality" that can be demonstrated for workload measurement techniques in various situations. Our approach to application and validation of a workload assessment technique was to seek and utilize any and all information that relates to the "meaningfulness" or operational reality of the OWL technique in question. The goal was to gather all this partial and uncertain information and put it together in a meaningful way.

General methods used for the OWL studies. A variety of OWL assessment techniques are available and most have been described in previous publications (e.g., Lysaght et al., 1989). These OWL assessment methods may be partitioned into two categories. The **empirical techniques** involve the assessment of workload while the operator is actually operating a simulator, prototype, or representative system, i.e., workload is assessed with the operator-in-the-loop. **Analytical or predictive** techniques, in contrast, may be applied early in the system design process, without an operator in-the-loop. The empirical techniques include those methods which measure the operator's performance, physiological responses, and reports of subjective experiences. The analytical techniques estimate workload through the methods of expert opinion, comparability analysis, task analysis, and simulation.

Based on our research review, we selected four different empirical techniques to use in our studies. They are: Task Load Index (TLX) (Hart & Staveland, 1987), Subjective Workload Assessment Technique (SWAT) (Reid, Shingledccker, & Eggemeier, 1981), Modified Cooper-Harper (MCH) scale (Wierwille & Casali, 1983), and Overall Workload (OW) (Vidulich & Tsang, 1987). Two of the scales (MCH and OW) are **unidimensional**, i.e., produce only an estimate of overall or global workload. The other two scales (TLX and SWAT) are **multidimensional**, i.e., provide information on the various components or sources of workload, as well as an estimate of global workload.

Based on our review of workload assessment methodologies, we selected different analytical techniques to use in two of the primary research studies: an expert opinion technique based upon the prospective application of the TLX method (Pro-TLX), and the task analytic and simulation methods incorporated in the Task Analysis/Workload (TAWL) and the TAWL Operating System Simulation (TOSS) methods (Bierbaum, Fulford, & Hamilton, 1990).

Systems used for the OWL studies. Five U.S. Army systems were selected for study: the Aquila Remotely Piloted Vehicle (RPV), the Line-of-Sight-Forward-Heavy (LOS-F-H) mobile air defense system, the UH-60 Blackhawk helicopter, the LOS-Rear Pedistal Mounted Stinger (PMS) mobile air defense system, and the Stingray system mounted on a Bradley Fighting Vehicle. Five separate studies were conducted using the LOS-F-H system, two with the Aquila RPV, and one each with the other three systems. With the exception of the UH-60 Blackhawk study, these workload studies were part of previously scheduled U.S. Army field tests or field exercises. The UH-60 Blackhawk study was conducted exclusively to assess workload factors using the UH-60 2B38 flight simulator. While each of the 10 studies had specific objectives, efforts were made to conduct the tests and data collection efforts in a similar manner so that comparisons could be made across studies.

Direct comparison of the empirical workload assessment techniques. For five of the studies, the four operating rating techniques -- TLX, SWAT, MCH, and OW -- were directly compared with one another along four dimensions: factor validity, operator acceptance, resource requirements, and special procedures (Hill, Iavecchia, el al., 1992). Principal component analysis (PCA) was conducted on all possible sets of workload measures within each study. Across all the studies, this analysis revealed a single component variable, hereafter termed the **OWL Factor**, which explained between 71 and 83 percent of the total variance in the data. Table 1 summarizes the results of the analysis of factor validities for all four rating scale techniques in each study for which the comparisons can be made. The table presents, for each study, the ordered mean factor loadings. Based on these and the other results of the direct comparison among rating scale techniques, it was concluded that the TLX technique is generally the preferred workload rating scale for all but screening applications, where it may be appropriate to use the OW technique.

Table 1
Factor Validity Scores Across Studies

STUDY	TECHNIQUE (Mean Factor Loading)			
LOS-F-H MOICE	TLX(.935)	OW(.927)	MCH(.862)	SWAT(.860)
LOS-F-H Generic	TL"(.926)	OW(.905)	MCH(.904)	SWAT(.778)
LOS-F-H Basic	TLX(.926)	SWAT(.900)	OW(.898)	MCH(.818)
Aquila FDTE	TLX(.910)	SWAT(.893)	OW(.869)	MCH(.833)
UH60A Simulator	TLX(.899)	OW(.872)	SWAT(.805)	MCH(.799)

Workload ratings and system performance. For the LOS-F-H and Stingray system, step-wise regression analyses were conducted to examine the relationship between operator workload and system performance. In two of the LOS-F-H studies the dependent variable was a measure of system performance scores (based on the number of targets destroyed during engagement opportunities) and for the Stingray system it was a measure of force effectiveness. In all three of these studies, the multiple correlations were significant: R = -0.66 and -0.65 in the two LOS-F-H studies, and R = -0.12 and -0.35 for the defensive and offensive Stingray missions, respectively. These results show that increases in operator workload ratings were associated with decreases in system performance. Hence, it is possible to demonstrate a meaningful quantitative relationship between workload ratings and system performance, even up to several months following the events to be rated. However, the presence of this relationship will depend upon the procedures used to measure both variables.

Sensitivity to expected variations in imposed OWL. The sensitivity of workload ratings to imposed workload was established in all but one of the ten OWL studies. For example, the results of one of the Aquila RPV studies revealed a significant interaction between mission segment and crew position, as illustrated in Figure 1. It may be seen that while the mission commander (MC) has the highest and relatively consistent OWL factor scores, the workload ratings of the other two crew members vary considerably and in opposite directions over segments. These results make sense. The workload of the MC is driven by a fairly constant level of responsibility over an entire flight of the RPV. The MPO who operates the mission payload has no direct responsibility during launch and recovery when the mission payload is not in use but higher than average workload during the flight when the mission payload is used to perform mission essential functions. On the other hand, the AVO who operates the air vehicle has the least workload in the flight segment of an RPV mission when flight operations are relatively routine but higher than average workload during launch and especially during recovery when various problems can and often do arise.

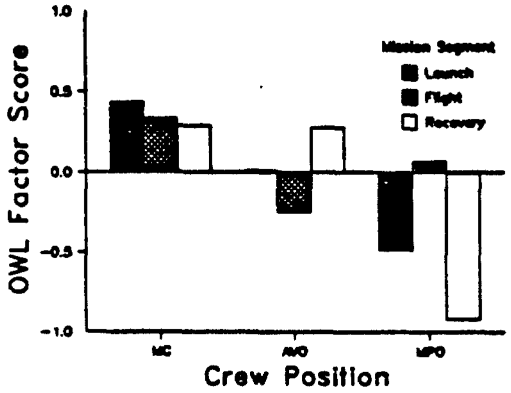


Figure 1. *The effect of mission segment and crew member position on workload in the Aquila FIREX 88 study.*

One of the goals of the OWL Program was to investigate how workload changes over an extended period of time. Figure 2 shows the mean workload rating of each of two crews as a function of time into their respective 48-hour missions. It may be seen that workload ratings generally increase across time for both crews. Since task demands were relatively constant over the duration of the 48-hour mission, the increase in workload over time may reflect a decrease in the resources the system operators have to commit to mission essential tasks.

Diagnosticity of multidimensional techniques. Diagnosticity refers to the extent to which a specific source or cause of workload is revealed by the measurement technique. The OWL Program focused on an analysis of the TLX subscales. The TLX subscales are: mental demand, physical

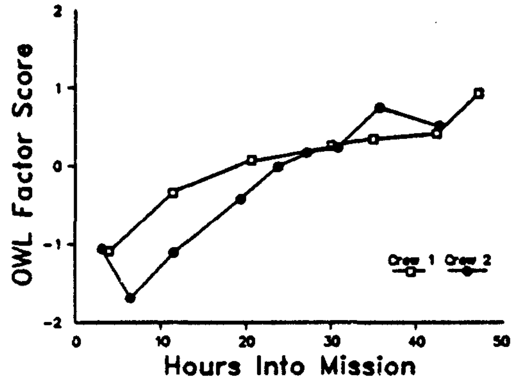


Figure 2. *The effect of an extended duration mission on workload in the LOS-F-H FDTE 48-hour mission study.*

demand, temporal demand, performance, effort, and frustration. An analysis of changes in the pattern of subscale values across key independent variables of a study can help to identify workload problems and their sources at a finer level of detail than can a global measure of workload. An example of a such an interaction effect was found for the UH-60A study and is illustrated in Figure 3. It may be seen that two subscales contribute to the higher mean workload observed for the LZ to PZ segment than for the other three segments shown; the LZ to PZ segment has larger effort and physical demand components than the other three segments. This result is reasonable since the LZ to PZ segment represents flying through hostile territory carrying a heavy load, a situation in which the platform is quite unstable.

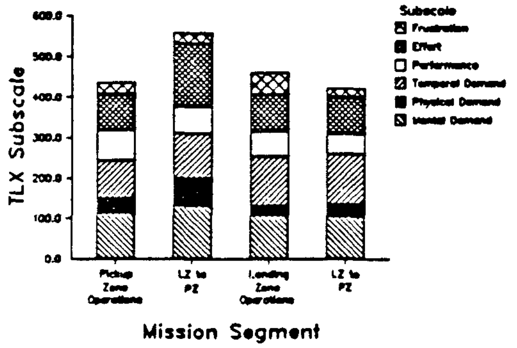


Figure 3. *The effect of mission segment and TLX subscale on weighted subscale scores in the UH-60 simulator study.*

Prospective TLX ratings. A major premise with regard to OWL measurement is the use of analytical or predictive workload assessment techniques allow the human factors analyst or practitioner to make meaningful contributions early in the design phase of an emerging system. The LOS-F-H Prospective study examined the workload ratings of operators for a variety of hypothetical situations. Figure 4 illustrates the prospective ratings associated with a more realistic configuration of several fire units. The "master fire unit" is the one with an active radar, which receives command and control data over an active radio network, and which determines the assignment of targets to fire units. The slave vehicle is responsible for engaging the assigned targets. Figure 4 illustrates a significant interaction of operation mode (Master, Slave, and Autonomous) and duty position (RO and EO). The overall workload of the RO and EO is rated about the same in the Autonomous Mode. However, the RO is projected to experience greater levels of workload than the EO in the Master Mode and the reverse is projected to occur in the Slave Mode.

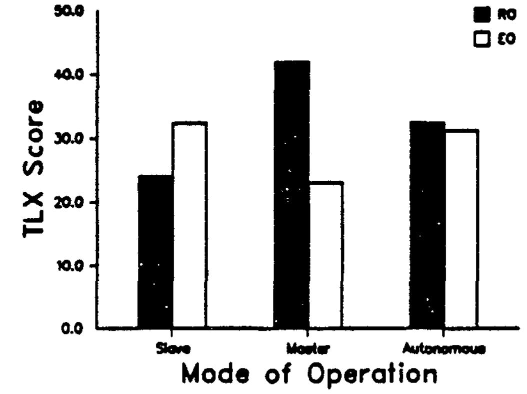


Figure 4. *The effect of proposed mode of operating multiple fire units and crew member position on TLX ratings in the LOS-F-H Prospective study.*

Analysis of the TAWL/TOSS Methodology. The validity of the TAWL/TOSS methods was analyzed for the UH-60 Blackhawk study. The approach used was to compare real-time operator ratings of workload with TAWL/TOSS-based predictions of workload. This technique proved to be quite reasonable. A significant correlation was found across crew members between TAWL\TOSS-

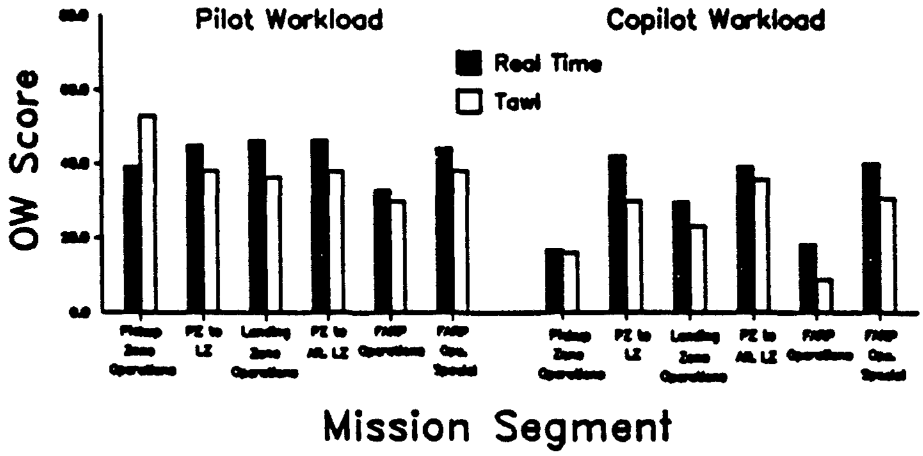


Figure 5. *The effect of mission segment and crew member position on real-time ratings and TAWL/TOSS model predictions of overall workload in the UH-60A simulator study.*

derived predictions of OW and the real-time OW ratings ($\underline{r}$ = 0.82). This high correlation suggest the validity of the underlying TAWL/TOSS data base and scenario generation techniques. Figure 5 illustrates this finding graphically by mission segment, separately for the pilot and copilot. As may be seen, TAWL/TOSS predictions track the relative overall workload between segments.

SUMMARY AND CONCLUSIONS OF THE OWL PROGRAM PRIMARY RESEARCH STUDIES

The empirical methods examined were four operator rating techniques: TLX, SWAT, OW, and MCH. In the studies reported, TLX was consistently highest in factor validity and operator acceptance. For these reasons, TLX is recommended for all but screening applications, where OW (because of its simplicity and convenience) may be used as a first step. The empirical workload ratings are shown to be sensitive to changes in system performance and in the expected levels of workload imposed upon the operator by the system, mission, and operational conditions. Additional analyses show that the TLX subscale ratings contain potentially useful information concerning the source or cause of experienced workload.

The analytical methods studied were prospective operator ratings using the TLX scale and the TAWL/TOSS task analytic and simulation model. The prospective rating technique shows promise as a method for identifying potential workload problems in emerging systems. The TAWL/TOSS model is shown to have a capability to track empirical workload ratings.

Future Research Directions. Based on accomplishments and lessons learned from the OWL Program and from other related research programs, several area for future work can be described. These include continuing work to generally improve our understanding of the concept of workload and its relationship to operator and system performance. However, it must also be recognized that the concept of workload is a critical variable in other domains of applied behavior and social science. Three areas of research which could benefit from a new or renewed consideration of workload issues and concerns are presented in succeeding paragraphs.

o Methods need to be developed to improve our ability to assess, understand, and utilize differences among individual soldiers in their reactions to workload extremes. It is generally understood that individual differences exist, but there is little research to relate them to workload.

o Alternative methods need to be developed for actually decreasing the extremes in workload imposed upon soldiers. Traditionally, these methods are based on the design and development of hardware or software systems and their interface with the soldier. Two alternative methods would address (a) the design and structure of the organizational unit within which the soldier/system is located and (b) the operational tactics, techniques, and procedures used during employment of the soldier/system.

o Methods are needed for increasing the soldiers' capability to successfully cope with extremes in operator workload. These methods may draw upon: (a) the identification, selection, and classification of soldiers whose performance is relatively tolerant to workload extremes, or (b) the design and implementation of training programs to develop effective individual and unit-level workload management strategies. In the latter case, there has been little meaningful research on the interaction of workload and training.

The Fort Bliss Field Unit of ARI has initiated work on two of these "new" areas of interest. We have launched a multi-year research program to develop methods for assessing and utilizing workload-related factors to improve the force design process. Our first area of concern in this regard is to address the problem of how to optimize the distribution of command and control across echelons in a combat environment; a problem related to the concept of span of control, clearly a workload-related matter. We have also initiated a program that will examine the impact of workload-related issues on collective multi-echelon unit training. We believe this latter work may facilitate the design of more effective training strategies and provide a basis for assessing the success of alternative training strategies.

ACKNOWLEDGMENT

This work was supported by the U.S. Army Research Institute under Comtract No. MDA903-86-C-0384. The views, opinions, and findings contained in this paper are those of the author and his colleagues, and should not be construed as an Official Department of Army position, policy, or decision.

REFERENCES

Bierbaum, C. R., Fulford, L. A., & Hamilton, D. B. (1990). Task Analysis/Workload (TAWL) user's guide - Version 3.0 (Research Product 90-15). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Christ, R. E., Bulger, J. P., Hill, S. G., & Zaklad, A. L. (1990). Incorporating operator workload issues and concerns into the system acquisition process: A pamphlet for army managers (ARI Research Product 90-30). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Harris, R. M., Hill, S. G., Lysaght, R. J., & Christ, R.E. (1992). Operator workload knowledge-based expert system tool (OWLKNEST) and an accompanying Handbook for operating the OWLKNEST technology (HOOT) (Research Note 92-49). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hart, S. G., & Staveland, L. E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Jr., Zaklad, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. Human Factors, 4, 429-439.

Hill, S. G., Byers, J. C., Iavecchia, H. P., Zaklad, A. L., Bittner, A. C., Jr., & Christ, R. E. (1992). Application and validation of workload assessment techniques (Technical Report - In preparation). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hill, S. G., Lysaght, R. J., Bittner, A. C., Jr., Bulger, J. P., Plamondon, B. D., Linton, P. M., & Dick, A. O. (1987). Operator workload (OWL) assessment program for the army: Results from requirements document review and user interview analysis (Technical Report 2075-2). Willow Grove, PA: Analytics, Inc.

Lysaght, R. J., Hill, S. G., Dick, A. 0., Plamondon, B. D., Linton, P. M, Wierwille, W. W., Zaklad, A. L., Bittner, A. C., Jr., & Wherry, R. J., Jr. (1989). Operator workload: Comprehensive review and evaluation of workload methodologies (Technical Report 851). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santà Monica, CA: Human Factors Society.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

Factors That Affect User Trust in Expert Systems

John L. Ward II, U.S. Army Natick RD&E Center
Larry L. Lesher, GEO-CENTERS, Inc.
Rhea Paniesin, Boston University

Sociological models of trust offer insight as to how trust is founded and nurtured between humans. However, with the proliferation of and reliance upon Artificial Intelligence systems, particularly in military administration, more understanding of how user trust affects the acceptance and use of intelligent computer systems is needed. The authors propose that system design features and demographic variables may help explain user confidence in such systems. An experiment was conducted with 87 soldiers at the U.S. Army Natick RD&E Center to determine if manipulation of two expert system variables (i.e. the Domain Expert and the System Developer) would influence soldiers' confidence in that system, and to determine which variable was more influential on user trust. Results indicate that subjects' confidence fluctuated as system variables were altered and that the Domain Expert had significantly more influence on user trust than did the system developer. This indicates that users' trust in intelligent computer systems can be manipulated by providing more specific system information to the user. Although demographic survey data showed a positive correlation between subjects' formal education and their reported comfort with computers, and between computer experience and such comfort, no variables were found to significantly influence the direction and magnitude of reported user confidence. A precedent has been established to further evaluate design strategies and demographic variables that are related to user trust, and to determine the criticality of user confidence in intelligent computer systems.

Expert systems are one of the most promising technological advances in recent years. The military is an environment that could receive enormous benefits from implementing these intelligent computer systems in critical areas like training, planning, logistics, battlefield operations and in weapons systems. Warfighting equipment is becoming extremely technologically sophisticated. Men and women who enter service today have much higher levels of computer skills than previous generations due to the proliferation of computers in everyday American leisure and education. Yet expert systems are rarely developed for or used by the majority of service members. One reason for this is that implementation of expert systems for service members must overcome not only those same human engineering problems that exist when expert systems are applied elsewhere, but they must also meet constraints imposed by military structure, culture and mission.

One feature that is essential to successful implementation of expert systems is user acceptance. There is reason to believe that the criteria for accepting an expert system differs from the criteria that determine acceptance of traditional computer programs because expert systems, in contrast, often *make* decisions verses simply organizing and computing data. This seems to add a new dimension to already existing problems in innovation acceptance, and we posit that the crux of this dimension is user trust and confidence. Throughout this paper the words trust and confidence are used interchangeably; they both refer to the expectations users hold that the expert system will provide a correct answer. Several researchers evaluating the acceptance (Mackie and Wylie, 1985; Reidel, 1988) and use (Aretz, 1987) of expert and decision support systems indicate that user confidence in the system is crucial, yet historically user trust has not been treated as a distinct variable upon which system acceptance is, at least partially, dependent. The focus of this paper is to develop a rationale for studying the trust that U.S. Army soldiers exhibit in expert systems, and to discuss an experimental approach for increasing user confidence.

Definitions and proposed sociological models of trust* between humans vary, and many constructs have been used in the attempt to illuminate critical aspects of trust that will enable us to better understand its etiology and nature. Many of these definitions and models (Barber, 1983; Rempel, Holmes and Zanna, 1985; Scanzoni, 1979; Deutsch, 1973; Rotter, 1980), although primarily oriented to interpersonal trust, are comprehensive and generic enough to be applied to the study of human trust in inanimate objects, processes or beliefs. Rempel, Holmes and Zanna (1985) offer the following definition of trust, one that is generic yet descriptive enough to encompass all of the characteristics of user trust in expert systems: "a generalized expectation related to the subjective probability an individual assigns to the occurance of some set of future events." Such background enables researchers in computer science and human factors to draw from existing models of trust and explore its component parts and its functions in relation to intelligent computer systems.

Expert systems used in military training or operations must not only be correct, they must be *perceived* as being correct and so trusted or they will not be used. Any benefits the military might derive from this technology is contingent upon soldiers using the expert system and thereby enhancing mission performance. If both acceptance and use of an expert system are so strongly affected by the amount of confidence a user has in the system, then research that focuses on increasing user confidence is needed to better leverage expert system technology. The following experiment was conducted to compare the relative effects of two expert system variables, the domain expert and the knowledge engineer, on soldier trust.

Methods

Subjects

Prior to the study, 135 soldiers stationed at the U.S. Army Natick Research, Development and Engineering Center completed a survey that provided pertinent demographic data as well as information on their life experience with and feelings toward computers. From this information, 120 subjects were selected to complete the study, but 33 of these subjects were transient and were reassigned prior to the study. The remaining 87 soldiers were blocked into two groups differentiated by two levels of computer experience, those with computer experience and computer novices. This determination was made following analysis of several questionnaire items dealing with the age at which they first used computers, total software packages used, familiarity with various computer environments and use of computers for work, education or leisure. The mean age of subjects was 27.5 yrs. (sd= 6.86) and the range in time in service was from 2.5 mos to 24.3 years (MD = 2.1 years). Their ranks ranged from E-1 (Private) to 0-5 (Lieutenant Colonel) with a total of 74 enlisted (85%) and 13 officers (15%) participating. There was an intentional use of this officer-enlisted ratio due to the lack of any expert system experimentation data with enlisted soldiers. The groups represented an extensive cross-section of (28) Military Occupational Specialties. Only one of the subjects had any prior knowledge of expert systems.

* For a comprehensive treatment of the origins and definitions of interpersonal trust the authors refer readers to The Logics and Limits of Trust, B. Barber (1983). An insightful review of several models of interpersonal trust, and an interpretation of how these might explain human trust in decision support systems, is found in B.M. Muir's paper Trust Between Humans and Machines, and the design of Decision Aids (see references).

Procedures

This study was conducted solely as a paper and pencil experiment. The experimental paradigm used was a 2x2x2 between subjects design. Treatment stimulus configurations are described in Table 1. The independent variables were the Domain Expert (DE), the Sytem Developer (DEV), and the level of computer experience of the groups. Two levels of the (DE) were created, one clearly superior (+DE) to the other (-DE), and two comparable levels of the System Developer (+ DEV, - DEV) were created. These descriptions were reviewed by three outside evaluators to ensure perceptual equality of the stimuli. Table 1 demonstrates the four possible pairs of stimuli which were combined to form vignettes about an expert system. The task of the expert system was foraging for wild edibles, taken from an actual field prototype that advises soldiers on what plants they can eat if in a survival situation. Subjects were given a six page packet that included an introduction to expert systems, clear definitions of each term or concept critical to the study, and instructions. They were given ample time to read the introductory and instruction pages and were encouraged to ask questions at any time during the study. The DE/DEV vignettes were then presented, randomly across the two user groups (High Computer Experience, Low Computer Experience), and subjects were asked to select the pair in which they had more confidence. They were then asked to indicate how much more confidence they had in the scenario that they selected by making a vertical mark on a 100 mm visual analogue scale anchored at either end with the numbers 1 and 100. The stimuus presentation order was counterbalanced to prevent extraneous ordering effects and the balanced design provided that all four DE/DEV pairs were presented an equal number of times to both groups. This presentation strategy created the conditions by which it would be possible to determine the effects on user confidence of providing normally unavailable information about two key expert system variables to the user.

TABLE 1: STIMULUS CONFIGURATIONS

		(DE)	
		POS	NEG
(DEV)	POS	+ +	- +
	NEG	+ -	- -

Results

A three-way ANOVA on the confidence ratings showed that there was no significant difference between the reported confidence of the High and Low Computer Experience groups. Thus the data from both groups were combined for the rest of the statistical analyses. Data were first analyzed to determine the rates of selection of the DE and DEV variables under each possible condition. A binomial test on conditions 1/2 (see Table 2) where the DE varied and the DEV was held constant showed that the selection of +DE over -DE was not due to chance (96.67%, p<.05). Similarly, in conditions 3/4, with the DE held constant and the DEV varied, a binomial test demonstrated that the

TABLE 2: EXPERIMENTAL CONDITIONS

DE	DEV	DE	DEV	CONDITIONS
+	+	-	+	(1)
+	-	-	-	(2)
+	+	+	-	(3)
-	+	-	-	(4)
+	+	-	-	(5)
+	-	-	+	(6)

selection of +DEV over -DEV was not random (79.31%, p<.05). However, the difference between 96.67% (rate at which the +DE was selected) and 79.31% (rate at which the +DEV was selected) was significant (X = 4.25, p<.05). In condition 5 subjects were simply comparing two positive levels of the DE and the DEV against the negative levels (+DE/+DEV vs -DE/-DEV). Here, too, the rate at which the positive stimuli were selected over the negative ones was statistically significant and not due to chance (92.86%, p<.05). Finally, in condition 6 neither variable was held constant, meaning subjects were comparing +DE/-DEV to -DE/+DEV. The scenario with +DE was selected a significant number of times over the scenario with the +DEV (85.71%, p<.05).

In order to assess the effect of manipulating the DE while keeping the DEV constant and vice versa, scores in conditions 1and 2 were grouped together, as were scores in conditions 3 and 4. A one-way ANOVA conducted on the magnitude of difference in confidence ratings between groups 1/2 and 3/4 was statistically significant (F = 10.9, df = 1, 57, p< .005) with the difference in confidence ratings being greater for the +DE condition than the +DEV condition. A two-tailed t-test comparing the condition 6 mean (32.00) of the +DE scores to 0 resulted in a significant difference between the ratings for the +DE and the +DEV variable (t = 2.81, p < .05); similarly, the means of the + DE scores from the conditions 1/2, 3/4, and 5 compared to 0 show a significant difference from the means of the +DEV (p < .05).

Linear regressions were performed on several questionnaire items to test for relationships between demographic variables and reported confidence. The correlation coefficients (Persons' first order correlations) indicated a strong relationship between reported comfort using computers and level of computer experience (r = .577, p< .001), and between overall positive experiences with computers and level of computer experience (r = .5644, p< .001). However, a multiple regression combining these variables, as well as others (education, age, rank, MOS) failed to predict expert system confidence ratings.

Discussion

Based on these analyses, we make the following points. The results showed that in all conditions the competence of the Domain Expert had a stronger positive effect on the reported confidence of the user than did the competence level of the System Developer. The first indication of this is the significant difference in the rate at which the +DE was selected over the +DEV in conditions 1/2 and 3/4. One might assume that holding DE constant while switching the DEV would have the same effect as holding DEV constant while switching the DE. However, the DE more frequently influenced the selection of +DE over +DEV. This reaction correlates with the magnitude of difference between confidence scores for conditions 1/2 and 3/4 in that the DE again pulled the actual values of the scores upward under conditions where the expected confidence ratings, if the independent variables had an equal effect, would fall within a close range. Additionally, in condition 6, when both levels of both independent variables systematically changed and created a much more complex decision for subjects, the +DE was clearly the variable that determined the rate of scenario selection and it significantly skewed the magnitude of the scores. These data reinforce each other and suggest that the Domain Expert has much more ability to raise the confidence level of a soldier than does the System Developer.

This experiment demonstrates that one expert system variable is more influential than another in influencing user trust in the system. It also demonstrates a method (i.e. providing influential information about the system to the user) that may be effectively applied with a number of variables to influence user perception of a novel decision-making system. Future research is needed to determine what other variables influence soldiers' trust in intelligent systems, and to systematically test these variables using the described methodology. Additionally, it is important that a metric be derived to determine baseline user confidence levels in an expert system in order to quantify increases in user trust. Then, true increases in soldiers' trust in intelligent systems can be quantified and directly attributed to system variables, like Domain Expert, and techniques that influence soldiers' trust, like the method reported here.

Acknowledgements

The authors would like to thank Mr. Rick Bell and Dr. Armand Cardello for their tireless efforts in support of this research.

References

Aretz, A.J. (1987). Dynamic Function Allocation in Fighter Cockpits. *Proceedings Of The Human Factors Society, 31st Annual Meeting*, (414-418).

Barber, B. (1983). *The Logics and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Deutsch, M. (1973). *The Resolution of Conflict: Constructive and Destructive Processes*. New Haven, CT: Yale University Press.

Mackie, R. R. and Wylie, C.D. (1985) *Technology Transfer and Artificial Intelligence: User Considerations in the Acceptance and Use of AI Decision Aids*. Technical Report for Naval Air Development Center. TR 51231-1.

Muir, B.M. (1987). Trust Between Huamns and Machines, and the Design of Decision Aids. *International Journal of Man-Machine Studies*, 27, (527-539).

Rempel, J.K., Holmes, J.G. & Zanna, M.P. (1985). Trust in Close Relationships. *Journal of Personality and Social Psychology*, 49, (95-112).

Rotter, J.B. (1980) Interpersonal Trust, Trustworthiness and Gullibility. *American Psychologist*, 35, 1-7.

Scanzoni, J. (1979) Social Exchange and Behavioral Interdependence. In: Burgess, R.L. and Huston, T.L., Eds. *Social Exchange in Developing Relationships*. New York: Academic Press.

WHAT ABOUT COMPUTER BASED ASSESSMENT?
a NATO-point of view

LtCol Drs. Paul B. Van Raay
Royal Dutch Army

1. INTRODUCTION

This paper presents a broad survey of computer-based assessment of military personnel from a NATO perspective. In the following chapter the reason why NATO is involved and the problems in how to cooperate are told. Chapter three deals with the different concepts in computer-based assessment. Chapter four follows with the implications for recruitment. Finally chapter five ends with the conclusion.

2. NATO

2.1 Why NATO ?

In general the armed forces in pratically all western countries are the biggest users of selection-systems. Moreover a considaribly number of military functions operate with highly advanced technical equipment. So due to the large number and the technical nature of the functions the computer was in a very early stage recognized as a big contribution to the selection of military personnel. This conclusion lead to the decision that computer-based assessment should be a NATO-topic.

Within NATO-PANEL-8, Research and Study Group-15 (RSG-15) was installed. It's main interest was the computer based assessment of military personnel. Participants were the USA, Germany, Great-Britain, Belgium, Canada, the Netherlands, France and Denmark. In this international forum one could more easily communicate, coordinate and exchange data, experiences and systems within NATO-countries. In fact RSG-15 has given us the possibility and the means to exchange testmaterial, testsystems and testdevelopment. Therefore forms and procedures have been developed.

2.2 Difficulties

A first stock-taking showed similairities as well as differences in the field of computer based military test-psychology within and between the different nations. Let us first review the differences.
In some cases is this a matter of competence and coordination. Between countries this can be ascribed to the autonomy and deviating responsibilities of each individual country. Although participating in NATO, each country has to recruit its own personnel to do the military job for national policy. So they can ask different qualities and quantities.
On a smaller scale, within countries it can be ascribed to the different tasks of army, air-force and navy. Not to mention rivalry and competition between the three. Each military force can develop seperate and individual automated test-systems.

But it is not only a matter of competence or coordination. It is also due to the fact that the armed forces need different kinds of personnel for the filling of there functions (for example: pilots vs infantry-soldiers). This means that different automated testsystems were developed on behalf of different predictive criteria.

Then there is also the question of "who" developes these systems. Is it done by the armed forces on there own or is it done partly or even completely by civil contractors. In this last example one can assume that contractors connect with there former and other work in this domain.

In this case the systems are probably very simular to those systems which are already used or those systems which will be used in the near future in the same domain. In case the army developes it on their own, this can mean considaribly difference from other developments. This can give a substential difference in the way systems operate in the same domain.

And then there is the problem between the different hard- and software. Let us first concentrate on the hardware. Most stand-alones and LANS are IBM-compatible nowadays. But there are still none-IBM systems operating. Also different operation-languages are used such as DOS vs UNYX. Compatibility is not the only problem. Construction-date is another one. Most machnes are from a different construction-date, and due to the gigantic and explosive on-going development, this has implications for technical possibilities like processing-capacity, clock-speed and so on. Normally this means at least calibration between machines.

The software is not always compatible. If the software is compatible a programm should operate on all compatible machines the same way. But it doesn't work that way. It often has to be adjusted. Mixing software isn't that easy either. One way or the other, there are often minor or bigger problems.

2.3 Organization and coordination

In order to organize the different tests among at least NATO-countries, Dr. Clessen J. Martin (1991) presented a taxonomy, based on the taxonomy from Fleishman (1975). This taxonomy does not imply either unidimensionality within a taxonomic category or independence among the domains, but organizes it in six test-domains (Spatial Orientation and Visualization; Numerical Facility; Time Sharing and Selective Attention; Reaction Time and Choice Reaction; Psychomotor; Complex Information Processing). It did permit an organization of the over 30 tests NATO-countries had developed. Let us therefore take a closer look at the concepts involved.

3. CONCEPTS

3.1 General

This section provides an overview of the concepts and philosophies underlying various NATO nations computer-based assessment systems. It reviews a broad range of conceptual and methodological issues ranging from cognitive and holostic approaches to the construction of assessments, to psychometric techniques used in adaptive testing, and the use of computerisation to assist vocational guidance.

3.2 Overview

Computers have changed psychological testing by opening up new possibilities for both test presentation and data recording. Computer technology affords the opportunity for interaction between the applicant and the computer; offers a mechanism for recording precise measures of reaction time and provides the means of presenting animation on the computer display, as only a few examples.

These new technological possibilities suggest novel testing concepts; for example, the possibility for interaction between the applicant and the computer enables adaptive testing, a procedure which customizes test difficulty level to the applicant (i.e. adapts the test to the ability level of the individual applicant). The possibility for interaction with the computer also provides the means for automated vocational counselling in which the applicant accesses relevant job information from the computer and provides information to the computer about his or her vocational interests and preferences.

Thirdly, interaction with the computer presents numerous possibilities for simulator-based testing, for example flight and tank simulator tests; the holistic approach to selection also relies on computer simulation. The ability of the computer to record precise measures of time brings mental chronometry within easy reach. Most theories about human information processing are built on reaction-time studies. With computer-based tests, it is now possible to measure specific aspects of human information processing.

This cognitive psychology approach enables the psychometrician to explore the "architecture" and function of the human information processing system in much more detail than was previously the case. There is a distinction between tests based on stage models and tests based on the differential ability literature. Computer-based testing is especially relevant for tests based on stage models of information processing, because these models rely heavily on precise time measurement and also on animation (e.g., for tracking moving stimuli). The cognitive approach is operationalized in two computer-based systems: The Netherlands Taskomat and the UK Micropat.

The holistic approach and the simulation-based approach to psychological testing have a great degree of similarity. The main difference between the two approaches is that, in the simulation-based approach, discrete measures of simple behaviours are typically combined, based on regression anlalysis; whereas the holistic approach attempts to measure the dynamic aspect of complex behavioural interaction. Different aspects of individual ability should be measured simultaneously, in terms of their interactions, rather than as independent, discrete measures. The holistic approach asserts that the prediction of performance requiring complex behaviours can only be accomplished by the measurement of complex behavioural interactions; discrete measures of simple behaviour, it is stated, can only be expected to predict performance on simple tasks. An example of the holistic approach, is the Belgian Gunner Testing System (G.U.T.S.).

Adaptive testing is represented by CAT-ASVAB, the computer adaptive testing version of the US Armed Services Vocational Aptitude Battery The adaptive nature of CAT offers numerous advantages over the paper-and-pencil version of ASVAB, such as a more efficient utilization of test items and a significant reduction in test administration time. This is achieved by focusing the difficulty-level of the items to the ability-level of the examinees (e.g. a high-ability examinee writing a conventional test, would receive all test items, regardless of difficulty level; under CAT, that examinee would receive only the relatively difficult items). On-going and future research, aim on enhancing the predictive capability of CAT-ASVAB, by introducing new types of computer adaptive tests which tap congitive domains not captured by the current version of CAT-ASVAB.

The AutomatedCounselling Component (ACC) is a prototype of a system where the computer can enhance selection by introducing a computer-administered interactive vocational counselling session as part of the selection process. It is designed to function as a component of the Canadian Forces Career Information System. The expected benefits are an increase in the amount of relevent vocational information provided to the applicant, the generation of more accurate expectations of military life, an increase in the applicant's level of confidence in occupational choice and ultimately a reduction in dissatisfaction/turnover during training. Computer interactive vocational counselling also promises a number of benefits to military recruiters and to the administrative staff at recruiting centres.

4. IMPLICATIONS FOR RECRUITMENT

4.1 General

In which way did the different perspectives adapted by the participating NATO-countries benefits military recruitment. Computerization of selection and classification tests offer many advantages to military recruiters. The first area is related to the processing efficiency of potential recruits. The second area relates to the use of adaptive testing techniques. The third area is related to a more technical consideration dealing with differentiation of human abilities. Lastly there is the area of costs-benefits, an area which is becoming increasingly important.

4.2 Processing efficiency

Processing efficiency is enhanced by fast and accurate scoring of candidate results. The computer has the capability of providing immediate results related to the applicants eligibility for military service. Also for those tests which determine job classification or job eligibility, composite scores are immediately available to recruiting personnel who assign individuals to jobs. Furthermore these test scores may be transmitted to training commands and enables recruiters to offer specific training assignments. Computer testing stations also offer the opportunity to use a flexible start-schedule. Whereas traditional paper-and-pencil group testing procedures require all examinees to begin at the same time, computer stations permit the testing of applicants whenever they arrive at the testing location. This is a particular advantage to recruiters especially if they have to travel fairly large distances to deliver applicants to the testing station. Also, if mobile testing units are available to recruiters, it enables them to deliver the tests to the prospective applicant. When telecommunication is available, test results then may immediately be transmitted to the recruiting station. Ease of administration and supervision of testing are enhanced by the use of computers. Simple instructions which are examinee paced facilitate understanding the requirements for the tests. Also Local Area Networks (LANS) offer the possibility for administrator test stations which monitor the progress of the examinee during the test session.

4.3 Adaptive testing

Adaptive testing offers the advantage of more precise measurement of applicant abilities. Because the test is tailor made for the individual, ability estimates are more reliable. This is especially important when bonuses are given for certain levels of test performance. Also precise measurement ensures better person-job match. The adaptive nature of the administration of test systems makes it possible to estimate the examinees abilities in shorter time periods than in conventional tests. Only those items which are at or near the examinees ability level are administered.

Test compromise or theft are less likely in computer testing programs. Because of encrypting techniques it becomes very unlikely that examinees can obtain test item pools and distribute them for their benefit. Another advantage of adaptive tests is that new items can be administered but not scored in order to obtain preliminary item statistics. While this may not be a direct advantage to recruiters, it does ensure that adaptive tests remain current.

4.4 New ability domains

Computer testing technologies also enhance the capability of assessing new ability domains. Psychomotor ability which is an important aspect of modern weapon systems is accurately measured in contemporary computer testing systems. Because the computer can track learning capability in a new or novel learning task, it is possible to plot learning rates for individuals during a testing session.

Dynamic spatial abilities can be measured in a computerized system. These spatial displays offer the capacibility of simulating important tasks which are those represented in a air traffic control system.

4.5 Benefits and costs

Finally in any new testing system it is necessary to consider the combined effects of benefits and costs.

Because new computerized tests are expected to enhance the predictive validity of selection and classification systems, individual job performance will result in a cost savings. It is expected that the utility of new types of computerized tests will offset the costs associated with implementation of a computerized system.

5. CONCLUSION

Although most systems are not operational yet, it seems clear that computer-based assessment can make a large contribution to the recruitment of armed personnel for the NATO-nations participating in RSG-15. But in order to secure that no work is done double or even triple, and therefore money is spent without real reason, there has to be cooperation in and between nations.

So the most important conclusion is that we must continue to work and coordinate in this field. The development is still going on and the technique is speeding ahead. Every day new possibilities present themselves. It is therefore a most promising area of research and constructing. Nevertheless one has to coordinate and cooperate more often and in an earlier stage. Only then countries can really work together and they will be able to exchange information, material and personnel in an smooth way.

AN INTRODUCTION TO THE HOLISTIC APPROACH

Agnès Kokorian
Direction des Armements Terrestres
Etablissement Technique d'Angers
France

INTRODUCTION

This paper is an introduction to a philosophy of testing that is widely used in Europe. This philosophy is based on the view that the individual is an indivisible whole that cannot be understood by simply considering its different components (physical, physiological and psychological) separately. So, rather than subdividing the individual into these separate components, this approach stresses the necessity of designing assessment tools that measure an individual's general capacity to respond to a complex situation.

I will describe the foundations of this philosophy through a systemic view of the individual. Then, I will define in more detail the holistic approach and the kind of tools that it uses. Finally I will describe the importance of this approach to military selection.

THE INDIVIDUAL AS A MULTI-DIMENSIONAL SYSTEM

The foundations of holistic view can be understood from a systemic perspective. A system is an organised global unit consisting of relations between components (Morin, 1990). Three characteristics of system, totality, interaction and organization will allow us to understand the complexity of an individual.

A system is composed of components but, from Gestalt psychology, we know that it is more than the simple sum of those components. The whole can not be simply reduced to its parts. This implies the emergence of new properties not discernible in the constituent components. It means that the independent assessment of individual components can not reveal those new dimensions. This is apparent for two reasons. The first one concerns the interactions between the different components : aptitudes do not act in isolation. The second reason concerns the relationships between these components within a given context : situational and environmental factors can modify the responses of an individual.

These interactions constitute an essential factor in performing the task and they express a structure or organization within the total system. In fact, it is in this organization of relationships between components that the system reveals its identity. We know that the same components organized in different ways will indicate different qualities. So, two people who have the same aptitudes will nonetheless perform differently in response to the same task because their individual performance is due to differences in the organization of those aptitudes.

We must also state that this organization is dynamic : aptitudes do not operate in a static or set manner during a task. As Relieu (1992) states "they form successive moments representing various constellations which accompany the moving environning situation captured in its totality." The efficiency of the system comes from this dynamic aspect and its flexibility. Efficiency can not be inferred from the isolated components because the organization does not exist prior to the task and can only appear through the complexity of a given situation.

THE HOLISTIC APPROACH AND ITS TOOLS

A systemic view of man thus emphasises that complexity can not be fully understood by a process of reduction to constituents components (Jacq, 1989). Rather, the holistic approach proposes that aptitudes should be measured simultaneously to evaluate general capacity to respond effectively and efficiently to complex situations.

Using the techniques of simulation based assessment, this approach puts the candidate into an assessment situation that is, psychologically, very similar to the real work situation. The assessment comprises several simple tasks presented simultaneously, and the number and difficulty of these tasks can be varied during the assessment process. The intention of this approach is not to attempt to replicate the job itself which would require that the individual had prior knowledge of the job. The intention is to reproduce the general factors that will be involved in future work situations. Two computer-based assessment (CBA) systems will now be described to provide a better understanding of how holistic approach gives a more detailed and realistic assessment.

ESPACE/JAMES (Relieu 1988 and 1992). This is a french CBA system developed by Direction Centrale du Service National for the selection and placement of conscripts. It consists of separate cabins which operate on a fully automatic basis. The candidate receives information through audio-visual equipment (screen and voice synthetizer board). Responses are made through a keyboard, two levers and two pedals. The tasks comprises logical reasoning, language knowledge, visual and auditory perception, psychomotor coordination, spatial orientation and decision making.

The assessment is divided into different phases to reproduce the complexity and the dynamic quality of a work situation. After learning a simple task, the individual's workload is first increased by adding another task. The rate of task presentation and the complexity of stimuli are then increased. Assessment of the individual is then made in the form of generals levels of adaptability both in terms of general predicted success in military life and with respect to specific job categories. However, this assessment is not based on a sum of individual ESPACE tasks, but on performance in the assessment as a whole.

GUTS (Lescreve 1991 and 1992). This is a Belgian CBA system developed for the selection of Tank Gunners. It reproduces tasks typical of those in tank gunnery including stressors in the tank environment. The work station is set up in a closed box that the individual has to enter through a small hatch. Assessment begins at this point as the individual has the choice to disengage from the procedure at any point, including refusal to enter the box (at any such refusal point the individual effectively de-selects from tank gunner selection). Inside the box, the individual has to perform a tank gunnery simulation while wearing combat clothes and an oxygen mask, and while performing a weight lifting exercise with one hand at regular intervals. The box itself is dark, warm and noisy as result of simulated combat environment. Thus, this system places the individual in direct confrontation with the context of the future work situation as well as the aptitudes required to perform the task. Factors such as response to the claustrophobic conditions of a tank would be impossible to assess effectively in the traditional aptitude testing approach.

These two examples show the basic principles of the holistic approach : the totality and simultaneity of measurement. Other systems that follow this approach are being developed by NATO nations particulary for pilot selection. A wealth of validity data on such system will become available over the next two or three years.

IMPORTANCE OF THE HOLISTIC APPROACH TO SELECTION

Through the representation of complexity and dynamism of work situations, it becomes possible to measure what Relieu call "the individual's psychic reactive complexity". This refers to the person's adaptability to complexity and change in the simulated situation. Earlier in this paper it was noted that this adaptability was due to the organization of components and the flexibility of this organization. Employing these principles, the holistic approach can take account of "vicariant" processes in performing a task. That is, individuals will approach a task with different modes of operating. Although there may be an optimum method of responding, this method may not be the one that the individual actually applies. Instead, an individual may substitute less optimal methods and still be successful in responding to the task. Also, an individual may compensate for a weakness in one specific area by substituting strength from another area.

Another advantage of the holistic approach is the combination of aptitude with personality assessment to measure overall adaptability. Furthe.more, this approach can measure personality related dimensions that can not be measured using classical methods or can not be measured with as much confidence. As in the GUTS system, the object of assessment is not only to measure the aptitude to perform tank gunnery (in simplistic terms a combination of perceptual and psychomotor abilities), but also whether the individual will "fit" in the physical context in which tank gunnery will be performed. Indeed, the growing interest in the holistic approach stems from concern over omitting important aspects of task performance in classical psychometrics (Lescreve, 1992). It is also of interest to note that the high apparent content validity of holistic tools has the positive effect of enhancing the candidate's motivation to participate in the assessment.

CONCLUSION

A systemic perspective emphasises the need to take into account the dynamic organization of the individual, and the necessity to design assessment tools that take into consideration both the complexity of the individual and work situations. By applying the principle of simultaneous measurement in a dynamic context, the holistic approach takes into account the interdependence of the factors involved in efficient and successful task performance. Although it may at first seem paradoxical, a more global approach to assessment design will provide a more detailed and realistic indication of the individual's suitability for military service. The effect on individual performance of the interaction between components can not be anticipated a priori. It is only by putting the individual in a simultaneous multiple task situation that this interaction can be assessed.

The depth of assessment possible and the individual's increased motivation to participate in the assessment process contribute to high levels of predictive validity for holistic tools. However, these tools are comparative'y new and there remains a need for research on their reliability and validity. Research is also needed to determine the degree of representation required in the use of simulator-based assessment in selection.

REFERENCES

Aschenbrenner H. (1990a). German update on progress in simulation based aptitude assessment. Paper presented at the London meeting of NATO RSG 15 (Computer-based Assessment). November.

Aschenbrenner H. (1990b). Simulation Assisted Aptitude Test (SEF) and Simulation Assisted Training (SGT). Paper presented at the London meeting of NATO RSG 15 (Computer-based Assessment). November.

Durand, D. (1979). La systémique. Puf, Paris.

Jacq, J. (1989). Contribution d'une approche holistique à l'évaluation des personnels militaires assistée par ordinateur. DRET.

Lescreve, F. (1991). GUTS - The Belgium Gunner Testing System. In Boer, L.C. (Ed.) NATO Research Study Group 15 Workshop : Computer based Assessment of Military Personnel. Report submitted to NATO Headquarters.

Lescreve, F. (1992). A holistic approach to psychometrics : some basic principles. In Burke, E.F. and van Raay, P.B. (Eds) Computer-based Assessment in NATO : Final Report of Research Study Group 15. Report submitted to NATO Headquarters.

Morin, E. (1989). De la complexité : complexus. Les théories de la complexité. Colloque de Cerisy, Seuil, Paris.

Morin, E. (1990). Le système, paradigme ou/et théorie. Science avec conscience, Seuil, Paris.

Relieu, P. (1988). Projet ESPACE. Paper submitted to Research Study Group 15 (Computer-based Assessment).

Relieu, P. (1992). Job Adaptation Measurement System (JAMES). Direction Centrale du Service National, laboratoire de recherches psychométriques. In Burke. E.F. and van Raay, P.B. (Eds) Computer-based Assessment in NATO : Final Report of Research Study Group 15. Report submitted to NATO Headquarters.

THE GUNNER TESTING SYSTEM (GUTS), AN APPLICATION OF THE HOLISTIC APPROACH.

Captain Lic Psy F. LESCREVE
CENTRE FOR RECRUITMENT AND SELECTION
BELGIAN ARMED FORCES

INTRODUCTION.

When in 1988, the 'Research and Study Group 15' on 'Computer based assessment of military personnel' (NATO AC 243/Panel 8) started its activities, it soon became clear that a number of very different approaches were used in the field of computer based assessment. One of them is called the holistic approach. Simultaneousness of measurement, anti-reductionism and globality are three important aspects of this approach. This article tries to explain some consequences of emphasizing those aspects when developing selection instruments. In a first part, some theoretical considerations are proposed and in a second part, a practical application of those principles is shown using the GUNNER TESTING SYSTEM (GUTS) developed by the Belgian Armed Forces.

THE HOLISTIC APPROACH.

When students in psychology are trained in psychometrics, they learn about measurement of different aspects of intelligence, personality, motivation and so on. Because people are different and they appear to behave in different ways when they are faced to psychometric instruments such as tests, it is possible to 'measure' their abilities. Classic psychometry however urges the psychologists to develop instruments which measure only one dimension.
When in a certain context it is necessary to measure several aspects of an individual, for instance in selection where it is necessary to predict his or her success in training, the classic psychometry tells us to combine different measurements of preferably independent dimensions. The multiple linear regression is one good example of that principle.
More recent developments in psychometrics, such as the item response theory, stick to the same principle; one test - one dimension.

The question I want to address here, is whether this approach is suitable to some specific problems we have to deal with. Let's take an example.
Imagine a psychologist has to develop a selection procedure for tank gunners. His first concern will be to identify which abilities are required to become a good gunner. He will perform a job analysis and as a result he will find out that for instance the gunner needs a good psycho-motor coordination, must be able to react according to strict procedures, must be able to take decisions quickly and must be resistant to stress.
According to classic psychometry, our psychologist then should take four different tests to measure the different abilities. We can easily imagine how the test battery would look like. There would be a test for psycho-motor coordination using a computer display to present a typical tracking task and two joy-sticks and two pedals as convenient interfaces. Secondly, which implies 'not at the same time', there would be a paper and pencil test in which the applicant first had to memorize some procedures and then has to reproduce them correctly or is asked to detect errors in some proposed items. Third, the applicant would have to perform another test to measure the speed of his decision making. And last he could choose the STROOP colour stress test.
The psychometrician probably would be happy because this battery allows the psychologist to have the relatively independent one-dimension measurement needed within each individual test.
The next step would be to collect test-results from a representative group of applicants and then have them trained as gunners. After that period, it becomes possible to calculate correlation coefficients between test results and training results. And at last we can combine the test results to predict the training scores using multiple regression.

Let us now take a closer look at the training of gunners. We can see that in the beginning things are kept quite simple and most of the trainees perform well. The big problems only occur from the moment all tasks have to be performed at the same moment for instance during a so called 'battle run'. To put it in an other way; the different aspects of the job are relatively simple when taken separately, but when they have to be performed simultaneously, performing well becomes very hard.

What conclusions can we make so far? First that performing well in the different dimensions separately is not a sufficient condition to perform the same dimensions well simultaneously. Second, when having done a job analysis, it's not only important to identify the basic dimensions of the job, but it also is necessary to assess the importance of their interactions to satisfy the job requirements.
To comply with the classic psychometric approach, it would be necessary to assess the applicants ability to organize his behaviour in order to perform well in the different tasks at the same time. That however can hardly been seen as a unidimensional ability!
The other solution we propose is to measure the different abilities needed according to the job analysis in a situation which can be compared to the most relevant criterion situation. When we go back to our example of the gunner selection, this would mean that we'll try to assess the four dimensions at the same moment in an environment which from a psychological point of view is very similar to a battle run. This approach inevitably leads to what is called simulation based assessment. Indeed, only a computer driven device can provide an environment sufficiently complex to be a true simulation of reality and yet being standardised and able to collect an impressive amount of data.
Assessing the different dimensions at the same moment, so that the candidate's ability to perform different single tasks simultaneously is taken into account is the basic principle of a holistic approach.

It is still possible to obtain the different measurements which were required according to the job analysis. The quality of those measures however may depend on the ability of the applicant to distribute his attention equally over the different tasks. That of course, is the reason why the classic psychometry looks for uni-dimensional tests. But the holistic approach provides us with another measure which could prove to be far more interesting than those measurements which were deteriorated due to the principle of simultaneousness. It is the measure of general efficiency. When during a job, one has to do different tasks simultaneously, it generally is done to achieve a complex goal. For instance, when our tank gunner needs his psycho-motor ability, his decision making, his sticking to procedures and his stress resistance, it all has to do with memorising weapon control orders, identifying targets, chosing ammunition, aiming and firing in order to destroy targets according to his weapon control orders. The measure of general efficiency in this example obviously will be the number of targets which have been destroyed correctly. According to the reasoning above, it makes sense to expect that the measurement of general efficiency would be more predictive than the measurements of the different dimensions taken separately and combined through regression.

A second important principle of the holistic approach is not to reduce measurement to a limited number of dimensions. The classic psychometric approach to the question of what needs to be measured when for instance we want to predict the training results of different applicants for a job, starts with a job analysis. Let us have a closer look to what a job analysis really is. Generally speaking, most jobs are not limited to repeating continuously the same single task but include a variety of behaviours that have to be done well. Typically, it appears to be possible to describe a specific (complex) task which can be used as a criterion to determine wether or not a person is good at his or her job. In our example, the way a tank gunner performs during battle runs will tell us whether or not he is a good gunner. Logically, the job analysis will focus on that typical task. Where things can start to go wrong, is when we try to reduce the reality of the task to identify dimensions which can be assessed easily. You remember the four dimensions which were identified as being the required ones to be a good gunner. Maybe you wondered why motivation was not one of them. Of course motivation is needed to perform well in the hard circumstances experienced by gunners even during peacetime. The reasons why an aspect can be omitted during a job analysis are manifold. In traditional job analysis, the analyst depends on information taken from those experienced in the job of interest. Those experienced in the job are often referred to as subject matter experts. The information given by the subject matter experts may exclude an aspect because they think that this aspect is obvious and self-evident. Another reason also can be that the classic psychometrician doesn't know a suitable way to assess a certain dimension and may simply choose to omit it.

Imagine for a moment that a military selection centre not only has to select applicants but also has to assign them according to their abilities. Such a centre would be tempted to develop a series of tests to assess separate dimensions. Applicants who happen to perform well in the tests concerning the four identified dimensions considered as essential for good gunners, could be assigned to a gunner course. But are they motivated for the job? We would only have an answer to that question if motivation was identified as being important and therefore assessed. If the job analysis omitted to identify an important dimension, the selection decisions will be taken without any influence of that dimension.

Here again, the use of simulation based assessment, which was a direct result of the principle of simultaneousness, can help us. It seems to be a good approach to try to develop a selection simulator, in which the applicant has to perform a task which from a psychological point of view is very close to the criterion task on the job. Doing so, the risk of omitting important aspects in the evaluation of a candidate will be much smaller than with the classic way of proceeding.

A third principle of the holistic approach is the use of the measurements to make decisions. For instance, the measurement of abilities is only one step in the process of personnel accession . The next step in larger organisations generally include the problems of ranking and assignment. Here also, the holistic approach can prove to be beneficial. When a number of applicants have passed a test battery, it becomes possible to calculate differential predictions for several available jobs. We then can make different ranking. We can look for a specific job, for which the applicant is best suited and we can decide to accept the best candidates for that job. On the other hand it is possible to look at an individual and determine for what job his probability of success is the highest and accept him for that specific job. Most of the classification and assignment methods used up to now however don't look at the problem as a whole. As a result, the assignments of candidates are rarely optimal from the point of view of the organization.

When we seek a global optimization of the assignment of candidates to the available jobs, it's interesting to take advantage of the developments of the operational research. To solve the problem, it is very well conceivable to determine a utility value to the assignment of every possible applicant to every available job. It then becomes possible to link each candidate to a job using for instance the so called Hungarian method, which is derived from the well known travelling salesman problem. This method will assign the candidates to the jobs in such a way that the sum of the so defined utilities is maximal. This example illustrates a third principle namely to use a global approach to the problems of classification and assignment.

Let us now have a closer look at one implementation of these principles.

THE GUNNER TESTING SYSTEM (GUTS).

The introduction of new types of equipment in the Belgian Army resulted in an increase of failures in training. Therefore, the Centre for Recruitment and Selection was asked to develop a new selection system for gunners in order to diminish the losses in the expensive courses. The answer got the name 'GUTS'.

The job-analysis.

During 1988 a job analysis was worked out for the different gunner jobs in the Belgian Army. Instructors and trainees were extensively interviewed as well as people actually doing the job. Different measurements were taken in the weapon systems and the users manuals were studied. Literature also was reviewed.

The job analysis concluded that four types of skills can be considered as essential to perform well as a gunner. The first ability won't surprise anybody. It's psycho-motor.

The second one was identified as 'sticking to procedures'. The sequence of the actions a gunner has to do could easily be represented by a flow chart. The different actions must be done in a strict sequential way and the choice between different paths is conditioned by specific orders or situations. Not respecting the procedures always results in poor performance.

The third skill is decision making. This ability is quite new for gunners. Older weapon systems tend to see the gunner as a man who only has to execute orders without having to take decisions. However, due to the reduced reaction time in modern weapon systems, procedures now often request the gunner to take decisions, especially concerning which target to engage.

The last ability which seems to be very important to gunners is their stress resistance. Physiological, physical and psychological stress all contribute to the difficulty of the job both in wartime and in training.

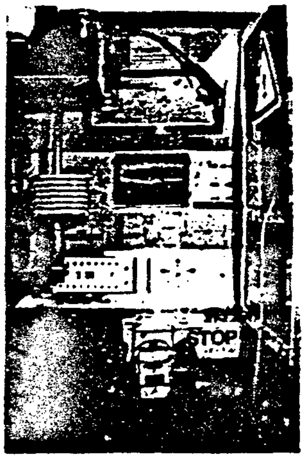
Description of the simulator.

We now would like to describe the test situation. But before doing so, it might be useful to give a description of the selection simulator itself.

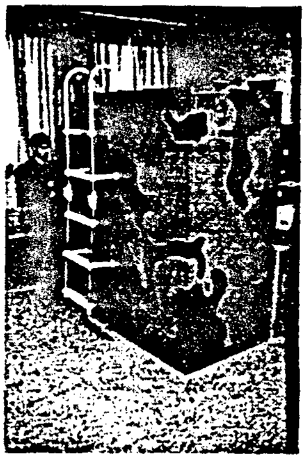
The cabin consists of two major parts. The working post for the subject is located at the back side. At the front side, the cabin provides the necessary space for the hardware.

Before the subject starts the test session, he has to climb upon the cabin using a small stairs. By opening the trap door he can enter the gunner post as in a real tank.

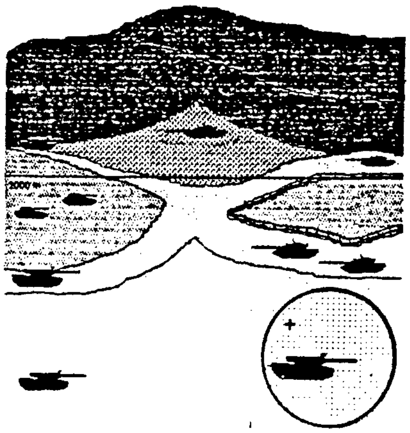
When the subject takes place on the seat, the lookhole reaches just as high as the eyes. With his hands, he can easily manipulate the "steering wheel" or handle, which has several functions. Just above this handle, at the left side, there is a distraction box, which has no particular function and serves merely to distract the subject during the task. At the right side of the distraction box is the identification box. The control box is situated between the top of the cabin and t: .ookhole. At the left side of the subject (at the same height as the lookhole), there is a load box and facing the latter, at the subjects right side, are the ammunition box (lower) and the radio box (higher). Under the handle, at the bottom of the cabin, two loudspeakers generate the distracting background noise. Finally, right under the seat, there is a heating device. In case the subject wants to stop (whatever the reason is), a emergency-exit can be pushed open. An emergency button, a smoke detector, emergency light and a little reading lamp make the whole complete.



GUTS : inside view



GUTS : outside view



GUTS : what the applicant sees

All the devices indispensable for controlling the process, for registering the actions of the subjects, printing out results and giving the test leader the

possibil:ty of following what is happening, were brought together in the front compartment.
On the left side in the front part, there's an output device for printing a subject's results (a trap-door provides easy access). On the right side there are the amplifier and the cassette-deck for the background-noises, in the front part the major control box (with the necessary electronics for interconnecting and controlling all the boxes). At the left side of this major cor.trol box is the heart of the system : the computer (Apple Mac II).
In the upper part of the front compartment, finally, there are two video displays, a small one for the test leader, a larger high resolution one for the subject. The latter project the landscape and the actions of targets to the subject, who can look at it through the lookhole.
Iet's now see what the subject has to do. The test consist of three important arts, the study of the instructions, a training period and the test itself.

The pretest-phase: learning the test instructions.
The first thing the candidate has to do after entering the testroom, is to study the test instructions. In a short introductory briefing, the booklet explains the main goal of the test, i.e. to examine whether or not the subject disposes of the required cognitive and psycho-motor skills, and whether or not his psychological fitness meets the needs of the gunner task.
A description of the instruments and their functions is the next step. The subject will be shown how friend and foe tanks or vehicles will look like and how they will apnear. The instructions tell the subjects about the possible ammunition types, which one to use where and when.
If all this is clear to the candidate, the instructions for executing the subtasks correctly are given: the instructions for engaging a target, identifying a target, the choice of th: right ammunition, loading, inscribing the ammunition, aiming and firing. Spe::ial attention goes to the weapon control orders which will tell the subject which targets he has to destroy and which he's not allowed to engage.
The total time spent for this pretest-phase is 30 minutes.

The training period.
A demonstration run precedes the test cycles. The candidate takes place in the working post via the emergency exit at the back. He can practice to see if he understood the instructions. When necessary, the testleader helps him to make sure there are no misunderstandings. This demonstration serves merely to show how the figures will move in the landscape and which effects are created by manipulating the instruments.

The testing period.
After the demo, the candidate comes out again and puts on the gas-mask, helmet with built-in headphones and battle dress. He climbs in the cabin and takes place in the working-post where he connects the air-supply and the headphones.
The actual test consists of three identical cycles of ten minutes each. During each cycle, the applicant will see 25 targets he has to engage and 25 he may not engage. Every cycle has 4 periods, one period for every weapon control order.
Once the subject makes the decision of trying to eliminate a certain target, he must follow strict rules. The first thing to do, is the engagement of the target. This can be done by steering the reticle such as to follow the target. When the reticle reaches the latter, the subject must push the engagement buttons on the handle.
The following step is the identification of the target. By positioning the selection button on the identification box, the subject selects what he thinks the target really is.
According to the type of target (tank or vehicle) and its distance (more or less than 2000 meters), a correct selection of the ammunition type must be made, using the ammunition box for both selecting and inscribing the ammunition after 'loading' it. To 'load' the gun physically, the subject has to push a handle with a weight (8 kilograms) until a load sign appears.
The final action to be performed is the aiming and firing itself, possible by manipulating the handle and pushing the fire buttons.
This engagement-firing cycle must be restarted for every target the testee wishes to eliminate.

The GUTS : a stress inducing device.
As the job analysis found out that stress resistance is an essential quality of a good gunner, it's important to assess if an applicant can cope when put in a stressing situation. Therefore some stressors were implemented in GUTS. Although we are aware that the stress we can induce cannot really be compared to actual combat stress, we're convinced that there is little chance that people who cannot resist the stress in GUTS will be able to resist combat stress well as gunners.

Let us now review the stressing elements in GUTS.

-Induction of claustrophobia : The very limited working space, having to go through a narrow hatch and the wearing of a gas-mask bring the applicant in a very claustrophobic situation. About 5% of the subjects show claustrophobic reactions.

-Heat and protective clothing : The subject has to wear a heavy protective clothing and is working in a heated cabin (30° C).

-Noise and information overflow : The headphones not only provide the weapon control orders which are needed to perform well, but also give disturbing radio communications. There are also loudspeakers producing very loud sounds of engines, tracks and explosions. In the visual domain, there are lots of LED's to disturb the candidate.

-Helmet and gas-mask : The wearing of a helmet and a gas-mask are also stressing elements. The gas-mask provides the subject with fresh air from outside the cabin.

-Interruption of air-supply : Every three minutes, the air-supply is cut off for five seconds to distract the subject and to induce more claustrophobia.

-Physical and mental workload : Performing well in the GUTS is a very demanding job both mentally and physically. The physical workload mainly results from 'loading' the gun in a very difficult position.

-Selection situation : Finally one should not forget that the subject applies for a job and this by itself already causes more arousal.

The measurements.
Since the GUTS is developed according to the holistic approach, it is quite evident that the most relevant measure of the applicant's ability will be a score of 'general efficiency'. This score reflects the number of targets which have been correctly destroyed.

The reaction time gives mean time and standard deviation needed for one sequence of engagement.

Three kind of errors also are registered : decision errors, manipulation errors and procedure errors. These data can be used for the assignment of the applicants to specific weapon systems.

Finally, as the test consists of three identical cycles, it is possible to give a graphical representation of the evolution of the performance.

CONCLUSIONS.

This article has not the pretension to present hard conclusions. The intention was to make the reader think about the usefulness of some principles of the holistic approach and to demonstrate the possibility of implementing these principles in a test instrument.

A Review of the Reliability and Validity of Computerized Tests Among RSG.15 NATO Countries

CLESSEN J. MARTIN
DEPARTMENT OF THE U.S. NAVY

PURPOSE

The purpose of this presentation is to review research results related to the reliability and validity of computerized selection and classification tests among member nations of Research Study Group (RSG .15), Computer-Assisted Assessment of Military Personnel. Member nations were requested to complete a test information form for each test included in this review. The test information form gives the nature of the test and the dependent variable. Also included on the form is a section requesting information on apparatus and equipment required for the administration of the test. A research results section requests information on the reliability and validity of the test as well as specification of the types of criteria used in the validation study. Also requested in this section is information on the type of reliability (e.g., retest, coefficient alpha, etc.) and whether any corrections were applied to the validity coefficients.

TAXONOMIC OUTLINE

The taxonomic outline used for organizing the tests is based on the taxonomy presented by Jones and Boer (1989) which was adapted from one originally proposed by Fleishman (1975). This taxonomy does not imply either unidimensionality within a taxonomic category or independence among the domains. However, it does permit an organization of the 30 tests reviewed and may provide greater insight concerning the test domains which offer the greatest payoff for future operational use. Only those U.S. tests which are presently being validated in the Joint-Service Enhanced Computer Assisted Testing (ECAT) program were included in this review. Many more U.S. tests are under development or have been developed in connection with the Army's Project Alpha (A) and the Air Force's Learning Abilities Measurement Program (LAMP). Also, some additional computerized test development is underway at the Navy Personnel Research and Development Center (NPRDC). Most of the tests submitted by the United Kingdom are those used in the MICROPAT system (Jones and Abram, 1990). According to the Jones and Abram (1990) report, the MICROPAT computer testing system was originally developed for selection of Army helicopter pilots. Since 1985, the British Navy has continued development of the system for Pilot and Observer selection.

RESULTS AND DISCUSSION

Only reliability information was available for the NAVOR test and Manikin test. For the NAVOR test, the lower internal consistency coefficients are associated with dependent measures associated with error scores while the higher coefficients are associated with inspection time scores. Likewise, the relatively high measures of internal consistency for the Manikin test are based on mean time per correct answer for the upright, inverted, or horizontal positions. Variability for the inspection time measure on the NAVOR test was one-half or less in the military trainee sample team than in the university sample. For the Target Identification test, internal measures of consistency tended to be higher than retest reliabilities. The criterion score is based on results from the tank gunnery simulator (Institutional Conduct of Fire Trainer (ICOFT)). The ICOFT records both speed and accuracy of firing. The criterion score is a composite of both the speed and accuracy measures. All validity coefficients are uncorrected and hence are underestimates of the true validity. Likewise, the validities for the Assembling Objects and Orientation tests are based on performance in the tank gunnery simulator. The Integrating Details validities are based on final school grades. It is interesting to note that the Integrating Details test correlates with school performance criteria at a higher level than the other three spatial tests correlate with hands-on criteria. The three Navy jobs in the study are Aviation Ordnanceman (AO), Avionics Technician (AT/AQ/AX), and Gunner's Mate (GMG). The AO personnel are aircraft armament specialists and are in charge of storing, servicing, inspecting, and handling all types of weapons and ammunition carried on Navy aircraft. AT/AQ/AX personnel are electronics technicians who maintain the advanced radio, radar and electronics equipment. Their work is in three basic categories: equipment testing and analysis, maintenance and repairs, and administrative tasks. GMG personnel are responsible for the operation and maintenance of guided missile launching systems, rocket launchers, gun mounts and other ordnance equipment. These persons also work with electrical and electronic circuitry. It is clear that spatial abilities as measured by the Integrating Details test is related to successful course performance in these jobs. The 0.44 validity coefficient for the Integrating Details test was obtaining in the AO rating. The Navy expects that people in this rating should have above average competence with tools, equipment and machinery.

The MICROPAT Subtract test validities range from 0.02 to 0.40. The criterion measure was based on a cumulative pass/fail training score. Two dependent variables were analyzed: a rate score based on time and accuracy and a percent accuracy score. Overall, higher validities were associated with the percent accuracy score with the exception of the 0.40 validity obtained for the Navy Observer sample. The ECAT Mental Counters test has been validated for Aviation Ordnanceman (r=.17), Gunner's Mate (r=.35) and Operations Specialist (r=.31). In addition, this test has been validated on a small sample (N=68) of Navy Electronics Technicians with hands-on performance measures. The resulting correlation was 0.22 and was significant at the 0.05 level. No validity coefficients are available for either the BARB Alphabet Forward and Backward test or the BARB Number Distance test. However, internal consistency coefficients are in the 0.90 range. While tests in this domain measure numerical facility, the Mental Counters test correlates 0.40 with the power tests of the Armed Services Vocational Aptitude Battery (ASVAB) and 0.50 with the Ravens Progressive Matrices.

Internal consistency reliabilities for the Schedule test were 0.53 and 0.77 while the validity coefficients were 0.00 to 0.09. These test validities were the lowest of the MICROPAT tests. The Schedule test correlated between 0.22 and 0.28 with the Minnesota Paper Form Board and approximately zero with Ravens Advanced Progressive Matrices. The results from this test are disappointing and it will not be used in any predictor composites for the prediction of aviator training. The Dual Task retest reliabilities were 0.22 and 0.67 for change in accuracy and average time per jump in the tracking task, respectively. Validities ranged from 0.01 to 0.14 and tended to be higher for the single task latency difference score. The Dual Task latency difference score had the lowest overall validity coefficients. Research results for the Pilot test are very promising. Retest reliability is 0.84 and the validities range from 0.55 for flight instructor rating to 0.74 for applicants with no prior flight training. For the Selective Listening test, the validity coefficient for air traffic control training was 0.43 and 0.36 for initial flight training. These results were based on 88 pilot applicants and 87 air traffic control applicants.

Only the Plane test has been validated. The validation samples were Army helicopter pilots and Navy Observers. The reliability samples were university students (N=53) Cathay Pacific applicants (N=302), and experienced Observers and Observers in training (N=202). All internal consistency measures except one were in the 0.60 to 0.80 range. Test-retest reliabilities were in the 0.42 to 0.72 range. The dependent variables for this test are: mean reaction time, probability of a hit over three test trials, probability of a false alarm over three test trials, a measure of discrimination sensitivity over three trials, and a measure of response bias related to tendency to respond. The discriminative sensitivity and response bias dependent variables were associated with the larger validity coefficients. For the Word Recognition and Meaning test and the Letter Checking test, no validity coefficients were available. However, retest reliabilities for the Word Recognition and Meaning test were 0.42 to 0.72; internal consistency reliability for the Letter Checking test was 0.80.

Internal consistency and retest reliabilities are available for the Adtrack 2 test. Highest validities were associated with a measure of mean difficulty level over the last 11 time periods of each trial. Much lower validities were associated with a difference score between the first and last trial levels. The Adtrack 3 test is similar to Adtrack 2 except the examinee can control the level of difficulty of the task. Internal consistency reliabilities are also quite high for this test ranging mainly in the 0.70 to 0.87 range. Validities for this test against pass/fail training criteria were mainly in the 0.10 to 0.21 range. Comparable results were obtained for the Landing test. For this test, test-retest reliabilities ranged from 0.19 to 0.73 while internal consistency measures ranged from 0.23 to 0.86. The most reliable dependent variable for this test was the score based on the accuracy over the last three flights. The lowest reliabilities tended to be associated with dependent measures based on difference scores from the first and last flights. These lower reliabilities undoubtedly reflect the fact that there are different learning rates within the task. Validity coefficients were based on pass/fail criteria at the end of training. The validities ranged from 0.01 to 0.18. The Comp 2D is the only multi-limb test included in this review. This test is a two-dimensional tracking task in which the x-axis is controlled by footpedals and y-axis by a joystick. The task is presented for three 2-minute trials and error scores for each axis are recorded. Because the errors are highly correlated, a composite error score is computed. Reliabilities for this test are the highest of any reviewed in this report. For both types of reliabilities, the coefficients are in the mid to high 0.90s. Validity coefficients are based on pass/fail criteria for Navy helicopter pilots. The correlations for the two samples were 0.22 and 0.19. It should be mentioned that in reporting the validity coefficients the algebraic sign is assumed to be positive. However, whenever the dependent variable is comprised of an error score of some type, the actual sign is negative. Of course, the value of the sign is irrelevant to the magnitude of the prediction. Data for the One-hand and Two-hand tracking tests are primarily form Busciglio (1990). He has evaluated the utility of the Army's Project A tests in terms of incremental validity over the ASVAB. The Busciglio (1990) report performed stepwise regressions using predictors and performance tests from the Army's 1985 concurrent validation study. This validation study was based on a total of 4,039 first-term enlisted personnel in nine Military Occupational Specialities (MOS). These MOS were: Infantry, Cannon Crew, Armor Crew, Single Channel Radio Operator, Light Wheel Vehicle Mechanic, Motor Transport Operator, Administrative Specialist, Medical Specialist and Military Police. The criterion measures were: School and Job Knowledge Tests, Hands-On test, General Soldiering Proficiency, Core Technical Proficiency, and Skill Qualification Tests.

The School and Job Knowledge tests are written, multiple choice items measuring technical information related to specific tasks in each MOS. The Hands-on tests are job sample measures requiring the soldiers to perform actual tasks pertinent to each of the nine MOS. For the Hands-on test, a total of 14 to 17 major job tasks were included in each of the MOS. These major job tasks were also included in the written job knowledge tests. The General Soldiering Proficiency test score is a composite score based on a variety of tasks common to many MOS as measured by both written and hands-on tasks. Examples of these common tasks are: determining grid coordinates on maps, recognizing friendly and enemy aircraft, and first-aid procedures. Core Technical Proficiency measures are MOS specific tasks and are based on a composite of both written test items and hands-on tasks. Finally, the Skill Qualification Test (SQT) score is based on results from paper-and-pencil tests of MOS specific technical knowledge. The SQT was developed by the United States Army Training and Doctrine Command for periodic assessment of MOS specific knowledge. The SQT is the only criterion in this series which was not specifically developed for Project A.

The validity coefficients presented are for the One-hand and Two-hand tracking tests and represent the variance in the criterion measures which is not accounted for by any other tests in the Project A battery or by the ASVAB. Thus, instead of the correlation representing the simple correlation between predictor and criterion, it is a partial correlation representing the unique variance of the predictor (Busciglio, 1990). At the time of preparing this report, the simple correlation coefficients were not available and had not been published. Reliability coefficients for both tracking tests ranged between 0.74 and 0.98. The validity coefficients for the anti-tank gunnery simulator were not partial coefficients but represented the simple correlation between predictor and criterion on infantry trainees. Two-hand tracking coefficients tended to be higher than One-hand tracking and were greatest for the Hands-on and General Soldiering criteria. It must be emphasized that the partial coefficients represent the unique variance not accounted for by ASVAB, six paper-and-pencil tests of spatial ability, computerized tests measuring target shooting, simple and choice reaction, short-term memory, perceptual speed and accuracy, and number memory. Thus, these validity coefficients are extremely constrained by the method of analysis.

The Risk test is a gambling type task in which the examinee may select up to eight keys. One key is a penalty key and results in a loss gained from the other keys. During the first set of trials, the penalty key operates 100% of the time but on the second set it ope-ates 50% of the time. A number of different dependent variables are computed for this task but the basic variable is the mean number of key presses in the 100% and the 50% conditions.

Internal consistency reliability measures range from 0.53 to 0.93. Again the dependent measures based on difference scores tend to have lower reliabilities than scores based on number of key presses. Validity coefficients based on a pass/fail training criterion ranged from 0.01 to 0.27 for samples of Army helicopter pilots and Navy Observers. The highest validity coefficient (r=.30) was obtained on a sample of 44 Observers for whom proficiency ratings were obtained. No validity results are yet available for the NAVCALC test. Parallel form reliabilities were 0.86 and 0.89 on a sample of 32 university students. Project A Figural Reasoning test was originally developed in a paper-and-pencil format. This test has been computerized for the ECAT battery but the validity coefficients are based on results from the paper-and-pencil version of Figural Reasoning. Again with the exception of the partial 0.15 coefficient for the 311 Army trainees in the anti-tank gunnery simulator, the validity coefficients are from the Busciglio (1990) report. The partial coefficient (r=.12) for the Hands-On test is in the same range as the partial coefficient for the anti-tank gunnery simulator. The Sequential Memory test internal consistency reliabilities are in the 0.85 range. This test has been validated in three Navy ratings: Aviation Ordnancemen, Gunner's Mate, and Operations Specialist. The validities for these three ratings were 0.19, 0.27, and 0.26, respectively. The criterion for each of the three validities was final training grades. No results are available for the BARB Ballistic Tasks. Only internal consistency data are available for the Who test and values exceeding 0.90 are reported.

CONCLUSION

In conclusion, the main features of computerized testing have been successfully developed and tested. While paper and pencil tests yield good measures of general G, and also good measures of the basic primary mental abilities which were determined several decades ago, computerized tests make it possible to develop new types of dynamic spatial and memory tests along with precise measurement of psychomotor abilities. The marriage of computer technology and cognitive psychology is and will continue to provide new approaches to the measurement of individual differences and will enhance our ability to measure those differential aspects of human abilities which relate to different types of jobs and job tasks. However, because computers cost more than paper and pencils, it is incumbent upon those of us in this field to develop new methods of showing the economic benefits of computers. Satellite transmission of data across large expansive geographic areas, item generation of new test items, shorter testing times, and more precise measurement of individual differences represent logistical advantages of computers in the operatonal testing environment. With advancement in modern weapons systems and the complexity of abilities represented in these systems, computer testing technology will be required for assignment of individuals to these jobs.

Differential Item Functioning Analysis for Computer-Adaptive Tests and other IRT-Scored Measures[1]

Rebecca Zwick, Dorothy Thayer, and Marilyn Wingersky
Educational Testing Service

1. Overview

The introduction of computer-adaptive tests (CATs) requires that new approaches be developed for analyzing item properties, including differential item functioning (DIF). The purpose of our project was to investigate whether existing DIF analysis methods could be modified to accommodate the data collected in a CAT. Our study also yielded information about DIF methods for nonadaptive tests that are scored using item response theory (IRT).

DIF detection may be *more* important for CATs than it is for nonadaptive tests. First, because fewer items are administered in a CAT, any item flaw may be more consequential for the examinee than it would have been in a nonadaptive testing format. Also, administration of a test by computer creates several potential sources of DIF that are not present in conventional tests, such as differential computer familiarity, facility, and anxiety, and differential preferences for computerized administration. Legg and Buhr (1992) and Schaeffer, Reese, and Steffen (1992) both report ethnic and gender group differences in some of these attributes.

In our study, we used simulated CAT data to investigate the feasibility of conducting DIF analyses using the Mantel-Haenszel (MH; 1959) approach of Holland and Thayer (1988) and the standardization method of Dorans and Kulick (1986) by matching examinees on expected true score for the entire 75-item CAT pool (computed using the estimated item parameters and the estimated ability from the 25 CAT items). To disentangle the effects of assigning items via the CAT algorithm on one hand and matching examinees on expected true score on the other, we also conducted an analysis in which expected true scores computed with the ability estimates from responses to all 75 items were used for matching prior to DIF analysis. The results of this analysis were compared to the results obtained by matching on the CAT-based score and to results obtained by matching on number-right score, as in conventional MH and standardization analysis. In this brief summary document, we focus on this portion of our analyses. In addition, we present only the MH results. In general, results from the standardization method were similar.

2. Simulation procedures

The design of the simulation had three main components: determination of the administration conditions, definition of the properties of the simulated CAT, and specification of item parameters. These components are described in the following sections.

[1] We would like to thank Paul Holland, Charles Lewis, Neil Dorans, Nancy Petersen and Ming Mei Wang for providing valuable consultation.

2.1 Administration conditions

Eighteen data sets were created, each corresponding to a CAT administration. The administrations were defined by the properties of the item pool, the ability distributions of the reference and focal groups, and the group sample sizes. These factors are described below. The number of levels of the three factors was 3, 3, and 2, respectively, resulting in 18 distinct data sets.

Item pool: Three item pools were included. Pool 1 had no DIF; its purpose was to allow investigation of the functioning of the DIF methods in the null case. Two types of DIF pools were included: Pool 2 had DIF that was uncorrelated with item difficulty, and Pool 3 had DIF that was positively correlated with item difficulty. Research has found that, for some pairs of ethnic groups, DIF is correlated with item difficulty, while for male-female analyses, it tends not to be. Pools 2 and 3 were created to allow investigation of the impact of this correlation. The item difficulty, discrimination, and guessing parameters were the same across all three pools of items; only the DIF properties varied.

Focal group ability distribution: Each DIF analysis involves two population groups--the group of primary interest, or *focal* group and the comparison, or *reference* group. The three possible focal group distributions were N(-1, 1), N(0, 1), and N(+.5, 1). In each case, the reference group had a N(0, 1) distribution.

Group sample size conditions: Two sample size conditions were included: $n_R = 500$, $n_F = 500$; and $n_R = 900$, $n_F = 100$, where n_R and n_F are the sample sizes for the reference and focal groups, respectively.

2.2 CAT simulation

In simulating the CAT data, item responses were generated based on the true item parameters, using the three-parameter logistic (3PL) item response function,

$$P_j(\theta) = c_j + (1 - c_j)\ (1 + \exp(-1.7a_j(\theta - b_{jG})))^{-1}, \qquad (1)$$

where $P_j(\theta)$ is the probability of answering item j correctly for examinees with ability θ, a_j and c_j are the discrimination and guessing parameters, respectively, b_{jG} is the difficulty in group G (G = reference or focal).

Our study was based on fixed-length CATs of 25 items, selected from one of the three pools of 75 items. The CAT simulation was designed as a simplified version of actual CATs being developed at Educational Testing Service. The algorithm that was used (implemented in a revised version of a program written by Martha Stocking based on the approach of Lord, 1976) selected as the next item to be administered the most informative item at the maximum likelihood estimate of ability computed from the items already administered. The item information function is defined as $P'_j(\theta)^2/P_j(\theta)Q_j(\theta)$, where $P_j(\theta)$ is the item response function (here, the 3PL function in equation 1), $P'_j(\theta)$ is the first derivative of $P_j(\theta)$ with respect to θ, and $Q_i(\theta) = 1 - P_i(\theta)$ (see Lord, 1980). Estimates of item information and examinee ability were computed using the estimated item parameters obtained from (nonadaptive) simulated reference group data using LOGIST (Wingersky, 1983; Wingersky, Patrick, & Lord, 1988). Most actual CATs under development select items on the basis of both information and other characteristics, such as item format and content.

2.3 Specification of item parameters

Within each of the 18 "administrations," the factors that were varied were the item discrimination (a) and difficulty (b) parameters, and the item d parameters, representing the degree to which the reference and focal group item difficulties differed. Based on analyses of actual SAT and GRE data sets, we selected the following values of the parameters for inclusion:

$ln\ a$: -.3, 0 (corresponding to a values of .74, 1)
b: -1.95, -1.3, -.65, 0 .65, 1.3, 1.95
d: -.70, -.35, 0, .35, .70 for Pools 2 and 3; $d = 0$ for all Pool 1 items.

Within each pool, the joint frequency distribution of these parameters was modeled using a multivariate normal distribution. Therefore, more extreme parameter values were less likely to occur than more central values. Item guessing (c) parameters were set to .15 for all items in all pools.

3. The Mantel-Haenszel DIF Procedure

In the MH method of DIF analysis, examinees are first grouped on the basis of a matching variable that is intended to be a measure of ability in the area of interest. In most DIF applications, the matching variable is a total test score. The score on the studied item, group membership, and the value of the matching variable for each examinee define a 2 x 2 x K cross-classification of examinee data, where K is the number of levels of the matching variable. This 3-way classification forms the basis of the MH procedure. Let T_k denote the number of examinees in the kth level of the matching variable. Of these, n_{Rk} are in the reference group and n_{Fk} are in the focal group. Of the n_{Rk} reference group members, A_k answered the studied item correctly while B_k did not. Similarly C_k of the n_{Fk} matched focal group members answered the studied item correctly, whereas D_k did not. The MH measure of DIF is defined as

$$MH\ D\text{-}DIF = -2.35\ ln(\hat{\alpha}_{MH}) \tag{2}$$

where $\hat{\alpha}_{MH}$ is the Mantel-Haenszel conditional odds-ratio estimator given by

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k\ D_k/T_k}{\sum_k B_k\ C_k/T_k}\ . \tag{3}$$

An estimated standard error for *MH D-DIF* is given in Holland and Thayer (1988).

Our matching variable for the DIF analysis of the CAT-administered items was obtained by (1) getting the examinee's maximum likelihood estimate (MLE) of ability, based on the responses to the 25 CAT items and (2) using this MLE, along with the estimated item parameters, to compute an expected true score on all pool items by summing the 75 valuus of the estimated item response functions. That is, the matching variable was

$$\textit{Expected true score based on CAT} = \sum_{j=1}^{75} \hat{P}_j\left(\hat{\theta}_{CAT}\right), \tag{4}$$

where $\hat{P}_j(\cdot)$ is an estimate of the function defined in equation 1 and $\hat{\theta}_{CAT}$ is the MLE of ability based on the CAT items. Examinees whose expected true scores fell in the same one-unit intervals were considered to be matched.

4. Comparison of CAT-based and Nonadaptive DIF Analyses

For simulation conditions 4, 6, 10, 12, 16, and 18 (see Table 1), we compared MH and standardization results from the CAT analyses, described above, to results of two nonadaptive DIF analyses. The first was a procedure ($\hat{\theta}$-75) in which all 75 pool items were "administered" and examinees were matched on expected true score calculated using the MLE of ability based on all 75 responses. That is, instead of the matching variable in equation 4, the matching variable was

$$\textit{Expected true score based on all 75 items} = \sum_{j=1}^{75} \hat{P}_j\left(\hat{\theta}_{75}\right), \tag{5}$$

where $\hat{\theta}_{75}$ is the MLE of ability based on all 75 items. The second approach (NR) was a conventional DIF analysis, in which all 75 pool items were adminis:ered and examinees were matched on number-right score.

Table 1
Description of Simulation Conditions 4, 6, 10, 12, 16, and 18

Condition	Focal population	Sample size per item		Pool
		Focal	Reference	
4	N(-1,1)	500	500	2
6	N(-1,1)	500	500	3
10	N(0,1)	500	500	2
12	N(0,1)	500	500	3
16	N(.5,1)	500	500	2
18	N(.5,1)	500	500	3

For each of the six selected simulation conditions, the correlation matrix was computed for four variables: the three types of DIF statistics and the "true DIF" for the item. For purposes of this analysis, true DIF was defined as the product of the item discrimination parameter (a) and the difference between the item difficulties for the reference and focal groups (d). (The theoretical basis for defining true DIF in this way is based on certain Rasch model findings, detailed in our full paper.) Each correlation matrix was based on the 71 items that were administered in the CATs. (Four of the 75 CAT items in each pool were never administered because, at every ability level, there were at least 25 items that were more informative than these items.)

Because of an estimation procedure we used for the CAT-based *MH D-DIF* statistics (detailed in our full paper), the CAT DIF statistics were much more precisely determined than the nonadaptive DIF statistics. To avoid giving a spuriously inflated impression of the performance of the CAT analyses, we computed correlations that were corrected for unreliability. These corrected correlations provide a more equitable way of comparing the CAT, $\hat{\theta}$-75, and NR analyses.

Both uncorrected and corrected intercorrelations of the three types of *MH D-DIF* statistics and the true DIF are given in Table 2 for each of the six conditions. The median across conditions is also given. The CAT, $\hat{\theta}$-75, and NR analyses produced results that were highly correlated with each other and with the true DIF values. In particular, the two analyses based on all 75 item responses produced virtually identical results. The median corrected correlation with true DIF was about the same for the CAT, $\hat{\theta}$-75, and NR analyses, which is somewhat surprising since the CAT DIF approach matches examinees on the basis of only 25 item responses. The near-unity correlations of the CAT DIF statistics with true DIF was a welcome finding.

High correlations alone, however, do not ensure the accuracy of the DIF methods. To determine whether the obtained statistics had the desired means, we computed, for each analysis strategy in each simulation condition, the mean *MH D-DIF* value across items. The means for the nonadaptive procedures were quite close to their nominal value of zero; the means for the CAT procedure were slightly inflated, ranging from .00 to .05, with somewhat larger means for the Pool 3 conditions than for the Pool 2 conditions. However, the practical implications of an inflation of .05 or less in the *MH D-DIF* statistic are small in that a difference this size is unlikely to have much effect on decisions about the item. Of course, it would be possible to rescale the statistics so that they would be centered on zero for a particular collection of items.

5. Summary and Discussion

The CAT-based DIF statistics were found to be highly correlated with true DIF and with DIF measures based on nonadaptive administration. Furthermore, the mean DIF statistics for each pool were close to their nominal value of zero, although the CAT-based statistics showed a slight inflation, particularly for Pool 3. Though some aspects of the results still need to be explored, our findings, in general, appear to provide good news for the testing programs that wish to establish DIF screening procedures for CATs.

Our study could not, of course, provide any data on the appropriateness of using item parameter estimates obtained through paper-and-pencil or nonadaptive computer administration to estimate item information and examinee ability in a CAT setting. If administration mode or item order and context affect the functioning of items (see Zwick, 1991), CAT-based ability estimation and hence DIF estimation will be impaired in this situation.

A useful finding concerning IRT-scored nonadaptive tests, was that in nonadaptive administration of 75 items, matching on the expected true score based on the MLE of ability led to essentially the same results as matching on number-right score. The similarity between these approaches, however, may be substantially less for shorter tests.

Table 2

Correlations for Three Types of *MH D-DIF* Statistics and True DIF (*ad*)

Matching Variables		Type of Correlation	Condition						Median
			4	6	10	12	16	18	
$\hat{\theta}$-CAT		Uncorrected	.83	.88	.89	.88	.91	.89	.89
	$\hat{\theta}$-75	Corrected	.93	1.00[b]	1.00	.97	1.00[b]	.99	.99
$\hat{\theta}$-CAT	NR	Uncorrected	.85	.87	.89	.86	.90	.90	.88
		Corrected	.96	.99	.99	.96	1.00	1.00[b]	.99
$\hat{\theta}$-CAT	*ad*	Uncorrected	.96	.95	.98	.96	.99	.96	.96
		Corrected	.97	.96	.99	.97	1.00	.97	.97
$\hat{\theta}$-75	NR	Uncorrected	.99	.99	.99	.99	.99	.99	.99
		Corrected	1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]	1.00[b]
$\hat{\theta}$-75	*ad*	Uncorrected	.84	.86	.88	.85	.90	.88	.87
		Corrected	.93	.97	.98	.93	.99	.98	.97
NR	*ad*	Uncorrected	.86	.87	.88	.84	.89	.89	.88
		Corrected	.95	.98	.98	.92	.99	.98	.98

[b] Corrected value was greater than unity. This can occur if reliability is underestimated.

References

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355-368.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test Validity*, pp. 129-145. Hillsdale, NJ: Erlbaum.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice*, 11, 23-27.

Lord, F. M., (1976). A broad-range tailored test of verbal ability. In C. L. Clark(Ed.). *Proceedings of the First Conference on Computerized Adaptive Testing*. Washington, DC.

Lord, F. M., (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.

Schaeffer, G., Reese, C., & Steffen, M. (August 3, 1992). *Field test of a computer-based general test*. Draft GRE Board Professional Report.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (ed.), *Applications of item response theory*. Vancouver: Educational Research Institute of British Columbia.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1988). *LOGIST user's guide: LOGIST Version 6.00*. Princeton, NJ: Educational Testing Service.

Zwick, R., (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, 10, 3, 10-16.

October 7, 1992

A Psychometrically Sound Cognitive Diagnostic Models: Rule Space

Kikumi Tatsuoka
Educational Testing Service

A Brief Summary of the Rule Space Model

An alternative approach to cognitive diagnosis -- contrast to the traditional bug analyses -- is the rule space model which is a probabilistic approach whose purpose is to identify the examinee's state of knowledge or cognitive states, based on an analysis of the task's cognitive requirements.

Having specified the task's cognitive requirement (also called attributes), an incidence matrix Q (K x n) (the number of attributes x the number of items) is constructed, which is an incidence matrix of item characteristics with respect to the underlying cognitive processes involved in each item. Cognitive patterns represented by K binary elements of unobservable atttributes that can be derived from the incidence matrix Q are called cognitive states (or attribute patterns). Boolean Discriptive Functions (BDFs) are used to systematically determine those cognitive states and map them into observable item score patterns (called ideal-item score patterns)(Tatsuoka, 1991; Varadi & Tatsuoka, 1989). It is assumed that an item can be answered correctly if and only if all the attributes involved in the item have been mastered. Unobservable performances on the attributes can be viewed analogously to an unobservable electric current running through various switches if they are closed. A closed switch corresponds to an attribute that has been mastered. All switches in a current must be closed in order for the current to go through. The cognitive states are represented by a list of mastered/not mastered (or "can/cannot") attributes. The increase of the number of states is combinatorial, but Boolean algebra is a useful tool for dealing with the problem of combinatorial explosion. Boolean algebra, which has been widely used for explaining various properties of electricity and combinatorial circuits have been utilized within the rule space framework for explaining the cognitive requirements underlying test performances.

Once the cognitive states (ideal-item-score patterns) are determined, the actual data are considered. The task now is to map the actual item response patterns of the examinees onto the cognitive states, i.e., to find the ideal-item-score pattern closest to the student's actual response pattern. Since the performance on test items usually includes slips or random errors, the observed item-response patterns are likely to deviate to some extent from the ideal-item-score patterns represented by the various cognitive states. Thus one is faced with a pattern classification problem which is handled by the rule space model (Tatsuoka & Tatsuoka, 1987). The model formulates the classification space and procedures. Item Response Theory (IRT) is utilized for formulating the classificaation space, which is a Cartesian product space of IRT ability θ and variable ζ which measures the unusualness of item score patterns (Tatsuoka, 1985) The cognitive states as well as the students' item response patterns are mapped as points in the classification space by computing their θ and ζ values. Tatsuoka (1990) has shown that the swarm of mapped "fuzzy" points of students'item-response patterns follows approximately a multivariate normal distribution with the centroid being a given cognitive state. Bayes' decision rules are applied for the final classificatiion and for computation of error probabilities.

Once this classification has been carried out, one can indicate with a specified probability level which attributes a given examinee is likely to have mastered or failed to master. If classification rates are as high as 90 % or above, then the attribute mastery patterns can be used for statistical analyses. For example, a factor analysis can be applied to examine the dimensionality of attributes, or a discriminant analysis can be used for investigating subgroup differences if the demographic information is available.

Illustration of the model with an Example

I. A Task Analysis and Attributes: SAT Mathematics

1.1. A mapping sentence. The cognitive requirements for solving the mathematics items were specified using data from two protocols. In order to summarize the content and process categories identified in the protocol analysis, a mapping sentence (Guttman, 1991; Tziner, 1987) was designed. The mapping sentence included 13 facets with a varying number of elements in each. Before presenting the mapping sentence, a word of caution is in order. The mapping sentence presented in Figure 1 is a preliminary one. By no means do we contend that it is complete or exhaustive. More insight into the cognitive requirement underlying the SAT-M items needs to be gained by a comprehensive protocol analysis on several forms of the SAT before a complete cognitive model can be constructed.

Insert Figure 1 about here

Every item in the test can be expressed as a combination of elements from the facets of the mapping sentence. For example: Item no. 1, "if 2x - 6 = 10, then 3x - 6 =____ , (A) 0, (B) 8, (C) 11, (D) 18, (E) 24 " can be expressed in terms of the above mapping sentence as the following combination of facet elements: A3.1.1, B1, C2, D1, E2, F1, G1, H1, I2.1, J2, L2, M1.

1.2. Making an incidence matrix. Fourteen elements from the mapping sentence were selected and expressed as attributes to be used in the rule space analysis. Their brief descriptions are: 1) Basic knowledge and skills in Arithmetic, 2) in elementary algebra, 3) in advanced algebra, 4) in geometry, 5) word problems, 6) comparison problems, 7) able to recall and apply knowledge and rules, 8) can solve equations, 9) can choose and apply theorems and properties, 10) reasoning skills, 11) analytic thinking skills, 12) can follow instructions and comprehend graphs, charts and figures, 13) practical and spontaneous wisdom, 14. can solve complex, multi-step problems.

An incidence matrix Q (60 items by 14 attributes) was constructed for SAT using the above mentioned attributes. Table 1a presents a part of Q matrix (25 items from the section 1) along with the percentage correct and IRT item difficulty parameters b's. Table 1b shows the results of regression analyses, predicting item difficulties for Items 1-25 from 14 attributes. R^2s are $.83(.59_{adj})$ for the percentage correct and $.91(.79_{adj})$ for IRT b-values.

Insert Tables 1a and 1b about here

A computer program RULESPACE (Tatsuoka & Varadi, 1989) produced 2461 ideal-item-score patterns for the incidence matrix of the order of 14x60 in

Table 1.2. Then the program classified 2335 examinees who took the SAT Mathematics test into one of 2461 cognitive states. Since the squared Mahalanobis distance in this case follows a Chi-square distribution with 7 degrees of freedom, $\chi^2 - 2.76$ (p-.01) is set as the first criterion for whether or not a student's response pattern can be classified into a cognitive state. It turned out that 98 % of the 2335 examinees qualified according to the first criterion, thus were classified into one of 2461 cognitive states. The examinees who were not classified are mostly very high scoring students and their θ values are larger than 2.5. After Bayes' rule was applied for the final classification, 387 cognitive states become non-empty, with 87 states having at least 5 examinees classified, 36 states having at least 10 classified. The Figure 2 shows the classification results with the most important states along the SAT Scale and the corresponding IRT θ values.

Insert Figure 2 about here

As can be seen in the results of the regression analyses and classification, Attributes 3 (advanced algebra), 9 (selection and application of theorems and properties), 11 (analytical thinking skills, cognitive restructuring) and 14 (complex problems) are difficult attributes. High ability students are loosing scores because of Attributes 3, 9 and 11 while many average students have trouble with complex problems.

References

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. Journal of Educational Statistics, 12, 55-73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnoses. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1991). Boolean algebra applied to determination of the universal set of knowledge states. Technical Report- ONR-1, (RR-91-4). Princeton, NJ:Educational Testing Service.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and pattern classification. Psychometrika, 52, 193-206.

Varadi, F. & Tatsuoka, K. K. (1989). BUGLIB. Unpublished computer program. Trenton, New Jersey.

Acknowledgements

This research was sponsored in part by the Cognitive Science Program, Cognitive and Neural Science Division of the Office of Naval Research, under the Contract No. N00014-90-J-1307, R&T 4421559, and was in part by the College Board, Educational Testing Service.

COMPUTERIZED MASTERY TESTING

Charles Lewis
Educational Testing Service

Abstract

Combining item response theory and Bayesian decision theory, it is possible to develop sequential rules for computer administration of mastery tests. In the framework adopted, a pool of parallel testlets is developed and randomly selected testlets are administered to examinees. After each testlet, depending on the examinee's cumulative score, one of three decisions is made: Pass the examinee immediately; Fail the examinee immediately; Administer another, randomly selected testlet. In a simulated application of this approach to a professional certification examination, it is shown that average test lengths can be reduced by half without sacrificing decision accuracy.

Introduction

Testing of examinees is carried out for a variety of purposes. One of these purposes is associated with mastery testing (see, for instance, Lord, 1980, Chapter 11), a procedure for deciding, on the basis of test performance, whether an examinee has attained a specified level of knowledge, or mastery, of a given subject. Mastery testing is applied to assist in licensing or certifying an examinee as being competent to perform some task or practice some profession. In an educational setting, mastery testing is used to help decide whether a student has learned the prescribed material sufficiently well to begin with subsequent study.

Item response theory (IRT) provides a framework within which we are able to define what is meant by a master or a nonmaster in terms of amount of knowledge, and to relate this to expected performance on the individual items making up a test. Following an approach described, for instance, by Lord (1980, p.163), two values on an IRT-based knowledge scale may be identified: θ_m, the lowest level at which an examinee would still be considered a master, and θ_n, the highest level at which an examinee would still be considered a nonmaster. The interval between these two values is a region where there is no clear consensus as to the mastery status of an examinee.

Bayesian decision theory (see, for instance, Lindley, 1971) provides a framework for actually classifying an examinee as a master (passing) or nonmaster (failing), based on test performance. In particular, numerical losses may be specified for each of the two types of classification error: passing a nonmaster (A) and failing a master (B). Decisions should then be made which minimize average loss, given the examinee's test score. Cronbach and Gleser (1965) were among the first to provide a decision theoretic framework for mastery testing.

If a mastery test is to be administered with the aid of a computer, an additional refinement is possible: The examinee's performance may be monitored as the test progresses, and early decisions made for those examinees with relatively unambiguous performance levels. This idea is referred to in the statistical literature as sequential testing (see, for instance, Wetherill, 1975) and has found wide application in fields such as quality control. Ferguson (1969) describes an early application of sequential procedures to mastery testing. More recent discussions of related methods are provided by Kingsbury and Weiss (1983) and by Reckase (1983).

Testing Design

Details of the approach described here are given by Lewis and Sheehan (1990). The crucial assumption we make is that it is possible to use the pool of available items 'o construct short, nonoverlapping, equivalent tests, each of which meets all content specifications for the test as a whole. Such short tests are sometimes referred to as testlets (Wainer and Lewis, 1990).

Computer-based test administration begins with the random selection of one or more of these testlets. When the examinee has completed all the initially selected testlets, one of three decisions is made. In addition to the usual pass and fail decisions, in line with the idea of sequential testing, we also consider the possibility of deferring these decisions and administering another, randomly selected testlet. If the decision is made to continue testing, the three options are considered once more when the examinee has completed the new testlet. This process continues until the examinee has passed or failed, or until a preset maximum number of testlets has been administered. In the latter case, a final pass/fail decision must then be made.

Although IRT allows optimal extraction of information regarding the amount of knowledge possessed by an examinee through the analysis of the examinee's complete pattern of item responses, the approach we adopt employs only cumulative number-correct scores for decision-making purposes. This choice was made primarily to simplify both our analyses and the final results. At each decision point there are two cut scores, expressed in terms of cumulative number-correct scores for all testlets completed so far. All examinees with scores greater than or equal to the upper cut score pass immediately. Those with scores less than the lower cut score fail the test. The remaining examinees are administered another testlet. The one exception to this rule is for examinees who have completed the maximum number of testlets allowed. For them, there is only a single cut score, which is used to distinguish passing from failing examinees in the usual manner.

Determination of Cut Scores

The basic principle behind Bayesian decision theory is to make the decision which minimizes the average loss. In order to apply this principle to sequential testing, we must associate a loss (or cost) with obtaining additional information, in our case, administering an additional testlet. (If there were no such loss, it would always be better to continue testing than to pass or fail an examinee on the basis of incomplete information.

Consequently, we are dealing with three numerical losses which must be specified before decision rules can be identified: A, the loss associated with passing a nonmaster, B, the loss for failing a master, and C, the loss for giving a testlet. The units in which these losses are specified are arbitrary, but it is often convenient to set C=1 and, thus, express A and B as multiples of the testlet administration loss.

In practice, we consider a range of choices for A and B, evaluating each choice through the use of simulations to obtain estimates of correct classification probabilities and average test length. Lower values of A and B result in more misclassifications and shorter tests. Higher values lead to more correct classifications and longer tests. The relative sizes of A and B are also of interest. When A is greater than B, for instance, it means that it is a more serious error to pass a nonmaster than it is to fail a master. This has the effect of raising cut scores and, thus, lowering passing rates.

For a given choice of A and B (with C=1), we work backward to obtain the set of cut scores. Starting with the last testlet that could be administered, we adopt a decision rule which passes an examinee whenever the probability that the examinee is a master, given their cumulative number-correct score on all testlets, is at least equal to the ratio A/(A+B). This choice is guaranteed to minimize the expected loss. The decision rule can be summarized with a single cut score and with the minimum expected loss associated with each possible score (sometimes referred to as a risk function).

The probability of mastery, given the examinee's score, which we use to identify the decision rule, is obtained via IRT (which is used to compute the probability of the score, given that the examinee is a master) and Bayes' Theorem. A prior probability of mastery is also needed to use Bayes' Theorem. In our work, we have used a prior probability of .5, based on fairness considerations, but historical data on the proportion of (classified) masters in the examinee population could also provide a basis for the prior. In any event, as with most Bayesian analysis, it is appropriate that the data (test scores) should dominate any prior information about the examinee's mastery status.

Proceeding with the cut score determination, we now take one step backward to the testlet preceding the final possible one. The expected losses for passing and for failing are determined as with the final testlet, except that the loss associated with testing is reduced by 1 (since one fewer testlet would be administered if a final decision were made at this point). In addition, we must now determine the expected loss associated with continuing to the final testlet.

To do this, we consider all possible scores on the last testlet, use the results from the last stage to compute the expected loss associated with the decision based on each such score, and average these values, weighting by the probability that the examinee would obtain that score, given the score achieved so far. For each of the possible scores achieved before the last stage of testing, the three expected losses are compared, and the minimum identified. This leads to the identification of an optimal decision for each possible score, and this set of decisions may be summarized with two cut scores: a higher value, the lowest score leading to a pass decision, and a lower value, the lowest score leading to a decision to continue testing. In addition, we retain the risk function, which gives the expected loss associated with the optimal decision for each score.

These same steps are repeated for each stage of testing, continuing to work backward until we reach the first stage at which a pass/fail decision may be made. We then have a complete summary of the decision process, consisting of a set of cut scores and the risk functions which indicate how well we have done in terms of minimizing expected loss.

Evaluation of the Sequential Testing System: An Illustration

The first application of the approach described above was to a component of the Computerized Architectural Registration Examination, developed under the auspices of the National Council of Architectural Registration Boards (NCARB). An available pool of multiple choice items testing Seismic Knowledge was used to construct six 10-item testlets and it was decided that a minimum of two and a maximum of six testlets would be administered to an examinee.

All items were calibrated using the three-parameter logistic item response model, and an estimated distribution of knowledge levels in the examinee population was obtained as well. Mastery and nonmastery levels were identified on the knowledge scale, based on information available from previously existing tests. After evaluating a number of alternatives, the loss associated with passing a nonmaster was identified as A=40 times that associated with administering a testlet, while the loss for failing a nonmaster was set at B=20 times the loss for a testlet. This asymmetry was judged to be appropriate, given the nature of the examination.

Using the procedure outlined in the previous section, a set of cut scores was determined. These values are given in Table 1. Examinees scoring at least 16 out of a possible 20 after the second stage of testing were passed immediately. Those scoring 11 or fewer correct were failed immediately. The remaining examinees continued to the third stage. For examinees completing the sixth stage, those with number-correct scores of 39 or higher, out of a possible total of 60 items, are passed while all others are failed.

Table 1
Number-Correct Cut Scores for a Test with Between Two and Six 10-Item Testlets

Stage	Items	Lower	Upper
2	20	12	16
3	30	18	22
4	40	24	28
5	50	31	34
6	60	39	39

Employing the IRT item calibrations and the estimated distribution of knowledge levels, examinee performance was simulated on the sequential test whose cut scores are given in Table 1, as well as on a conventional test consisting of all six testlets and using the final cut score for the sequential test to make pass/fail decisions. The average test length, pass rate, and proportion of examinees misclassified (both false positives and false negatives) were obtained for each test. The results are summarized in Table 2

Table 2
Simulated Performance Characteristics for a Sequential and a Conventional Test

Test Type	Average Length	Pass Rate	False Positive	False Negative
Sequential	27.7	.640	.026	.082
Conventional	60.0	.658	.027	.066

In the simulated examinee population, approximately 70 percent were masters. Note that both tests have pass rates below that level and that both have more false negative classifications than false positives. This is a result of using an asymmetric loss specification, in which passing a nonmaster was judged more serious than failing a master.

The classification performance of the two tests is quite similar, with a slight advantage in favor of the conventional test. The most striking difference is in average test length. The sequential test is able to identify nonmasters equally well and masters almost as well as a test using, on average, more than twice as many items. This finding is in general agreement with the statistical literature on sequential testing.

To summarize, the research described above suggests that, when a computer-based test delivery system is available, it is possible to construct a variable-length sequential mastery test which will substantially reduce average testing time and item exposure rates with a minimal loss in classification accuracy compared to a conventional fixed-length test.

Referencea

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana IL: University of Illinois Press.

Ferguson, R. L. (1969). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburgh PA: University of Pittsburgh Learning and Research Development Center. (ERIC Documentation Reproduction No. ED 037 089)

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-283). New York: Academic Press.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386.

Lindley, D. V. (1971). *Making decisions*. London and New York: Wiley-Interscience.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wetherill, G. B. (1975). *Sequential methods in statistics*. London: Chapman & Hall.

Some Impacts of Downsizing on Recruiting

by

W. S. Sellman
Office of the Assistant Secretary of Defense
(Force Management and Personnel)

Because of the end of the Cold War and the demise of the Soviet Union, the United States is reducing the size of its military by 25 percent. The stated plan is to go from 2.2 million Service members in Fiscal Year 1987 to 1.6 million in Fiscal Year 1995. The United States has reduced force size in the past, but generally at the end of an armed conflict and when the force was composed largely of conscripts. This downsizing is a first in that today's Services consist of volunteers, men and women who wanted to serve and most of whom planned to make the military a career.

To drawdown the force, the Department adopted a four-part strategy: (1) Reduce the number of new accessions to levels required to sustain the force at the post-Fiscal Year 1995 level. As a result, the Services will recruit about 225,000 young men and women in Fiscal Year 1993 instead of the 330,000 of 5 years ago. (2) The Department will control reenlistment into the career force to ensure high quality people in shortage skills move successfully into their second term. (3) For members with 6-19 years service, the Department has offered monetary incentives for them to separate voluntarily. These incentives are available to people with skills or in year groups that are overstrength. (4) For members with more than 20 years service, the Department will identify individuals for early retirement based on past performance, professional attributes, and potential for future service.

Today, I want to briefly discuss the first part of the drawdown strategy--reduced recruiting. The size of our recruiting cohorts is shrinking. In Fiscal Year 1989, we recruited 294,000 people; last year, in Fiscal Year 1992, we recruited 203,000. We believe the Fiscal Year 1992 number was too low so we have encouraged the Services to increase their recruiting levels to about 225,000 people in Fiscal Year 1993. Given their budgetary problems, it is tempting for the Services to under recruit because that saves money and allows a gentler policy for reduction of careerists. Of coarse, if you do not recruit enough people up front, it creates a shortage which becomes a bow-wave which moves through the career force and within a 5-year period the Services will not have the NCOs and mid-level officers needed.

Congress also sees the reduction in recruiting as a target of opportunity for budget cuts. Since Fiscal Year 1989, overall recruiting resources have been reduced by 20 percent. Advertising is down by 55 percent, production recruiters by 10 percent with another 10 percent cut scheduled over the next 2 years, and recruiter offices by 20 percent. Congress almost prohibited the Services from advertising on television. The only Services to be on television this year will be the Army and Marine Corps because the Navy, Air Force, and the Joint Recruiting Advertising Program do not have enough money to be on that medium.

How has the recruiting decline affected recruit quality and social representation? Because we are recruiting fewer people, recruit quality has reached historic highs. In Fiscal Year 1992, 99 percent of new recruits were high school diploma graduates, and 75 percent were graduates who scored at the 50th percentile or above on the Armed Forces Qualification Test. Can we sustain those levels of quality? That depends on Congressional budget cuts. Fiscal Year 1993 cuts were relatively benign, at about $25 million. This is an election year, and there were a lot of programs that were virtually untouched because of Congressional need to demonstrate concern about military people leaving service as well as DoD civilians and Defense contractors facing layoffs. Fiscal Year 1994 cuts, no matter who wins the election, could be much worse, and recruit quality could be affected.

There has been a decline in enlistment of Blacks. In Fiscal Year 1989, 22 percent of all new recruits were Black; and in Fiscal Year 1992, that number was only 17 percent. The 17 percent has been steady over the last 18 months. We believe the drop in Black enlistment is a result of several factors. During Operations Desert Shield and Desert Storm, there was much written and discussed in the media about Black overrepresentation, and how young Black Service members were bearing disproportional burdens during war. Much of this was triggered by the President's recent veto of the civil rights act. Black leaders around the country were angry. Overrepresentation was used as an issue to get back at the Administration for failure to pass civil rights legislation. Many prominent Black leaders used it to advance their own agendas. In particular, there was an NBC news broadcast which stated that the Department was targeting Black high schools for over recruitment. That allegation was not true.

Jesse Jackson wrote the President about the disproportional number of Blacks who were in Reserve units being sent to the Persian

Gulf. The facts were that those Selected Reserve units were assigned by mission capability, not social composition. As a result of Reverend Jackson's letter, we provided demographic information on all Reserve units in the United States to the Military Assistant to the President so he would have it available, should more inquiries be received. There also has been negative publicity for the military by a rap group called the Ghetto Boys. This group produced a tape that can be purchased in any music store in the country that has several songs that are profanely anti-military.

Another problem is that young men and women do not believe that the Services still are hiring. People have seen so much publicity about the drawdown that they do not know that we are recruiting. This is not just the American public; this also includes a number of members of Congress. Last year, we received hundreds of letters about why are the Services still recruiting and advertising. It never occurred to us that Congressmen would not understand that we are a closed personnel system and have to recruit from the bottom and train people for senior leadership roles. Had I been more perceptive, we would have changed our advertising campaign to something such as, "Yes we are getting smaller but we still need to hire." Maybe that would have precluded much of the criticism directed at the Department. Today, there also is the belief that the military no longer offers opportunities for upward mobility because of the force drawdown.

Recruitment of other minorities (Hispanics, Asian-Americans, Native Americans) has remained steady over past 5 years at about 5.5 percent. The Services also have had no trouble recruiting women--in Fiscal Year 1992, percentages of new accessions who were women were Army-16 percent, Navy-14 percent, Marine Corps-5 percent, Air Force-22 percent, Department of Defense-15 percent. In terms of officers, there has been a drop in the number of new accessions from 27,000 in Fiscal Year 1989 to 20,000 in Fiscal Year 1992. This number will probably remain constant in Fiscal Year 1993 until we see what the ultimate size of the base force becomes. Down sizing is a growth industry in the Department of Defense and is likely to be so for the next several years.

In sum, by the end of Fiscal Year 1992, the Services have already reduced recruiting by 25 percent. They plan to recruit to sustainment levels over the next 3 years. Minorities are still overrepresented in the recruit cohort when compared to the youth population, but they will not be as seriously overrepresented as in the past. Recruitment of women is at satisfactory levels. Recruit quality is at historic highs, although I expect it to drop somewhat. Finally, if Congress provides adequate resources, we can continue to recruit well qualified men and women to be the leaders of our future military.

Session 15
Symposium: The Incredible Shrinking Force: Issues in Downsizing
CBT: The Time Has Come
(MTA 1992)

Dr. Frank Leo Vicino

1. INTRODUCTION

The present force restructuring calls for revisiting traditional approaches to personnel selection. Wally Sinaiko stated "the restructuring of the military forces is the most profound change of its kind in forty years." This restructuring, that will be causing many ripples in the military personnel world, will also continue to be quite dynamic. This character of change will be with us for some time. <u>Any "Final" Selection and Classification System design, therefore, is made elusive by the ambiguity and changes I foresee in future force makeup</u>.

2. THE SITUATION

These profound changes will be affecting the structure of any future military selection and classification system. You have heard today about many of these changes. Let's take another look at some of them and their possible effect on any future Selection and Classification System and what we hope to do to soften the blow.

New roles and missions are developing and they will demand people who can operate in a more complex and rapidly changing environment than we have ever seen before. The order will be flexible and creative leadership.

Personnel, in a downsized military, will be older and more experienced. In the last three years the average age of naval personnel has increased 10 percent. With this higher overall experience level it should be possible to increase productivity and operate, for example, the same number of ships and squadrons with fewer people.

I agree with the former Army DCSPER when he stated that we have not recognized the inherent versatility and capability of our young service members. In this shrinking force it will be necessary to expand our human capital, to recognize the capability of our people to do more with less. We need to find ways to expand our human capital.

Traditional occupational structures are too narrow to be effective in a leaner force. The Navy has about 100 ratings and there are even more MOS designations in the Army.

3. WHAT CAN WE DO ?

In the absence of anything specific about the composition of future forces we can still anticipate perturbations to our selection and classification practices. For personnel research the implication is clear: any selection and classification system will accommodate rapid changes and, possible, radically new demands. Traditional selection and classification processing, like the traditional roles they represent, may need altering.

Computerized testing is an obvious way to go, given the following expectations. Computers make it possible, among other things, to integrate selection procedures with military personnel data bases, and with manpower decision models.

More specifically, computerized testing can manage testing sequences, administer adaptive tests, change scoring protocols, present dynamic and interactive test scenarios, and more. At present there is in place a major component of computer based testing: The CAT-ASVAB has been operational in several enlistment processing sites for about a year.

New areas not yet integral to computer-based testing are creativity and leadership. Computers make it possible to measure reaction time, complex composite responses, trends in test time responses and responses to complex and stressful situations.

Another major advantage of computer-based testing is its amenability to revising tests. Proposed new measures can be imbedded in a standard test battery to supply researchers new data that does not become part of the applicant's score. The new item data can be assessed and, if found to be useful, can be rapidly incorporated in the score.

Selection and classification needs to have added the outcomes of work in job analysis, manpower modeling, and manpower requirements estimation. Another suggestion by the Air Force ROADMAP program that team and individual performance should be part of selection and classification research appears promising. Army's Projects "A" and "B" and Navy's New Measures programs point to the importance of interactive tests and other new measures

.

TAPSTEM, which is defining the Service's MPT research direction, has made researchers more aware of selection and classification areas that need further resolution. The Navy's CAT-ASVAB and ECAT programs, the Air Force's ROADMAP and LAMP, the Army's projects "A" and "B" and the Marine Corp's job performance measurement research are all providing excellent answers and, at the same time, are helping to define future selection and classification research. The Navy's new SYMONAC program and its proposed STAR program offer promise for integrating findings from TAPSTEM and other programs into a fully computerized personnel system.

STRATEGIC ACCESSION PLANNING
AT THE U.S. ARMY RECRUITING COMMAND (USAREC)

Peter McWhite, Ph.D

HumRRO
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314
(703) 549-3611

McWhite Scientific
451 Hungerford Drive, Suite 600
Rockville, MD 20850-4151
(301) 251-0975

USAREC's Recruiting Operations (RO) controls day-to-day fill of nearly 280 Military Occupational Specialties (MOS) while meeting policy and resource constraints. RO controls the day-to-day MOS eligibility of applicants by their Armed Forces Qualification Test (AFQT) Category (TC) as well high school graduation status. Table I. gives TC ranges used in recruiting.

Table I. AFQT Score Ranges

TC	I	II	IIIA	IIIB	IVA	IVB	V
AFQT Percentile Score Range	93-99	65-92	50-64	31-49	21-30	20-10	1-9
Remarks	Quality Applicants			Needed to fill large, unpopular programs	Entry restricted		Entry prohibited

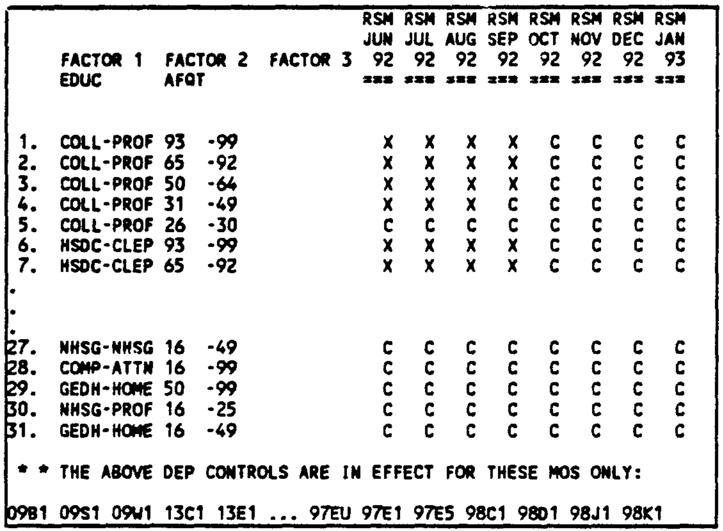
TC I is the top 7 percent of the youth population. They are most readily trained to perform the most complex tasks. TC V represents the most difficult to train. By law, all TC V applicants, and those in TC IV who have not graduated from high school, are disqualified from military service. TC I-IIIA are called quality applicants because they rank in the top half of the population. Some MOS accept only quality applicants.

Accession Timing Control

This control is possibly the most complex in the recruiting process. RO must consider simultaneously *when* and *to what MOS* recruits of a given TC should access. It must look ahead at MOS requirements and recruiting propensities for the current and future FYs.

Timing control is accomplished through the Delayed Entry Program (DEP) using REQUEST's[1] Report/Update of DEP (RUDEP) program. RUDEP acts like a gatekeeper to nongraduate and TC IIIB and IV Army accessions. It is the only process that can have total control of mission ceilings. No applicant will be offered an MOS unless he or she meets its RUDEP accession window and its requirements for education and quality.

Seventeen RUDEP tables control most nonprior service accessions. Figure 1 is a typical RUDEP Table. The *X* and *C* mean that the above Recruit Ship Month (RSM) for an MOS is open or closed, respectively, to applicants with the indicated education and quality attributes.



```
                                           RSM RSM RSM RSM RSM RSM RSM RSM
                                           JUN JUL AUG SEP OCT NOV DEC JAN
    FACTOR 1  FACTOR 2  FACTOR 3            92  92  92  92  92  92  92  93
    EDUC      AFQT                         === === === === === === === ===


 1.  COLL-PROF 93   -99                      X   X   X   X   C   C   C   C
 2.  COLL-PROF 65   -92                      X   X   X   X   C   C   C   C
 3.  COLL-PROF 50   -64                      X   X   X   X   C   C   C   C
 4.  COLL-PROF 31   -49                      X   X   X   C   C   C   C   C
 5.  COLL-PROF 26   -30                      C   C   C   C   C   C   C   C
 6.  HSDC-CLEP 93   -99                      X   X   X   X   C   C   C   C
 7.  HSDC-CLEP 65   -92                      X   X   X   X   C   C   C   C
.

.

.
27.  NHSG-NHSG 16   -49                      C   C   C   C   C   C   C   C
28.  COMP-ATTN 16   -99                      C   C   C   C   C   C   C   C
29.  GEDH-HOME 50   -99                      C   C   C   C   C   C   C   C
30.  NHSG-PROF 16   -25                      C   C   C   C   C   C   C   C
31.  GEDH-HOME 16   -49                      C   C   C   C   C   C   C   C

 * * THE ABOVE DEP CONTROLS ARE IN EFFECT FOR THESE MOS ONLY:

09B1 09S1 09W1 13C1 13E1 ... 97EU 97E1 97E5 98C1 98D1 98J1 98K1
```

Figure 1. A RUDEP Table.

Each Week the RUDEP Noncommissioned Officer (NCO) modifies the timing and MOS content of the RUDEP tables. Some considerations are:

- Current MOS accessions plus DEP fill and fill rate.
- Accession windows for High School Senior (HSSR) and Currently in College (CC) prospective graduates consistent with balancing quality over the summer months.
- Differential and projected DEP loss rates.
- Expected recruiting performance. A small window can inhibit market penetration.
- A consistent and equitable policy that helps recruiters and guidance counselors maintain credibility with their community.
- Small windows that encourage near-term fill for large programs, since they have frequent class starts.

[1] REQUEST is a nationwide time-sharing computer service. It provides centralized recruit qualifications data, training space, and unit reservations. It is USAREC's official management information reporting system.

- Large windows that ensure .ill of small programs that have few class starts.
- Longer DEP periods (from large windows) have high DEP loss rates.

The above process used to require 3 days effort each week by the RUDEP NCO.

The RUDEP Expert System

Since June 1991, RO has developed strategic plans with the microcomputer-based RUDEP Expert System. It addresses the above considerations while developing a weekly RUDEP strategy in a few hours. Tables II and III show the Table configuration and strategy that is the basis for the RUDEP Expert System. (*CAF* is command average percent fill; *Target RSM* is the month(s) planned for most of the accessions that will support that RSM's accession mission.)

Table II. RUDEP Tables and Assigned MOS.

Table	Description	MOS Criteria
1	Seldom Start (ST)	Small programs with < 10 AIT start dates. Supersedes all other categories. AMB selects.
	09S1 09W1 13C1 13E1 13M1 14D1 16D1 16E1 16J1 16R1 23R1 24C1 24G1 24H1 24K1 24M1 24N1 25L1 25P1 25Q1 25R1 25S1 27B1 27E1 27F1 27G1 27H1 27J1 27K1 27L1 27M1 27T1 29M1 29Y1 31Q1 33P1 33Q1 33R1 33T1 33V1 33Y1 35H1 35Y1 39B1 39C1 39D1 39E1 39G1 39L1 41C1 42C1 42D1 43M1 45E1 45N1 45T1 46R1 51G1 51M1 52F1 52G1 55D1 55G1 55R1 57F1 62G1 62H1 63N1 67A1 67H1 67S1 68B1 68D1 68G1 68H1 68Q1 68R1 71M1 73D1 75F1 76J1 77L1 81B1 81C1 81Q1 82B1 82D1 83E1 83F1 88K1 88L1 91C1 91D1 91F1 91G1 91H1 91J1 91L1 91M1 91N1 91P1 91Q1 91R1 91S1 91T1 91U1 91V1 91Y1 92B1 93D1 93F1 96D1 96R1 98C1 98D1 98J1 98K1	
2-3	At CMF or better. (Quick sellers responding to open months.)	Not ST/OSUT/HTQ.
	16P1 16X1 24T1 31M1 31N1 43E1 45G1 45L1 51B1 51R1 52C1 62J1 63D1 63E1 63H1 63J1 63T1 63W1 67N1 67R1 67T1 67U1 67V1 67Y1 68F1 68J1 68N1 74D1 75C1 76V1 76X1 77W1 88H1 88M1 88N1 91A1 93B1 93C1 96B1 96F1 96H1 97B1 97EL 97G1 98C3 98H1 98XL	
4	Selling below CMF.	Not ST/OSUT/HTQ.
	13F1 13M1 13P1 13R1 31K1 38A1 44B1 44E1 45B1 45D1 52D1 62B1 62E1 62F1 63Y1 73C1 75B1 75D1 76Y1 82C1 91B1 91E1 93P1 96U1 98G3	
5-6	OSUT	Designated OSUT.
	11C1 11H1 11M1 11X1 12B1 12C1 13B1 19D1 19K1 54B1 95B1	
7	Hard to Qualify For (HTQ)	Difficult qualifications: APT scores ≥ 100/110, 2 APT scores ≥ 100
	29E1 29J1 29N1 29S1 29V1 31C1 31D1 31F1 35G1 36L1 36M1 42E1 46Q1 68X1 71D1	
8	Extremely Behind CMF. (EBCF)	Available to all categories.
	12F1 14S1 16S1 31L1 31V1 37F1 45K1 51K1 55B1 57E1 63B1 63S1 68L1 71G1 71L1 74C1 75E1 76C1 76P1 77F1 94B1	

RUDEP Strategies

Table II. summarizes RUDEP strategies for controlling the RSM that MOS types are open to recruit TC and education types. (F is MOS fill, ∆ Fill is program - current fill, One Station Unit Training (OSUT).)

Table III. Recommended RUDEP RSM.

Table	Description	Months Open	Market Strategy
1	Seldom Taught (ST)	**Current RSM to Target RSM(s) (will vary with seat openings)**	TC I-IIIA '92 HSSR
2	Attractive & DQ Tolerant (Quick sellers) F > CAF, ranked by ∆ Fill, stop at 10% available program.	**Target RSM(s) Shift to Next month when Target at 75% of Mission plus excess for DEP loss.**	Closed to HSSR-92. DQ used to stop fill. Save for Fall HSDG.
3	Attractive & DQ Tolerant F > CAF; not on Table 2.	**Target RSM(s)**	Closed to HSSR-92. Fill target month. Open to TC I-IIIA
4	Attractive & DQ Tolerant (Slow fill) F ≤ CAF	**Target RSM(s) + 1 (A's) Target RSM(s) (B's)** (B's may be limited by policy)	Closed to HSSR-92. Open to TC IIIBs
5	OSUT (5 & 6 permit different OSUT fill rates for e.g., 11X1)	**Target RSM(s)** HSSR-92 May-June until 50%, then July, etc.	Push HSSR-92. Only A's.[2]
6	OSUT	"	"
7	Hard to Qualify For (HTQ)	**Target RSM(s) + 1**	Push HSSR-92 Accept all qualified.
8	Extremely Behind F ≤ CAF and: Program ≤ 200 then Fill/CAF ≤ .85 or Program ≤ 100 then Fill/CAF ≤ .7	**Target RSM(s) + 2 HSSR-92 open entire summer (< 12 mos. in DEP).**	"

[2] High quality means DEP losses in summer can be replaced by B's.

Summary of RUDEP Expert System Features

(1) As the RUDEP System directs applicants to enlist for training opportunities in the RSM when are needed, its constraints on DEP are not so restrictive that they limit accessions overall.

(2) The System avoids being so restrictive that it creates an environment that will fail to attract enough recruits. On the other hand, it ensures that RO will meet accession requirements.

(3) The DQ process supports MOS-specific quality allocations; however, it cannot control the annual quality mission. Also, DQ cannot accomplish monthly quality leveling. The RUDEP process accomplishes this by directly controlling entry to the Army, by month, of TC IIIB-IV accessions.

(4) Mission categories need strategic management. Because their availabilities to access are determined by graduation dates, HSSRs and CCs have specific accession windows. Other windows provide accessions during *slow* periods, such as March and April.

(5) Level loading training seats provides applicants with varied opportunities.

Disclaimer.

The assessments of recruiting strategies and policies are entirely the author's. No endorsement by the U.S. Army Recruiting Command is claimed or implied.

ALCOHOL AND DECISION MAKING COMPETENCE

Siegfried Streufert

Pennsylvania State University
College of Medicine, Hershey, PA, 17033

Alcohol consumption during work hours, especially when higher intoxication levels are reached, can have serious consequences for task performance (e.g. Heishman, Stitzer and Bigelow, 1988; Streufert, Pogash, Roache, Gingrich, Landis, Severs, Lonardi and Kantner, 1992). We know quite well that alcohol impairs the competence of truck drivers, airline pilots and operators of machinery, resulting in an inordinate number of accidents. For example, 33% of fatally injured truck drivers tested positively for alcohol and other drug abuse (NTSB, 1990). Our knowledge about deleterious alcohol effects is based on years of research on traffic safety and other settings where intoxication may produce personal injury and/or property damage. Yet we know relatively little about the effects of alcohol (or for that matter the effects of other psychoactive substances) on more complex functioning, including the impact of alcohol on decision making (cf. Streufert et al, in press).

There are at least two reasons for our lack of knowledge: (1) Alcohol abuse at senior levels often remains hidden (Pace, 1989) and may be ignored by the drinker's associates and supervisors. (2) Prior researchers have had difficulty developing research methodologies that allow reliable and valid measurement of complex decision making competency. Both issues will be considered below:

(1) Limits on Knowledge about Alcohol Effects

Some limited information about the impact of alcohol (or other psychoactive substances) upon decision maker performance might be derived from observations of the excessive drinker on the job. Yet, the incidence and severity of alcohol effects upon decision makers are difficult to estimate. Generally, data on rate of abuse and on the effects of abuse are not made available. In some cases, alcohol abuse at very senior levels may even be supported and/or exploited by the afflicted individual's staff (Trice and Belasco, 1970). Lowered performance at senior levels rarely result in formal disciplinary action (Beyer and Trice, 1984). Of course, whenever alcohol abuse is tolerated or remains hidden, performance data will not be collected or, if available, will not be released.

As a result, we must generally rely on conjecture in estimating the impact of alcohol on decision making performance. Observations and a very few experiments seem to suggest that alcohol consumption may diminish performance (e.g. Beeman, 1985; De Clifford, 1985; Helm and Gaffney, 1986; McCann, 1983; Trick, 1984). Risk taking and information recall (Jobs, 1989), reasoning (Beeman, 1985), decision quality (DeClifford, 1985; Jobs, 1989; McCann, 1983; O'Broin 1985) and tendencies toward aggressiveness (Bushman and Cooper, 1990) might be affected. Intoxicated individuals may respond to their task environment with a "myopic" orientation (Steele and Josephs, 1990), i.e., might restrict their focus to simpler and immediately obvious demand characteristics.

Unfortunately, limited opportunity to engage in observation and the few research studies that employed college students rather than experienced decision makers as subjects are hardly adequate sources of needed knowledge. If we wish to persuade the potential drinker to decrease alcohol intake or to abstain during working hours, we need to demonstrate the deleterious effects of alcohol, if any, beyond a reasonable doubt. After all, some researchers (albeit employing tasks much simpler than complex decision making) have been unable to demonstrate performance decrements due to alcohol, especially at lower levels of intoxication (e.g., Lukas, Lex, Slater and Greenwald, 1989). Others have even reported data suggesting improved performance (Mann, Cho-Young and Vogel-Sprott, 1984). When an intoxicated aircraft cockpit crew not long ago flew their plane without incidence to its destination, some argued that alcohol would not affect performance seriously, especially when no emergency is encountered, i.e., as long as "standard operating procedures" remain effective. The present paper considers whether such arguments apply to complex decision making settings.

Any effort to investigate the impact of alcohol on senior level decision making and any associated effort to persuade (or treat) personnel toward greater abstinence during working hours would hardly be worth while if alcohol abuse were relatively rare. For example, one could argue that the ethic of military services to limit alcohol use might be adequate toward preventing excessive drinking. Nonetheless, despite ethical constraints, problem drinkers do exist in higher level job categories, including in the military. Use and abuse of alcohol, in effect, may be more widespread than thought (Martin, 1990). Alcohol consumption is especially extensive among segments of the population (Hilton, 1991; Shore, 1985a,b) that include the majority of senior level decision making personnel: managers in the private sector as well as officers in the military. In fact, access to alcohol is also easier at higher job levels and, by some, the excessive alcohol consumption may even be considered somewhat "justified" as a reducer of "job stress.' As a result, intoxication may be actively supported by an offender's peers (Roman, 1974).

(2) Methodological Limits on Data Collection

Data on the effects of alcohol abuse have also been limited by methodological problems. As Aluisi (1982) suggested, "assessment of human performance is often difficult at best (e.g., in assembly line settings) and almost impossible at worst (in white collar jobs)." Both reliable and valid applied data on senior level functioning have been hard to come by. Such data (with some partial exceptions, e.g., Jobs, 1989) on alcohol effects upon decision making have been rare. Fortunately, Streufert and associates (Streufert, Pogash and Piasecki, 1988a) have developed a validated quasi-experimental simulation methodology (Streufert and Swezey, 1985), that permits the application of experimental procedures to the assessment of decision maker performance. Their simulations, employing up to three parallel scenarios and yielding more than forty validated performance measures that correlate highly across the scenarios (Streufert et al, 1988a), have been used to measure the effects of various drugs on managerial and/or decision maker performance (e.g., Streufert, DePadova, McGlynn, Pogash and Piasecki, 1988b). Predictive (criterion) validity for the simulation system has been demonstrated both in North America and in Europe, obtaining correlations above .6 with decision maker success (e.g., job level at age, number of persons supervised, income at age and number of promotions over a ten year period; Berndt, 1991; Schopf, 1990; Streufert et al, 1988a).

Over the last two years, we have collected data on three different samples of senior decision making personnel. This paper summarizes the results of research concemed with (a) the effects of alcohol intoxication at the .05 and .10 breath/blood alcohol levels on complex decision making performance in general, (b) the effects of alcohol intoxication on competency to make decisions in the face of a serious emergency and (c) effects of a hangover (intoxication at the .10 level the night before task performance) on decision maker competence.

METHOD

Subjects: For all three investigations, managers were recruited via newspaper ads with the offer of a free major medical examination and payment of a stipend ranging from 250 to 350 Dollars. Volunteers with medical morbidity, psychiatric problems, addictions, current drug abuse, excessive weight or inadequate decision making experience were eliminated from the samples. A total of (a) 44, (b) 25 and (c) 20 males completed the three research projects. All procedures followed a double blind placebo-controlled research methodology.

Alcohol Treatment: Alcohol drinks were provided by mixing 10% w/v solution of 95% USP ethanol with tonic water. Drinks contained .05g or 1.0g ethanol (alcohol) per kg bodyweight toward attaining a .05 or .10 breath alcohol level (BAL). Placebo drinks contained only tonic water, however both placebo and alcohol drinks were lightly sprayed with a 10% ethanol solution to provide the impression of an alcoholic beverage. In fact, participants in all three research projects believed that they had been drinking alcohol on both alcohol and placebo treatment occasions. For experiment (a) BALs of .05 and .10 were maintained over several hours; experiment (b) was designed to generate intoxication at the .10 level immediately prior to the experience of a serious (simulated) emergency and experiment (c) employed treatments to attain a .10 intoxication level during the evening hours prior to participants' arrival in the simulation lab on the following moming.

Simulation: Subjects participated as individuals in two scenarios of a quasi-experimental simulation technique. Both simulation scenarios, each lasting an entire day, employ a complex computer program that presents events, collects performance data, calculates more than forty performance scores and generates graphic representations of performance. Obtained data load on 13 different performance factors.

The simulation room is equipped as a senior level office. Decision makers are provided with an assistant (actually an experimenter) who facilitates the decisions made by participants. The assistant enters decision texts and associated codes into a computer. Information arrived via video screen and hard copy: 50% of that information was pre-programmed, representing independent variable manipulations (this constancy is needed to permit comparisons among performance data from diverse participants as well as to permit comparison of each person's performance to a criterion of excellence). Additional information is partly responsive to each decision maker's prior actions, providing the impression of reality. Participation resulted in considerable motivation. Fcr the data collection effort involving emergency responding, a serious emergency was introduced at the beginning of the fourth hour of participation. The emergency resolved itself at the beginning of the fifth hour, irrelevant of the actions of the participant (even though participants had the impression that they influenced the outcome).

RESULTS AND INTERPRETATION

1. Effects of Two Levels of Alcohol on Complex Decision Making

Alcohol intoxication at the .05 breath level had limited effects upon Speed of Response to incoming information and upon Decision Making Activity. However, the capacity to plan for future actions (ANOVA F = 6.32, 1/46 df, p = .015) and the capacity to employ strategy (F = 4.07, p = .049) was diminished.

Intoxication at the .10 level affected decision making to a greater extent. Speed of Responding to incoming information slowed (F = 5.29, p = .026). While the frequency of decision making did not decrease, the resulting decisions became myopic, i.e., were generally restricted became widely restricted to "reacting" to information. Initiative was diminished (F = 4.24, p = .045). Planning and strategy diminished sharply; even the capacity to develop a coherent approach within relatively limited spheres of activity decreased (F = 12.11, p < .001).

The simulation technique is able to distinguish between strategic activities at several levels of complexity. At the most basic level, this technique focuses on single actions that create the preconditions for later (different) actions. Where both actions are carried out, credit for one simple strategic effort is given. In contrast, the simulation calculates different scores for "Advanced Strategic Competence." The latter measure indicates whether decision makers are able to interrelate multiple strategic steps across diverse areas of operation over time. Considering that a sharp decrease in planning and basic strategy, both at the .05 and the .10 level did occur, one might expect that Advanced Strategic Competency would also diminish. In fact it does. However, this finding deserves a closer look: The diminished basic strategic competence (discussed above) necessarily limits the basic strategic raw material that might become a comoonent of more advanced strategic efforts, necessarily resulting in a lower score for Advanced Strategic Competence. However, if one corrects for the fewer basic strategic actions that remained, one finds that those (fewer) actions were interrelated into complex patterns just as effectively as was the case when the decision makers were not intoxicated. Interestingly, decision makers were quite aware of their apparently intact capacity to develop complex strategic plans and, as a consequence, tended to believe that alcohol did not have a significant impact on the quality of their decision making. In other words, the data indicate that decision making competence is diminished by alcohol. even where task demands are not unusual or unexpected.

2. Effects of Alcohol (.10) upon Decision Making during an Emergency

Alcohol intoxication resulted in a less broad approach to the simulated emergency (F = 3.28, p =

.046). Use of Planning and Strategy in handling the emergency decreased ($F = 3.35$, $p = .044$). If plans were made, they were not likely carried through ($F = 14.13$, $p < .001$). Initiative, i.e., the capacity to derive novel solutions to the serious problems, declined ($F = 5.22$, $p = .009$). In sum, the capacity to deal with an emergency during intoxication diminished considerably.

3. Hangover Effects

Research participants who had been intoxicated at the .10 level during the prior evening, in most cases, felt miserably during the next day. Nonetheless, the simulation measures did not indicate that differences between placebo vs. alcohol treatments exist.

Conclusions

Clearly, alcohol has considerable effects on decision making competency. At the .05 level of intoxication, simple responsiveness appears unaffected: decision makers are still responding with the same frequency and speed as they do under placebo treatment. A tendency toward myopic action is, however, evident, likely resulting in diminished planning and strategic ability. At the .10 level of intoxication, deterioration appears to be more general. In other words, decision makers, for example in the military, who are merely implementing the decisions made by their superiors might still be able to do fairly well at the .05 intoxication level. However, personnel who must develop their own plans and must translate these plans into their own effective strategies, are no longer effective. Worse, these individuals may not even recognize their diminished competence since they are still combining a few remaining basic strategic efforts into relatively complex patterns.

The military ethic against drinking more than one or, at most, very few alcoholic drinks appears then well justified. However, a higher level of intoxication during the week-end (for those who need not be concerned about a sudden call to duty) does not appear to be especially harmful to subsequent decision making competence, even if it might produce an "unpleasant" hangover.

REFERENCES

Aluisi, E.A. (1982). Stress and stressors: Commonplace and otherwise. Human Performance and Productivity, 3, 1 - 10.

Beeman, D.R. (1985). Is the social drinker killing your company? Business Horizons, 28(1), 54 - 58.

Berndt, G. (1991) Entscheidungsfahigkeit bei komplexen Problemlosungen. Bereitschaftspolizei heute: Einsatz und Fuhrung, 11, 15 - 22.

Beyer, J.M. and Trice, H.M. (1984). A field study of the use and perceived effects of discipline in controlling performance. Academy of Management Journal, 27, 743 -764.

Bushmamn, B.J. and Cooper, H.M. (1990). Effects of alcohol on human aggression: An integrative research review. Psychological Bulletin, 107, 341 - 354.

DeClifford, M. (1985). Executive health. Australian Accountant, 55, 20 - 22.

Heishman, S.J., Stitzer, M.L. and Bigelow, G.E. (1988). Alcohol and marijuana: Comparative dose effect profiles in humans. Pharmacology, Biochemistry and Behavior, 31, 649 - 655.

Helm, L. and Gaffney, C. (1986). The high price the Japanese pay for success. Business Week, April 7, 52 - 54.

Hilton, M.E. (1991). The demographic distribution of drinking patterns in 1984. In W.B. Clark and M.E. Hilton (Eds.): Alcohol in America. Albany: SUNY Press, 73 - 86.

Jobs, S. (1989) Impact of moderate alcohol consumption on business decision making. Paper presented to the NIDA Conference on Drugs in the Workplace.

Lucas, S.E., Lex, B.W., Slater, J.P. and Greenwald, N.E. (1989). A microanalysis of ethanol, induced disruption of body sway and psychomotor performance in women. Psychopharmacology, 98, 169 - 175.

Mann, R.E., Cho-Young, J. and Vogel-Sprott, M. (1984). Retrograde enhancement by alcohol of delayed free recall performance. Pharmacology, Biochemistry and Behavior, 20, 639 - 642.

Martin, J.K. (1990). Jobs, occupations and patterns of alcohol consumption: A review of the literature. In P.M. Roman (Ed.): Alcohol problem intervention in the workplace. New York: Quorum Books, 45 - 66.

McCann, P. (1983). A cure for drinking directors. Chief Executive, December, 16 -17.

National Transportation Safety Board (1990). Fatigue, alcohol, other drugs and medical factors in fatal to the driver heavy truck crashes. US Government publication.

O'Broin, C. (1985). Drink and drugs in the workplace. Personnel Management, 17, 28 - 30.

Pace, L.A. (1989). When managers are substance abusers. Personnel Journal, 68, 70 - 73.

Roman, P.M. (1974). Settings for successful deviance: Drinking and deviant drinking among upper level employees. In D. Bryant (Ed.): Deviant behavior. Chicago: Rand McNally.

Schopf, K. (1990). Validierung der Strategischen Management Simulation Shamba in deutscher Fassung. Universitat Paderborn: Diplomarbeit.

Shore, E.R. (1985a). Alcohol consumption rates among managers and professionals. Journal of Studies on Alcohol, 46, 153 - 156.

Shore, E.R. (1985b). Norms regarding drinking behavior in the business environment. Journal of Social Psychology, 125, 735 -741.

Steele, C.M. and Josephs, R.A. (1990). Alcohol myopia: Its prized and dangerous effects. American Psychologist, 45, 921 - 933.

Streufert, S., DePadova, A., McGlynn, T., Pogash, R. and Piasecki, M. (1988b). Impact of beta blockade on cognitive functioning. American Heart Journal, 116, 311 - 315.

Streufert, S., Pogash, R. and Piasecki, M. (1988a). Simulation assessment of complex managerial performance: Reliability and validity. Personnel Psychology, 41, 537 - 557.

Streufert, S., Pogash, R., Roache, J., Gingrich, D., Landis, R., Severs, W., Lonardi, L. and Kantner, A. (1992). Effects of alcohol intoxication on risk taking, strategy and error rate in visuomotor performance. Journal of Applied Psychology, 77, 515 - 524.

Streufert, S., Pogash, R., Roache, J., Severs, W., Gingrich, D., Landis, R., Lonardi, L. and Kantner, A. (in press). Alcohol and managerial performance. Journal of Studies on Alcohol.

Trice, H.M. and Belasco, J.A. (1970). The aging collegian: Drinking pathologies among executive and professional alumni. In G.L. Maddox (Ed.): The domesticated drug. New Haven: College and University Press, 218 - 233.

Trick, K.L. (1984). Drink: Concern and control. Director, 38, 68.

INTELLIGENT SCAFFOLDING FOR DESIGNING INSTRUCTION

J. Michael Spector
Daniel J. Muraida
Armstrong Laboratory, Brooks AFB, TX

Introduction

Designing effective computer-based instructional materials (courseware) is a problem that is becoming more difficult for three reasons: (1) rapidly changing interactive multimedia technologies, (2) growing complexity of highly technical subject matter, and (3) increasing demand for computer-based training due to downsizing and other constraints. The Air Force Armstrong Laboratory's Advanced Instructional Design Advisor (AIDA) project is an exploratory research and development effort which addresses this problem (Spector, 1990). The project has been underway since 1989 and is currently aimed at providing subject-matter experts (SMEs) with a set of powerful tools and techniques for use in designing and developing interactive courseware.

The AIDA project is currently in the process of evaluating an experimental prototype (XAIDA) designed around the concept of transaction shells (intelligent lesson frameworks that are accessible to SMEs) developed at Utah State University by M. David Merrill and colleagues (Merrill, Li, & Jones, 1990). In addition to transaction shells, the AIDA project involves two additional approaches to automating instructional design: (1) Robert M. Gagné's intelligent advisor (1992), which couples high-level development guidance with completely worked examples for a variety of instructional objectives, and (2) Robert D. Tennyson's (1991) intelligent tutoring system for instructional design.

All three of these approaches to providing instructional design assistance can be considered intelligent scaffolding. They are intelligent in the sense that they incorporate various techniques and methodologies used in artificial intelligence (e.g., expert systems, case-based reasoning, etc.). They are also intelligent in the sense that they capture some genuine instructional design expertise (see Gagné & Merrill, 1990).

Intelligent Scaffolding

Thinking About Design Guidance

In order to provide a context for offering instructional design expertise to novice developers of CBT, let's first consider an imaginary situation with regard to architects. Suppose a great many architects and builders have moved to some remote location thereby creating a sudden shortage of architectural and home construction expertise in another part of the country. A forward thinking firm there has decided to respond to this crisis by building an architectural advising program. This program will be made available to recent graduates of a nearby university who are being hired by local home builders. These graduates all have a basic understanding of the uses of hammers, nails, floor joists, etc. However, these novice builders/architects obviously lack the kind of expertise in design and construction that comes from years of experience.

In the design phase of this effort, this firm has hired an artificial intelligence expert and an experienced architect. They have provided the following advice regarding the home design advisor program:

1. There are four kinds of scaffolding that can be provided in an automated home design advisor:

a. High level reminders. These are things that should have been learned in school. An example might be this: If the house is to be located near the base of a hill, then an underground ditch should be constructed to divert run-off. It will be important to make this kind of advice easily accessible and context-sensitive so as to avoid information overload. On-line help systems provide a reasonable model in this regard.

b. Example plans. These should be based on previous cases that are known to provide aesthetically and economically attractive solutions. This case-based advice should also be context-sensitive and easily accessible. In addition, these examples should be modularized so that portions could be extracted and inserted into a current plan under development.

c. Configurable shells. These provide a ready-made framework for the kind of home design being developed. These shells would come complete with default settings which would determine such things as the placement of doors, windows, and electrical outlets. Of course, the defaults would automatically satisfy existing building codes since they would be generated by a rule-base. These shells would provide the designer with a rapid prototyping capability so that clients could be shown the consequences of early design choices such as one- vs. two-story, adobe vs. wood, etc.

d. Consistency checks. Since the shells can be altered by the novice designers, there would need to be some consistency checking of proposed designs to insure that building codes had not been violated, that original design constraints (e.g., cost, square footage, etc.) had been satisfied, and that basic design principles had been followed.

2. In order to provide this kind of scaffolding, the program would require the following:

a. A set of rules representing existing building codes.

b. A set of rules representing basic design principles.

c. A set of examples representing a variety of building designs.

d. A representational scheme which allows the previous three items to be related.

e. A mechanism which could interpret a given design goal, select an appropriate shell, and allow changes that did not violate any rules.

3. It would be desirable to embed this design advice in a program which automates the more tedious aspects of design (e.g., a computer-aided design (CAD) program).

4. Providing this type of design advice is possible, but it is also complex and requires a significant initial investment. The minimum that should be attempted is a set of worked examples along with the high level reminders. The configurable shells can be developed one at a time starting with the most frequently used shell (e.g., one-story, single-family, three bedroom, two bath, brick, etc.).

The point of this imagined architectural design advisor is that there are a variety of kinds of scaffolding that can be provided. To the extent that these four (and others) methods incorporate artificial intelligence methodologies (such as expert systems, case-based reasoning, fuzzy logic, etc.) in order to capture and provide design expertise, they can be considered intelligent.

Instructional Design Guidance

The previous scenario was introduced because home designs result in familiar physical structures, which makes it easier to imagine what design expertise is essential and how it might be represented and made available to novices. Instructional designs result in plans of instruction which are used to create or modify mental structures and associated behaviors in learners. Mental structures are not so readily accessible for evaluation, which makes assessment of instructional designs more difficult. Nevertheless, we argue that the situation is similar to the home design situation sketched above and that it is useful to think of automated instructional design advice in an analogous fashion.

We shall assume that our novice instructional designer either has subject matter expertise or has access to a subject matter expert (SME). Furthermore, we shall assume that our novice designer has had at least a short course (perhaps the equivalent of one college semester's work) in instructional design. We are now positioned to consider whether and what instructional design expertise can be automated and offered to such users.

It is certainly possible to offer high level reminders. In the case of instructional design, these amount to something like Gagné's nine events of instruction (1985). These events can be provided in a context sensitive manner. If the user is involved in a particular event (e.g., gain attention -- event #1) and asks the system for help, then the system would provide elaboration of techniques for gaining attention. If the user has just completed an event and is wondering what to do next, then the system could provide an overview of the next event. If something other than Gagné's nine events of instruction were used as an organizing schema for the instructional design advisor, then that scheme would have useful high level reminders which could be elaborated at the user's request.

To illustrate these high level reminders, it is desirable to have fairly complete examples in a variety of domains so that an example sufficiently close to the case at hand can be offered to the user. Several recent studies argue that expert instructional designers in fact work from a mental model of a previous relevant case (Perez & Neiderman, in press; Rowland, 1992; Tessmer, 1992). These examples or cases can be cross-indexed by instructional objective, instructional event, and subject domain. In short, context-sensitive example plans can also be incorporated into an instructional design advisor. In fact, the Guided Approach to Instructional Design Advising (GAIDA) is an existence proof of the possibility of accomplishing these first two kinds of intelligent instructional design advice.

It would certainly be useful to provide novice designers with the equivalent of configurable shells -- intelligent design frameworks for lesson development. Merrill *et al.* are attempting to do just this with the notion of a transaction shell in their second generation instructional design theory (1991). An evaluation of an early prototype of an instructional development shell for teaching nomenclature (name, location, and function of parts of a device) clearly indicates that in some cases it is possible to provide meaningful shells (Spector & Muraida, 1991).

The difficulty or challenge with regard to configurable shells is in generating sufficient rules and details to support a much wider variety of learning objectives and subject matter. This is the challenge addressed by the Experimental Advanced Instructional Design Advisor (XAIDA). It is still too early to say to what extent we will achieve success in this area, but four shells designed by Merrill and colleagues at Utah State

University and engineered by Mei Technology, Inc., are currently being field tested. These shells are: Identify, Classify, Interpret, and Decide. The first two concern mainly declarative knowledge involving objects or entities; the second two involve processes and procedures. The subject domain chosen for experimental purposes is electronics maintenance.

The analog of a consistency checker is conceivable for an instructional design advisor (Duchastel, 1990). Such a system would act as a critic and inform the novice designer when a basic rule or principle had been violated. However, this type of system has yet to be prototyped, so an evaluation of its possible success is at this time premature.

An intelligent tutoring system for the domain of instructional design is also clearly a possibility (Tennyson, 1991). However, it would appear that until there is a well-established record with the previous methods that the knowledge required to construct such a system will be incomplete and arbitrary.

Conclusions

In general it appears that some instructional design expertise can be captured in automated systems. In addition, both case-based and rule-based systems have a role to play in this enterprise. In short, the techniques of artificial intelligence do have a legitimate application in the domain of instructional systems design and development.

Both GAIDA and XAIDA provide the potential to conduct a great deal of empirical research in the area of instructional design. GAIDA can be used as a mechanism for decompiling instructional design expertise. XAIDA can be used to systematically alter specific instructional parameters in a shell (e.g., sequencing of events, placement of objects on the screen, time allotted for learning activities, etc.) in order to determine optimal default settings for a range of instructional situations (different learning objectives, learner profiles, delivery media, etc.).

Finally, these and other automated instructional design systems and tools will provide the means for determining whether and to what extent instructional design is a science or art. Our belief is that there are scientific aspects of instructional design; for example, designing materials so as to minimize the load on working memory appears to be empirically well-established. However, there are also artistic aspects of instructional design; for example, designing a specific event that is relevant to the instructional task at hand which gains and is likely to maintain attention appears to be a task for a creative artisan.

References

Duchastel, P. C. (1990). Cognitive designs for instructional design. *Instructional Science*, 19(6), 437-444.

Gagné, R. M. (1985). *The conditions of learning* (4th ed.). New York: Holt, Rinehart and Winston.

Gagné, R. M. (1992). *Tryout of an organizing strategy for lesson design: Maintenance procedure with checklist* (AL-TP-1992-0016). Brooks AFB, TX: Armstrong Laboratory, Technical Training Research Division.

Gagné, R. M., & Merrill, M. D. (1990). Integrative goals for instructional design. *Educational Technology Research and Development*, 38(1), 23-30.

Merrill, M. D., Li. Z., & Jones, M. K. (1990). Second generation instructional design (ID-2). *Educational Technology*, 30(2).

Perez, R. S. & Neiderman, E. C. (in press). Modelling the expert training developer. In R. J. Seidel & P. Chatelier (Eds.), *Advanced training technologies applied to training design*. New York: Plenum.

Rowland, G. (1992). What do instructional designers actually do? An initial investigation of expert practice. *Performance Improvement Quarterly*, 5(2), 65-86.

Spector, J. M. (1990). *Designing and developing an advanced instructional design advisor* (AFHRL-TP-90-52). Brooks AFB, TX: Armstrong Laboratory.

Spector, J. M. & Muraida, D. J. (1991). Evaluating transaction theory. *Educational Technology*, 31(10), 29-35.

Tennyson, R. D. (1991). Framework specifications for an instructional systems development expert system. In Gagné, R. M., Tennyson, R. D. Tennyson, & Gettman, D. J. Designing an advanced instructional design advisor: Conceptual frameworks (AL-TP-1991-0017-Vol-5). Brooks AFB, TX: Armstrong Laboratory, Human Resources Directorate.

Tessmer, M. (1992, April). *The practice of instructional design: A survey of what designers do, don't do, and why they don't don't do it.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

AN ALTERNATIVE INSTRUCTIONAL METHODOLOGY
BY

IMELDA IDAR
BRUCE MCDONALD

In fiscal years '88 and '89, first term attrition showed a disturbing trend. Despite the high recruit quality, four of every ten sailors were leaving the Navy before the end of their first enlistment. This high attrition rate cost the Navy an approximate $750 million dollars a year. An analysis of first term attrition indicated that it occurred in two distinct periods. The accession phase, encompassing recruit training and all subsequent training prior to the sailors reporting to their first duty station, accounted for 40 percent of attrition through the first 12 months of service for four-year, non-prior service enlistees. The second phase of first term attrition occurred when sailors reported to their first command. This period accounted for the remaining 60 percent of total first term attrition, and losses were spread evenly over the remainder of enlistments. While both the accession and first assignment shared many of the same underlying reasons for attrition, the predominant causal factors were different.

This paper discusses only accession losses and focuses on research effort to design an instructional package which specifically enhances student motivation and maximizes training.

Accession Losses

During the accession phase, attrition is attributed to factors of sociological adaptability, educational deficiencies and medical or physiological problems. From a sociological viewpoint, a generally permissive society contributes to a lack of adolescent discipline, low expectation, and a desire for instant gratification. Such tendencies conflict with the discipline of Navy life and the standards at recruit training. Training and other schooling lead to frustration, reduced motivation, and failure among young sailors. These trends are further exacerbated by the poor fundamental skills carried by each recruit from the U. S. school systems and their less then optimal physical condition.

An in depth analysis of technical training attrition conducted in December 1990 verified these major reasons for student discharges. Academic factors accounted for 50 percent of the separations in fiscal year '89 and 43.8 percent in '90. Non-academic reasons contributed to 50 percent of attrition in fiscal year '89 and for 56.2 percent in fiscal year '90. Figure 1 is germane.

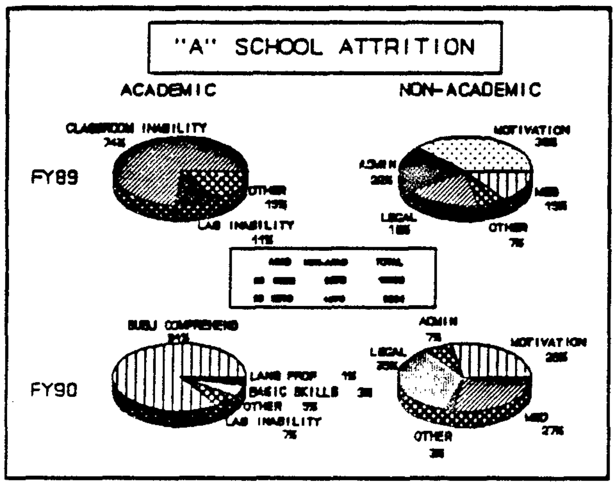


Figure 3: "A" School Attrition

Classroom inability, specifically a lack of comprehension and retention, reflected through testing was the most salient reason for academic attrition. Motivation at 38 and 28 percent was the second most significant cause of non-academic attrition. Separations for legal reasons, 18 percent in '89 and 35 percent in '90, were due to misconduct which closely correlated to lack of motivation and frustration with the then current instructional practices. Clearly, lack of motivation among our trainees contributes significantly to the discharge of Navy sailors.

Counter-attrition Initiatives

To stem this high rate of separation of sailors, the Navy developed and executed an aggressive counter-attrition campaign. This campaign initially focused on the training arena since it was an area where control could be exerted and monitored. The counter-attrition efforts included the following actions:

1. Improved Company Commander Screening. This process was expanded to include a sending command's interview of the individual. It also ensures that candidates meet the Navy's physical fitness and body standards.

2. Cessation of Separation for Sailors not Achieving Swimmer Fourth Class. Instead of separation, sailors are provided with up to three week swim remediation after boot camp. If no success is achieved during this period, the sailor loses his technical school assignment and continued remediation through their first enlistment.

3. Discontinuance of All Attrition for the First Three Weeks of Boot Camp - except for those sailors seperated during the Moment of Truth and Psychological Screening. This policy allows for acculturation to military life.

4. Stoppage of All Academic Attrition at Recruit Training. Since all of recruit training is orientation to military life, remediation is a better path to achieving the required behavior modification.

5. Gradual Phasing of Recruits Into Boots. For the first three weeks of boot camp, the recruits wear a sport shoe to avoid incurring or exacerbating orthopedic injuries.

6. Implementation of Two Screening Processes at Recruit Training. The first is Moment of Truth which is designed to resolve any service contract irregularities. The second is a psychological screening which identifies and discharges those recruits most at risk for separation due to psychological reasons. Savings in training costs are significant.

7. Revision and Sequencing of Basic Skills Programs. The objective of this action is to provide training to requirements only with maximization of instruc^ion.

8. Extensive Review of the Technical Training Process. This policy dictates the identification and validation of training requirements which lead to improved training curricula.

9. Modification of the Integrated Training Battalion. This concept was brought in consonance with the technical training goals and objectives in order to eliminate the conflict between the academic and military aspects of the technical training process.

10. Research Efforts. These efforts focus on the innovative application of instructional technology and advances to improve current instruction.

Skill Enhancement Initiative

Having addressed most of the causes for attrition with aforementioned policy actions, the Navy next focused on the issue of the lack of motivation among its trainees. The decision was made to attack this issue through the medium of research and development in order to provide the greatest latitude with which to mitigate it. Significant consideration was given to the extensive research done on motivation. Most of it addressed behavior modification techniques to extrinsically motivate students to master instructional material. The major finding of this work (Thorensen, 1979) (Lepper and Greene, 1978) is that extrinsic motivation frequently undermines intrinsic motivational accomplishments. Upon removal of the external rewards, the goal behavior ceases. Students experience considerable deterioration of knowledge once they pass the test.

Research on intrinsic motivation, on the other hand, focused on cataloging various needs that people are motivated to meet and showed that measures of those needs are correlated to behavior. For example, McClelland and Winter (1969), Atkinson (1964) and deCharms (1984) provided data on how to increase the need to achieve. The work of Weiner (1972), Dweck(1986) and Ames and Ames (1984) reported on types of goals (e.g. performance versus learning goals and competitive versus cooperative versus individualistic goals). They examined the relationship of the various goal types to behavioral outcomes. They found that if a student sees an instructional situation as involving performance or competitive goals, then that student is likely to attribute a negative feedback as indicating that he/she is not good at the subject matter rather than viewing it as a chance to learn something new as in the case of learning or individualistic goals.

Keller (1983), in his model of motivational design for technical instruction, identified four motivational factors integral to learning:

1. attention - the arousal and sustaining of the learner's curiosity;
2. relevance - the correlation of instruction to the learner's personal goals;
3. confidence - control of successful outcomes
4. satisfaction - the integration of extrinsic with intrinsic motivation.

Keller also provides instructional strategies to address these factors.

Malone (1981) assessed the impact of computer games on motivation and proposed three categories of instruction to enhance intrinsic motivation: (a) fantasy, (b) challenge and (c) curiosity.

Technological Advances

In light of the aforementioned research and prior to initiating the research and development effort, a thorough review of the current instructional computer software was conducted. Lee (1989) explains three major approaches to Computer Based Training (CBT) - in this case Intelligent Tutors or ITs: framed-based, Artificial Intelligence (AI) based, and mental-modeled based. He states that framed-based focuses on the narrative exposition of facts and concepts. It views knowledge as a "knowing what." Its governing metaphor is that of the slideshow/examination. The governing metaphor of most AI-based Intelligent Tutors (ITs), on the other hand, is that of the coach. They are often called coaching systems. These systems emphasize learning by doing, and view knowledge more as "knowing how" than "knowing what". They identify knowledge with skill and skill with procedure. The reduction of knowledge to procedure logically implies a view of instruction which conceptualizes the learning process as practice - i.e., the repeated execution or performance of that procedure or method. Hence, most ITS coaches tend to focus on trainee performance and to assess that performance from the standpoint of how an expert would perform when tackling the same problem. Lee (1989) contrasted the framed-based, AI based and model-based intelligent tutoring systems.

Lee concluded that AI-based ITs are shallow and brittle because they neglect the role of and changes mental model effect in learning. In this respect, they are similar to framed-based CBT. Lee characterizes frame-based CBT as failing because its emphasis on "knowing what" leads to a straight jacket instructional presentation where the trainee is viewed as a passive receptacle into which knowledge is poured. Lee avers that AI-based ITs systems also fail because their emphasis on "knowing how" leads to equally limited and superficial instructional interactions in which the trainees are viewed as imperfect expert systems that have buggy rules and procedures. The model-base ITs are successful because they uncouple knowledge from the constraints of instructional sequencing. Strength in a system can be found in the following properties:

"a. To accommodate the diversity of paths that may be taken in learning. It must uncouple the representation of knowledge from the constraints and requirements of instructional sequencing;

b. While incorporating and exploiting the "knowing what " and the "knowing how" views of knowledge, it must also permit the representation and expression of the "knowing why" view of knowledge;

c. While supporting learning of facts and concept, and facilitating learning by doing, it must also encourage the construction of mental models in order to facilitate understanding of how facts, concepts, and procedures fit together into a mutually reinforcing and sustaining whole;

d. In order to sustain properties (b) and (c) the system must be able to exploit the figurative resources of discourse - metaphor and metonymy" (Lee, 1989 p. B-11).

Skill Enhancement Project (SEP)

Using this information, the Chief of Naval Operations (OP-11) formed the SEP Working Group. The SEP Working Group, in turn, commissioned a consortium of experts to design the research. The consortium included computer and programming experts from the Institute of Simulation and Training (IST), gaming specialists from Atari, and authorities on individual cognitive learning strategies from Pangaro Inc. The SEP Working Group proposed to the consortium the use of alternative instructional technologies, and specifcally, use of computer gaming and ITs, to enhance motivation, attention, and qualitative understanding in an "A" school environment.

The Aviation Technician Technical School was chosen as the test site for the Skill Enhancement Project because, in fiscal year '92, it became one of the top fifteen technical school experiencing high increases of attrition.

The Naval Training Systems Center (NTSC) contracted with the Institute of Simulation and Training (IST) to develop a computer-based game that augments instruction on the following topics:

1. Basic Theory of Capacitors and Capacitance;
2. Alternating Current (AC) Capacitative Circuits
3. Direct Current (DC) Resistor Capacitor (RC) Circuits;
4. AC RC Circuits.

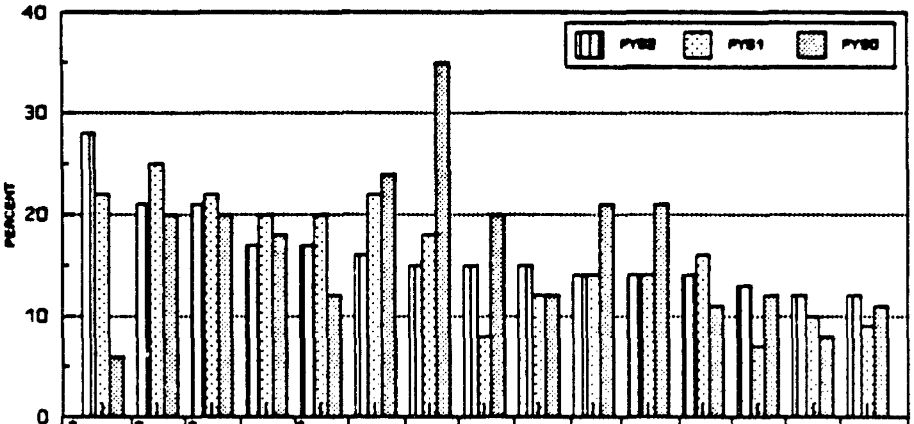
The computer-based game will be integrated into the current instructional program at the Aviation Technician (AV) "A" school under the purview of the Chief of Naval Technical Training (CNTECHTRA) at Memphis. Figure 4 depicts its increase in attrition from eight percent in fiscal year '90 to 12 percent in fiscal year '92. Furthermore, the school is sited at Memphis and it can be closely monitored by CNTECHTRA during tne project.

The overall objective of the portion of the A course instruction selected for the project is to provide students with the ability to solve RC circuit problems and explain or describe the functional aspects of RC circuits. The instruction developed for this purpose are designed to support these specific instructional objectives by assuring that game objective are congruent with learning objectives.

Currently, the Electronics Training course at CNTECHTRA is a six week program and covers the basics of electricity and electronic circuitry. Existing methods of presentation media consist of lecture, text books, and overhead transparencies. While these methods can be effective if used properly, they must compete with visual and auditory stimulation that students receive outside the classroom.

TOP FIFTEEN "A" SCHOOL PIPELINES

ATTRITION RATE AS OF APR 92



	*STS SS	*RM SS	*ET SS	FC	*FTG SS	AC	CTR	BT 6YO	SWGE	ET AEF	AW	AG	MM 6YO	AW	CE
FY92	28	21	21	17	17	16	15	15	15	14	14	14	13	12	12
FY91	22	25	22	20	20	22	18	8	.12	14	14	16	7	10	9
FY90	6	20	20	18	12	24	35	20	12	21	32	11	12	8	11

12 MONTH MOVING AVERAGE

*High pipeline attrition due to BERT segment.

Figure 2: Top Attrition "A" Schools

Student Population

Students enrolled in the electronics training course have undoubtedly come in contact with both video and computer games not to mention television. They often seek the same type of instant gratification in their instruction. Instructors must continually devise new and exciting methods of instruction to maintain their attention and keep them motivated. Note that the existing course includes a laboratory section which, for obvious reasons, can not be replaced by the video game format.

Upon completion of the electronics training course, the students are expectsd to have obtained enough skills and knowledge to perform at a 70 percent level on exams administered at the end of the course. Specifically, the students receive instruction and are tested on DC and AC circuits and their electrical applications.

The students who are enrolled in the Electronics Training program are of above average intelligence. They must score 218 or higher on the Armed Services Vocational Aptitude Battery (ASVAB) subtest comprised of the following factors:

Math Knowledge (MK) + Electronic Information (EI) +
General Science (GS) = 156 + Arithmetic Reasoning (AR) = 218.

However, most of these students have not progressed beyond high school education. They range in age from 18 to 25 years of age. These demographics directed the selection of a game which can be categorized as fantasy.

Instructional Design

The instructional design was based on an analysis of the existing course material. Its 46 instructional objectives were best addressed through an adventure game format. Accordingly, the game "Electro Adventure" was created as the vehicle of instruction. It simulates a futuristic ship called "Electro" that was struck by an enemy laser and sent spinning in a ten dimensional time warp. The Electro arrived in the year 1992 badly damaged and in need of repair. The student's mission for this adventure is to restore the Electro to working order and send it back to the time from which it came. The trainee must solve basic electronic circuit problems as he/she moves from one ship compartment to the other. The terminology of this ship mirrors Navy lexicon.

Each compartment has an assigned time limit within which the student should easily complete the required activities. If the student becomes confused or starts to exceed the given time limit, he or she will receive a verbal prompt from the Sage - a figure who provides instruction and help throughout the game as required. No prompts will be given if the student completes the activity within the allotted time. However, as time elapses, the student is given increasingly more forceful prompts or clues. The program analyzes the student's time spent in the adventure as a whole. If the student completes a compartmental activity before the allotted time, the residue time can be utilized in a later activity.

The gaming aspect of this instructional design reflects the strategies first used by Crawford and Hollan (1983). Each instructional activity presented to the students consists of a series of problems embedded within a gaming scenario. The student must solve the problems to progress through the game scenario. The scenario for each compartment is designed primarily to increase extrinsic motivation, i.e., there are external rewards to be gained in each compartment. The student is allowed to gain points or pick up key objects such as maps, ropes, or tools to continue the game.

A variable which contributes to extrinsic motivation is goal. In Electron Adventure, the goal is to return the spaceship to its original time frame. In accordance with the research of Driskell and Dwyer (1984) and Malone (1980), the game also provides a challenge. The student is always in a quandary as to what he/she will be facing in each compartment. But, the student is also assured that help will be provided to complete the tasks. The games's variable difficulty levels, multiple level goals, cumulative score keeping and informational feedback also contribute to keeping the student challenged.

Finally, Electro Adventure provides the fantasy by which a trainee might fulfill his or her wishes. It furnishes the medium by which a learner may fulfill such needs as power, success, and wealth.

There is a classroom aspect to the instructional design as well. This aspect of the design addresses the remedial needs of the student. As required, the student is transported to a separate classroom on the ship to receive the necessary instruction to solve the problems in the compartments. It employs speech, color animation and visualization. More significantly, the instruction is based on computer based training (CBT) principles which are mental-modeled based. Its algorithm assess learning strategies and uses pedagogy which accommodates them. This particular algorithm is the hyper-media based THOUGHTSTICKER. This dynamic system classifies trainees as either serialist or holist learners or versatile - that is a learner who uses a combination of the two styles depending on the environment. A learner is classified as a serialist if he/she processes information piece meal - concentrating on specific minute details. A holist on the other hand requires a macro view of the problem and processes information in relation to that view. THOUGHTSTICKER through conversations with the learner continually updates learning strategy data and adjusts it instructional presentation accordingly. This aspect of the instruction is designed to address the intrinsic motivation of a student by incorporating and exploiting the "knowing what" and "knowing why" views of knowledge. Further, it permits the representation and expression of the "knowing why" view of knowledge - a facet of learning integral to intrinsic motivation. Not only is the program supporting the learning of facts and concepts, it is also encouraging the construction of mental models in order to facilitate the understanding of how facts, concepts and procedures fit together into the mutually reinforcing whole.

Conclusion

The Skill Enhance Project is still in the development stage. Its completion and implementation are scheduled for the latter part of fiscal year '92. By providing both aspects of extrinsic and intrinsic motivation with innovative addressal of cognitive learning strategies, the Navy endeavors to maximize its instruction. Thus, as side effects, training length will reduced as well as produce a decrease in attrition due to lack of motivation. Further, this particular research design will provide acturial data on addressal of individual cogntive learning strategies. While there has been much debate and discussion on the merits and the impracticality of this practice, little research has been done on it. Actual statistics on the program with be forthcoming in fiscal year '93.

REFERENCES

Ames, C. and Ames, R. (Eds.) (1984) Research on motivation in education. New York: Academic Press.

Atkinson, J.W. (1964) An introduction to motivation. Princeton, N.J. Van Nostrand.

Crawford, A.M., & Hollan, J.D. (1983). Development of a computer based tactical training system system (Spec Rep. NPRDC-SR-83-13). San Diego, CA: Navy Personnel Research & Development Center.

deCharms, R. (1984) Motivation enhancement in educational settings. In R. Ames and C. Ames Research on motivation in education. New York: Academic Press.

Driskell, J.E. & Dwyer, D.J. (1984). Microcomputer videogame based training. Educational Technology, 24, 11-16.

Dweck, C.S. (1986) Motivational processes affecting learning. American Psychologist, 10, 1040-1048.

Keller, J.M. (1983) Motivational design of instruction. In C.M. Reigeluth (Ed.) Instructional design theories and models: An overview of their current status. Hillsdale, NJ; Erlbaum.

Lee, W. (1989) Business plan. Cambridge, MA: Soliton Inc.

Lepper, M. R. and Greene, D. (Eds.) (1978) The hidden costs of reward. Hillsdale, NJ; Erlbaum.

Malone, T.W. (1980). What makes things fun to learn? A study of intrinsically motivating computer games (Doctoral dissertation, Stanford University, 1980). Dissertation Abstracts International, 41(5-B), 6603.

Malone, T.W. (1980). Microcomputers in education: Cognitive and social design principles. Sigue Bulletin, 17(2), 6-20.

Malone, T.W. (1981) Toward a theory of intrinsically motivation instruction. Cognitive Science, 5, 333-369.

McClelland, D.C. and Winter, D.G. (1969) Motivating economic achievment. New York: Free Press.

Thoreson, C.E. (Ed.) (1979) Behavior modification in education. National Society for the Study of Education.

WHY SOME INSTRUCTORS DON'T LIKE COMPUTER-BASED TRAINING

Stanley D. Stephenson
Southwest Texas State University

Teacher resistance to the introduction of computers into the classroom is not unknown (Wedman & Heller, 1984). For instance, Marche (1987) reported that 54 percent of the survey respondents agreed that the people who work in their school system felt that they were threatened by the actual, or expected, impact of information technology on them and their jobs.

There have been many attempts to try to capture why this resistance exists or why computers are not being used more extensively. Ross (1991) reported that non-use of computers in the classroom is tied to both a lack of experience with and training on computers. Moreover, he reported that non-users expressed concern over the supervision and organization of the computer-assisted learning environment. Dupagne and Krendl (1992) found that teachers were generally in favor of computers but were concerned with the availability, choice, and evaluation of appropriate software. Farina et al. (1991) found resistance to the use of computers to be tied to such factors as trait anxiety, mathematics anxiety, and the impact of computers on society. The relationship between math anxiety and computer non-use is interesting in that it indicates that non-users see computers as being primarily related to numbers and calculations; computer users do not necessarily view computers in this manner. Violato et al. (1989) focused on uncovering the underlying factors associated with the use or non-use of computers and found four factors: Gender; Comfort (e.g., feeling of powerlessness); Liking (e.g., previous experience); and Value (e.g., what will it do for me?). Kay (1989) found three factors: Cognitive (e.g., will a computer make me more creative); Affective (e.g., good-bad); and Behavioral (e.g., previous experience with computers).

A problem with these approaches is that they treat resistance to computers both in somewhat of an objective light and also as being perhaps easily overcome. However, I believe that the resistance issue is more complex than this. I sense that there are four dimensions; these dimensions contain the ideas presented above plus other, more pragmatic issues. Some of these dimensions can be modified relatively easily. Others are more permanent and can only be gradually and perhaps partially modified. But it is my impression that teacher resistance to the use of computers cannot be eliminated easily or quickly and instead must be attacked in a variety of ways over time.

Dimension 1
Fear of Computers, Programming, and Math

I do not like computers
I do not want to become computer literate
I do not want to learn BASIC or other computer language
I may not be capable of operating a computer
I (secretly) hate math
I will have to become an expert in the operation of computers as well as my own discipline

The basis for Dimension 1 is fairly simple: fear of computers and math specifically and new technology in general. Will a teacher have to become a computer expert if computers are introduced into the curriculum? This fear is not unfounded; some people simply do not feel that they have the ability to learn how to operate a computer. Moreover, in many teacher computer education courses, the acquisition of programming skills is the primary course objective (Troyer, 1988). Also part of Dimension 1 are the fears a teacher may have about having either to learn how to type or to demonstrate math proficiency, fears even more basic than the fear of computers.

Dimension 2
Fear of Change in Teaching Style

Computers will take away what I like to do - teaching in the traditional manner

I will become more of a counselor than a teacher

I will not know what to do with my time if computers are introduced

I do not know how to use computers in my discipline

I am responsible for my students' performance; I do not want to turn that responsibility over to a computer

I will lose control of my classroom

Forcing me to use computers implies that I need to improve my teaching performance

I will be evaluated differently

Computers cannot do as good a job as I can do

My job cannot be computerized; I am unique

Dimension 2 is based on the fear of no longer being able to teach in the manner a teacher may have come to know and enjoy. Teachers know what they do now; they may not know what they will be doing if computers are implemented. Teachers also fear that they will lose some degree of control over classroom behavior. Tied to these uncertainties is the belief that somehow current teaching performance is being questioned. I.e., if I am doing a good job now, why are computers replacing me? Teacher evaluation may also be different. During a two year observational study of the impact of computers on secondary math education, Schofield et al. (1989) reported that "Interestingly, the chairman of the math department mentioned to our project staff that he could not evaluate teachers using the intelligent tutors very well since those classes were run so differently from ordinary ones and different teaching styles were needed (p. 14)."

Dimension 3
Fear of Losing Job or Money

I may lose my job

Money spend on computers may eliminate pay raises

Dimension 3 is the pragmatic fear of job loss, an outcome which is always denied by administrators but which is typically suspected by many teachers. Education costs are usually under constant scrutiny. How will computer purchases be funded? A teacher might well wonder how school administrators are going to find funds to purchase computers when these same administrators have consistently stated that there are no funds available for pay increases. A natural perception for teachers is that computer purchases will be funded by either a reduction in the number of teachers or by a reduction in the salaries of the existing teachers; neither alternative is very appealing. Computer advocates are a major contributors to this fear because they tend to promote the cost saving features of computers. If a student can learn faster and better on a computer, why not just replace some teachers (and their salaries) with computers? This dimension is rarely mentioned in articles on computer resistance.

Dimension 4
Fear of Change

I do not like change

I did not have a voice in the decision to implement computers

I have seen these state-of-the-art devices come and go before

There may not be enough resources to support this change

Dimension 4 is the normal fear-of-change, especially forced change, found in most people. Why change? Experienced teachers have seen other technologies come and go (Cuban, 1986), and computers have not always produced a consistently positive effect on learning (Kulik & Kulik, 1987). Moreover, there is the very real fear that there will not be enough resources to fully support the change; e.g., there may be only one computer per classroom versus a computer for every student. Related to this dimension is the fact that the changes which result from the adoption of computers can be far greater than the system developers ever imagined (Whitney & Urquhart 1990).

It is important to note that teacher resistance may never be verbalized. Teachers may not be fully aware of why they resist computers. Also, teachers know that, for evaluation and political purposes, it is not to their advantage to state their resistance, especially if the decision to adopt computer-based learning technologies has already been made. Also, it should be emphasized that teacher behavior will change with implementation of computers. The issue is how it will change, not if or will it change.

RECOMMENDATIONS

It is safe to assume that in any large scale computer implementation program there will be some degree of teacher resistance. Rather than ignoring it or treating it as an issue which will simply go away, program administrators should treat resistance as a variable which must be included in the system design. Although resistance cannot be totally eliminated, it can perhaps be reduced by attacking it before and during the implementation process.

Dimension 1: Basic Fear of Computers, Programming, and Math. Instead of initially giving the computers to the students, the first computers on campus should be given to the teachers. Let them use the computers without any pressure to immediately implement them into the curriculum. If most teachers do not want to learn how to program, then do not teach them or require that they learn programming skills. Instead, teach them how to use existing, user-friendly software which is not specific to a discipline. For example, classroom management has been shown to be a critical factor in academic success (Brophy, 1986). Why not use in-service computer education sessions to teach teachers how to use a computerized classroom management package? This would serve to both familiarize teachers with the operation of a computer and also provide them with a management tool which can increase their teaching effectiveness. This approach could both lessen fears that the computer are either a threat or math-related and also increase the perception that the computer can be useful without even directly impacting on the discipline.

Dimension 2: Fear of Change in Teaching Style. This dimension cannot be easily addressed. If computers are to be implemented into the curriculum, at some point they must be introduced at the discipline level. School administrators should make it known that they encourage the use of computers and also why they encourage the use of computers. They should also offer the use of school facilities for individual disciplines to hold work- shops, etc. Since most academic departments have a teacher who is either interested or already knowledgeable in computers, that person should be appointed as a focal point and tasked with proposing how computers could be introduced into the curriculum.

Is computer usage going to be part of teacher evaluations? Remove the mystery by publicizing school evaluation policy well in advance of the implementation date. Once teachers realize that computer usage will be part of the evaluation system, usage should increase. However, school administrators need to publicize that at some point in the future some or all of the faculty will be expected to be using computers to some degree.

Including computer use in teacher evaluation raises a critical validation question. Can the school system demonstrate that those teachers who use computers are "better" teachers than those who do not? If such a claim can not be supported with a validation study, then computer use can not be included in the evaluation system. The problem becomes obvious. If teachers are not going to be evaluated on use of computers, why should they use computers and can they be forced to use computers?

A teacher's fear that he or she will lose some degree of classroom control is a real concern. However, the fear could be countered by demonstrating how teachers can actually increase the control of their classroom through the use of computer classroom management or some other classroom planning software. Under this scenario, a computer is not being introduced into the classroom, a better method of instruction is. Another critical factor in the success of a computer-based educational system may be how well the role of the computer-based teacher is specified. A well-defined teacher role will help reduce teacher resistance.

Dimension 3: Fear of Losing My Job. This fear can be modified only by the actions of school administrators. Teacher retention policy must be well defined and frequently publicized. However, actions speak louder than words, and regardless of what is said, many teachers will adopt a wait-and-see attitude. To counter the perception that computers are being adopted to primarily decrease teacher costs, school officials could adopt a pro-learning attitude. That is, officials could offer the policy that computers are being adopted to increase student learning opportunities. This approach would create the perception that school administrators are more interested in raising the academic achievement of their students than they are in cutting teacher salary costs. Another alternative explanation for adopting computers is cost avoidance; i.e., introducing computers now will reduce costs in the future. This approach is appealing to currently employed teachers because it removes the threat of immediate job loss. School officials could also publicize where the computer program money is coming from rather than remaining silent and having teachers speculate about the fund source.

Dimension 4: Fear of Change. Fear of change is not easily overcome. Teachers typically have had little experience with change, and actually little change has occurred in educational technologies (Cuban, 1986). This dimension may be one which can only be modified over time. That is, if computers are introduced, if the transition goes smoothly, and if computers prove to be useful additions to the classroom, then this dimension will become a non-issue. Obviously, if any of these three results do not occur, teacher resistance to the use of computers may actually increase after adoption. However, school officials can reduce this fear of change by being constantly aware that it is an issue.

CONCLUSIONS

Perhaps the most important message in this paper is that teacher resistance to computers is to be expected and is in some instances justified. But teacher resistance is multidimensional and cannot be prevented or eliminated with a single approach. Some resistance dimensions can be modified through education prior to implementation. Others can only be modified over either time or successful program implementation. The suggestions offered here are but a framework to use to help address the resistance issue. If school officials have a better understanding as to why teachers resist the introduction of computers into the classroom, then perhaps the anxiety and fear which accompanies computer introduction may be lessened. If anxiety and fear can be lessened, then computers should be integrated into the curriculum more quickly and more efficiently.

REFERENCES

Brophy, J. E. (1986). Teacher influences on student achievement. American Psychologist, 1069-1077.

Cuban, L. (1986). Teachers and machines. New York: Columbia University Teachers College Press.

Dupagne, M, & Krendl, K. A. (1992). Teachers' attitudes toward computers: A review of the literature. Journal of Research on Computing in Education, 420-429.

Farina, F., Arce, R., Sobral, J., & Carames, R. (1991). Predictors of anxiety towards computers. Computers in Human Behavior, 7, 263-267.

Kay, R. H. (1989). A practical and theoretical approach to assessing computer attitudes: The computer attitude measure (CAM). Journal of Research on Computing in Education. 456-463.

Kulik, J. A., & Kulik, C. C. (1987). Review of recent research literature on computer-based instruction. Contemporary Educational Psychology, 12, 222-230.

Marche, M. M. (1987). Information technologies in education:

The perceptions of school principals and senior administrators. Educational Technology, 28-31.

Ross, E. W. (1991). Microcomputer use in secondary social studies classrooms. Journal of Educational Research, 85, 39-46.

Schofield, J. W., Evans-Rhodes, D., & Huber, B. R. (1989). Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students. (Technical Report No. 3). Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Stephenson, S. D. (Technical report, in press). The use of small groups in computer based training. Brooks Air Force Base, TX: Armstrong Laboratory, Technical Training Research Division.

Troyer, M. B. (1988). Issues and problems in teacher computer literacy education. Journal of Research on Computing in Education, 141-154.

Violato, C. Marini, A., & Hunter, W. (1989). A confirmatory factor analysis of a four-factor model of attitudes toward computers: A study of preservice teachers. Journal of Research on Computing in Education. 199-213.

Wedman, J., & Heller, M. (1984). Concerns of teachers about educational computing. AEDS Journal, 31-40.

Whitney, R. E., & Urquhart, N. S. (1990). Microcomputers in the mathematical sciences: Effects on courses, students, and teachers. Academic Computing, March, pp. 14, 53.

Investigation of a Non-Punitive Discipline System in a DoD Organization*

Joyce Shettel Dutcher
Navy Personnel Research and Development Center
Russell S. Cooke
San Diego State University
John P. Sheposh
Navy Personnel Research and Development Center

Research on disciplining employees in organizational settings has been notably lacking (Barrick & Alexander, 1987; Arvey & Ivanecivich, 1980). This is particularly true for milder forms of formal discipline (e.g. reprimands) and alternative approaches to traditional discipline (Beyer & Trice, 1984). Several factors contribute to the low level of research activity on this subject. First, there is a prevalent belief that the use of discipline could result in acts of aggression, negative feelings toward the punishing agent, and withdrawal. Katz and Kahn (1978) and other psychologists, therefore, have typically emphasized the use of rewards and downplayed discipline. Second, the bulk of the research that explores discipline and its effect on the individual and the organization is in clinical, educational, and family settings (Cameron & Dupuis, 1991). The results from these studies have yielded useful information concerning the effective modification of behavior within the teacher/student or parent/child relationships, but, for the most part, have limited applicability to the supervisor/subordinate relationship and offer even less in resolving other discipline problems associated with large organizations. Furthermore, the research on discipline in industrial settings has been limited primarily to traditional forms of discipline rather than new alternatives.

Obviously disciplinary actions are a basic part of organizational life and the appropriate application of discipline and rewards can be instrumental in increasing performance (Podsakoff, 1982). Inaction in the face of a problem that requires discipline can produce feelings of inequity and dissatisfaction in employees, which affect productivity. Failure to apply discipline in one case is tantamount to the supervisor waiving enforcement in future cases involving the same issue (Klaas, 1990).

The purpose of the present study was to examine and describe the effects of a non-punitive approach to discipline that was introduced in a DoD organization. It was one of the first such alternative systems introduced in a public sector organization. A major component of the alternative discipline (AD) process was coaching on the part of the supervisor and open communication between employee and supervisor. Supervisors were expected to identify and clearly describe the problem behavior to the employee. Together they would analyze the problem and agree to a strategy for improvement, including a timetable for completion. AD was expected to be less negative and adversarial than the traditional system, provide a more effective way of correcting employee problems, reduce the number of repeat offenders thereby eliminating lost production time due to suspensions, and maintain a high level of morale.

Method

Site

The site at which AD was implemented is located on the east coast and provides supply support for the U.S. Navy. The work force of approximately 3,500 was predominately blue-collar, with many supervisors and managers promoted from blue-collar positions. Two groups of organizational activities were selected to serve as experimental and control groups with approximately 900 employees in each group. The experimental activities--those using AD for disciplinary actions--included the storage, packing, and data processing departments, and the

* The opinions expressed in this paper are those of the authors. They are not official and do not represent the views of the Navy Department.

control activities--those using the traditional discipline system--included the receiving, contracting, and administrative departments.

Intervention

AD was developed by the Labor and Employee Relations Division of the Personnel Office. It was designed to foster behavior change through the use of a positive approach to discipline. The AD approach involved a series of actions progressing in seriousness from the giving of an oral notice to removal. AD differed from the traditional system by establishing a counseling role for the supervisor and the use of letters of warning in place of formal reprimands and suspensions. AD was developed to instruct and provide opportunities to help an individual overcome weaknesses and correct his or her behavior. Effective coaching on the part of supervisors, therefore, consisted of mutual respect, supervisor commitment to develop employees, supervisor focus on improvement, clear communication, and a supportive environment.

Assessment

Several methods were used to assess how well AD was being implemented and its effect on the organization. On-site interviews were conducted with groups of 10 to 15 individuals per session, of similar job status (i.e., managers, supervisors, and personnelists). Union representatives were also interviewed. The interviews were to be conducted each year of a three year test; however, only two of the three originally planned sets of interviews were conducted because of the project's early termination. The first set of interviews was held one year after the introduction of AD, and the second set was conducted the following year. Quarterly progress reports, supplied by the organization, provided information regarding the use of AD and the traditional approach in the form of numbers of reprimands, suspensions, disciplinary actions, repeat offenders, removals, and grievances. Surveys measuring general organizational climate (Gordon & Cummings, 1979) and the perceived usefulness and acceptance of AD were administered to managers, supervisors, employees, and personnelists after the first year AD was in use. A random sample of all non-supervisory employees (N=121, approximately 15% of the experimental and control groups), and nearly all of the managers (N=27), supervisors (N=99), and personnelists (N=51) were surveyed for a total 298 individuals.

Results and Discussion

Table 1 reports the organizational climate means for AD and control activities. Survey respondents at both AD and control activities rated their organization as slightly positive on the climate dimensions.

Table 1
One-Way ANOVAs for AD and Control
Groups on Climate Measures

	AD (133)	Control (106)
Job Satisfaction	4.67[a]	5.03*
Organizational Effectiveness	5.51	5.91*
Org. Accommodation of Change	4.41	4.77*
Resistance to Change	3.93	3.94
Organizational Clarity	4.38	4.59
Decision-Making Structure	4.38	4.58
Organizational Integration	3.66	4.12*
Management Style	4.32	4.53
Human Resource Development	4.05	3.99
Organizational Vitality	4.40	4.67
Organizational Commitment	4.93	5.23*

[a] Responses based on 7-point scales, higher scores are more positive.
*Significant at .05 level.

AD respondents, however, saw their organization as significantly less positive on job satisfaction, organizational effectiveness, accommodation of change, organizational integration, and commitment than control activity respondents. The fact that perceptions of AD respondents were only mildly positive may indicate that the organizational context was not sufficiently suitable for or supportive of an initiative that was a significant departure from the traditional approach.

Statistics for the number of reprimands, suspensions, disciplinary actions, repeat offenders, grievances, and removals for the year prior to AD and during the two years AD was used are presented in Table 2 for AD and control activities. Data from the third year of the test were not available because AD was terminated early. Reasons for the termination will be discussed later.

As can be seen in Table 2, the total number of disciplinary actions taken in the AD and control groups varies from year to year. Under AD reprimands and suspensions were not used, whereas the control group had a fairly large number of both actions in the two years of the test. The elimination of suspensions and reprimands under AD obviously produced savings in terms of processing time required, suspension time, and associated costs. The statistics for Year 1 look positive for the AD group, which had fewer removals, repeat offenders, and grievances than the control group. This trend reverses in Year 2, however, during which many more removals and repeat offenders are seen in the AD group.

Tabie 2
Frequency of Disciplinary Actions Taken Under the Alternative Discipline System (AD)
and Traditional Discipline System (Control)

	Baseline		Year 1		Year2	
	AD	Control	AD	Control	AD	Control
Disciplinary Actions	79	58	74	90	42	46
Reprimands	--[a]	17	0	37	0	24
Suspensions	--	--	0	47	0	19
Removals	6	5	3	5	5	0
Repeat Offenders	7	13	9	23	25	12
Grievances	12	9	5	10	26	28

[a] Not reported

Table 3 presents the evaluations of managers and supervisors from AD and the control activities of their respective discipinary systems. Both systems are seen as effective and fair in dealing with disciplinary problems.

Table 3
Comparison of Mean Evaluations of the Alternative Discipline System
(AD) to the Mean Evaluations of the Traditional Discipline System
(Control) by Managers and Supervisors

	AD	Control
The Discipline System ...	(68)	(52)
is effective in handling disciplinary problems	4.93[a]	4.48
interrupts daily work	3.59	4.48*
is easy to carry out	4.96	4.13*
is fair	5.12	4.73
creates union friction	3.77	4.38
has a negative effect on overall productivity	3.15	3.45
helps employees	5.18	4.54*
creates distrust	3.19	3.46

[a] Responses based on 7-point scales; higher scores are more positive.
*Significant at .05 level.

AD managers and supervisors regarded their system as significantly less disruptive, easier to carry out, and more helpful to the employee. On all other items, AD ratings are similar to traditional discipline system ratings.

In the survey administered in Year 1, managers, supervisors, and personnelists involved with AD were asked to estimate the effect of AD on such things as worker complaints. Table 4 reports the estimated amount of change, reported as increases or decreases, for various organizational factors. Some questions were considered more relevant to one group than another, therefore, questions were tailored to the three groups responding.

Table 4
Managers', Supervisors', and Personnelists' Perceptions of the Consequences of AD

	Managers	Supervisors	Personnelists
AD has increased/decreased...	(19)	(49)	(51)
number of repeat offenders	3.41[a]	3.61	4.90
worker complaints	3.67	--[b]	3.44
supervisory complaints	3.83	--	4.13
union complaints	4.12	--	3.74
production time lost	3.39	3.61	--
time to carry out disciplinary actions	3.61	--	--
problems	3.39	4.09	--
number of disciplinary actions	--	3.60	--
employee responsibility for conduct	--	4.60	--
need to monitor employees	--	4.00	--
time spent advising supervisors	--	--	4.61
number of job responsibilities	--	--	4.35
number of different tasks	--	--	4.30
responsiveness to managers	--	--	4.33

[a] Scores below "4" indicate a decrease, "1" being the most extreme, "4" indicates no change; scores above "4" indicate an increase, "7" being the most extreme.

[b] A dash (--) in this table indicates that respondents were not asked this question.

The responses of the managers, which are presented in the first column of Table 4, form a positive picture of AD. According to managers, since the introduction of AD, the number of repeat offenders, worker complaints, production time lost, and problems associated with the dispensation of discipline have all been reduced. Similarly, supervisors perceive a reduction in repeat offenders and production time lost. They also reported a reduction of disciplinary actions and an increase in employee responsibility for personal conduct. Personnelists, similar to managers, saw a reduction in worker complaints, but they perceived an increase in repeat offenders and greater demands on their job in terms of increased advising time, number of job responsibilities, and variety of tasks.

Table 5 presents the feelings of managers, supervisors, employees, and personnelists concerning AD, which system they preferred, and the extent to which they wanted AD to continue. A score of 5 denotes the highest possible positive score. Clearly managers were very receptive to AD. Supervisors were also positive but to a lesser degree; nonsupervisory employees and personnelists were neutral to slightly negative. Simple effects tests showed significant differences between managers and both employees and personnelists ($p<.05$). The feelings about AD were mixed after the first year, but there was not strong support for continuing the project.

Table 5
Evaluations of the Alternative Discipline System by Managers, Supervisors, Employees, and Personnelists

	Managers (19)	Supervisors (53)	Employees (61)	Personnelists (51)
Positive/Negative feelings about AD?	4.00[a]	3.46	3.05	3.22*
Which system do you prefer (AD or traditional)?	4.12	3.22	3.07	2.95*
Should AD be continued?	3.83	3.27	3.02	2.76*

[a] Responses based on 5-point scales; higher scores are more positive.
*Significant at .05 level.

In the second year of the test, AD was terminated based upon a formal request from the union and concurrence from the Personnel Office. Information obtained from interviews conducted after AD was in operation for one year, and at the end of the second year, provide some understanding as to why the AD experiment was prematurely ended. The majority of managers interviewed were favorable toward AD in both interview periods. Among its beneficial effects were: (a) the avoidance or loss of an employee as the result of a suspension, (b) a less time consuming and cumbersome process, (c) a better way to identify the problem and to involve the employee in dealing with the problem, and (d) a more consistent, progressive series of disciplinary steps, building a more solid basis for future disciplinary action. Among the difficulties associated with AD were: (a) the more complex role of the supervisor, requiring counselling as well as disciplining, and (b) the unwillingness of some employees to communicate in the coaching sessions, which rendered the approach less effective.

Supervisors were less positive, and felt that in some instances AD was too lenient and had no appreciable effect on the employee. While many supervisors liked the system, some felt that they were burdened with additional work (e.g., more writing), and felt that they were not fully prepared to conduct face-to-face coaching sessions with employees.

Union representatives felt that conceptually AD was an extremely good idea, but they were opposed to the way it was put into practice. Because employees were not given the type of feedback they were accustomed to, they did not clearly understand that they were actually being disciplined. The union representatives felt disciplined employees were not given adequate opportunity to deny wrongdoing and appeal actions taken against them. Furthermore, the interviewees felt that union representatives should be, but had not been, available at the earliest stages of the discipline process. They also cited inconsistencies in the application of AD.

Personnelists interviewed after the first year of AD felt that it was an excellent idea but that it was not working, could not be fixed, and should be scrapped. These same sentiments were expressed even more forcefully in the second interviews a year later. Personnelists believed that AD was given a fair test and it just did not work. They felt it had started out as an innovative program but it evolved into the same old way of dealing with disciplinary problems. They cited several reasons for the ineffectiveness of AD: (a) the organization had not really bought off on it; there was little evidence of interest or support from top management, (b) a large segment of the supervisors was not effective in face-to-face meetings, (c) a great amount of time and effort was required of the personnelists to coach and advise supervisors, and (d) blue collar workers were not accustomed to responding to this type of discipline and had a difficult time understanding that they were being disciplined. The absence of constructive communication in many AD sessions, union disruption, the continuing need for heavy personnelist support, and the absence of strong management interest and support, they felt, overrode the benefits of time and money savings achieved through AD.

Conclusions

Despite the early promise of the AD system, by the second year of its use employees were increasingly dissatisfied with it and supported its termination. Those most removed from conducting AD--the managers--saw it most positively. Those who were more directly involved with carrying it out--supervisors and personnelists--were less supportive of its continued use. Interestingly, the factors that were instrumental to the unsuccessful test of AD are those discussed by authorities of implementation and institutionalization of change (Roberts-Gray & Scheier, 1988, Goodman and Bazerman, 1930). Among the factors they identify are: (a) congruence of the new approach with existing organizational values, policies, and structure, (b) effective intergroup dependencies, (c) union-management involvement in all phases of the test, and (d) congruence between expected and actual outcomes. All of these is^ues were raised in the interview sessions. The AD system was seen as positive in isolation, however, the other values and policies in the organization remained consistent with a hierarchical organization rather than supportive of a cooperative relationship among supervisory and nonsupervisory employees. The key participants in the efforts to improve behavior--the employee and supervisor--were not comfortable in nor adept at the new roles prescribed by the system, and continuing facilitation of the counseling sessions by personnelists was required. The inability of supervisors to administer the new system without extensive support from personnelists and the exclusion of the union in the development and implementation of the project led to intergroup friction regarding discipline. The actual outcomes of the effort fell far short of expectations. Perceptions of the extent of support activities needed to improve the system (e.g., training), and the lack of strong support from top management for the project's continuation, led to its discontinuance.

The implications of these findings are that consideration of adoption of an organizational change should not be determined solely on the merits or properties of the change, but also should address the appropriateness of the change to the organization, how accommodating the organization is to the change, and the chances for a successful implementation. If the change is adopted, potential problem areas can be addressed in the implementation effort. In conclusion, a deeper understanding of the impact of these factors at the outset is required if innovative changes in such critical areas as discipline are to be given a fair and full test.

References

Arvey, R. D., & Ivancevich, J. M. (1980). Punishment in organizations: A review, propositions, and research suggestions. *Academy of Management Review*, *5*(1), 123-132.

Barrick, M. R., & Alexander, R. A. (1992). Estimating the benefits of a quality circle intervention. *Journal of Organizational Behavior*, *13*, 73-80.

Beyer, J. M., & Trice, H. M. (1984). A field study of the use and perceived effect of discipline in controlling work performance. *Academy of Management Journal*, *27*(4), 743-764.

Cameron, J., & Dupuis, A. (1991). The introduction of school mediation in New Zealand. *Journal of Research and Development in Education*, *24*(3), 1-13.

Goodman, P. S., & Bazerman, M. (1980). Institutionalization of planned organizational change. In L. L. Cummings & B. M. Staw (Eds.), *Research in Organizational Behavior, Vol. 2*. Greenwich, CN: JAI Press Inc.

Gordon, G.G., & Cummings, W. (1979). *Managing management climate*. Lexington, MA: Lexington Books.

Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations*. New York: Wiley.

Klaas, B. S. (1989). Managerial decision making about employee grievances: The impact of the grievant's work history. *Personnel Psychology*, *43*, 53-68.

Podsakoff, P. M. (1982). Determinants of a supervisor's use of rewards and punishments: A literature review and suggestions for future research. *Organizational Behavior and Human Performance*, *29*, 58-83.

Roberts, Gray, C., & Scheire, M. G. (1988). Checking the congruence between a program and its organizational environment. In K. J. Conrad and C. Roberts-Gray (Eds.), *Evaluating program environments: New directions for program evaluation*, *40*, 63-81.

Characteristics and Organizational Orientation of Empowered Employees*

Mark B. Rosenthal
California School of Professional Psychology

John P. Sheposh
Joyce Shettel Dutcher
Navy Personnel Research and Development Center

Abstract

Interest in the concept of empowerment has grown in the last decade. However, organizational research has only recently begun to identify the cognitive elements of empowerment (Thomas and Velthouse, 1990) and its relationship to organizational or other individual characteristics (Conger and Kanungo, 1988). The present research was an effort to further contribute to this understanding by identifying some of the specific organizational and individual factors which distinguish employees who differ on empowerment. A survey instrument was used as the primary means of data collection. Two U.S. Navy engineering facilities served as the sites at which the survey administrations were conducted. The results, based on responses from 368 supervisory and nonsupervisory employees, revealed that empowerment was significantly related to such demographics as supervisory level, education, and age. In addition, employees who differed on relative levels of empowerment differed significantly on perceptions of their organization, job satisfaction and job stress. Specifically, employees who scored high on a measure of empowerment reported greater communication, cooperation, openness, customer-orientation, and opportunities for creativity within the organization, greater job satisfaction, less constraining work conditions and less job stress than employees who scored lower on empowerment. Significant differences were also noted in perceptions of a Total Quality Leadership (TOL) improvement effort, with highly empowered employees reporting more involvement in and greater personal and organizational acceptance of such an effort as well as fewer perceived impediments to its implementation than less empowered employees. It was concluded from the results that certain contextual factors consistent with Conger and Kanungo's (1988) conceptualization contributed to the perceived characteristics of the organization.

Background

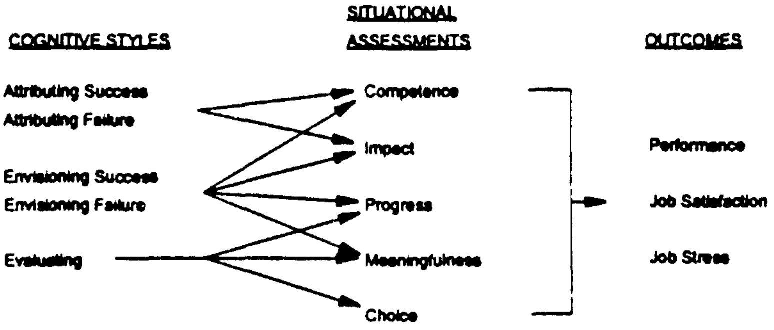
Competitive pressures on American organizations have brought attention to the need for self-motivated and innovative or empowered organizational members. Interest in the concept of empowerment has grown steadily in the last decade. Organizational researchers have given prominence to three theoretical perspectives: the structural perspective, the leadership perspective, and the individual (or self-empowerment) perspective (Tymon, 1988). The structural perspective focuses on organizational mechanisms or processes which ensure that people can get the power they need to innovate. The leadership perspective views empowerment as a product of the interaction between a leader and his or her subordinates. Through the delegation of authority and the sharing of resources leaders foster empowerment. The individual perspective conceptualizes empowerment as an intrapersonal motivational state through which individuals empower themselves. A basic assumption underlying this perspective is that empowering experiences are a product of not only organizational structures and delegation, but also of various other factors, such as cognitive styles and attribution processes which moderate feelings of self-efficacy.

With the recent advancement of the individual perspective, organizational research has begun to identify the cognitive elements of empowerment (Thomas and Velthouse, 1990) and their relationship to organizational or other individual characteristics (Conger and Kanungo, 1988). Using an adaptation of Thomas and Velthouse's (1986) cognitive model, Tymon (1988) proposed an integrative approach to empowerment which serves as the theoretical basis for the current study (see Figure 1). According to Tymon, cognitive styles directly influence

* The opinions expressed in this paper are those of the authors. They are not official and do not represent the views of the Navy Department.

situational assessments which directly influence performance, job satisfaction, and job stress. Situational assessments are viewed as the psychological components of empowerment, while the cognitive styles represent critical intrapersonal processes influencing these situational assessments.

Figure 1
Tymon's Cognitive Model of Empowerment



Preceding the cognitive interpretation of Tymon was the work of Conger and Kanungo (1988). They identified antecedent conditions and practices that affect empowerment. Lack of meaningful tasks, low task variety, and poor organizational communication are scme of the factors that they cite as lowering empowerment.

In this study, an adaptation of Tymon's (1988) conceptualization is used as the basis for operationalizing empowerment. The purpose of the present study was: 1) to assess the level of empowerment among employees at two DoD engineering facilities; 2) to determine the extent to which some of the contextual factors identified by Conger and Kanungo (1988) are related to empowerment, i.e. do differentially empowered employees perceive specific organizational and individual factors differently?; and 3) to assess differences in perceptions of factors relating to the implementation and use of a Total Quality intervention among differentially empowered employees. The selected individual and organizational characteristics were all expected to differ among employees who felt differentially empowered in their jobs.

Method and Procedures

Overview

The current study was designed to assess the relationship between empowerment and selected individual and organizational characteristics. An administration of a survey instrument was conducted at two U.S. Navy engineering facilities. The workforce at each site is predominantly technical. Approximately 40% of the workforce is comprised of scientists and engineers, 30% logisticians and administrators, 20% technicians, and 10% clerical.

Materials

A survey instrument, which was used as a primary means of data collection, included the following content areas: 1) general organizational climate characteristics such as openness, communication, and cooperation (Gordon & Cummins, 1979; Klauss & Bass, 1982); 2) constraints on the immediate work situation (Peters & O'Conner, 1980), task preparation, relevant information, materials and supplies; 3) perceived relationship between rewards and performance; 4) perceived involvement and recognition in the conduct of work, which is an indicator of the integration of employees into the workings of the organization (Hatvany & Pucik, 1982); 5) motivating potential of one's job based on specific job characteristics (Hackman & Oldham, 1980); 6) empowerment, which included an assessment of cognitive styles (e.g., attributing success) and situational assessments (e.g., competence, impact) (Tymon, 1988); 7) job satisfaction (Young, Reidel, & Sheposh, 1975);

8) involvement and participation in TQL activities; 9) impediments to the successful implementation and operation of TQL (e.g., fear, lack of adequate training, lack of support); and 10) personal and organizational acceptance of TQL. The groupings of the items for each of these areas were confirmed by means of factor analyses using varimax rotation. All non-TQL items employed a 7-point scale ranging from 1: "Strongly Disagree" to 7: "Strongly Agree." TQL-related items used a 7-point scale ranging from 1: "Not At All' to 7: "A Very Great Extent." In addition, survey respondents were invited to provide written comments on the issues addressed in the survey.

Subjects

Three hundred and sixty-eight randomly selected employees (approximately 15%) across all organizational levels were selected from one site and all 73 employees from the second site participated in the survey. In all, 260 nonsupervisory and 108 supervisory employees completed and returned the survey. Based on the survey results of the empowerment scores for the total sample, respondents were split into three equal-sized groups differing on relative measures of empowerment. For purposes of reporting, the groups will be referred to as "Low", "Medium", and "High" with the understanding that these titles are somewhat of a misnomer--a majority of even the "Low" group indicated empowerment scores above the scaled mid-point of 4.

Results

Preliminary results revealed an overall mean of 5.57 (SD=.865) on a 7-point scale for the empowerment scale across the two sites. The sample of respondents in general perceived themselves as highly empowered. Chi-square analyses were conducted on each demographic variable from the questionnaire to determine its relationship to reported empowerment. These analyses revealed that age ($X^2(14, \underline{N}=336) = 24.24, p<.05$) and level of education ($X^2(16, \underline{N}=336) = 35.27, p<.005$) were related to empowerment. Older workers were more likely to report higher levels of empowerment than younger workers. More specifically, while 41% of workers over 40 reported empowerment scores falling within the High group only 24% of workers under age 40 fell into this category. A significant X^2 was also obtained for education level. Those reporting the most empowerment tended to be employees with some college and accompanying technical training or employees having earned a graduate degree. These findings indicate that for those positions which require a college degree it would appear that individuals with only a bachelor's degree felt the least empowered. Similarly, for positions that do not require a bachelor's degree, individuals with less education were less empowered. In addition, when empowerment was analyzed in comparison to supervisory level, it was found that over 75% of the supervisors who responded to the questionnaire reported moderate to high levels of empowerment while only 62% of nonsupervisory employees reported moderate to high empowerment scores. A chi-square analysis indicated that in comparison to nonsupervisory employees, a significantly higher proportion of supervisors fell into the High group relative to the other empowerment groups ($X^2(1, \underline{N}=181) = 4.46, p< .05$). Among the remaining demographic variables, e.g. gender, ethnicity, length at present pay grade, etc., none were significantly related to empowerment.

A series of 1-way analyses of variance was also performed to assess the differences in employees' perceptions on selected organizational factors across the differentially empowered groups (see Table 1). As can be seen in Table 1, employees in the High group reported more favorable perceptions of their organization and their jobs than those in either the Medium or Low groups. Particularly impressive were the differences obtained for the motivating potential of one's job (MPS). The most highly empowered employees were clearly more positive on this index than less empowered employees. Furthermore, these employees expressed greater job satisfaction and less job stress than workers in the Medium or Low groups.

Table 1
Mean Responses to Organizational and Individual Job Factors across Groups

	Low	Medium	High	F Ratio[d]
Overall Organizational Climate	3.77[a]	4.22	4.81	22.58
Communication	3.69	4.32	4.87	24.88
Openness	3.66	· 4.09	4.72	14.86
Cooperation	3.36	3.51	3.87	3.29
Customer Orientation	4.40	4.72	5.22	12.11
Creativity	3.90	4.61	5.27	22.88
Performance Appraisal	4.00	4.71	5.18	22.37
Lack of Control Over Work	3.95[b]	3.65	2.88	24.30
Lack of Sufficient Support	3.64[b]	3.25	2.74	14.48
Motivating Potential of the Job (MPS)[c]	97.17	152.78	220.34	112.68
Job Satisfaction	4.46	5.42	6.15	53.55
Job Stress	4.36[b]	3.99	3.42	20.37

[a] The higher the value, the more positive the response.
[b] The lower the value, the more positive the response.
[c] $MPS = \frac{[Skill\ Variety + Task\ Identity + Task\ Significance]}{3} X\ [Autonomy]\ X\ [Feedback].$
Scores can range from 1 to 343.
[d] All F values are significant at the p < .0001 level, except for "Cooperation" for which the F value is significant at the p < .05 level.

Table 2 represents responses concerning TQL-related factors across the three groups. Employees in the High group were more personally involved in TQL implementation, reported higher personal acceptance, and perceived greater organizational acceptance of TQL. In addition, they saw factors potentially impeding the successful implementation of TQL as less severe than employees in the Medium or Low groups.

Table 2
Mean Responses to TQL Factors across Groups

	Low	Medium	High	F Ratio[c]
Organizational Acceptance	3.46[a]	4.05	4.34	13.46
Personal Acceptance	4.46	4.98	5.21	5.69
TQL Involvement	3.84	4.31	4.83	8.25
Impediments to Implementation	3.73[b]	3.38	3.18	6.09
Fear	2.98	2.43	2.35	7.53
Lack of Knowledge	4.48	3.89	3.55	7.11
Lack of Support	4.27	3.85	3.43	7.05

[a] The higher the value, the more positive the response.
[b] The lower the value, the more positive the response.
[c] All F values significant at the p < .005 level.

Summary and Conclusions

The current findings have implications for research and practice. They provide strong theoretical support for the cognitive model of empowerment. The results show that the application of the cognitive model of empowerment as conceptualized by Thomas and Velthouse (1986) and Tymon (1988) successfully differentiated individuals in terms of their perceptions of relevant organizational factors. The most highly empowered employees differed from less empowered employees on a number of perceived organizational and job dimensions. These include some of the conditions proposed by Conger and Kanungo (1988), hypothesized to foster empowerment. Of particular interest was the differential response to TQL based on empowerment. Employees in the High group were favorably disposed and more involved in the implementation of TQL. Consistent with the cognitive theory it would appear that individuals who see themselves as having higher competency levels are more willing to be involved in such efforts as TQL.

The cognitive approach employed in this study provides an alternative to the leadership perspective which places major focus on what leaders can do to enhance empowerment. The cognitive model suggests a more complex process. It is one that takes into account the effects of interventions and/or contextual factors that encourage individuals to act in accordance with their intrinsic motives. But the model also focuses upon the cognitive styles and global assessments of individuals. By including individual differences in terms of cognitive styles this model is general enough to ascertain the effect of empowerment as it relates to different types of interventions or changes. Thus in most cases when a change is introduced, empowered individuals as identified by the cognitive model can be selected for involvement in the adoption and implementation of that change.

Continued research into empowerment requires longitudinal studies so that the causal relationships between organizational factors and empowerment are more clearly established. Such studies could also help to determine the extent to which TQL--which ideally should provide workers the opportunity to participate in decision making and enhanced responsibility over job processes--heightens intrinsic motivation and empowerment.

References

Conger, J.A., & Kanungo, R.N. (1988). The empowerment process: Integrating theory and practice. Academy of Management Review, 27, 454-489.

Gordon, G.G., & Cummins, W. (1979). Managing management climate. Lexington, MA: Lexington Books.

Hackman, J.R., & Oldham, G.R. (1980). Work Redesign. Reading, MA: Addison-Wesley.

Hatvany, N., & Pucik, V. (1982). Japanese management in America: What does and doesn't work. National Productivity Review, Winter, 61-74.

Klauss, R., & Bass, B. (1982). Interpersonal communication in organizations. Orlando, FL: Academic Press.

Peters, L.H., & O'Conner, E.J. (1980). Situational constraints and work outcomes: The influences of a frequently overlooked construct. Academy of Management Review, 5, 391-397.

Thomas, K.W., & Velthouse, B.A. (1986). Cognitive elements of empowerment. Paper presented at the meeting of the National Academy of Management, Chicago, IL.

Thomas, K.W., & Velthouse, B.A. (1990). Cognitive elements of empowerment: An "interpretive" model of intrinsic task motivation. Academy of Management Review, 15(4), 666-681.

Tymon, W.G. (1988). An empirical investigation of a cognitive model of empowerment. Published dissertation. Ann Arbor, MI: U.M.I.

Young, L.E., Reidel, J.A., & Sheposh, J.P. (1979). Relationship between perceptions of role stress and individual, organizational, and environmental variables. (NPRDC TR 80-8). San Diego, CA: Navy Personnel Research and Development Center.

MOSKOS' INSTITUTIONAL AND OCCUPATIONAL ORIENTATIONS: A TRACKING THROUGH THE 1980'S[1]

Trueman R. Tremble, Jr., & Gerald F. Goodwin

U.S. Army Research Institute for the
Behavioral and Social Sciences

Beginning in 1976, Moskos advanced the notion that the military organization is moving from an institutional (INS) to an occupational (OCC) model (Moskos, 1976; 1981; 1983; 1986). In the INS model, individual interests and competencies are subordinate to the organization, and organizational values and norms have broad implication for the lives of military members and their families. In the OCC model, individual interests and competencies are supported by their prevailing values in the larger economic market, and organizational control tends to be limited to the specifics of the work place. This shift from INS to OCC was viewed as a change in organizational structure which, Moskos implied, would ultimately be represented in the organizational commitments or role orientations of individual members of the military.

As part of a quality of life survey, Stahl, Manley, and McNichols (1978) derived and validated questionnaire measures of the INS and OCC orientations of Air Force members. Higher ranking Air Force members were found to have higher INS and lower OCC scores than junior members. INS scores were positively correlated with time in service, career intent, and satisfaction. Negative relationships were found for the OCC orientation.

The Air Force scales are relatively short. In 1981, these scales were adapted for research on the organizational commitment of members of the U.S. Army. From 1981 through 1990, the adapted INS and OCC scales were administered as part of surveys that differed in purpose and in the demographic characteristics of the soldiers surveyed. Tremble and Brosvic (1987) summarized findings of the six surveys conducted from 1981 through 1986. This paper integrates results of the 1990 survey to track three general issues through the decade of 1980:

(1) The reliability or robustness of the adapted INS and OCC scales for groups of soldiers that, across surveys, differed in the maturity of their military careers.

(2) The validity of the scales for the Army samples.

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the position or policy of the U.S. Army Research Institute or the Department of the Army.

(3) The transition of role orientations of soldiers from INS to OCC.

METHOD

The sample was 523 soldiers assigned in two U.S. Army infantry battalions in spring 1990. Of this sample, 73.1% were in the grade of E4 or below, and only 2.8% were commissioned officers. Average time in service was approximately 20 months. With respect to career intentions, the modal response was "undecided". The 1990 sample was relatively more junior than the samples in the earlier surveys in terms of both grade (median of the sample percents in grade E4 or less was 58%) and average time in service (median of the sample averages was about 32 months).

The 1990 survey was administered to soldiers in groups that typically consisted of all platoon members and leaders assigned to a company. The soldiers first responded to a larger questionnaire that measured organizational conditions (leadership, cohesion, motivation, etc.) under investigation for their effects on unit performance. The larger questionnaire yielded data on sample characteristics and scale scores for job satisfaction, organizational identification, and job involvement (see Tremble & Alderks [1992] for a description of scales). The soldiers then completed a supplementary questionnaire and used 5-point scales ranging from strongly disagree (1) to strongly agree (5) to respond to the eight items modified to measure the INS and OCC orientations in Army samples.

RESULTS AND DISCUSSION

Robustness of the INS-OCC Scales

To assess robustness, the approach used by Stahl et al. for scale derivation was followed. That is, responses to the eight INS-OCC items were factor analyzed (principal components with varimax rotation). Table 1 summarizes the three factors obtained in the 1990 survey.

The first factor had items with strong, positive loadings by only the four INS items, and this pattern replicated the INS factor originally obtained in the Air Force sample. Responses to the four items defining this INS factor indicated agreement that: soldiers should have more interest in mission accomplishment and less interest in their personal concerns (mission accomplishment); more soldiers should really care about national security (national security); lower ranking soldiers need to be supervised more (more supervision); and there is not enough discipline in the Army (more discipline).

The four OCC items had highest loadings on the second and third factors instead of defining a single factor representing the expected OCC scale. One factor suggested satisfaction with the securities of the military: disagreement that a person can

Table 1

Summary of Rotated Factor Loadings and Correlations with Validation Variables

	Sample														
	1981a		1981b		1983a		1983b		1984		1986		1990		
	INS	OCC	INS	OCC	INS	OCC	INS	OCC	INS	OCC	INS	OCC	INS	SAT	FRUS
Rotated Loadings															
Mission Accomplishment	.62	-.05	.60	.22	.57	-.27	.56	-.29	.62	-.27	.59	-.27	.57	.40	.05
National Security	.67	-.04	.60	.04	.66	-.01	.66	.09	-.43	-.26	.59	.11	.59	.21	-.15
More Supervision	.66	.24	.65	.03	.68	-.04	.65	-.10	.74	.01	.67	-.19	.76	-.02	-.17
More Discipline	.73	.07	.66	.03	.69	.02	.63	.15	.75	.07	.74	-.06	.81	.00	.02
Relative Equity	-.12	-.63	.03	-.69	-.08	.73	.05	.74	-.02	.72	-.06	.77	-.07	-.47	.56
Job Opportunities	-.02	.75	.10	.69	.08	-.67	.07	-.61	.24	-.67	.22	-.55	.01	.68	-.26
Non-Job Activities	.27	-.49	-.09	-.34	-.01	.62	.01	.55	.09	.61	-.02	.62	-.13	.04	.88
Post Good	.31	.59	.05	.70	--	--	--	--	.31	-.63	-.24	-.62	.20	.79	.13
% Variance Accounted	45%		40%		46%		42%		47%		45%			58%	
Scale Score Properties															
Mean	3.44	2.84	3.04	2.45	3.36	3.09	3.22	3.49	3.17	2.96	3.27	2.92	3.38	2.47	3.34
Standard Deviation	.79	.75	.75	.75	.70	.82	.69	.74	.73	.59	.77	.60	.77	.82	.87
Sample Size	301		1672		1531		2466		416		4807			523	
Correlations with Validating Variables															
Career Intent	.18	.51	.21	.40	.24	-.29	--	--	.45	.17	.32	.02	.26	.46	-.30
Grade	.39	.02	.26	.06	.39	-.22	--	--	.38	.02	.25	-.12	.31	.19	-.24
Time in Service	.47	.32	.19	.07	.33	-.14	--	--	.42	.12	.31	-.08	.17	.04	-.10
Job Satisfaction													.38	.46	-.34
Organizational Identification													.45	.43	-.25
Job Involvement													.43	.38	-.28

Notes. SAT="satifaction". FRUS="frustration". The "post good" item was not included in the 1983 surveys, and correlations were not computed for the 1983b sample since it was in the process of exiting service. Correlations greater than .05 were statistically significant given sample sizes.

get more of an even break as a civilian than as a soldier (relative equity), agreement that it would be hard today to find a civilian job as good as the current job (job opportunities), and agreement that an Army post is a good place to live (post good). The other factor suggested the types of job frustrations expected with an OCC orientation, that is, agreement that: a person can get more of an even break as a civilian than as a soldier (relative equity) and my supervisor makes me do too many things that are not related by my job (non-job activities).

This factor structure fits results of the earlier surveys (see Table 1). In all surveys, the INS factors was obtained, and the relative levels and directions of factor loading on the INS factors have presented only minor deviation from expectation. Across surveys, the OCC orientation, as measured by the four items used here, has been less robust. In particular, the two 1981 surveys produced a factor with highest loadings by the four OCC items, but the factor loadings of those items were algebraically reverse from expectation. The reserve loadings made the 1981 OCC factors generally similar to the 1990 factor suggesting satisfaction with the military.

A major difference between the surveys that did and the surveys that did not produce the expected OCC factor was the career maturity of the samples. The two 1981 and the 1990 samples were composed of soldiers relatively lower in rank and time in service than the samples in the other surveys. This raises the possibility that an orientation like INS is broadly applicable in the Army. The OCC orientation, however, may be viable for only those soldiers who have been in the Army--or in a career--for some period of time.

Validity of the INS-OCC Scales

For subsequent analyses, scale scores (ranging from 1 to 5) for the INS factor and the other two factors (labelled "satifaction" [SAT] and "frustration" [FRUS]) were formed by averaging responses to the items highlighted earlier as having highest loadings on the factors. Responses to items with negative factor loadings were reverse scored before averaging.

While only moderately strong at best, the positive correlations between INS and the originally used validation variables of rank, career intent, time in service, and job satisfaction were statistically significant. In contrast, the correlations between FRUS scores and the original validation variables were negative and statistically significant. With one major exception, the correlations for SAT were positive, statistically significant, and comparable in magnitude to those obtained for INS. SAT also had the relatively strongest correlation with career intentions.

These results for INS and FRUS (as a component of OCC) fit both expectations and the patterns obtained in the earlier Army surveys. The 1990 survey also included measures of organizational identification and job involvement that, given the INS-OCC definitions, should have positive correlations with INS and negative or no association with OCC. Results (Table 1) supported those expectations as well.

Transition from INS to OCC

Findings on changes in the levels of INS scores across the 1980's bear on Moskos' hypothesis of a transition of the military role orientation from INS to OCC. For this issue, comparison of all surveys except the 1983b survey is pertinent since all samples except the latter consisted of active duty soldiers. In contrast, the 1983b sample consisted of soldiers leaving service after having successfully completed their service terms.

As summarized in Table 1, the obtained mean INS scores demonstrated remarkable stability across the decade of 1980, despite differences in the demographic compositions of the samples.

The issue of a shift from INS to OCC can also be addressed by comparing the INS and OCC orientations of soldiers exiting service with those remaining in service. The argument of an unidirectional shift in role orientation suggest no difference in the INS and OCC orientations of these two groups of soldiers, in contrast to the traditional expectation of continued ascendance of INS in soldiers remaining in service.

As indicated earlier, the two 1983 surveys differed in sample composition. The 1983a sample consisted of those making a permanent change of duty station and remaining on active duty. Soldiers in the 1983b samples were voluntarily exiting service at the end of their terms of service. Significant one-way analyses of variance confirmed that the soldiers remaining in service had significantly higher INS scores ($\underline{F}$ (1,3925) = 36.06, $\underline{p} < .0001$) and significantly lower OCC scores ($\underline{F}$ (1,3924) = 239.50, $\underline{p} < .0001$) than soldiers exiting service. These differences were such that even if the Army's organizations had become more OCC in nature, the INS role orientation continued to be more characteristic of individual soldiers remaining on duty than of those soldiers disengaging from the Army. In fact, the OCC orientation tended to differentiate the 1983 soldiers remaining and exiting service to a greater extent that did the INS orientation.

SUMMARY AND CONCLUSIONS

Results generally support the reliability and validity of the adapted INS-OCC scales, especially for U.S. soldiers with some organizational experience. Across the 1980's, no dramatic decline in the INS orientations of soldiers was apparent; and, based on the 1983 surveys, the U.S. Army continued to retain soldiers with relatively higher INS and relatively lower OCC orientations. The emergence of an OCC orientation for only those soldiers with some amount of organizational experience and the robustness of the INS orientation support Moskos's contention that organizational structure--or the socialization experiences engineered by that structure--can influence the organizational commitments or identities which its members develop. Such results also suggest that the organizational commitment judged important for military effectiveness is a candidate criterion for decisions about the structure of the future U.S. military.

REFERENCES

Moskos, C. C. (1977) From institution to occupation: Trends in military organization. Armed Forces and Society, 4 (1), 41-50.

Moskos, C. C. (1981) Institution versus occupation: Contrasting model of military organization. Final Report, AFOSR-TR-81-0295. Bolling AFB, DC: Air Force Office of Scientific Research.

Moskos C. C. (1983) Contrasting models of military social organization. Unpublished Technical Report. Evanston, IL: Northwestern University.

Moskos, C. C. (1986) Institutional/Occupational trends in armed forces: An update. Armed Forces and Society, 12 (3), 377-382.

Stahl, M. J., Manley, T. R., & McNichols, C. W. (1978) Operationalizing the Moskos institution-occupation model: An application of Gouldner's cosmopolitan-local research. Journal of Applied Psychology, 1978, 63 (4), 422-427.

Tremble, T. R., Jr., & Alderks, C. E. (1992) Measures for research on small unit preparedness for combat effectiveness. Research Note 92-03. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Tremble, T. R., Jr., & Brosvic, G. M. (1987) Moskos's institutional-occupational orientations: A review of research from 1981-1986. In Proceedings, Third Annual Leadership Research Conference. Fort Leavenworth, KS: Center for Army Leadership.

CRITERION-REFERENCED TESTING'S ROLE IN NAVY ENLISTED PROMOTIONS

Don C. Phillips
Naval Education and Training Program Management Support Activity
Pensacola, Florida

INTRODUCTION

The U.S. Navy requires enlisted advancement candidates in paygrades E-4 through E-7 to pass a criterion-referenced test. This test is the Military/Leadership (Mil/Lead) examination. The Mil/Lead examination is a result of the Chief of Naval Operations' emphasis on Navy tradition, integrity, and professionalism.

Mil/Lead exams are qualifying exams. Commanding officers use them as part of the overall qualifying process to determine a sailor's qualifications for advancement. Passing of the Mil/Lead exam shows minimum NAVSTD knowledge of the next higher paygrade. They aid commanding officers in identifying candidates for advancement who are knowledgeable in military subject areas. The individual command administers and grades the Mil/Lead exams.

The basis of the Mil/Lead exams is the naval standards (NAVSTDs). While occupational standards (OCCSTDs) are the Navy's minimum requirements for enlisted occupational skills, NAVSTDs are a systematic listing of those minimum capabilities the Navy expects and requires of individuals within each rating (occupation). NAVSTDs represent the abilities, skills, and knowledges, other than those defined by occupational standards (OCCSTDs), that are essential to the overall effectiveness of enlisted personnel in the performance of their duties. NAVSTDs express the nonrating-specific skill and knowledge requirements for enlisted personnel in paygrades E-2 through E-9. They are written in the form of skill or knowledge statements to aid enlisted personnel, commanding officers, and personnel managers in identifying the basic military requirements for enlisted personnel. These personnel must show that they have the capability to perform a skill or that they have attained the knowledge described by a NAVSTD as part of their advancement process. NAVSTDs are cumulative; that is, personnel are responsible for the NAVSTDs of the paygrade they are trying for, of their present paygrade, and of all lower paygrades. NAVSTDs encompass military requirements; essential virtues of pride of service in support of the oath of enlistment; maintenance of good order and discipline; and basic skills and knowledges about the well-being of Navy personnel, which directly contribute to the mission of the Navy. (Manual of Navy Enlisted Manpower, 1991)

Candidates gain their knowledge by studying those publications listed as supporting bibliography for the NAVSTDs (Military Requirements training packages). This bibliography, like the NAVSTDs, is cumulative.

BACKGROUND

The Navy has tested military subject areas since 1958, although early testing was part of the Navywide enlisted examination and, as such, was a norm-referenced test. The Mil/Lead exam used today began in the Bureau of Naval Personnel in late 1964; however, actual criterion-referenced testing did not begin until July of 1966. Early testing was for E-4 and E-5 personnel. Men and women candidates took different exams. This system remained in effect until 1975 when E-4 and E-5 exams were changed to today's form. In 1984 E-6 and E-7 personnel became part of the testing scheme. Since 1975 the only significant change in testing was the addition of a separate 25-question section for personnel in the Construction (SeaBee) ratings.

Mil/Lead exams are an annual evolution. Writing teams develop pretests in three forms (A, B, and C). Personnel at select fleet locations take these pretests. After the pretest validation process, statistics become available and preparation begins on two final form (A and B) exams, consisting of 100 items each, for each paygrade.

PURPOSE

Mil/Lead examinations are criterion-referenced (qualifying) rather than norm-referenced (discriminating). The purpose of the examination is to determine if a candidate has attained a prescribed minimum amount of knowledge of specific subject matter as stated by the Chief of Naval Personnel.

Examination writers must understand the difference between qualifying and discriminating examinations, the latter of which presupposes the qualification of all candidates and strives toward the goal of discriminating between the best qualified and those less well qualified.

The distinction between the two types of examinations and the purpose of a qualifying Mil/Lead exam must remain paramount in the examination writer's mind throughout the development of the examination series. The writer's goal is to be sure that in developing items, he or she does not exceed the minimum knowledge, or level of difficulty, specified by the qualifications (NAVSTDs).

DEVELOPMENT POLICY AND GUIDELINES

In developing Mil/Lead exams, writers follow the policy and guidelines of the Chief of Naval Education and Training (CNET) and the Naval Education and Training Program Management Support Activity (NETPMSA) directives; textbooks/handbooks concerning testing techniques, educational measurement, and personnel assessment; and references on punctuation and English grammar and usage. They also ensure the Mil/Lead exams comply with the Bibliography for Advancement Study, NAVEDTRA 12052; Examination Writing Standards Manual, NETPMSAINST 1552.1A; Advancement Examination Development Manual, NETPMSAINST 1418.1A; and the Manual of Navy Enlisted Manpower and Personnel Classifications and Occupational Standards, Volume I, NAVPERS 18068F.

QUALIFICATIONS TO BE TESTED

The NAVSTDs section of NAVPERS 18068F specifies the minimum qualifications that a candidate taking the Mil/Lead exam must achieve. The Manual of Navy Enlisted Manpower (1991) explicitly defines NAVSTDs as "the minimum capabilities required of naval personnel"; "subjects of which [those] personnel should have skill or knowledge"; "[subject] universal to all rates and ratings"; and "basic things which should be known, but not necessarily done as a matter of routine."

EXAMINATION PROFILE

NAVSTDs are a specific profile (brief description of the duties of each paygrade within the U.S. Navy) of the minimum knowledge that a candidate must have. Therefore, NAVSTDs are equivalent to the occupational profile used in developing the discriminating (rating) examinations and formulate the Mil/Lead exams.

TEST PLAN AND TEST OUTLINE

A test plan and a test outline (TP & TO) for a Mil/Lead exam are developed from the NAVSTDs specified for the paygrade we are to examine.

The TP is a statement that identifies the knowledges, skills, and abilities associated with the subject matter that the particular exam will test. The TP sets up the exam's rationale. As a minimum the TP describes the topics we will test; specifies the degree of knowledge and skills for which candidates will be responsible; and outlines the flexibility we are allowed in testing the subject matter.

The TO is the working document from which we construct the exam. The TO shows specifically how we carry out the TP.

We use the same TP & TO to develop the three forms of the pretest. TP & TOs for the final form exam may be different from those used in the pretest, however both forms of the final form exam will have the same TP & TO.

The TP & TO must include all the NAVSTDs specified in NAVPERS 18068F for the paygrade for which we are developing the examination. Because these qualifications are the minimum requirements for advancement, they represent the minimum skills and knowledges that we must test in a Mil/Lead exam. Once the exam writer has exhausted all chance for further development of items based on the qualifications for the paygrade we are testing, he or she begins developing items based on standards from lower paygrades. Candidates are equally responsible for the knowledge of "all preceding requirements" (Manual of Navy Enlisted Manpower, 1991).

SECTION DEVELOPMENT AND WEIGHTS

The broad subject areas we must test in a Mil/Lead exam are military requirements, personnel safety, material condition, military conduct and justice, professional development, naval tradition, leadership, international agreements, security requirements, and programs and policies.

Weights (number of questions that will adequately test a section) assigned to examinations are not found through arbitrary assignment. The purpose of a qualifying examination is to test all subject areas specifically, as stated by the NAVSTDs, not to test all subject areas equally. Specific NAVSTDs may limit the depth of coverage (number of items) of an examination section and, therefore, the size of the supporting bibliography; or they may allow a broad depth of coverage and, therefore, an extensive supporting bibliography.

The minimum section and subsection weights used in developing discriminating (rating) exams are not applicable to Mil/Lead exams, which deal with explicit minimum qualifications. Writing items to fill an arbitrary section weight detracts from the intent of the NAVSTDs to examine minimum qualifications; it also forces the writer into the bibliography in depth in search of information from which to develop items. That increases the relative degree of difficulty of the items and also the prospect of the candidate's incorrect response to such items. At this point, the exam, having exceeded the minimum standards, begins to become a discriminating vice a qualifying exam.

EXAMINATION EVOLUTION

Chief petty officer, subject matter experts (SMEs), start the annual Mil/Lead exam evolution by working with an Instructional Systems Specialist to develop a pretest exam TP & TO. During this planning stage, we group the NAVSTDs within their functional areas to form the skeleton of the exam.

Section weights of a Mil/Lead exam must develop naturally, following the context of the standard and the supporting bibliography. Proper planning leads to effective item development through the following steps: (1) We begin with the most explicit NAVSTD and review the bibliography that supports this standard, estimating the potential number of items that we may develop on that subject. (2) Proceeding toward those subject areas broadest in scope, definition, and bibliography, we use the same procedure to determine the number of potential items available in each of the other subject areas. (3) Should the total of potential test items fall short of the desired total for the examination, we begin at the first step again. We use similar subject area qualifications from the next lower paygrade(s) to compile the required number of items. (4) After exhausting all similar subject matter qualifications from lower paygrades, should the total of potential test items still fall short of the number required, then we test other qualifications. Developing items in this manner ensures we test those sections with the least potential for the development of exam items as thoroughly as the minimum standards allow. Similarly, we do not test sections that have the greatest potential for the development of items beyond the minimum prescribed by the NAVSTDs. The bulk of the examination consists of items based on the NAVSTDs specified for the paygrade we are testing and on similar qualifications from lower paygrades. Subject matter areas and qualifications unrelated to those specified for the paygrade we are examining are given the lowest possible priority for inclusion in the examination.

ITEM DEVELOPMENT

The examination writer must exercise extreme care in developing items for a qualifying examination. The writer must keep the following points in mind: No potential item is too simple to include in a Mil/Lead exam, **provided** it conforms to the intent of a qualification as stated in the NAVSTDs. A writer must be acutely aware that the simplicity of an item depends on the degree of exposure the individual has had to the subject matter we are testing. What may appear simple, routine, or matter-of-fact to the SME writer (having considerable exposure to the subject by virtue of time in service) may be difficult and unfamiliar to the candidates of a petty officer third class (PO3) Mil/Lead exam, who are in the learning and developmental stages of their naval careers.

The writer should make sure candidates can clearly understand the intent of each question. While we may expect the candidate to have knowledge associated with specific military terminology required in identifying certain elements within an examination item, all other wording used within the stem of the item should be as simple as possible. We can test a knowledge of terminology—we may not test a knowledge of vocabulary!

PRETESTING

The Mil/Lead pretest, like the Mil/Lead final form, is made up of 100 multiple-choice questions. Candidates have a maximum of 2 hours to complete either exam. The pretest has three forms; its function is to get item performance statistics. These statistics are the basis for developing the two Mil/Lead final forms. Each exam has two sets of items for questions 76 through 100. Since SeaBees) are exempt from several NAVSTDs (e.g., shipboard damage control), they complete their last 25 questions in the SeaBee portion of the exam.

Pretest exams are given at select fleet pretest sites to all available personnel. All exam administration follows strict examination guidelines (verbatim) laid out in an administrative instruction.

Once commands complete pretesting, NETPMSA grades and validates all exam answer sheets.

When NETPMSA completes the grading and validation process, participating commands receive pass/fail lists.

VALIDATION PROCESS

The NETPMSA Data Analysis Branch scores pretests and prepares item and test performance statistics.

Final form Mil/Lead exams have 100% validation control. We use items from the three pretests (100 items each) to develop the two final form exams (100 items each) in each paygrade. All final form items fall within the validation model.

Passing score is 63 items correct out of 100. A score of 63 is 1 standard deviation (SD) below the mean and is significantly above chance (guessing).

After statistical analyzation of each exam item, the Data Analysis Branch sends performance statistics of each item and of the entire exam to the Military Leadership and Basic Training Branch. Item statistics include the average difficulty index (p-value) for the overall item and each alternative, the size of the sample, and the number of examinees omitting (not answering) the item. Additionally, although of no statistical significance in a criterion-referenced test, the Data Analysis Branch furnishes a discriminatory index (r-value). The r-value helps the writing branch to red flag items with possible problems; for example, a correct answer keyed wrong.

FINAL FORM CONSTRUCTION

Once the team leader approves a final form TP & TO, the writer uses pretest exam items falling within the validation model (p-value of .50 to .90) to complete a P-Value Distribution Worksheet.

The worksheet is an item inventory sheet. When the worksheet is complete, the p-values identify a group of p-values whose average is as close as possible to the target p-value. That enables selection of specific exam items that both satisfy the exam's p-value requirement and test the subjects desired.

We test the same minimum qualifications in both forms of each exam. That is why we use the same TP & TO for both exams.

Each final form Mil/Lead exam has parameters with a p-value average of .72 ± .5. Each section (about 6 to 9 per exam) has a p-value average range of .68 to .76.

After assembly, exams pass through several quality control checks before delivery to the Defense Printing Services Office (DPSO).

Mil/Lead exams are sent to all Navy ships and stations and all grading is done locally.

EXAMINATION ADMINISTRATION

All candidates for advancement in rate must pass the Mil/Lead exam before competing in the fleet-wide exam.

Educational service officers control and administer exams to the candidates following strict administrative instructions. The form "A" exam is administered first to the candidate. Should the candidate fail the form "A" exam, the form "B" exam is given later. We continue to alternate forms as later failures occur.

Commands send their completed answer sheets to NETPMSA quarterly, and the Data Analysis Branch provides statistical data on the Mil/Lead final form exam.

REFERENCES

Advancement Examination Development Manual (NETPMSAINST 1418.1A). Pensacola, Fla.: Naval Education and Training Command, 1990.

Bibliography for Advancement Study (NAVEDTRA 12052). Pensacola, Fla.: Naval Education and Training Command, 1992.

Examination Writing Standards Manual (NETPMSAINST 1552.1A). Pensacola, Fla.: Naval Education and Training Command, 1990.

Manual of Navy Enlisted Manpower and Personnel Classifications and Occupational Standards (Vol. 1, Part A) (NAVPERS 18068F). Washington, D.C.: U.S. Navy Bureau of Naval Personnel, 1991.

Cross-Validity Analyses of a Multilevel Model of Job Performance

Rodney A. McCloy, Dickie A. Harris, Jeffrey D. Barnes
Human Resources Research Organization

The chief goal of the Linkage project (Harris et al., 1991), conducted for the Office of the Assistant Secretary of Defense as part of the Joint-Service Job Performance Measurement/Enlistment Standards (JPM) Project, was to model the relationship between recruit characteristics and on-the-job performance so that a cost-performance tradeoff model could be developed that permits the determination of the most cost-effective recruit quality mix predicted to meet a specified performance goal. Data were available for 8,464 individuals from 24 JPM jobs.

A multilevel regression approach (Bock, 1989) was chosen to model job performance (operationalized as a hands-on work sample test) because of the nature of the JPM data (i.e., a nested design—people within jobs) and the possibility that both individual and job characteristics may affect job performance. Multilevel regression allows simultaneous consideration of the effects of individual characteristics (e.g., cognitive ability, experience), job characteristics (e.g., cognitive complexity, difficult working conditions), and their interaction (Hedges, 1988). Specifically, the multilevel performance equation estimated was the following (Harris et al., 1991; McCloy, Harris, & Hedges, 1991; McCloy, Hedges, & Harris, 1991):

$$P_{ij} = \alpha_j + \beta_j A_{ij} + \phi_j T_{ij} + \gamma E_{ij} + \delta_j X_{ij} + \rho T_{ij} X_{ij} + \varepsilon_{ij} \tag{1}$$

where α_j, β_j, ϕ_j, γ, δ_j, and ρ are model parameters, and ε_{ij} is the error term. In words, this equation says that the hands-on performance test score for person i in job j (P_{ij}) depends on an individual's aptitude test scores (A_{ij}, the ASVAB AFQT composite score; and T_{ij}, the ASVAB Technical composite score, TECH), education (E_{ij}), and time in service (X_{ij}) (see Harris et al., 1991, and McCloy et al., 1992, for a description of these variables and their development). An interaction between the Technical composite score and time in service is also included. The subscripted model parameters α_j, β_j, ϕ_j, and δ_j indicate that the intercept and the effects of AFQT, TECH, and time in service can, in principle, vary across jobs.

The variation is addressed by assuming that the parameters themselves have a stochastic structure:

$$\alpha_j = \alpha + \pi_\alpha M_j + \eta_{\alpha j} \tag{2}$$

$$\beta_j = \beta + \pi_\beta M_j + \eta_{\beta j} \tag{3}$$

$$\phi_j = \phi + \pi_\phi M_j + \eta_{\phi j} \tag{4}$$

$$\delta_j = \delta + \pi_\delta M_j + \eta_{\delta j} \tag{5}$$

where (1) α, β, ϕ, and δ are the mean values for the parameters across all jobs (note the lack of the j subscript); (2) M_j is a vector of four job-level component scores (working with things, cognitive complexity, unpleasant working conditions, and fine motor control) that represent characteristics of the military jobs in the JPM Project data; (3) π_α, π_β, π_ϕ, and π_δ are vectors of coefficients describing the degree to which the variance in the job-specific parameters is due to the M_j variables; they are constrained to be the same across jobs (i.e., they are "fixed" coefficients); and (4) η_α, η_β, η_ϕ, and η_δ are random errors that may covary. Because the effect of education was found not to vary across jobs,

$$\gamma_j = \gamma \ . \qquad (6)$$

Cross-Validity Analyses

Including job characteristic information is our attempt to generalize from our small sample of jobs (the 24 JPM jobs having hands-on criterion data) to the population of military jobs. Such generalization is crucial, because only by yielding performance predictions for all Service jobs is it possible to identify the most cost-effective mix of recruits that meets a specified performance goal for each Service. As long as job characteristic information is available for a job, its job-specific parameters can be derived (cf. equations 2 through 5; see also McCloy, Harris, & Hedges, 1991, for an example). These parameters, together with the fixed effects of education and the interaction between the Technical composite score and time in service, constitute job-specific linkage equations.

The capacity to generate job-specific prediction equations whether criterion data are available or not (given job-characteristic data) is a very attractive feature of the performance equation. Nevertheless, the issue of how well the job-specific equations predict performance in out-of-sample jobs remains. The goal of this research was to investigate the validity of the job-specific prediction equations generated by the performance equation. This information is vital because these situations reproduce the scenario in which the model will be implemented by manpower planners. Addressing this question requires jobs that have criterion data but that were not part of the estimation sample for the performance equation. There are essentially two ways such a situation could arise: (1) Manufacture such a situation out of the extant sample by using a holdout procedure, or (2) Obtain relevant data on one or more new jobs after estimating the original performance equation. Both conditions obtained in the present analyses.

Method. Two types of analyses were performed. First, each job was withheld from the sample and a performance equation estimated on the remaining 23 jobs. Each of the 24 "reduced" performance equations was used to generate a job-specific equation for the corresponding holdout job. The observed performance scores for each holdout job were then correlated with the performance scores predicted by the corresponding job-specific prediction equation. Second, the 24-job performance equation was used to generate job-specific equations for two Navy ratings (EM and GSM) and five Marine Corps Military Occupational Specialties (3521, 6112, 6113, 6114, and 6115) for which performance data became available after the 24-job equation was estimated. As in the holdout analyses, the correlation between the observed and predicted performance scores was obtained.[1]

Results. Table 1 contains the results of the cross-validity analyses, including (1) the sample size for each job (N), (2) the squared multiple correlation for the least squares job-specific regression equations (R^2_{OLS}), (3) the squared multiple correlation for the job-specific linkage equation generated from the reduced 23-job or full 24-job performance equation (R^2_{cv}), (4) the difference between the R^2 values from the two types of equations, and (5) R^2_{OLS} values adjusted using various shrinkage formulae (R^2_{adj}). (Note that the R^2 values for the job-specific least-squares and linkage equations given in Table 1 have not been corrected for range restriction or criterion unreliability.)

Two features of the first two columns of R^2 values are of note: (1) The values are quite variable, ranging from .065 to .508 for R^2_{OLS} and .031 to .461 for R^2_{cv}, and (2) $R^2_{OLS} > R^2_{cv}$. This latter finding is expected, given that the least-squares equations are optimal for the samples upon which they were derived; the job-specific linkage equations are not. The largest differences between R^2_{OLS} and R^2_{cv} occur primarily in the jobs having the fewest observations (e.g., EM, GSM, 328X0). The absolute magnitude of the differences is not particularly large, however, ranging from .006 for 11B to .083 for 328X0. The question remaining is what to make of this difference in R^2 values.

Comparison of adjusted and cross-validity R^2 values. Because the job-specific least-squares equations are optimal for the samples on which they were developed but the job-specific linkage equations are not, the comparison of R^2_{OLS} to R^2_{cv} is not exactly fair. A more equitable comparison obtains through adjustment of the R^2_{OLS} values for shrinkage. Perhaps the best known shrinkage formula is one developed by Wherry (1931):

$$R^2_{adj} = 1 - \left(\frac{N-1}{N-k-1}\right)(1-R^2_{yx}) \qquad (7)$$

where N is the size of the sample used to estimate the equation, k is the number of predictors, and R^2_{yx} is the sample coefficient of determination (R^2_{OLS} from Table 11). Wherry's formula gives the value for R^2 expected if the equation were estimated in the population rather than a sample.

Because the population will virtually never be at the researcher's disposal, Wherry's formula is of little practical value. As noted by Darlington (1968) and Rozeboom (1978), the Wherry formula does not answer the more relevant question of what the R^2 would be if the sample equation were applied to the population. Both Cattin (1980) and Campbell (1990) recommended a formula developed by Browne (1975), on the basis of its desirable statistical properties. Browne's formula, appropriate when the predictor variables are random (as opposed to fixed), is

[1] Although space considerations do not permit their publication here, the parameters for the 24-job performance equation and the 24 23-job "reduced" equations are given in McCloy et al. (1992).

Table 1. R^2 Obtained for the Job-Specific Least-Squares and Linkage Equations[a] and Shrinkage Expected Using 4 Formulae

JOB	N	R^2_{OLS}	R^2_{cv}	Difference[b]	Shrinkage Formula (R^2_{adj})			
					Wherry	Browne[c]	Rozeboom	Lord-Nicholson
11B	663	0.086	0.080	0.006	0.080	0.076	0.075	0.073
13B	597	0.065	0.038	0.027	0.059	0.054	0.052	0.051
19E	465	0.141	0.126	0.015	0.133	0.128	0.126	0.124
31C	346	0.140	0.103	0.037	0.130	0.123	0.120	0.117
63B	594	0.076	0.057	0.019	0.069	0.065	0.063	0.062
64C	646	0.108	0.089	0.019	0.103	0.099	0.097	0.096
71L	490	0.127	0.110	0.017	0.120	0.114	0.112	0.111
91A	483	0.117	0.056	0.061	0.110	0.104	0.102	0.100
95B	657	0.102	0.056	0.046	0.097	0.093	0.091	0.090
ET	136	0.081	0.056	0.025	0.053	0.039	0.025	0.018
MM	178	0.154	0.120	0.035	0.135	0.122	0.116	0.111
RM	224	0.154	0.099	0.054	0.138	0.128	0.123	0.119
EM	80	0.348	0.281	0.067	0.313	0.288	0.279	0.270
GSM	88	0.140	0.077	0.063	0.098	0.076	0.058	0.047
112	166	0.141	0.106	0.034	0.119	0.105	0.098	0.093
272	171	0.077	0.031	0.046	0.055	0.043	0.033	0.027
324	124	0.224	0.181	0.042	0.198	0.180	0.172	0.165
328	83	0.223	0.140	0.083	0.183	0.157	0.144	0.134
423	216	0.173	0.149	0.024	0.157	0.146	0.142	0.138
426	188	0.088	0.050	0.038	0.068	0.056	0.049	0.043
492	120	0.216	0.178	0.038	0.189	0.170	0.162	0.155
732	176	0.226	0.198	0.028	0.208	0.195	0.190	0.185
031	940	0.324	0.314	0.010	0.321	0.319	0.318	0.317
033	271	0.358	0.297	0.061	0.348	0.341	0.338	0.336
034	253	0.379	0.366	0.013	0.369	0.362	0.360	0.357
035	277	0.238	0.230	0.008	0.227	0.219	0.216	0.213
3521	907	0.240	0.176	0.064	0.237	0.233	0.234	0.232
6112	152	0.464	0.461	0.003	0.449	0.435	0.438	0.431
6113	93	0.508	0.453	0.055	0.486	0.464	0.469	0.458
6114	190	0.187	0.167	0.020	0.169	0.152	0.157	0.148
6115	113	0.237	0.203	0.034	0.209	0.181	0.189	0.174

[a] All job-specific equations derived from 23-job performance equations except EM, GSM, 3521, 6112, 6113, 6114, and 6115 (derived from the 24-job equation).
[b] Differenco = R^2_{OLS} - R^2_{cv}
[c] Formula for predictors as random effects.

$$R^2_{adj} = \frac{(N-k-3)\rho^2+\rho}{(N-2k-2)\rho+k} \tag{8}$$

where ρ is the adjusted R^2 from the Wherry formula; N and k are defined as above. (Browne also provided a formula for fixed predictor variables.)

A second formula for estimating the validity of the sample equation in the population was provided by Rozeboom (1978):

$$R^2_{adj} = 1 - \left(\frac{N+k}{N-k}\right)(1-R^2_{yx}) \tag{9}$$

with N, k, and R^2_{yx} defined as above.

The shrinkage formulae just described allow one to estimate the population multiple correlation for the full sample equation. If the average sample cross-validity coefficient (i.e., the R^2 expected if the sample equation were applied to another sample) is of interest, Lord (1950) and Nicholson (1960) independently developed a shrinkage formula for estimating this value:

$$R^2_{adj} = 1 - \left(\frac{N+k+1}{N}\right)\left(\frac{N-1}{N-k-1}\right)(1-R^2_{yx}) \tag{10}$$

with N, k, and R^2_{yx} defined as above.

The four shrinkage formulae were applied to the R^2 values from the least-squares job-specific regression equations (i.e., R^2_{OLS}). The values of R^2_{adj} and R^2_{cv} were then compared (see Table 1). In general, the decrease in R^2 associated with the job-specific linkage equation vice the least-squares equation is virtually identical to that expected based on the Browne, Rozeboom, and Lord-Nicholson formulae (i.e., $R^2_{cv} = R^2_{adj}$) —the unweighted and weighted (by sample size) average differences ($R^2_{cv} - R^2_{adj}$) being -.008, -.003, .002; and -.014, -.011, and -.008; respectively. In contrast, R^2_{adj} as given by the Wherry formula is typically larger than R^2_{cv} (unweighted and weighted differences of -.019 and -.021, respectively), but this comparison is not particularly appropriate because no population equation exists. For the present analyses where sample equations are applied to a new sample (the holdout jobs), the Lord-Nicholson formula is perhaps the best standard of comparison. Nevertheless, to the extent that job-specific equations will be generated for the entire population of military jobs, the Browne adjustment arguably provides an additional, viable referent.

The conclusion is the same regardless of the comparison one chooses: The negligible differences between the R^2_{cv} and R^2_{adj} values demonstrate that the linkage methodology provides a means of obtaining predictions of job performance for jobs without criterion data that are as valid as predictions obtained when (1) criterion data are available for the job, (2) a job-specific least-squares prediction equation is developed, and (3) the equation is applied in subsequent samples. Thus, the linkage methodology has yielded a performance equation that generates job-specific equations providing predictions for out-of-sample jobs that are not terribly below the best one could expect. The present analyses strongly suggest that predictions may be made with reasonable confidence for jobs devoid of criterion information, with predictions generally better for high-density jobs than for low-density jobs. We view these results as positive and supportive of the basic linkage methodology.

References

Bock, R. D. (1989). Multilevel analysis of educational data. San Diego, CA: Academic Press, Inc.

Browne, M. W. (1975). Predictive validity of a linear regression equation. British Journal of Mathematical and Statistical Psychology, 28, 79-87.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette and L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 1, pp. 687-732). Palo Alto, CA: Consulting Psychologists Press.

Cattin, P. (1980). Estimation of the predictive power of a regression model. Journal of Applied Psychology, 65, 407-414.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. Psychological Bulletin, 69, 161-182.

Harris, D. A., McCloy, R. A., Dempsey, J. R., Roth, C., Sackett, P. R., Hedges, L. V., Smith, D. A., & Hogan, P. F. (1991). Determining the relationship between recruit characteristics and job performance: A methodology and a model (FR-PRD-90-17). Alexandria, VA: Human Resources Research Organization.

Hedges, L. V. (1988). The meta-aralysis of test validity studies: Some new approaches. ·In Wainer, H. & Brawn, H. I. (Eds), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample. In Research Bulletin (50-40), Princeton, NJ: Educational Testing Service.

McCloy, R. A., Harris, D. A., Barnes, J. D., Hogan, P. F., Smith, D. A., Clifton, D., & Sola, M. (1992). Accession quality, job performance, and cost: A cost-performance tradeoff model (FR-PRD-92-11). Alexandria, VA: Human Resources Research Organization.

McCloy, R. A., Harris, D. A., & Hedges, L. V. (1991). A multilevel model of job performance: The primary linkage equation. Proceedings of the 33rd Annual Conference of the Military Testing Association (pp. 126-131). San Antonio, TX.

McCloy, R. A., Hedges, L. V., & Harris, D. A. (1991). Development of a methodology to link recruit quality requirements to job performance: Estimation of model parameters (IR-PRD-91-07). Alexandria, VA: Human Resources Research Organization.

Nicholson, G. E. (1960). Prediction in future samples. In I. Olkin et al. (Eds.), Contributions to probability and statistics (pp. 424-427). Stanford, CA: Stanford University Press.

Rozeboom, W. W. (1978). The estimation of cross-validated multiple correlation: A clarification. Psychological Bulletin, 85, 1348-1351.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. Annals of Mathematical Statistics, 2, 446-457.

Preliminary Identification of Dimensions of Rated Effectiveness in Desert Storm[1]

Margaret M. Matyuf and Leonard A. White
U.S. Army Research Institute

Walter C. Borman
University of South Florida

A primary role of soldiers is to perform effectively in combat situations. However, the opportunity to evaluate soldier performance in combat situations is rare. Various attempts have been made to identify dispositional determinants of combat effectiveness such as cognitive ability and temperament either through simulation studies or by assuming that indicators of non-combat performance are also predictive of performance in combat.

Cognitive ability has been linked to combat success. In their simulation study of infantry soldiers, Whitmarsh and Sulzen (1989) found the General Technical (GT) composite of the Armed Services Vocational Aptitude Battery (ASVAB) to be positively related to performance. Mahan and Clum (1971) found education level and scores on the General Classification Test Battery to correlate positively with performance in both combat and non-combat situations. However, in a more recent study of combat effectiveness during Operation Just Cause, the ASVAB did not differentiate soldier performance in combat situations (Dover, unpublished manuscript). Dover suggests that abilities needed to succeed in Army training may not be the same abilities needed to succeed in combat situations.

Soldier's temperament has also been implicated in combat effectiveness (Egbert et al.,1958). Egbert et al. concluded that the better soldier during the Korean War was more masculine, socially mature and emotionally stable.

The research presented here examines predictors of Desert Shield/Storm combat performance for soldiers in the Army's longitudinal validation study, Project A. Measures of soldiers' temperament and cognitive aptitude were obtained in FY86/87 when the soldiers entered the Army as new recruits. Supervisor and peer combat performance ratings were obtained for soldiers in the LV sample who participated in Desert Shield/Storm. Hence, this research attempts to identify the individual characteristics of the better combat soldier.

Method

Sample and Procedure

The predictor measures were administered to a longitudinal validation (LV) sample of 49,108 soldiers in 21 MOS who entered the service in FY86/87. Testing occurred during reception station processing. More detail on the predictor measures and the data collection can be found in Campbell (1989).

Supervisor and peer ratings were made once soldiers had returned from Desert Shield/Storm. Ratings were made on 262 soldiers. Of these, 142 soldiers were matched with the LV predictor data. Approximately 90% of the soldiers were Specialist 4s, Corporals, and Sergeants. Fifty percent of the

[1] Presented at the meeting of the Military Testing Association, October, 1992. All statements expressed in this paper are those of the authors and do not necessarily reflect the official opinions or policies of the U.S. Army Research Institute or the Department of the Army.

soldiers were from combat MOS. Males accounted for 98% of the sample and 50% of the sample was white.

Predictor Measures

ASVAB. Cognitive ability was measured by the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is used for the selection and classification of applicants for the Services. It consists of 10 subtests that can be represented by four factors: Quantitative, Speed, Technical and Verbal aptitude. Parallel forms reliabilities range from .78 - .92 (Kass, Mitchell, Grafton, & Wing, 1983).

A composite of the ASVAB called the Armed Forces Qualifications Test (AFQT) is used for selection. It is reported as a percentile score with a mean of 50. The current AFQT composite (AFQT89) was implemented in January 1989. It is a composite of four subtests; Word knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge.

Temperament. Measures of temperament were obtained from the Assessment of Background and Life Experiences (ABLE). This instrument consists of 11 temperament scales which are organized into seven factors: Internal Control, Cooperativeness, Dominance, Dependability, Physical Conditioning, Stress Tolerance and Work Orientation. The ABLE has been used to predict a wide variety of criteria, including leadership potential, job performance, and first term attrition (White, Nord, Mael, & Young, in press). Five year test-retest reliabilities for the factors range from .33 to .59.

Criterion Measure

Army-Wide Performance Ratings Peer and supervisor ratings were collected on a set of twelve 7-point scale dimensions called the Army-Wide Behaviorally Anchored Rating Scales (BARS). The scales were originally developed to assess second tour soldier effectiveness in any Army job. Each dimension was given an overall definition and contained three scaled behavioral summary statements reflecting low, middle-, and high-level performance in the category. (See Campbell (1992) for more information on the development of the rating scales). These scales were slightly modified to reflect the somewhat different job performance requirements in Desert Shield/Storm.

Results

Rated Familiarity

The majority of raters reported that they had worked with the ratees for at least 7 months. Raters also reported that they had sufficient opportunity to observe ratee job performance; 85% of the peers and 94% of the supervisors indicated they observed performance several times a week.

Factor analysis of ratings

Principal factor analysis with varimax rotation was used to explore the dimensionality of the Army-wide BARS for the combined peer and supervisory ratings. Prior to the factor analysis, a mean peer/supervisor rating was computed for each ratee giving equal weights to each ratee/rater pair. A three factor solution based on 11 items was chosen as the most meaningful (Maintaining Assigned Equipment was removed due to low reliability, $r=.142$). Six items formed the first factor which was labeled Supervision/Technical Skill and is summarized as showing technical knowledge, exerting effort and demonstrating leadership abilities and support to other soldiers. The second factor was labeled as Personal Discipline and represents following regulations and orders, integrity and self-control. The third factor, Physical Fitness/Bearing is defined by maintaining an appropriate military appearance and staying in good physical condition. The three factor solution for combat performance is similar to the non-combat performance solution (Campbell, 1992). The factor loadings obtained for the combat ratings are displayed in Table 1.

Composite scores were computed by summing the items that had the highest loadings on each of the three factors. Interrater reliabilities for the three factors were as follows: .453 for Supervision/Technical Skill, .345 for Personal Discipline, and .487 for Physical Fitness/Bearing. Correlations among the factors ranged from .527 to .748.

Table 1

Principal Factor Analysis of Ratings on Army-Wide BARS

Behavior Scales	Supervision/ Tech. Skill	Personal Discipline	Physical Fitness/ Bearing
Technical Knowledge/Skill	.67	.25	.22
Effort	.67	.39	.29
Supervising	.64	.37	.34
Following Regs/Orders	.46	.64	.24
Integrity	.47	.53	.27
Training/Development	.67	.20	.26
Physical Fitness	.28	.19	.51
Self-Development	.63	.42	.26
Consideration for Subordinates	.60	.41	.31
Military Appearance/Bearing	.28	.46	.52
Self-Control	.21	.59	.20
% Variance Explained	28.64	18.55	10.73

Note. Principal Factor Analysis with varimax rotation (n=204)

Mean performance ratings for each factor are given in Table 2. Mean ratings were significantly higher for Physical Fitness for the minority group as compared to whites, F(1,215)=17.60, $p < .05$. This finding has also appeared in peacetime ratings (Pulakos, White, Oppler, & Borman, 1989).

Table 2

Mean Performance Ratings

	Supervision/ Tech. Skill		Personal Discipline		Physical Fitness/ Bearing	
Ratee Race	M	SD	M	SD	M	SD
White (n=111)	4.388	.993	4.469	1.138	4.405*	1.177
Other (n=105)	4.410	.894	4.620	.990	5.040*	1.056

*$p < .05$

Correlations

Correlations between dimensions of rated effectiveness and the ASVAB and ABLE are shown in Table 3. The zero-order correlations were corrected for range restriction due to multivariate selection on ASVAB (Lawley, 1943).

The Quantitative, Technical and Verbal ASVAB factors and AFQT are predictive of higher ratings for the Supervision/Technical Skill performance factor. The ASVAB factors and AFQT were not significantly correlated with performance on the other factors.

As expected, ABLE Physical Conditioning was positively correlated with ratings of Physical Fitness/Bearing. Unexpectedly, the correlation between Physical Conditioning and Personal Discipline was also significant.

Table 3

Correlations of ASVAB and ABLE with Combat Performance Rating Dimensions

Predictors	Supervision/ Tech. Skill	Personal Discipline	Physical Fitness/ Bearing
ASVAB (n=128)			
Quantitative	.253* / .40[a]	.165 / .21	-.003 / .03
Speed	-.031 / .26	-.002 / .09	.089 / .05
Technical	.216* / .30	.091 / .12	-.153 /-.15
Verbal	.224* / .39	.090 / .12	-.107 /-.07
AFQT89	.273* / .42	.143 / .18	-.033 / .01
ABLE (n=120)			
Internal Control	-.089 /-.04	.020 / .04	.042 / .05
Cooperativeness	-.102 /-.18	.026 / .00	-.089 /-.08
Dominance	-.061 / .00	-.123 /-.09	-.038 /-.03
Dependability	-.033 /-.01	.047 / .05	.056 / .04
Physical Conditioning	.068 / .01	-.176* /-.17	.231* / .21
Stress Tolerance	-.111 /-.03	-.113 /-.08	-.074 /-.06
Work Orientation	-.048 / .02	-.129 /-.09	-.063 /-.07

[a] Correlations were adjusted for range restriction. Format is uncorrected/corrected.

* $p < .05$

Discussion

Principal components factor analysis revealed three dimensions of combat performance; Supervision/Technical Skill, Personal Discipline and Physical Fitness/Bearing. Mean ratings on these factors were correlated with cognitive ability and temperament as measured by ASVAB and the ABLE, respectively. The ASVAB results indicate that cognitive ability is an important predictor of combat performance.

Although the present sample is too small (n=27) to allow within-subject comparisons of combat performance and non-combat performance, comparisons can be made across investigations. The factor analysis results are similar to previous findings performed on non-combat ratings in which factor analysis performed on 19 Army-wide BARS yielded a four factor solution labeled Supervision, Technical Skill, Personal Discipline and Physical Fitness/Military Bearing (Campbell, 1992).

Research examining the relationships of temperament and cognitive ability to ratings of non-combat performance has shown the ABLE factors to be more strongly related to performance. Borman, White, Pulakos and Oppler (1991) showed Achievement orientation and Dependability to be predictive of performance in non-combat situations while smaller relationships were shown to exist between cognitive ability and performance ratings. In the combat environment, we found this pattern was reversed with ASVAB strongly predictive of Supervision/Technical Skill performance and ABLE showing near-zero correlations with ratings.

The research presented here is preliminary. More combat performance data is being collected on cases in the Project A LV sample using the Army-Wide BARS. In addition, performance data has also been collected using another set of rating measures called the Combat Performance Scales (CPS). The CPS consists of 27 items developed specifically to examine combat performance. Analyses using these scales will allow more insight into the characteristics of the effective combat soldier.

References

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. Journal of Applied Psychology, 76, 863-872.

Campbell, J. P. (1989). Improving the selection, classification, and utilization of army enlisted personnel: Annual report, 1987 fiscal year (Report No. 862). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. P., & Zook, L. M. (1992). Building and retaining the career Force: New procedures for accessing and assigning Army enlisted personnel: Annual report, 1990 fiscal year (Report No. 952). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Dover, S. H. The characterization and the prediction of US Army combat soldier performance in view of their performance in operation "Just Cause". Unpublished manuscript.

Egbert, R. L., Meeland, T., Cline, V. B., Forgy, E. W., Sprinkler, M. W., & Brown, C. (1958). Fighter I: A study of effective and ineffective combat performers. (Special report 13). Presidio of Monterey, CA: U.S. Army Leadership Human Research Unit (Human Resources Research Office).

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factor structure of the Armed Services Vocational Aptitude Battery (ASVAB) Forms 8, 9, and 10; 1981 Army applicant sample. Educational and Psychological Measurement, 43, 1077-1088.

Lawley, D. A. (1943). A note on Karl Pearson's selection formulas. Royal Society of Edinburgh Proceedings, Section A, 62, 28-30.

Mahan, J. L., Jr., & Clum, G. A. (1971). Longitudinal prediction of marine combat effectiveness. The Journal of Social Psychology, 83, 45-54.

Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. (1989). Examination of race and sex effects on performance ratings. Journal of Applied Psychology, 74, 770-780.

White, L. A., Nord, R. D., Mael, F. A., & Young, M. C. (in press). The Assessment of Background and Life Experiences (ABLE). In T. Trent and J. H. Laurence (Eds.), Adaptability screening for the armed forces. Washington, DC: Office of Assistant Secretary of Defense (Force Management and Personnel).

Whitmarsh, P. J., & Sulzen, R. H. (1989). Prediction of simulated infantry-combat performance from a general measure of individual aptitude. Military Psychology, 1, 111-116.

A Dose Equivalency Strategy to Index Human Performance Degradation

Robert S. Kennedy, Essex Corporation
Janet J. Turnage, University of Central Florida
William P. Dunlap, Tulane University

ABSTRACT

Toxic and controlled substances, stress, sleep loss, and environmental effects can render an individual unfit for duty. Bioassays of body fluids or hair clippings can establish the presence of some substances, but beg the question of whether work performance will be degraded and are not useful for identification of agents which have no chemical trace. A performance-based method would permit more direct assessment of fitness for duty. While the most obvious way to measure job performance is on the job, such measures are not usually sufficiently reliable, not stable, and uneconomical to collect. Standardized performance test batteries can be employed to screen workers. But it is not enough to show sensitivity of a test battery; what is required is a quantitative standard or yardstick which will permit changes that occur on performance tests to be translated into the more meaningful contexts such as operational performance. Alcohol was examined for feasibility for this standard. A series of metrically-sound microcomputer-based tests were administered before, during, and after graded dosages of alcohol were applied. The ascending and descending limbs of the alcohol concentration curves were then followed, and multiple regression analyses were calculated. Multi-variate predictor equations of performance deficit were generated. To demonstrate generalizability of the dose-equivalency method, an algorithm, empirically derived from the descending limb in one sample, was cross-validated to predict both ascending and descending limbs in another sample ($p < .01$). The feasibility of using such a methodology to aid interpretation of military stressors and their effects on operational performance is discussed.

INTRODUCTION

Elsewhere, we hypothesized that it would be possible to create tests which: (1) are sensitive to the same agents as those which affect fitness-for-duty, and (2) tap the same mental faculties that are demanded in job performance. We have named such an approach surrogate testing (Turnage, Kennedy, Gilson, Bliss, & Nolan, 1988) and offered a set of criteria for surrogate test development. In addition to sensitivity and relevance, surrogate tests should also: (3) take less time than real-world tasks to reach stable levels of performance; and (4) be administered in a short period of time. To this end, we have developed a menu of microcomputer-based tests which achieve stability quickly (< 15 minutes of practice) (Kennedy, Baltzley, Wilkes, & Kuntz, 1989), are reliable (retest reliability for each test > .707 for 3-minute tests) (Kennedy, Wilkes, Dunlap, & Kuntz, 1987), and are factorially rich (3+ factors in < 10 minutes) (Kennedy, Baltzley, Turnage, & Jones, 1989). Scores on combinations of subtests show reasonably high multiple regressions with global measures of intelligence (Kennedy, Wilkes, Dunlap, & Kuntz, 1987), and with tasks related to job performance (Kennedy, Baltzley, Turnage, & Jones, 1989). Several of the tests have been employed in validation studies and have proven to be sensitive to drugs (Parth, Dunlap, Kennedy, Lane, & Ordy, 1989) and treatments, such as hypoxia (Kennedy, Dunlap, Bandaret, Smith, & Houston, 1989). Recently, these validation studies were reviewed and performance deficits were compared to changes induced by alcohol intoxication (Kennedy, Dunlap, & Turnage, in preparation). From these relations, we propose that alcohol concentration should be investigated to determine psychometrically whether it would be suitable as a "gold standard" to index performance deficits. Then, regardless of whatever agents were reducing performance on the job, the change could be reflected in equivalent personal dose equivalency alcohol concentrations (DEAC's). The availability of such a system could provide additional guidance to an employeee or his or her supervisor and could provide feedback via self-testing prior to work.

The following logic forms the basis for the formal argument for a dose equivalency model or strategy. First, a set of target performance tasks and an indexing agent are selected. Then, graded "dosages" of the indexing agent are administered and performance decrements are marked against the various dosages. One is then left with a functional relationship between an agent and performance(s). This relationship can become the dose response "yardstick" against which other agents and treatments of general fitness are marked. Such an approach, using alcohol, has been anticipated by at least three other research groups (Klein, 1972; O'Hanlon, Brookhuis, Louwerens, & Volkerts, 1986; Billings, Demosthenes, White & O'Hara, 1990). The research questions are: "How strongly are performance and alcohol dosage related? Can the relationship be mathematically expressed so that it can be applied automatically? Wnat is the rate of error?"

After selection of alcohol as the indexing agent, our next step was to set up conditions to answer research questions. For example, we wanted to determine when an individual would be unfit to perform normally. We had already obtained a regression equation from a preliminary alcohol consumption study where performance was also measured with nine computerized cognitive tests (Kennedy, Wilkes, Dunlap, Fowlkes, & Smith, 1990). Although statistically significant findings were found, that work suffered because: (1) alcohol dosages were abruptly ingested; (2) statistical significances of dosage effects were marginal at alcohol concentration (AC) .05 and below; and, (3) only the descending limb of alcohol concentration was followed. However, in that pilot study, all test performances were correlated with alcohol concentration. Furthermore, a regression analysis yielded an equation which made use of only four tests and produced a multiple correlation of R = .77. From this outcome, we theorized that it should be possible to develop algorithms which would permit the selection of a specific alcohol concentration limit, and "Back-Solve" the equation to establish prescribed combined limits on these four tests. If these limits were exceeded, it would permit identification of that individual as prospectively unfit.

In order to pursue further this strategy for research, it was necessary to determine whether the findings would generalize beyond the sample and specific conditions of the pilot experiment. The primary purpose of the present research was to expand on the original work with new research that also followed both the ascending and descending limb of the alcohol content curve and then compared multiple regression equations for both studies. It was also an opportunity to study procedural variables where little baseline testing was established, and to examine the effect of more realistic (less abrupt) alcohol ingestion which more nearly represented social drinking conditions. Based on the outcome from this experiment, decisions about subsequent experimental work could be made to continue with alcohol and to investigate other agents.

ALCOHOL EXPERIMENT

Subject Solicitation and Selection. Adult male University of Wyoming students were solicited for research participation with informational posters and advertisements in the University newspaper. Females were not employed because of human use restrictions. The candidate subjects completed several questionnaires and were informed that notification of selection would be completed within the following 2-week period. Information from the Personal Information Questionnaire (PIQ), Iowa Scale of Preoccupation with Alcohol (IS), and Cahalan Volume-Variability Scale (Cahalan, Cisin, & Crossley, 1969) was then reviewed and assessed concerning subject selection criteria. Potential subjects were dropped from further consideration based on the criteria from these preselection questionnaires. The final sample consisted of 30 subjects.

Automated Performance Test System (APTS) Tests. The mental acuity tests selected for inclusion in this study were from a portable microcomputer-based menu of tests developed by us (Kennedy, Baltzley, Wilkes, & Kuntz, 1989). The battery selected for study consists of subtests requiring approximately 12 minutes of real-time testing. The individual subtests used in this research are described more completely elsewhere (Kennedy, Baltzley, Wilkes, & Kuntz, 1989). The tests for this study were implemented on a portable, battery-operated laptop computer (NEC PC8201A), although versions are available for IBM-compatible PC's and some normative data for stabilized performance are available.

Data Collection. Subjects were requested to not ingest alcohol, other drugs or solid food between the training session and the data collection session. Upon arriving at the data collection site, the subjects filled out a Current Health State Questionnaire (CHSQ) and then performed two additional APTS battery baseline sessions (ca. 20 minutes). During the computer testing the CHSQ's were reviewed for subject suitability for research participation. In particular, health status, medications, suitability statement, and weight were noted. After the APTS tests the subjects were given a breath test. "Warm-up" testing insured that all subjects were at an alcohol concentration (AC) = 0.000% and they were well practiced and performing at asymptotic levels as well as providing the prealcohol baseline data necessary for post-alcohol comparison.

Upon completing the "warm-up" session, subjects then began the alcohol consumption portion of this experiment which consisted of eight cycles of testing (approximately 80 minutes per cycle) and involved the following procedures: During cycle one, subjects consumed one third of the grain alcohol mixed with citrus punch and ice. The quantity of grain alcohol usage was calculated for each subject based on his reported weight. After consumption, subjects waited for a minimum of 12 minutes before testing began. Following this waiting period, a breath test was administered to measure blood alcohol content. Next, subjects took the APTS battery on NEC laptop computers. These two tests, (breath and APTS), were administered in that order throughout the entire experiment. During cycle two, subjects again consumed one third of the grain alcohol mixed drink with a minimum 12-minute waiting period followed by the breath test and APTS battery in that order and another breath test after completion of the APTS battery. This same sequence was repeated for cycle three. During cycle four, subjects ate lunch and were administered the series of tests previously described. During cycles five and six, breath tests and the APTS were administered. During cycle seven, subjects ate dinner after which the APTS was administered. Finally, in cycle eight, subjects were administered their last APTS battery and breath test, and, if they achieved acceptable scores on the breath test given after dinner, or on a breath test given after the battery, they were transported home.

DOSE EQUIVALENCY FINDINGS

Regression equations were calculated using APTS scores as independent variables and AC as the criterion. The regression analyses obtained in previous work (Kennedy, Wilkes, Dunlap, Fowlkes, & Smith, 1990) were then applied to these data. In that work, the same nine tests from the core battery were administered, but only the descending limb of AC curve was monitored. Because four tests (Code Substitution, CS; Math Processing, MP; Two-Hand Tapping, THT; and Grammatical Reasoning, GR) were as effective in a multiple regression coefficient (R = .75) as all nine (R = .77), these four were again used to create an algorithm for predicting alcohol concentration.

Validation of Algorithm. The manner in which the algorithm is constructed is as follows:

First, percent decrements for each of the four core subtests of the battery were computed relative to baseline.

$$\text{Percent Decrement} = 100 \times (\text{Baseline} - \text{Score})/\text{Baseline}. \qquad (1)$$

Next, a composite score (S) from CS, MP, TFT, and GR was computed at

$$S = (9CS + 6THT + 5MP + 2GR)/1000 \qquad (2)$$

If S was less than zero, S is set equal to zero. Estimated Alcohol Concentration (AC) is then computed as

$$AC = 0.2\ S^{1/2}. \qquad (3)$$

Fitting the Algorithms to Previous Experimental Findings. Using the warmup trials as the baseline, the algorithm was seen to fit the ascending limb of the alcohol intoxication curve quite well. However, preliminary fits to the subsequent descending limb of the curve were found to badly underestimate actual AC. This underestimation may have resulted from one or both of two possible processes. First, it is known that performance during the descending curve of alcohol intoxication curve may be influenced by a short-term tolerance phenomenon, such that less intoxication is seen during the falling curve than on the rising part of the curve, even at the same actual AC (Taylor, 1986; Wilson, Erwin, & McCleran, 1984). The other more likely problem with the current data is that practice on the eight trials prior to alcohol intoxication was insufficient to reach asymptotic levels of performance. If this were the case, additional practice under alcohol treatment would continue to cause the performance curve to rise, owing to learning, such that baseline performance at the end of the session would logically be higher than baseline performance at the beginning of alcohol testing.

To test the latter hypothesis, the following strategy was adopted. A first approximation to a learning curve (Lane, 1987) was fit to each subject's first eight (nonalcohol) trials of performance on the target tasks:

$$\text{Performance} = B \times \text{Log}_e\,(\text{Trial Number}) + A. \qquad (4)$$

This learning curve was then used to project baseline scores for trials 9 through 15, during which subjects performed under various degrees of alcohol intoxication. It should be remembered that in actual implementation of an automated performance fitness test station, workers would be tested daily; therefore, the question of the stability of asymptotic performance would be moot, since after 20 or more sessions, the learning curve can be expected to be moving more slowly.

The learning curves calculated for the group data for four tests are depicted in Figures 1a-d, where it can be seen that the learning function appears to fit well. Then, using the predicted scores from individually fitted learning curves as estimated baselines, the predicted ACs from Equations 1 to 3 were computed again, and were found to fit both ascending and descending limbs of the actual AC curves much more adequately. Predicted AC scores were correlated to actual AC scores, and the correlation was 0.722, which is nearly as high as the Multiple R of 0.75 in the fit to the previous data set and strongly argues for a predictively useful dose equivalency model. It should be remembered that this model was derived from an entirely different data set, therefore, the above analysis establishes the validity of that prediction algorithm.

Analysis of Predicted AC. Repeated-measures analysis of variance was applied to the predicted AC scores in order to compare baseline to alcohol intoxication trials, in a subjects by treatments ANOVA. A highly significant effect of trial was found [F(7, 154) = 69.36, p< 0.001]. This analysis was followed by Dunnett's test to compare each alcohol intoxication trial to the baseline predicted score; predicted AC on all alcohol trials differed significantly from the warmup baseline, except for the final trial (p < .06) on which actual AC was only 0.038. This indicates that the AC prediction algorithm is sensitive at 0.05 AC and somewhat below; more detailed investigation at lower AC levels is clearly warranted.

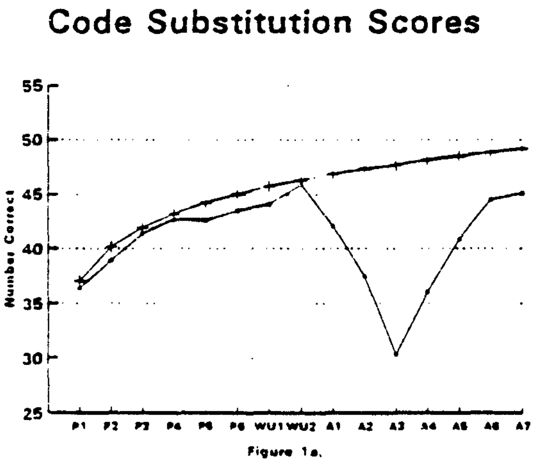

Code Substitution Scores

Figure 1a.

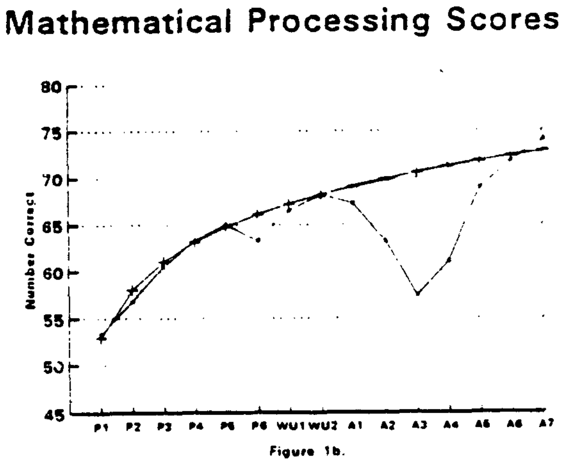

Mathematical Processing Scores

Figure 1b.

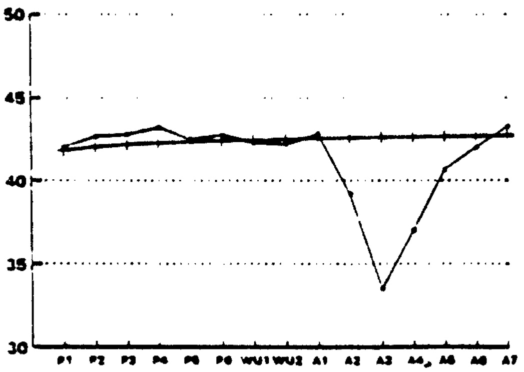

Two Handed Tapping Scores

Figure 1c.

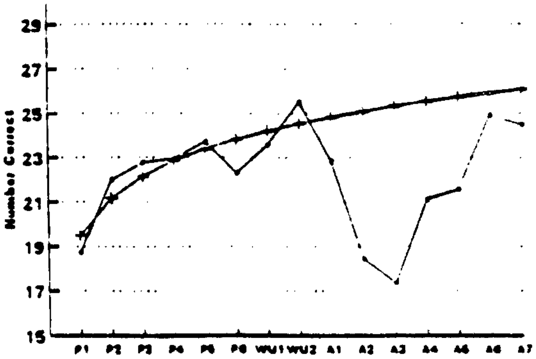

Grammatical Reasoning Scores

Figure 1d.

In summary, the research described above provided considerable encouragement for continued exploration of the dose equivalency methodology and no insurmountable technical difficulties were surfaced. It was seen that incremental changes in baseline performance due to learning could be fitted and covaried by following well-established equations (e.g., Lane, 1987). Using this information, a new multiple regression equation was derived for the present experiment and compared (cross-validated) with one obtained previously. The additional variance-accounted-for was very small and, because of the generalizability of the cross-validation outcome, the original dose equivalency algorithm was retained.

DISCUSSION AND CONCLUSIONS

A series of metrically sound microcomputer tests were administered before, during, and after graded dosages of alcohol were applied. The ascending and descending limbs of the AC curves were followed, and multivariate predictor equations of performance deficit were generated. To demonstrate generalizability of the dose-equivalency method, an algorithm, empirically derived from the descending limb in one sample, was cross-validated to predict both ascending and descending limbs in another sample ($p < .01$).

Preliminary work has been performed to test the logic of the dose equivalency model. In one study, the same cognitive and motor tasks as were used here were studied in tracking cognitive performance decrements occasioned by systematically decreasing atmospheric pressure applied to seven subjects who lived for 40 days in a hypobaric chamber to simulate a climb to the height of Mount Everest (Kennedy, Dunlap, Bandaret, Smith, & Houston, 1989). In a second study, the feasibility of calibrating performance decrements in terms of alcohol dose was demonstrated in a study conducted at a treatment center for leukemia patients, where toxic chemoradiotherapy was used to destroy diseased bone marrow prior to implanting donor bone marrow (Parth et al., 1989). (Note that it is not suggested that there is a direct pharmacological linkage between chemoradiotherapy and CNS depression, nor that performance changes are due to the chemicals, the radiation, or some other factor governed by emotion. These results are merely expositional.) Impacts of the chemoradiotherapy were monitored via performance on the cognitive tests. From these data sets, one might equate the effect of the simulated Mount Everest climb to the summit (25,000 ft) from a cognitive standpoint to be roughly equivalent to the early stages of chemoradiotherapy, or about a .11 BAL, which is equivalent to driving under the influence (DUI) in all states in the country.

These two examples appeared to support the feasibility of continued development of the dose equivalency concept, using larger and more heterogenous samples of subjects in order to broaden the applicability and refine the predictor equation. It should also be remembered that each particular toxic agent is likely to produce its own particular spectrum of performance deficits, and, in factorially rich batteries, one cannot expect the constellation of deficits to exactly parallel that of alcohol. Future work should also concentrate on the uses of dose equivalency to situations where the treatment condition would be expected to produce similar pharmacological responses to the responses produced by alcohol. At this point, the findings in this study cannot be extended to exposures other than situations where central nervous system depression is the primary toxic effect. Scaling to the alcohol metric, however, does permit comparisons across studies and across toxic conditions in terms of a common metric that has both intuitive and scientific appeal, as well as fairly well established safety limits.

REFERENCES

Billings, C. E., Demosthenes, T. A., White, T. R., & O'Hara, D. B. (1990). Effects of ethyl alcohol on pilot performance in simulated flight. Paper presented at the 61st Annual Meeting of the Aerospace Medical Assn.

Cahalan, D., Cisin, I. H., & Crossley, H. M. (1969). American drinking practices: A national study of drinking behavior and attitudes (Monographs of the Rutgers Center of Alcolol Studies No. 6). New Haven, CT: College University Press.

Kennedy, R. S., Baltzley, D. R., Turnage, J. J., & Jones, M. B. (1989). Factor analysis and predictive validity of microcomputer-based tests. Perceptual and Motor Skills, 69, 1059-1074.

Kennedy, R. S., Baltzley, D. R., Wilkes, R. L., & Kuntz, L. A. (1989). Psychology of computer use: IX. A menu of self-administered microcomputer-based neurotoxicology tests. Perceptual and Motor Skills, 68, 1255-1272.

Kennedy, R. S., Dunlap, W. P., Bandaret, L. E., Smith, M. G., & Houston, C. E. (1989). Cognitive performance deficits occasioned by a simulated climb of Mount Everest: Operation Everest II. Aviation, Space, and Environmental Medicine, 60, 99-104.

Kennedy, R. S., Dunlap, W. P., & Turnage, J. J. (In preparation). Indexing performance decrements to alcohol concentration: Dose equivalency. Orlando, FL: Essex Corporation.

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., Fowlkes, J. E., & Smith, M. G. (1990). Characterizing soldier responses to irritant gases (Final Rep., Contract DAMD17-89-C-9135). Fort Detrick, Frederick, MD: U.S. Army Medical Research Acquisition. (NTIS No. AD A220 429)

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987). Development of an automated performance test system for environmental and behavioral toxicology studies. Percept. and Motor Skills, 65, 947-962.

Klein, K. E. (1972). Prediction of flight safety hazards from drug induced performance decrements with alcohol as reference substance. Aerospace Medicine, 43(11), 1207-1214.

Lane, N. E. (1987). Skill acquisition rates and patterns: Issues and training implications. New York: Springer-Verlag.

O'Hanlon, J. F., Brookhuis, K. A., Louwerens, J. W., & Volkerts, E. R. (1986). Performance testing as part of drug registration (pp. 311-330). In J. F. O'Hanlon & J. J. deGier (Eds.), Drugs and driving. London: Taylor & Francis.

Parth, P., Dunlap, W. P., Kennedy, R. S., Lane, N. E., & Ordy, J. M. (1989). Motor and cognitive testing of bone marrow transplant patients after chemoradiotherapy. Perceptual and Motor Skills, 68, 1227-1241.

Taylor, L. (1986). Drunk driving defense. Boston: Little, Brown.

Turnage, J. J., Kennedy, R. S., Gilson, R. D., Bliss, J. P., & Nolan, M. D. (1988, December). The use of surrogate measurement for the prediction of flight training performance. Orlando, FL: Institute for Simulation and Training, University of Central Florida.

Wilson, J. P., Erwin, V. G., & McCleran, G. E. (1984). Effects of ethanol: I. Acute metabolic tolerance and ethnic differences. Alcoholism: Clinical and Experimental Research, 3(2), 228-232.

THE ROLE OF THE INSTRUCTOR IN PAIRED COMPUTER-BASED TRAINING

Stanley D. Stephenson
Southwest Texas State University

Although the effects on achievement of non-content training system factors in computer-based training (CBT, used here as a generic term for all computer-based, computer-assisted training) has not been extensively studied, McCombs et al. (1984) did find that two non-CBT content factors were critical to the success of CBT courses. They were: (a) adequate opportunities for student-instructor interactions, and (b) the incorporation of group activities with individualized training.

The student-instructor interaction result is a significant finding since one of the most consistently reported positive traditional instruction (TI) instructor behaviors is frequent but short student-instructor interactions; i.e., an increase in student-instructor interactions produces an increase in achievement (Brophy & Good, 1986). Stephenson (1991) manipulated this variable and found that in individual CBT settings (one student-one computer) student-instructor interaction had a positive effect on achievement in CBT even when the interactions were not related to the CBT content.

The second McCombs dimension, group activities, is a dimension frequently not found in CBT, perhaps due to the fact that CBT is typically conducted in a one student-one terminal environment. Group activities in CBT can occur in more than one way. One way is to have students work CBT in small groups, and the results of work done in this area are fairly consistent. "No study has reported significantly greater learning when students work alone (Webb, 1987, p. 195)."

Consequently, there is some evidence that both student-instructor interaction and working CBT in dyads/ triads can increase achievement in CBT. However, there is no available research on the interaction between these two variables. Therefore, a 2 x 2 factorial design field experiment was conducted. Based on the literature, it was hypothesized that both student-instructor interaction and paired learning would have positive effects on achievement.

METHOD

Subjects

Eighty four business statistics students completed a field study exercise on how to use a computer spreadsheet package to perform statistical calculations. Ss' prior experiences on a personal computer (PC) and spreadsheet were assessed.

Experimental Materials

The CBT software was the spreadsheet tutorial portion of a larger commercial software tutorial package designed for an integrated spreadsheet-word processing-database program. The tutorial is basically linear and learner-controlled.

The larger tutorial was modified to include just the introduction to the integrated package plus that portion of the tutorial software devoted to the use of the spreadsheet. The introduction portion (Part A) contained four lessons, and the spreadsheet portion (Part B) contained eight lessons. The tutorials were run on Tandy 1000SX PCs.

A statistical exercise (calculate means and standard deviations) designed to evaluate mastery of the spreadsheet tutorial commands was added to the experimental software. Therefore, the total experimental material consisted of a CBT spreadsheet tutorial modified to include a statistics-based exercise. The exercise was also worked on the computer.

Experimental Design

Ss were arranged in a 2 x 2 x 2 factorial design. Main effects were learning setting (dyad/individual), student-instructor interaction (present/absent), and spreadsheet/PC experience (high/low).

Ss were randomly assigned by spreadsheet/PC experience to one of four groups. Group I (n=20) worked CBT in dyads and received instructor-initiated interactions. Group II (n=25) also worked CBT in dyads but did not experience student-instructor interactions. Group III (n=20) worked CBT individually with interactions, while Group IV (n=19) worked individually without interactions. Ss were assigned to teams based on grade point average, college major, and gender.

Procedure

All groups worked the CBT tutorial in three sessions. In session one, all groups started on lesson A1 and worked in the tutorial for 70 minutes. In the second session, all groups started on lesson B1 and worked for 70 minutes. In the third session, all groups started on lesson B3 and worked for 35 minutes. Therefore, all Ss had a single exposure to lessons A1 though A4 and repeated exposure to lessons B1 through possibly B8, the spreadsheet portion of the tutorial. Since teams and individuals went at their own speeds, total individual subject time on task varied. After 35 minutes on day 3, Ss worked individually on the statistics exercise for 30 minutes.

At the beginning of day 1, the instructor interacted with all teams/individuals to insure that all Ss were properly logged into the tutorial. The instructor also responded to all subsequent student-initiated interactions with one or more of three responses: (1) "Try pushing the [ESCAPE] key;" (2) "Try pushing the [SPACE] bar;" or (3) "Re-boot the system and start over." These suggestions were given in sequence; e.g., if "Try pushing the [ESCAPE] key," did not correct the problem, then the S was told to "Try pushing the [SPACE] bar." For the groups not receiving interactions, Ss, these suggestions were the only instructor interactions experienced after the startup on day 1.

In addition to the startup instructions interactions listed above, two groups also received subsequent instructor-initiated interactions which lasted 5 - 10 seconds. In the first session, the instructor initiated four interactions with each team/ subject. In sessions two and three, the instructor initiated three and one interactions, respectively. These interactions were related to location of keys on the Tandy keyboard. E.g., shortly before the Back Slash (\) key was needed in the tutorial, the instructor would tell the Ss where that key was located on the Tandy keyboard. Key location was explained and diagrammed in instructions given to all

Ss, but for most Ss key location on the Tandy keyboard was a minor problem due to previous exposure to an IBM keyboard.

It should be noted that in no instance did the instructor provide information which was not available to all Ss elsewhere in the instructional materials. Also, in no instance did the instructor provide feedback on any S's CBT performance.

Dependent Measures

Two dependent measures were recorded: performance on the statistics exercise and number of spreadsheet commands used while working the exercise. Since most spreadsheet procedures can be performed in more than one way (e.g., a cell entry can be changed via an EDIT command or by simply re-typing the entry), this second measure was recorded to assess how many different spreadsheet commands were actually used during the exercise.

RESULTS

Means and standard deviations for Exercise Performance and Use of Spreadsheet Commands are given in Tables 1 and 2. Analysis of variance results are presented in Tables 3 and 4.

Table 1
Statistical Exercise Performance
Means and Standard Deviations

SETTING	EXPERIENCE	INTERACTION	
		Yes	No
Paired	Low Experience	44.50/18.17 (n=10)	42.86/14.10 (n=14)
	High Experience	77.00/12.52 (n=10)	71.36/16.45 (n=11)
Individual	Low Experience	42.73/14.55 (n=11)	31.67/13.92 (n=9)
	High Experience	63.89/14.09 (n=9)	63.50/11.56 (n=10)

Table 2
Use of Spreadsheet Commands
Means and Standard Deviations

SETTING	EXPERIENCE	INTERACTION	
		Yes	No
Paired	Low Experience	17.00/6.75 (n=10)	18.57/6.91 (n=14)
	High Experience	29.44/4.52 (n=10)	23.64/9.51 (n=11)
Individual	Low Experience	16.36/4.52 (n=11)	19.44/5.83 (n=9)
	High Experience	21.11/6.51 (n=9)	24.50/6.85 (n=10)

Table 3
Analysis of Variance
Statistical Exercise Performance

Source	SS	df	MS	F-Ratio	P-Value
Setting	1389.61	1	1389.61	6.624	0.012
Interaction	507.59	1	507.59	2.420	0.124
Experience	17087.60	1	17087.60	81.454	0.000
Dyad x Inter	35.97	1	35.97	0.171	0.680
Dyad x Exper	61.40	1	61.40	0.293	0.590
Inter x Exper	39.97	1	39.97	0.191	0.664
Dyad x Inter x Exper	236.91	1	236.91	1.129	0.291
Error	15943.44	76	209.78		

Table 4
Analysis of Variance
Use of Spreadsheet Commands

Source	SS	df	MS	F-Ratio	P-Value
Setting	34.57	1	34.57	0.745	0.391
Interaction	23.72	1	23.72	0.511	0.477
Experience	822.61	1	822.61	17.728	0.000
Dyad x Inter	152.45	1	152.45	3.285	0.074
Dyad x Exper	79.93	1	79.93	1.723	0.193
Inter x Exper	67.56	1	67.56	1.456	0.231
Dyad x Inter x Exper	40.90	1	40.90	0.881	0.351
Error	3526.63	76	46.40		

The hypothesis that Ss arranged in dyads would outperform Ss working CBT individually was supported, while the hypothesis that student-instructor interaction would produce higher performance received moderate support. For Statistical Exercise Performance (Table 3), Ss who worked CBT in pairs (dyads) outperformed Ss who worked individually (p < .02). Also, Ss who interacted with the instructor tended to outperform Ss who did not interact with the instructor (p < .13). However, Table 1 shows that the No Interaction/Low Experience Ss had the disproportionately lowest average score of all groups; this group's score tended to pull down the overall average score of the No Interaction Ss. In two of the four comparisons between Interaction/No Interaction Ss, the No Interaction Ss performed equally as well as the Interaction Ss. As expected, High Experience Ss outperformed Low Experience Ss.

For Use of Spreadsheet Commands (Table 4), High Experience Ss used more commands than did Low Experience Ss, also as would be expected. There was also a significant (p < .074) interaction between Setting (dyad/individual) and Instructor Interaction. The Use of Commands means presented in Table 2 suggest that the basis for this interaction was the No Interaction Ss. No Interactions Ss in the Individual Setting tended to use more commands than did their interaction/individual counterparts; this was not true in the Paired Setting in which Interaction Ss used more commands on average.

DISCUSSION

Stephenson (1991) found that instructor interaction had a positive effect on achievement and suggested that student-instructor interaction met certain social needs frequently not considered in CBT. However, Ss worked CBT individually in that study. In the present study instructor interaction had little or no effect on achievement when Ss worked the CBT tutorial in pairs. Moreover, instructor interaction did not have a disproportionately larger effect on those paired Ss without prior spreadsheet experience, a result opposite to that reported by Stephenson (1991). In this study low experience Ss working in pairs without instructor interaction performed equally as well as low experience Ss working in pairs with instructor interaction. Evidently, a dyad partner can provide the feedback, support, and social facilitation usually provided by the instructor in a more traditional classroom setting.

Ss who both worked CBT individually and received instructor interaction performed slightly (but not significantly) lower than Ss who worked CBT in pairs with instructor interaction. This result suggests that there is a social dimension to learning which can be provided by the instructor in an individual setting and by a team partner in the dyad setting. The fact that the lowest scoring Ss were those who worked CBT individually without instructor interaction, a result also reported by Stephenson (1991), suggests that weak students are more impacted by lack of social interaction than are strong students. Conversely, weak students benefit more from social interaction.

Ss with prior PC/spreadsheet experience tended to perform equally within learning setting (dyad/individual), suggesting that high skill students have less need for social support. However, prior experience Ss in the dyad setting did perform higher than their counterparts in the individual setting. This result suggests that, besides the social benefit derived from interacting with the instructor, working CBT with another person provides an additional benefit of added knowledge. Based on observations of the instructor in this study, this finding is not surprising. Student-instructor interaction is a sometimes thing, while interactions between paired students is an on-going event. In the dyad setting, social facilitation is constantly occurring in the form of immediate feedback, two person trial-and-error, and consensus behavior. While high skill dampens the impact of social interactions, it does not eliminate the positive influence of working CBT with another person.

The significant Use of Spreadsheet Commands interaction between Setting (dyad/individual) and Instructor Interaction can not be easily explained. Stephenson (1991) had previously reported that Use of Commands did not vary between Ss receiving/ not receiving instructor interaction; i.e., all Ss learned the commands equally well, but some used the commands better. To a large degree, this was also true in the present study. However, High Experience Ss used more spreadsheet commands than did Low Experience Ss, which may simply reflect the fact that these Ss brought more skill and knowledge with them into the session.

These results on the positive effect of working CBT in pairs support the results reported by Dalton (1990) and others and reemphasize the social nature of learning. For some students, learning is simply a social event. In the traditional classroom the instructor may provide most of the social functions. In individual CBT situations, the computer can not provide these functions. Consequently, when Ss run CBT individually, student interaction with a human instructor has a measurable effect. However, when social functions can be provided by a team partner, the need for interacting with the instructor is reduced. It also appears that interactions between students positively affect achievement above and beyond the affect produced by just interacting with the instructor.

Even though the Ss felt that the information gained from this CBT program would be valuable to their careers, this study's relatively short tutorial limits the generalization of the results. That limitation not withstanding, the results do suggest that the best CBT environment may not be the one in which students work individually. Rather, higher achievement is found when students work CBT in pairs. Whatever it is about the effect of social facilitation on learning is served quite well by the partner in a study team and, in fact, may be served better than by an instructor.

These results have several implications. First, since the results suggest that CBT should be worked in pairs, CBT software should be written for a small group setting. Second, the results suggest that the instructor should be properly trained to work with students who are arranged in such an environment. Third, the entire approach to CBT should incorporate social facilitation in the planning and implementation stages.

Overall, it does appear that the social aspect of learning needs to be considered in CBT. If students are working CBT individually, the instructor (or course administrator) must provide the social functions. However, if students are working in pairs, the instructor's social role is reduced. In the latter situation, the instructor's role may become very specialized; e.g., to provide social support for weaker students.

REFERENCES

Brophy, J. E. & Good, T. L. (1986) Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Third Handbook of research on teaching: 328-375. New York: Macmillian.

Dalton, D. W. (1990) The effects of cooperative learning strategies on achievement and attitudes during interactive video. Journal of Computer-Based Instruction, 17, 8-16.

McCombs, B. L., Back, S. M., & West, A. S. (1984) Self-paced instruction: Factors critical to implementation in Air Force technical training - A preliminary inquiry. (AFHRL-TP-84-23). Lowery Air Force, Base, CO: Air Force Human Resources Laboratory, Training Systems Division.

Stephenson, S. D. (1991) The effect of student-instructor interaction on achievement in computer-based training (CBT). (TP-91-0002). Brooks Air Force Base, TX: Armstrong Laboratory, Training Systems Division.

Webb, N. M. (1987) Peer interaction and learning with computer in small groups. Computers in Human Behavior, 3, 193-209.

Development of a Foreign Language Tutor Incorporating Intelligent Computer-Assisted Training

Richard E. Maisano and Cathie E. Alderks

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia

Introduction

The U.S. Army Research Institute and the U.S. Army Intelligence Center and School (USAICS) have undertaken an effort to apply advances in Intelligence Computer-Assisted Training to the problem of second language acquisition and maintenance for Army Military Intelligence (MI) linguists, specifically tactical interrogators, MOS 97E. The area .of intelligent computer-assisted language learning (ICALL) seeks to incorporate advances in Natural Language Processing (NLP) into new training systems. These advances permit training devices to be built that can analyze students' free language input, detect subtle errors made by the student, and formulate realistic responses for students in dialogue-like interactions. To develop effective ICALL requires a series of design decisions that exploit the capabilities of NLP tools while recognizing their limitations. This paper will describe the process and the design decisions involved in developing a recent ICALL system - the BRIDGE Tutor, intended to teach job language skills to MI linguists.

The problem. MI linguists receive second language (L2) training in global language skills: reading, listening, and speaking, at the Defense Language Institute (DLI). The course is intensive, but little time is spent on job-specific language training. This problem is compounded when the DLI graduate does not immediately rotate to a position requiring or using language skills. There can be a considerable time between when a soldier receives language training and when these skills are job requirements.

The skills required in interrogation are complex, even when conducted in the soldier's first language (L1) as taught at USAICS. The added burden of needing to master these skills in a foreign language can be very demanding. However, no formal L2 training is given in interrogation skills at USAICS. Often, L2 interrogation skills are acquired by apprenticing to a senior staff member once the student has left USAICS and is assigned to an operational unit. But again, there is no formal L2 instruction in these skills. This lack of formal L2 interrogation training recommended the MOS 97E MI interrogator as a good candidate for a job-oriented ICALL tutor. The first step was to identify the L2 skills most in need of instruction. Next, these findings were considered in relation to what was known about the current state of the Intelligent Computer-Assisted Instruction (ICAI) and NLP to determine what material would be presented by the tutor and how it would be presented.

Job Language Needs Analysis

A job language skills needs analysis was conducted to define in greater detail the foreign language skills required by MI interrogators for direct military interrogation. The analysis was conducted on German. German was chosen to be the language of the first prototype tutor because at the inception of the project German was one of the languages of primary interest at USAICS. An extensicn to Arabic is currently under way.

Information required for design decisions included what constructions (sentences, questions,) the NLP module would have to be able to analyze and what words needed to be included in the lexicon. If a word is not in the lexicon, the NLP module would not be able to analyze it. The study included identification of the German grammar characterizing the soldier with DLI proficiency levels 1 through 2, typical of MOS 97E. Identification of grammar elements would allow the design of the tutor to respond to those areas where the need was greatest.

Method. First, general materials on intelligence interrogation techniques and training were obtained from USAICS and reviewed. These materials provided an overview and some detail on the process of interrogation. Materials reviewed included FM 34-52, Intelligence Interrogation, (Department of the Army, May 1987), STP 34-97E1-SM, Soldier's Manual Skill Level 1 MOS 97E Interrogator, (Department of the Army, December 1986) and SIS 65820, (Interrogation Guide, Department of the Army, December 1983). In addition, three videotapes were obtained showing interrogations by USAICS students in L1 (English).

All the above materials dealt with L1 interrogation. This situation is not likely to occur very often in the field. Information on L2 interrogation, German in this case, was obtained from videotapes of six staged German prisoner of war (POW) interrogations. These interrogations were staged using full props and costumes. The German POWs were native German speaking professional actors. The interrogators in four of the interrogations were MI interrogation instructors, fluent in L2. The interrogators in the other two interrogations were MI interrogation students, not as fluent in German. In addition, a videotape of a staged Skills Qualification Test (SQT) with a fluent German MI interrogator and source (POW) was obtained.

Analysis. The L1 interrogation materials and manuals provided a framework for considering the videotapes. All the videotapes of both the German and English interrogations using both USAICS instructors and students were reviewed by a research team including a German teacher. This review concentrated on identifying categories of words used in interrogation, especially military words, and the types and forms of sentences used during an interrogation. These categories and constructions provided information for the development of the tutor's size limited lexicon and the syntactic parser, the heart of the tutor.

Besides reviewing the videotapes, semi-structured interviews were conducted with USAICS instructors and an analysis was made of the instructional materials. Visits were arranged to USAICS where L1 interrogation practice was observed. A list of questions was used to help gather needed information during the interviews.

Results. An analysis of the L1 materials used to teach interrogation led to the identification of what information is important in an interrogation. This was the basis for the review of the videotapes. This information included lexical, grammatical features, sentence level and structure, and discourse features.

Job specific military vocabulary needs to be learned. It is not currently taught in the general language courses soldiers take at DLI. Categories of words identified were: personal particulars, documents and equipment, missions, compositions of units, strength, dispositions, tactics, training, combat effectiveness, logistics and miscellaneous other military words. Some words will overlap with common words and help limit the size of the lexicon needed to support the tutor. The estimate on the size of the vocabulary needed to accomplish the job of L2 interrogation is 3000 miliary and 3000 common words. It can be assumed that most of the common words are taught at DLI.

Six grammatical features were found to have high priority for MOS 97E interrogators: question formation; past and present tenses; use of location and time phrases; use of modals; use of passives; and proper forms of address. Much of the information on students' weaknesses was obtained from the videctapes of the less fluent L2 student interrogators.

The information obtained from the needs analysis was then used to help guide the design decisions for the tutor. These decisions concerned the curriculum, the NLP formalism, and the tutor environment and approach.

Design Decisions

The curriculum. The first design consideration was what to teach the students, the curriculum. The decision was to focus on grammar. There were two reasons for this. The analysis of the video tapes, as well as the interviews with instructors, indicated that students made a number of systematic grammatical errors. The second reason is that syntactic parsers are one of the best developed NLP technologies. This decision meant that less emphasis would be placed on the development of a comprehensive semantics module that could support dialogues and simulate authentic task interaction. This was considered outside the scope of a first attempt at the development of a practical ICALL tutor.

The results from the analysis and curriculum decisions were used to identify the words and constructions the NLP module must be able to interpret and what could be expected as inputs from students. These student inputs feed the intelligent teaching modules designed to meet the instruction goals of the lesson. Also, the teaching modules make decisions on remedial or additional practice in problems areas identified from student inputs.

The ability to form questions involving spatial information was discovered to be of critical importance to a military interrogator. The job needs analysis revealed that students had great difficulty with the grammar of prepositional phrases that required distinguishing between direction and place. These phrases require selecting between the dative and accusative cases in German.

Agreement rules between subject and verb and between modifiers and nouns were another area shown cause students problems. Additionally, in German, all parts of sentence, subject, verb, and object, must agree in gender, number, and case. In choosing these areas in which to concentrate the curriculum, the emphasis of the tutor was on improving students' overall language ability.

The criteria used for selection of curriculum have applicability to other intelligent computer-assisted training situations. Curriculum was selected to address the gap between the job requirements and what students could do. An additional consideration was what could be represented by the tutor. Having these criteria, we arrive at a set of grammatical errors on which to concentrate instruction. These five errors are: subject-verb agreement, subject not in nominative case, verb-preposition agreement errors, preposition-noun agreement errors, and modifier- noun agreement errors.

Lexicon. The lexicon is the list of words the tutor can interpret. It is more than a dictionary. A lexicon contains information on what parts of speech and parts of a sentence words can assume, as well its relationships with other words. This information is used by the NLP module to identify legal inputs and errors. If students enter words not in the lexicon, the tutor can not handle the input. The results of the needs analysis were used to build a 5000 word lexicon that could the expected student inputs.

NLP Formalism. The next major decision in the design of the tutor was the NLP formalism appropriate for working with a computer tutor and how could it be operationalized in the tutor. Once the curriculum to be emphasized in the tutor was determined, this information helped guide the development of the NLP module. The need to be able to handle certain kinds of errors and a wide variety of sentence constructions helped identify the characteristics the NLP module would have to possess. The criteria for the NLP module were: compactness of representation, accuracy within the specified domain, error tolerance, and extendibility across constructions and languages. Some representations of the domain of a NLP module list every construction-specific rule to characterize all syntactic possibilities in a language. This exhaustive approach is not a very efficient, compact, or economical way to address the required constructions, especially when the tutor must run in a PC environment. Other formalisms seek to apply a limited set of rules, governed by setting parameters for each rule, to account for all legal constructions of a language. The approach chosen for the tutor was one of these latter formalisms.

The chosen formalism is called Government Binding (GB) theory. A GB approach allows modules to be developed which can generate all possible constructions in any language, not just the legal constructions. But the illegal constructions in a language are weeded out by other modules that use a limited set of rules to define legal representation. Another advantage of GB theory is that the modules can be turned on or off independently. Therefore, if certain types of errors are not of interest, the parser will not report them and the tutor will continue to function. This approach answers the need to have the tutor expendable to other languages. Unlike exhaustive formalisms, this approach should be adaptable to a second language by switching the parameters and setting the limiting rules in the modules for the legal constructions of the second language. Currently, we are in the process of extending the tutor to Arabic.

Tutor Frame. Another important decision was the instructional environment design. The problem was how to support the development of language skills while keeping the student interested and simulating as much as possible the environment in which the language would be used. The decision was to use a communicative approach to tutor environment design. This approach is based on mimicking the real occasions and consequences of language use.

Most communicative environments attempt to immerse the student in the language. Language production is stimulated by interactions and language acquisition is accomplished within the framework of real communication instead of repetitive exercises. The ideal would be to simulate an interrogation dialogue within the tutor. This was considered ambitious for a first attempt. A compromise solution was developed that had the student work through a lesson using a map and text about the area represented on the map. The tutor asks questions about the map and the student responds, in many instances the responses are in the form of free text.

Tutoring rules. In the current system, tutoring rules, written in Prolog, determine exercise sequencing and feedback based on student performance. Rules used to select exercise sequences reflect what linguistics principles are to be taught. The rules can also determine remedial activities and link feedback to student input. Often these rules will reflect the pedagogical theory surrounding the tutor. The current pedagogical theory for L2 is a communicative approach developed for the classroom, not an ICALL system. Its assumptions may not work in the same way on a computer tutor as they do with an instructor. To resolve this issue for the BRIDGE tutor an attempt was made to extract rules from a communicative approach oriented German expert. Second, an authoring interface was developed that allows an instructor or researcher to change which linguistic principles are being taught, change feedback received by the student, and to manipulate branching. These features will allow exploration of different rules to determine the optimal rules for use in the tutor.

A major issue, amenable to exploration using the authoring interface, was the degree and type of feedback the student would receive during the exercise. A communicative approach would advocate limiting the feedback. Information on factual accuracy as it relates to the scenario used in a lesson fit in a communicative approach. This could be presented in a conversational form. But feedback for every grammatical error clearly violates the assumptions of the communicative approach. One solution to this dilemma might be to save grammatical feedback until the end of a lesson and provide it as a form of overall comment. This is one of the issues for research to resolve.

Error handling. The BRIDGE tutor detects errors and can inform the student about the type and location of the error. But the information on errors from the analysis of the input performed by the present tutor is insufficient for diagnosis of systematic problems and guidance for remediation. Nevertheless, it is felt that this system can provide more information to the student on errors than non-parser-based systems.

Conclusions. The issues considered in the design of the BRIDGE tutor are generalizable to other tutoring environments. A compromise between parsing and communication had to be reached. It was beyond the scope of this project to develop a truly communicative environment that might better support the acquisition of critical language skills for MI linguists. But, the information gained from this project have moved the technology closer to fully integrating NLP into a CALL system.

REFERENCES

Criswell, E., Byrnes, H., Rapaport, B., Dukes, M., Miller, 1., Blascak, D., (1989), Analysis of Foreign Language Requirements for Military Interrogators, Task 2 Technical Report, U.S. Army Research Institute, Alexandria, VA.

Holland, V. Melissa, Issues in Developing an intelligent Tutor for Second Languages, (1991), Conference on Intelligent Computer-Aided Training, 20-22 November, NASA/Johnson Space Center, Houston, TX.

Miller, L., Criswell, E., Weinberg, A., Byrnes, H., Groundwater, E., Blyskal, J., (1989), Design Document for the BRIDGE Tutor, Task 1 Technical Report, U.S. Army Research Institute, Alexandria, VA.

Evaluating a Foreign Language Tutor: Extendibility and Usability[1]

Cathie E. Alderks and Richard Maisano

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia

There is an ever demanding need for Military Intelligence linguists to maintain their language Proficiency. Many times, these linguists are assigned duty responsibilities that either do not require their linguistic skills and/or leave little opportunity to maintain or improve their abilities. In addition, in school, there are few resources to teach job specific language skills. Recognizing the needs of the intermediate/advanced linguist, the Army Research Institute (ARI) developed a unique product: the BRIDGE Foreign Language Tutor.

Important attributes of the tutor include 1) the incorporation of a natural language processing (NLP) parser that allows for the analysis of freely-typed input by the students in the target language, 2) the detection and categorization of the type of error made by the student, and then the selection of additional instruction in areas where the students needs more work, 3) extension to other languages, and 4) a completely authorable interface that permits changing and/or adding new exercises without the necessity of being knowledgeable of specific programming languages.

There are advantages for using parsers in foreign language tutors. There is tremendous freedom of expression. Because the parser operates on rules for word combinations and not on specific strings of words, a student can choose how something is to be said. No one correct answer is required. The instructor doesn't need to second-guess the students and the students don't need to second-guess the instructor. For example, any number of correct answers could answer the question "How did the dog get the bone?". Possible correct answers could be "The boy gave the dog the bone.", "The dog got the bone from the boy.", "The bone was given to the dog by the boy.", or another grammatically correct sentence. If a response is grammatically incorrect, the parser will indicate the problem.

Natural Language Processing

The broad goal of natural language processing is natural language understanding, i.e., to have a computer analyze and interpret free-form text and discourse as a human would. This is a bit ambitious for present technology. A more modest goal is parsing, to have a computer analyze free-form text and discourse into the basic grammar structures of the language. Thus, an NLP parser takes language input and then analyzes it according to rules and patterns (syntax) to determine if it is grammatical, to translate to a second language, to extract information, to draw inferences, to act, or to respond in a

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

dialogue. Ideally, the NLP also incorporates semantics, fact knowledge, and knowledge about discourse interaction in its analysis. In actuality, parsers rarely go beyond syntax analysis; semantic analysis is uncertain and necessarily narrow.

The parser in the BRIDGE tutor is composed of several main components: a preprocessor, a morphological analyzer, a lexicon, a syntactic parser, an error handling facility, and a semantic interpreter. The morphological analyzer and lexicon work in conjunction in order to minimize the size of the lexicon (dictionary) and yet have it as expansive as possible. The words are broken into segments of roots and affixes. Subparts of the words are then retrieved when needed, their features merged and then they are passed to the syntactic parser for further analysis. The BRIDGE lexicon contains approximately 5000 entries. If morphology is determined to be correct (words in lexicon, no misspellings, etc.) the syntactic parser then determines phrase structure and syntactic correctness and parse trees are constructed. Errors are detected through the error handling facility.

Type of Error Made by Student

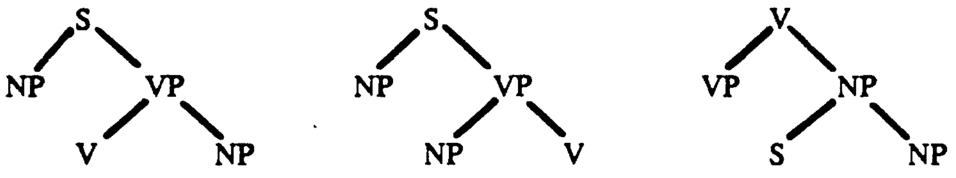
The present NLP parser is constructed for the German language. Requirements for the parser include being able not only to handle correct sentences and phrases, but also to tolerate, detect, and diagnose errors without being hung-up. Likely types of errors in the input include words not in the lexicon, misspellings, semantic errors, and various grammar errors. The present NLP parser diagnoses five classes of grammar errors: 1. subject-verb agreement, 2. subject not in the nominative case, 3. verb-preposition agreement errors, 4. preposition-noun agreement errors, and 5. modifier noun agreement errors. Errors the parser does not handle include word order errors (the verb must be in the second position in the main clause or final in the subordinate clause in German), sentences that begin with conjunctions, compound verbs and sentences, relative clauses, and grossly ungrammatical sentences or fragments. Additionally, the semantic analyzer, which is organized into an "isa hierarchy", has not been incorporated into the tutor. The implications of this are that the tutor can give feedback only about form, not about the correctness of the factual information. For example, if a question asked was "Where is Oregon?", answers of "Oregon is north of California." and "Oregon is south of California." would both be considered correct because they are both grammatically correct. The BRIDGE parser, at present, can be used for grammar instruction, not for factual knowledge.

Extension to Other Languages (German to Arabic)

The parser was constructed according to the principles of Government-Binding (Chomsky, 1981). Briefly, the underlying principles include the idea that a small number of abstract principles apply to many different constructions. Independent formulae are grouped into interacting modules. In turn, these modules are parameterised such that by modifying them to a small degree, patterns associated with a variety of languages can be generated. Put glibly, by switching a few switches, the basic constructions of another language can be handled.

In a test of this extendibility, a language unlike German was chosen: Arabic. Using those modules that captured the structural similarity across languages, a parser in Arabic was developed in less than 1/3 the time that was required for the German parser. This was all the more impressive because no comprehensive syntactic analysis of the Arabic language was available, whereas German had been extensively analyzed. It must be noted that development was still costly in that it took several months, but it did help to have the tools in place.

Examples of interacting grammar modules include phrase structure and case assignment. All languages have some sort of phrase structure. For example, German is Head-Final (the structure that determines the identity of a phrase generally appears on the final edge of the phrase), and English is Head Initial (the identifying structure appears at the beginning edge of the phrase). Arabic allows the English order, but also allows the verb to appear at the head of the sentence. The head of the category is that which determines that category's type, the verb of a verb phrase, the noun of a noun phrase, etc. A language that has a verb appearing before its objects is head initial and will also have a noun appearing before its objects. Thus, individual language phrase structure rules for each category type are not needed as the general form for the module can be initialized as head-initial or head-final, depending on the requirements of the language--in this case head final for German (SOV) and head initial for Arabic and English (VSO and SVO). A general transformation allows the movement of the verb from the second position to the initial position for Arabic and from the final to the second position for German. In a simplified example, for the English sentence "Dogs eat meat.", the German sentence would be with an embedded clause "[I know that] dogs meat eat.", and the Arabic sentence would be "Eat dogs meat.". Figure 1 further illustrates this structure.



(SVO)
Dogs eat meat.

(SOV--embedded clause)
...dogs meat eat.

(VSO)
Eat dogs meat.

Figure 1. Phrase Structure Rules Generated by a Government-Binding Grammar Approach.

According to Government-Binding Theory, every lexical Noun phrase must receive a grammatical case. Examples of cases are nominative, accusative, dative, genitive. Case is assigned by predicates (verbs, prepositions, tense or agreement marker) to particular positions. For example, in the following sentences, the required pronoun marking is determined by whether the noun phrase is pre- or post-verbal.

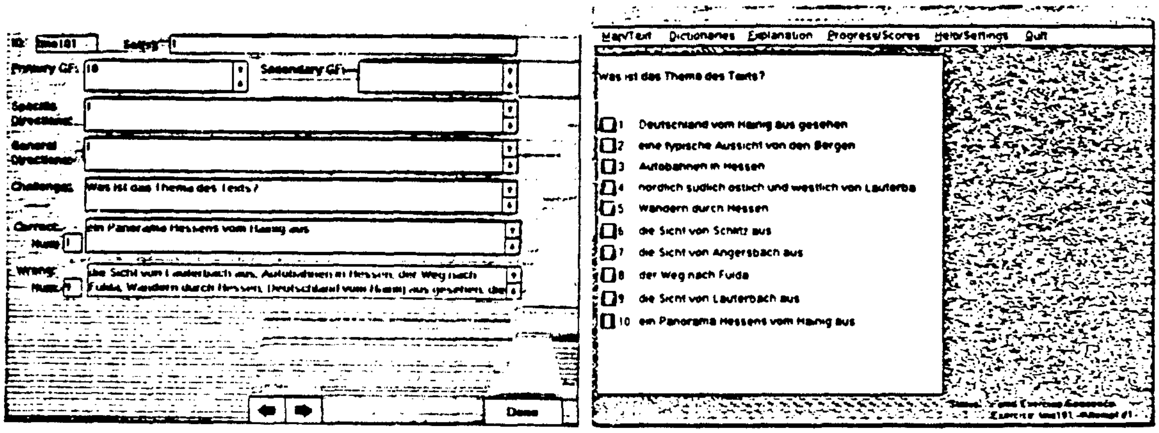
1. He(NOM) looked at him(ACC).
2. *Him(ACC) looked at he(NOM).

In the correct sentence, sentence 1, the subject "He" is in the nominative case and the object of the preposition "him" is in the accusative case. The incorrect example, sentence 2, has the subject incorrectly in the accusative case and the object of the preposition incorrectly in the nominative case. Within the parser, modules that assign case can be switched as appropriate according to the language in question.

Completely Authorable Interface

An authorable interface is desirable for both theoretical and practical reasons. As everything is not known in second language acquisition, we want to author lessons to perform research that will determine ideal instructional design and investigate the manner in which particular languages are learned. Examples of research questions might be: Does stopping the student for errors aid in language learning, or is it best to allow certain errors to go uncorrected or unsignalled? Are certain constructions always learned before other constructions? On a more practical nature, an authorable interface allows instructors to make or modify their own lessons.

Several ways in which the interface is authorable will be discussed. Each type of exercise has an authorable template. Figures 2 and 3 show two of these templates and the corresponding student exercises. Types of exercise include multiple-choice, pointing-to-locations-on-a map, classification, fill-in-the-blank, and full-sentence. Challenges may be either written or pre-recorded. Data for the exercises may come from maps, written text, or pre-recorded spoken text. New exercises may be constructed by simply pulling up a template and then filling in the new question and where applicable as in the case of multiple-choice, the correct and incorrect possibilities. It should be noted that multiple-choice responses need not be in any particular order as the computer randomly places them in the exercise each time the question is used.



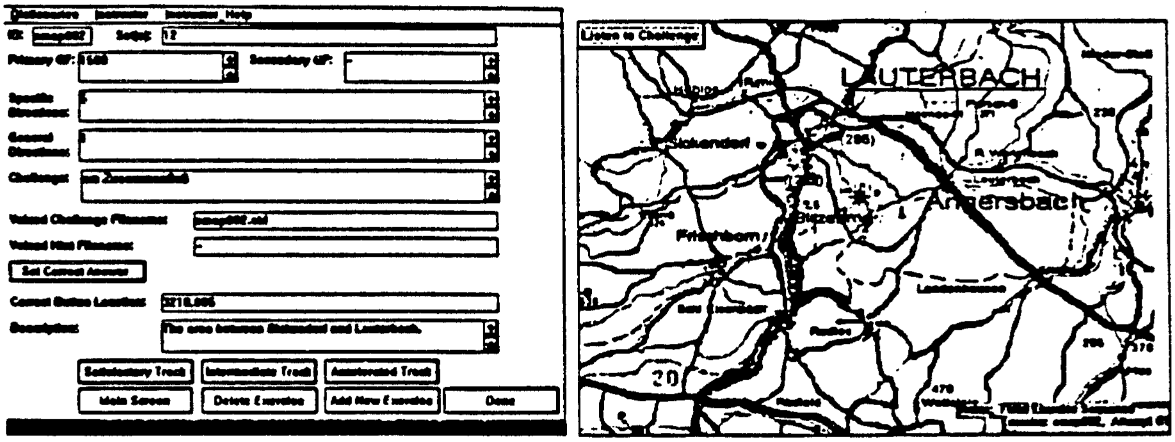
ID: ime101
Subject: 1
Primary GF: 10
Secondary GF:
Specific Directions: 1
General Directions: 1
Challenge: Was ist das Thema des Texts?
Correct: ein Panorama Hessens vom Hainig aus
Num: 1
Wrong: die Sicht von Lauterbach aus, Autobahnen in Hessen, der Weg nach Fulda, Wandern durch Hessen, Deutschland vom Hainig aus gesehen, die
Num: 9

Was ist das Thema des Texts?

□ 1 Deutschland vom Hainig aus gesehen
□ 2 eine typische Aussicht von den Bergen
□ 3 Autobahnen in Hessen
□ 4 nordlich sudlich ostlich und westlich von Lauterba
□ 5 Wandern durch Hessen
□ 6 die Sicht von Schlitz aus
□ 7 die Sicht von Angersbach aus
□ 8 der Weg nach Fulda
□ 9 die Sicht von Lauterbach aus
□ 10 ein Panorama Hessens vom Hainig aus

Exercise ime101 -Attempt #1-

Figure 2. Multiple-choice authoring template and student screen.



ID: map002 Seq#: 12

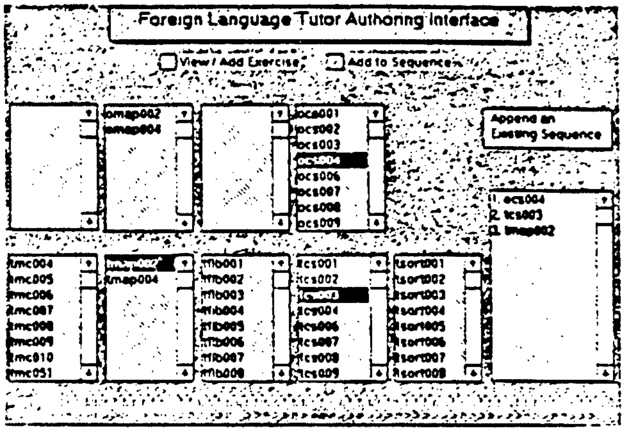
Valued Challenge Filename: map002.aid

Correct Button Location: 218,005

Description: The area between ... and Lauterbach.

Figure 3. Map exercise authoring template and student screen.

Sequencing of exercises into lessons is equally easy. Figure 4 shows the template for sequencing. All the instructor needs to do is click on an exercise and it appears in order on the list. Once the list of exercises is formalized, it can be saved. The organization of the lesson then is finished.


Foreign Language Tutor Authoring Interface
□ View / Add Exercise
□ Add to Sequence
Append an Existing Sequence

omap002
omap004

ocs001
ocs002
ocs003
ocs004
ocs006
ocs007
ocs008
ocs009

1. ocs004
2. tcs003
3. tmap002

tmc004
tmc005
tmc006
tmc007
tmc008
tmc009
tmc010
tmc051

tmap002
tmap004

tfib001
tfib002
tfib003
tfib004
tfib005
tfib006
tfib007
tfib008

tcs001
tcs002
tcs003
tcs004
tcs006
tcs007
tcs008
tcs009

tsort001
tsort002
tsort003
tsort004
tsort005
tsort006
tsort007
tsort008

Figure 4. Instructor's template used for sequencing exercises.

Errors may be classified as either primary or secondary depending upon the objectives of the lesson. For example, in an exercise involving location and direction, case errors following prepositions may be determined to be the main objective, therefore, only these specific errors "count" in determining "correctness". Others, considered secondary, are de-emphasized and will be noted, but will not stop the progression of the lesson. These secondary errors may be investigated should the student desire, but it is not required. Therefore, the lesson progression is not stopped for every little error. Any number of captured errors may be classified as primary or secondary, as the instructor desires.

Usability of the Tutor

Initial usability of the tutor was evaluated by four students and three instructors of military interrogation (MOS 97E) at Ft. Huachuca. Selected usability concerns included 1) the tutor's capability to correctly parse freely-typed, original sentences in German and 2) interfaces, help, and feedback for each type of exercise presented. Many insightful suggestions and comments were provided amidst unanimous endorsement of the tutor. Illustrative comments by instructors included: the tutor is "leaps and bounds over what we've had before," "Production skills are important...it is good to have [students enter] full sentences," and "Instructors could develop lessons on the tutor as part of their own language maintenance programs." Students likewise were enthusiastic with comments such as "It's nothing like we've ever had before," A "really good program". The dictionary is "kind of cool," I like "not having to read instructions in a book", The tutor is "better than 'string-matching' programs which are frustrating" because of the requirement to exactly match the canned answer. The students also liked getting grammatical feedback, even the "parse trees". They tolerated some uncertainty and gaps in the parser. These observations were surprising because 1. the parser had not yet been incorporated into the tutor such that students were typing in sentences without the framework of the lesson and 2. the students seemed to turn the experience into a "game" to try to determine the parameters of the parser, i.e., what it could do and what the boundaries were. Their dissatisfaction was that they wanted more--more maps, more scenarios, more graphics. The instructors liked the ease with which lesson exercises could be changed. However, opinion was split on whether other demands on their time would limit lesson development; there was some feeling that native speakers with military and instructional design knowledge should be developing the lessons. It was also felt that lesson development could aid in their own language maintenance program for instructors. Further tests of the tutor's usability and effectiveness are planned.

Summary

The authorability of the tutor provides many opportunities for research in second language acquisition as well as providing excellent opportunity for language maintenance and learning. The ability to freely input self-generated statements allows for practice in production skills with immediate correction if needed and desired. Students are able to "try out" constructions with which they may feel "shaky" to gain confidence and further increased proficiency.

References

Chomsky, N. 1981. Lectures on Government and Binding. Dordrecht: Foris Publications.

BATTLEFIELD OUTCOMES PERCEIVED AS A FUNCTION OF ATTACK STRATEGY, DEFENSIVE POSTURE, AND SURPRISE

Phillip L. Vandivier and Raymond O. Waldkoetter
U. S. Army Soldier Support Center
Fort Benjamin Harrison, IN 46216-5530

Stella Vandivier
Psychological & Educational Publications
Indianapolis, IN 46256-4653

INTRODUCTION

We usually perceive organizational variables such as attacker strategy, defensive posture, and the element of surprise have important impacts on battle outcomes. However, does objective, quantifiable evidence exist which indicates the superiority of certain attacker and defender strategies/postures over others? Nearly everyone assumes the element of surprise is conducive to success on the battlefield. But just how important is it? How much difference in battlefield outcomes does the element of surprise make?

The purpose of this study was to investigate how battlefield outcomes vary when perceived as a function of initial attacker strategy, defensive posture, and the element of surprise.

Initial Attack Maneuvers

A frontal attack that a leader may order strikes the enemy across an area over a direct approach. It exposes the attacker to concentrated fire while simultaneously limiting the effectiveness of the attacker's own organizational firepower. Hence it is the least economical form of maneuver (FM 100-5, 1986). Nevertheless, it has been used by an overwhelming 73% of attackers in battles since the inception of World War II (McQuie, 1988).

Envelopment, on the other hand, has units strike at the flanks and rear of the enemy while a diversionary attack is staged in the forward area. Single and double envelopments are directed against one and two enemy flanks, respectively (FM 100-5, 1986). Approximately 13% of attackers since World War II have employed this initial strategy (McQuie, 1988).

River crossing attacks require the quick crossing of bodies of water without losing momentum (FM 100-5, 1986). The goal is to facilitate organizational movement across the water in a way that minimizes the attacking force's ability to project firepower (FM 100-15, 1989). Approximately 10% of attackers have used this strategy (McQuie, 1988).

Other leaders' initial attack strategies include breakthrough and pursuit, which together only comprise about 4% of attacks (McQuie, 1988).

Initial Defense Postures

Fortified defense for any position is a defense system prepared with sufficient material and time to complete entrenchments, fortifications, and obstacles. About 39% of defenders since WW II have used this posture (McQuie, 1988).

Prepared defense is a defense prepared in one day to improve a given position. It lacks the time and material investment that is necessary to establish a fortified position. Approximately 30% of defenders have employed this posture (McQuie, 1988).

Hasty defense is normally organized while in contact with the enemy or when contact is imminent and time for battle preparation is limited. Foxholes, emplacements, and obstacles are used. With enough time (approximately one day), a hasty defense can be improved to a prepared or fortified defense. Approximately 24% of defenders have used this posture (McQuie, 1988).

Delay is a retrograde movement in which the defender organization slows down and damages the enemy to gain time, but avoids decisive engagements and tries to keep from being outflanked. About 5% of defenders have used this posture (McQuie, 1988).

The remaining 2% of defenders utilized a mobile defense or withdrawal (McQuie, 1988).

Element of Surprise

Surprise is achieved by a leader and organization when their force is able to confront the enemy with tactical circumstances that the opponent did not anticipate or adequately prepare for. Surprise often is accomplished with respect to time, location, firepower, or maneuver (McQuie, 1988). When it is psychologically successful to a high degree confusion and broken courage in the enemy's ranks occur (Clausewitz, 1962). Less than 15% of the battles fought since the beginning of WW II had the element of surprise (McQuie, 1988).

METHOD

Data were obtained from the Historical Characteristics of Combat for War Games (Benchmarks) published by the U. S. Army Concepts Analysis Agency (McQuie, 1988), which includes 260 conventional land warfare battles conducted between 1937 and 1982. Battle is defined as "a significant combat encounter between hostile forces at various echelons of aggregation up to and including corps, army, and army group" (McQuie, 1988).

Data were collected by a cross section of military historians from archives, interviews, and books, and reviewed by an independent panel of experts. Data utilized included the following (see above for definitions):

Initial Attack Maneuver--Frequencies of battles in which the attacker organization used frontal attack, single or double envelopment, or river crossing on the initial attack.

Initial Defense Posture—Frequencies of battles in which the defender organization used delay or hasty, prepared, or fortified defense.

Surprise—Frequencies of battles in which the attacker organization achieved or failed to achieve a significant element of surprise.

Attacker wins—The frequency with which attacker and defender resolutions matched the definitions provided in Figure 1. Attacker wins are considered effected with the resulting defensive failures to win.

Attacker failures to win—The frequency with which attacker and defender resolutions matched the definitions provided in Figure 1. Failures of attackers to win are considered effected with the resulting defender wins.

Only battles with matching echelons (e.g., divisions vs. divisions) were used to maintain relative equivalence of force strength of attackers and defenders. A total of 131 battles were used for the analysis.

Chi-squared analyses were conducted to test several hypotheses (H) to determine the extent to which attacker: Win vs. failure-to-win rates would differ for all 131 battles (H1); and Win vs. failure-to-win rates would vary across attack strategies (H2), defensive postures (H3), and the presence or absence of surprise (H4).

RESULTS

Overall chi-squared results for H1 indicated significantly more of the attacking forces won than failed to win battles, $X^2(1)=4.77$, $p < .03$. A total of 59.5% (n=78) of the attacking forces won their battles; 40.5% (n=53) failed to win.

Organizational results for H2, H3, and H4 are provided in Figures 2, 3, and 4. Of these only H3 achieved an acceptable level of statistical significance, indicating that attacker organizational results do differ across defensive postures, $X^2(3)=8.21$, $p < .05$. Attackers achieved 100% success against delay, followed by 63.2% and 55.8% for prepared and fortified, respectively. The hasty defensive posture results were obviously of minimal effect for the attacking organizations and slightly favored the defender (Figure 3).

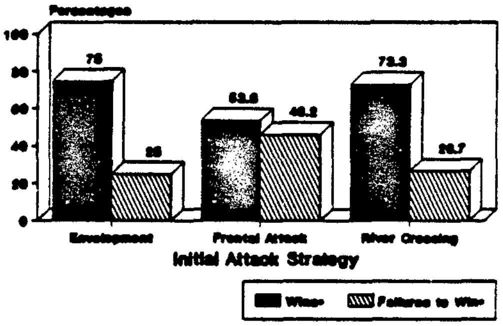
The H2 organizational results approached, but did not meet statistical significance, $X^2(2)=4.45$, $p < .11$. The attack win rates for envelopment and river crossings are rather evenly matched at 75.0% and 73.3%, respectively, while frontal attack is only 53.8% (Figure 2).

The H4 results indicated no significant differences exist in attack win/fail-to-win rates as a function of the element of surprise (Figure 4), $X^2(1)=.12$, $p = .73$.

Figure 1. Definitions of Battle Outcomes

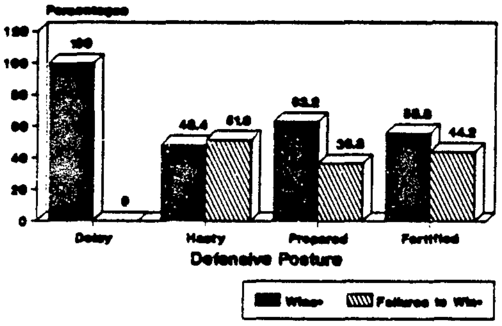
Attacker Wins Defined By Following Attacker/Defender Resolutions:		Attacker Failures to Win Defined By Following Attacker/Defender Resolutions:	
Attacker	Defender	Attacker	Defender
Penetration	Withdrawal	Repulsion	Pursuit
Breakthrough	Withdrawal	Repulsion	Withdrawal
Breakthrough	Annihilation	Repulsion	Stalemate
Breakthrough	Surrender	Withdrawal	Stalemate
		Withdrawal	Pursuit
		Penetration	Stalemate

Figure 2. Percentages of Attacker Wins Across Different Attack Strategies



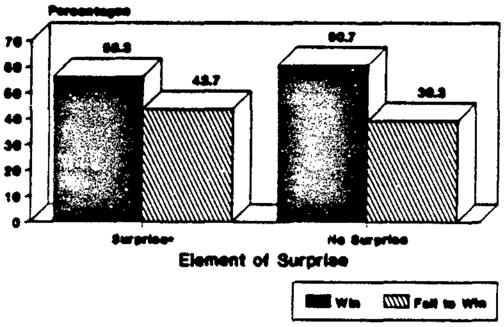
*From standpoint of attacker.

Figure 3. Percentages of Attacker Wins Across Different Defensive Postures



*From standpoint of attacker.

Figure 4. Percentages of Attacker Wins As a Function of Surprise



*Surprise achieved by attacker.

DISCUSSION

The organizational results can be applied usefully by leaders and planners of future conventional land battles or those battles fought under similar conditions.

Attackers generally have an advantage over defenders. While this conclusion probably will not amaze most battle strategists or experienced leaders, results go somewhat further in suggesting that such an advantage applies across most defensive postures and attack strategies. Differences in wins across defensive postures seem to be modulated somewhat by an extreme disparity of 100% attacker wins under a delayed posture--which, by definition, is a retrograde defensive maneuver.

Attackers should probably consider other attack strategies than frontal assault. A casual observer might question, "Why is this the most popular attack strategy when it has resulted in the lowest win rates?" Although river crossing attacks are not always feasible, one wonders why envelopment is not more readily embraced in view of its somewhat higher win rate. However, it may be that envelopment demands greater organizational force and material and takes longer to maneuver.

Time to prepare for attack does not appear as decisive as previously thought for defenders, since they successfully held off more attacks under hasty preparation than under prepared and even fortified postures. But in the battle sequence a hasty organizational posture will occur most often when an attacker loses momentum.

The element of surprise might not be as decisive for the overall outcome of battle as for the initial stages of conflict. The attackers performed only slightly better than defenders when the former had the advantage of surprise (Figure 4). The multifaceted nature of perceived surprise and leaders intentions (Reeves & Hansen, 1989) also suggests the possibility that some surprise might be more effective than others. Nevertheless, attack strategy with these findings suggests initial advantages attackers had from the element of surprise were apparently confounded prior to the final outcomes of a large proportion of battles involving differing attack strategies and defensive postures.

REFERENCES

Headquarters, Department of the Army (1986). Field Manual 100-5: Operations. Department of the Army, Washington, D.C.

Headquarters, Department of the Army (1989). Field Manual 100-15: Corps Operations. Department of the Army, Washington, D.C.

Leonard, R. A., Editor (1967). A short guide to Clausewitz on war. New York: Putnam's Sons.

McQuie, R. (1988). Historical characteristics of combat for wargames. U.S. Army Concepts Analysis Agency, Bethesda, Maryland.

Reeves, D. T. & Hansen, R. J. (1989). Development of the human dimension combat readiness index-experimental (HDCR1-X). Proceedings of the 31st Annual Conference of the Military Testing Association, San Antonio, TX, 6-9 Nov.

Performance Appraisals: An Employee Perspective

Joyce Shettel Dutcher
Navy Personnel Research and Development Center
Cynthia B. Heller
California School of Professional Psychology, San Diego
John P. Sheposh
Navy Personnel Research and Development Center

Abstract

Although performance appraisals have received much attention from I/O psychologists, employee opinions of appraisal systems and their relationship to aspects of jobs and organizations have been relatively unexplored (Sulsky & Balzer, 1988). Examined in this research were the opinions of employees from a DoD activity who were exempt from performance appraisals (PAs) by waiver of Federal Law, and employees from comparable DoD activities who continued to receive PAs over a four year period. Results indicated a majority of exempt employees reported that, since the elimination of PAs, there was (a) greater cooperation and less competition between employees, (b) less friction between supervisors and subordinates, and (c) increased team orientation in work units. Both exempt employees and those who continued to receive PAs increasingly preferred not to receive appraisals. Also investigated was whether employees who preferred appraisals differed in their view of the organization from those who preferred not to receive appraisals. Employees who received PAs but preferred not to receive them were less positive regarding such issues as the quality of feedback and supervisor-subordinate relationships than were employees who wished to receive PAs. Thus, those who were less positive about their work experiences preferred not to receive PAs. Surprisingly, exempt employees, who preferred not to receive appraisals, were initially less positive than were employees who preferred appraisals, but over time, became significantly more positive. This may reflect a growing realization that effective feedback and rewards can be accomplished in the absence of PAs. Alternatives to PAs are discussed in light of the results.

Background

Performance management and, in particular, performance appraisal in the public sector have been the target of much discussion, analysis, and criticism (Alexander & Wilkins, 1982; Stephan & Dorfman, 1989; Sulsky and Balzer, 1988; and Balzer & Sulsky, 1990). Organizations want a way to distinguish between good and bad performers. Unfortunately, most appraisal systems are seen as being influenced by political considerations which lead to inflation or deflation of appraisals (Longenecker, Sims, & Gioia, 1987). The conclusion of 50 years of research on performance appraisal has shown that there is no perfect, objective appraisal system because one human being is required to evaluate another (Bradshaw, 1989).

An alternate to individual appraisals is recommended by W. E. Deming (1986). He argues for the removal of performance appraisals that focus on the individual's performance because they lead to a competitive atmosphere rather than a cooperative one. The individual focus not only creates negative effects (e.g., builds fear, nourishes rivalry and politics) but it inhibits the cooperation that is essential to encourage the dedication needed for constant improvement (Deming, 1986; p. 102). Scholtes (1987) extended Deming's definition of the failure of performance appraisals by stating that performance appraisal encourages mediocrity by rewarding those who set safe goals, by putting pressure on employees to work around systems rather than improve them, and "by demoralizing employees by creating either losers or cynics" (p. 6). In place of individual performance ratings, Deming (1986) recommends the monitoring of individual performance through organizational performance measures using industrial work measurement systems and statistical process control (SPC). These processes can be used not only to empower the employees but also to allow the group to follow their own progress against objective standards.

The present research reports the opinions and perceptions of employees from a DoD activity that had implemented a series of changes including Total Quality Management (TQM), productivity gainsharing, and the elimination of performance appraisals. Individual performance appraisals were eliminated and performance was monitored through organizational effectiveness indices (e.g., personnel costs relative to output). Furthermore, the rewards for organizational effectiveness were shifted from individual performance to group performance by means

The opinions expressed in this paper are those of the authors. They are not official and do not represent the views of the Navy Department.

of organization-wide productivity gainsharing. These initiatives can be viewed as a radical departure from the appraisal and reward systems that are employed in Federal Government organizations. Specifically, this study compares the opinions of employees who are exempted from personal appraisal by waiver of Federal Law and those from similar DoD activities who continued to receive individual performance appraisals.

Method

Sites

The present study was conducted at five sites within a Department of Defense agency, one experimental site and four sites which served as comparisons. The five sites provide logistic support to the armed services. Three sites are located in the West and the remaining sites are located in the South and the Midwest. At the experimental site, a set of initiatives which included TQM, gainsharing, and the elimination of individual performance appraisals was in its fourth year of implementation. Organizational performance measures and process improvement techniques such as SPC procedures (Deming, 1986) were used to provide performance feedback to employees exempted from individual performance appraisals at the test site. Employees at the control sites continued to receive performance appraisals. Employees from all sites came from both the blue- and white-collar sectors.

Subjects

A total of 2716 supervisory and nonsupervisory employees, 895 from the experimental site and 1821 from the combined control sites. The sample consisted of 54 percent males and 41 percent females. The majority of the sample ranged from 30 to 49 years of age (63%).

Measures

A questionnaire was designed to measure respondents' perceptions of the organization. Items bearing directly on the performance appraisal system were examined in the present study. Items employed 5 point Likert-type scales.

Procedure

The fourth annual questionnaire administration took place approximately three years after implementation of the organizational changes. Subjects were randomly selected from the five separate work forces, representing experimental and comparison groups. Questionnaires were given on site by researchers from the Navy Personnel Research and Development Center. In addition, interviews with groups of 10-15 randomly selected employees were conducted. Approximately 200 supervisory and nonsupervisory employees were involved in the group interviews.

Results

The responses of the workforces at both the experimental and control sites to the question concerning preference for performance appraisals (PAs) are presented in Table 1. Both exempt employees (experimental site) and those who continued to receive PAs increasingly preferred not to receive PAs.

Table 1
I would prefer not to receive an annual performance appraisal from my supervisor.

	Experimental Site			**Control Sites**		
	Disagree	Undecided	Agree	Disagree	Undecided	Agree
Baseline						
1988	44.2%	17.3%	38.5%	44.8%	12.8%	42.4%
1989	46.9%	18.8%	34.3%	38.4%	12.0%	49.6%
1990	38.1%	19.4%	42.5%	33.8%	12.1%	54.1%
1991	33.8%	18.9%	47.3%	27.5%	12.6%	59.8%

Supervisors' attitudes toward managing their employees under their respective systems are shown in Tables 2 and 3. The pattern of responses were similar for supervisors from the experimental and control sites for the fourth year. It is of interest to note that supervisors at the experimental site do not see a lowering of their ability to improve performance of subordinates (Table 2) as a result of the elimination of performance appraisals. In fact, they agreed more over time with this statement whereas supervisors at the control sites agreed less. The results in Table 3 show that a large percentage of supervisors disagreed that it was difficult to reward or discipline employees without PAs. In 1992 several questions were asked to directly assess the effect of not having PAs. The responses to these items shown in Table 4 indicate that the majority of the workforce at the experimental site believe that there is less competition among employees, less friction between supervisors and subordinates, and increased team work in the absence of appraisals.

Table 2
The current system enables me to help the people I supervise improve their performance.

	Experimental Site			**Control Sites**		
	Disagree	Undecided	Agree	Disagree	Undecided	Agree
Baseline						
1988	30.4%	9.5%	60.1%	21.0%	7.0%	72.0%
1989	27.6%	18.8%	61.3%	26.3%	13.3%	60.4%
1990	24.0%	12.4%	63.6%	21.4%	13.2%	65.4%
1991	21.5%	15.4%	63.1%	25.7%	11.8%	62.5%

Table 3
Without performance appraisal it is more difficult to reward or discipline employees

	Experimental Site			**Control Sites**		
	Disagree	Undecided	Agree	Disagree	Undecided	Agree
Baseline						
1988	48.1%	12.8%	39.1%	36.1%	10.8%	53.2%
1989	44.0%	15.9%	40.1%	47.6%	8.9%	43.5%
1990	58.7%	17.5%	23.8%	55.7%	7.6%	36.7%
1991	65.2%	9.1%	25.8%	65.8%	9.6%	24.6%

Table 4
Elimination of Individual Performance Appraisals has...

	Disagree	Undecided	Agree
created less competition among employees.	26.5%	18.2%	55.3
created less supervisor/employee friction.	18.8%	17.5%	63.8%
helped my unit work as a team.	27.9%	28.5%	43.6%

To better understand differing attitudes of employees who did or did not prefer performance appraisals, respondents were split into two groups on the basis of their response to the item asking them if they wanted a performance appraisal (Table 4). Respondents who preferred appraisals and those who did not prefer appraisals were compared on issues dealing with feedback and supervisor-subordinate relationships. Tables 5 and 6 present the results for the baseline year and the fourth year. As can be seen, employees who received PAs but preferred not to receive them were less positive regarding feedback and supervisor-subordinate relationships than employees who wished to receive PAs. Thus those who were less positive about their work experiences preferred not to receive PAs. Employees who were exempt from receiving appraisals and who preferred not to receive them were initially less positive than were employees who preferred appraisals. Over time, however, they became significantly more positive. Overall, in comparison to the other three, this group (those who preferred not to receive PAs and did not) showed the most positive or least negative shift over time with respect to these questions.

Statistical analyses performed on the individual items in Tables 5 and 6 produced statistically significant Preference X Site X Time interactions for every item.

Table 5

	Experimental Site		**Control Sites**	
	Prefer PAs	Do Not Prefer PAs	Prefer PAs	Do Not Prefer PAs
High performers tend to stay with DS.				
1988 (Baseline)	2.60	2.54	2.94	2.69
1991	2.42	2.54	3.05	2.79
I am satisfied with my opportunities for advancement.				
1988 (Baseline)	2.04	1.88	2.74	2.10
1991	1.73	2.10	2.24	1.99
Promotions depend on how well a person performs his/her job.				
1988 (Baseline)	2.27	1.99	2.81	2.14
1991	1.72	1.94	2.52	2.07

Table 6

	Experimental Site		**Control Sites**	
	Prefer PAs	Do Not Prefer PAs	Prefer PAs	Do Not Prefer PAs
I usually know whether or not my work is satisfactory.				
1988 (Baseline)	3.90	3.80	4.04	3.74
1991	3.66	3.76	3.98	3.71
My Supervisor gives me information on how well I am performing.				
1988 (Baseline)	2.80	2.37	3.45	2.67
1991	2.56	2.68	3.23	2.73
My Supervisor keeps informed about the way subordinates think and feel.				
1988 (Baseline)	2.78	2.58	3.22	2.84
1991	2.71	2.83	3.16	2.84

Interviews conducted at the experimental site revealed that interviewees were nearly unanimous in favoring the elimination of PAs. The interviewees, however, cited some problems. Some stated that the issue of poor performance had not been adequately handled. Employees felt that the system in place - use peer pressure and public criticism of fellow employees' performance to resolve performance deficiencies - was not workable. Interviewees were also concerned that top performers were not receiving the attention and rewards they felt were necessary to foster their continued outstanding performance. Further, SPC and related systems and tools that were supposed to become the focal point for process improvement and monitoring had not materialized to a substantial degree and, in combination with no individual performance appraisals, led to insufficient performance feedback in the organization.

Conclusions

In summary, employees who did not receive PAs compared favorably with employees who did receive PAs in terms of performance feedback and supervisory-subordinate relationships. The elimination of individual performance appraisals was seen as beneficial in reducing competition and promoting team work. Over time, employees at all sites expressed an increasing desire *not* to receive performance appraisals. Those employees who did not want to receive performance appraisals and who did not, demonstrated improved attitudes toward their organization over time. To be more effective, the intervention to eliminate individual performance appraisals must be strengthened in two areas: 1) the establishment of process monitoring (e.g., SPC) and performance feedback systems to provide employees with information necessary to improve performance; and, 2) the development of approaches to reward top performers (e.g., non monetary rewards) and to deal with poor performers.

References

Alexander, III., E.R. & Wilkins, R.D. (1982). Performance Rating Validity: The relationship of objective and subjective measures of performance. Group & Organizational Studies, 7, 485-496.

Balzer, W.K., & Sulsky, L.M. (1990). Performance Appraisal Effectiveness. In Murphy, K.R., & Saal, F.E. (Eds). Psychology in Organizations: Integrating Science and Practice. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bradshaw, P. (1989). Performance Management: Small group discussion (report). 1989 Personnel Research Conference Proceedings. Chevy Chase, Md.: United States Office of Personnel Management.

Deming, W.E. (1986). Out of the Crisis. Cambridge: MIT Center for Advanced Engineering Study.

Ilgen, D.R., and Feldman, J.M. (1983). Performance appraisal: A process focus. Research in Organizational Behavior, 5, 141-197.

Longenecker, C.O., Sims, H.P., and Gioia, D.A. (1987). Behind the mask: The Politics of employee appraisal. The Academy of Management EXECUTIVE, 1, 183-193.

Scholtes, P. R. (1987). An elaboration on Deming's teachings on performance appraisal. ??? Madison, WI: Joiner Associates Inc.

Stephan, W.G., & Dorfman, P.W. (1989). Administrative and developmental functions in performance appraisals: Conflict or synergy? Basic and Applied Social Psychology, 10, 27-41.

Sulsky, L.M., & Balzer, W.K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 497-506.

U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

(703) 274-8293
DSN 284-8293

INTEGRATING DATA ACROSS MULTIPLE SOURCES: LESSONS LEARNED

Alma G. Steinberg, Beverly C. Harris, & Jacquelyn Scarville

Presented at the 34th Annual Military Testing Association Conference, 28 October 1992, San Diego, California.

INTEGRATING DATA ACROSS MULTIPLE SOURCES: LESSONS LEARNED[1]

Alma G. Steinberg, Beverly C. Harris, & Jacquelyn Scarville
U.S. Army Research Institute for
the Behavioral and Social Sciences

Background

Traditionally, survey researchers have identified a problem for research, reviewed relevant findings in the literature, and then proceeded to design surveys to address the issue, collect and analyze their data, and report findings from the data they collected. The research design seldom called for the incorporation of data that might be available from other past or current surveys (designed for different purposes by others inside or outside the organization) because each research effort was seen as a separate entity. Although findings from previous research might be discussed, data from other efforts were not typically integrated directly into or reanalyzed in conjunction with the current study.

However, with the continued use of large surveys, the general increased awareness of computer capability on the part of sponsors and budget watchers, and declining resources for conducting research, a new demand is being made. Researchers are being asked to integrate current and past results found by a number of other researchers (giving them full credit, of course) and to present them in conjunction with their own research findings. The reasoning is that if other surveys contain relevant questions to the issue at hand, there is no need to repeat the questions because one can integrate the results into the current effort. Also, if similar survey items have been used in past research, one should present current results in conjunction with past ones to provide a broader perspective by: (a) supplementing the current data, (b) giving the data context, and (c) showing trends over time.

Note that this type of data integration is different from the meta-analytic approach to data integration. It serves a totally different purpose and the method is different, as well. Meta-analysis uses quantitative methods to summarize different studies that address a common hypothesis. In a meta-analysis, each study as a whole is the unit of analysis and study results are stated in a common quantitative form such as effect size or statistical significance (Green & Hall, 1984). Examples of hypotheses that have been tested using meta-analysis are the: (a) statistical validity of Fiedler's contingency model of leadership effectiveness (Strube & Garcia, 1981), and

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army

(b) relationship between behavioral intentions and employee turnover (Steel & Ovalle, 1984). In contrast, in the multiple-source data integration approach addressed here, data within a variety of studies conducted for a variety of purposes are used to supplement the current data collection and are incorporated in the presentation of results. In order to facilitate this data integration, there may be a need for reanalyses of portions of the existing data. For example, a particular subpopulation may need to be analyzed separately or responses to an item may need to be recoded in order to integrate results. Also, data integration need not be limited to survey data. Non-survey information may be integrated as well. For instance, relevant demographic and test score or performance data may be available on personnel files which can be match.d by social security number to enhance the existing research data.

This paper presents the lessons learned from a successful U.S. Army research project which required the review and integration of data from multiple attitude surveys. One survey was designed specifically for the project and the remaining surveys had been designed and administered over the past several years by different groups of researchers, for a variety of purposes, on different topics, and for different sponsors. Because the data from the surveys that had already been administered were available on tapes and/or in written reports, it was assumed that findings on a number of different topic areas could be integrated into a coherent whole. Also, since the surveys were administered at different points in time, it was anticipated that the integration would result in valuable trend data not otherwise available. The paper also includes a discussion of the approaches to the data analysis, some of the many problems one may encounter in the process, and suggestions about the design of surveys to facilitate the integration of survey data in the future.

Some Lessons Learned

Obtain organizational support. Top-level organizational support, from the start, is crucial to the success of a multiple-source data integration project. It is needed to overcome any resistance to the concept and to support the sharing of data across internal organizational groups. This is a particularly sensitive issue if research results have not been published prior to the integration project. Top-level organizational support is also needed to authorize the appointment of knowledgeable points of contact (POCs) for each data source and POC time for further data analyses.

Identify the topics to be addressed and the audience. At the very beginning, it is important to clearly identify the topics and subtopics to be addressed, in writing. Without a written outline as a guide, it is easy to get off track or overwhelmed by the many topics included in multiple surveys and lose focus. It is also important to identify the audience for the final product to determine the level of information and analyses needed.

Identify suitable data sources. Suitable data sources are ones that address the topics of interest and have the relevant data available for retrieval and/or reanalyses. In determining data relevancy, pay special attention to the sampling (which impacts on the generalizability of the findings) and the format of the data. Thus, for example, if all you have is a written report, group means for each item, and respondents all grouped together, then you cannot reanalyze the results for subgroups (e.g., for males vs. females) and the results cannot be easily integrated with other data that is in the form of percents.

Obtain knowledgeable POCs for each data source. Because there is no one code or set of rules for establishing databases, defining and naming variables, and documenting data sets with comments, it is difficult for even the most experienced individual to interpret another researcher's database without a well-written codebook or extensive instructions. For this reason, and in order to ensure maximum accuracy and efficiency, it is best to work with a knowledgeable POC for each data source. This POC is responsible for providing information about the data source (e.g., procedures used, sampling information, existing documentation, information about the reliability and validity of the items), outputs of requested analyses, and assistance with interpretation of results--all in a timely fashion. Since the POCs are likely to have other ongoing projects which are probably of more immediate importance to them, priorities may have to be established by top-level management.

Assemble a data-integration team. Data integration as presented here is not a one-person job. In addition to POCs for each data source, there needs to be a core dedicated team that: (a) identifies the topics, sources, POCs, the analyses required, and the format for the results; (b) coordinates collection of information from the POCs and checks the accuracy of the analyses; (c) assembles and analyzes the tons of computer printouts; and (d) converts all this information into a coherent, organized whole. It is important that team members be familiar with the statistical packages used by the POCs so they can detect coding discrepancies or errors in the computer runs and correctly interpret the computer printouts.

Establish common format for data analyses. In order to use POC time efficiently and to obtain data in a form which allows them to be integrated, there is a need to establish written specifications of the variables or survey items, sample cuts for the analyses (e.g., rank, specialty), and the specific analyses required from each POC (e.g., frequencies, crosstabs, means, correlations). Written specifications are needed to clarify and document the request. They also facilitate discussion of idiosyncracies in the databases prior to analyses. Be as specific as possible (e.g., response categories to be collapsed, number of places for rounding, rules about inclusion of response categories such as "missing," "don't know," "does not apply") and include instructions for labeling variables on the computer output so they can be readily identified. The goal should be to ask for analyses in such a way as to eliminate further need for new runs or hand calculation.

Become thoroughly familiar with the method and instrument for each data source. Accurate interpretation of data provided by POCs is vitally important. Be warned that this is difficult to do when the data and research design are not yours. It is easy to make faulty assumptions, especially when data are not documented in a technical report or codebook. Be sure to review relevant documentation/publications for each source and have them handy when questions arise. If the documentation does not answer your questions or does not exist, go back to the POC. Assume nothing.

It is especially important when conducting the analyses to pay attention to the context of survey items and the sampling. Sometimes, especially when the item is ambiguous, the previous questions and the section headings and instructions affect the way respondents interpret the questions. Also, pay particular attention to skip patterns which eliminate certain respondents from answering a set of questions. They are easy to lose sight of when they go on for more than a page.

Establish a tracking system for each source, from the beginning. Using multiple sources requires tracking and filing systems for the source survey and variables, the data, and the related references. Also, each time the data are incorporated in a presentation, the source must be documented. Thus, if you present the results from five different sources in one graph, you need to be sure to document each source and be able to find the exact printout, publication, or other source to verify the variables, population, and analyses that led to the inclusion of the data points. You also may need to include documentation for your sources in the final product. This can be in the form of an appendix which provides the reader with the relevant background and references for each source used.

Design a compatible survey. If a new survey is to be developed for inclusion in a specific integration project, it should take into account the areas that are already covered by other databases. It is important, though, to repeat relevant background data (e.g., rank, type of unit assigned to) and to be sure that any questions included to obtain trend data have identical wording and response categories to the original source. Even small changes can have a major impact.

Develop systems to facilitate data integration in the future. The following are some organizational suggestions to facilitate conducting data integration projects in the future:

(a) Develop a set of standardized demographic items. It is difficult to integrate data for subgroups when the items used to define these subgroups are worded differently and/or have different response alternatives. Problems also arise when demographic items are not specific enough. For example, Army branch can be an important variable for defining subgroups, but merely asking for "branch" may not be sufficient. An officer's branch could be infantry, but he may not be assigned to an infantry unit. Similarly, asking Army soldiers if they are "Active" is confusing because they could be on "Active" duty but not be "Active" Component (i.e., they are Reserve Component).

(b) Develop a permanent file of surveys that are developed in the organization. At a minimum, include in the file a copy of the survey itself, a crosswalk for each item in the survey that indicates whether the item is brand new or where it appeared in other surveys, where the data tapes are, what documentation exists for the survey and the results, and the POCs for the project.

(c) Document surveys with a description of the purpose, methodology, samples, and findings. Also, document items that don't work (e.g., are found to be ambiguous, have incomplete response categories) so that these are not used again uncorrected.

(d) Consider a survey database with the above information and more to allow computerized searches of topics and items. In order to be maximally useful, it needs to be continually updated.

Conclusion

The kind of results that are possible to obtain from this approach depend on the number and type of relevant data sources (e.g., with respect to the overlap of topics and populations, the time periods covered, sample sizes), the availability of the data, and knowledgeable POCs for the sources. To obtain the best results, it is necessary to have (a) clear criteria for screening items, (b) a thorough understanding of the issues to be examined, the audience for the information, and the form of the final product (e.g., briefing, report, etc.) and (c) the management and administrative support to ensure cooperation across research and organizational lines. If done well, using multiple sources results in a broader view and deeper understanding of many issues. It also provides additional insights over and above the answers to the specific targeted topics. For example, in our particular integration project, we observed how consistent attitudes were across time and we were able to obtain insights on changes in meaning when small changes in item wording occurred. Finally, data integration projects can be facilitated by organizational policies which require keeping data tapes available for future analyses, fully documenting survey procedures and data tapes, and maintaining survey items (coded by topic) on a database.

References

Green B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. Annual Review of Psychology, 35, 37-53.

Steel, R. P., & Ovalle, N. K. (1984). A review and meta-analysis of research on the relationship between behavioral intentions and employee turnover. Journal of Applied Psychology, 69, 673-686.

Strube, M. J., and Garcia, J. E. (1981). A meta-analytic investigation of Fiedler's contingency model of leadership effectiveness. Psychological Bulletin, 90, 307-321.

A Study of The
Perceived Impact of Women in Combat

Louis M. Datko
Air Force Military Personnel Center
Directorate of Personnel Operations

Abstract

In response to congressional legislation repealing the restrictions of assigning women to duty in aircraft engaged in combat missions, the Air Force initiated a number of studies. This study looked at the assignment of women to combat units and the impact, if any, on unit performance, readiness, morale, and cohesion. Review of available literature raised additional issues suggesting a broader consideration of prevailing stereotypes and begged the "bottomline" question -- Should women be in combat? To this end, two parallel paper and pencil survey instruments were developed: one for service group members and one for commanders. The stratified random sample of officer and enlisted personnel, men and women, personnel deployed and not deployed also included those who would be directly and indirectly affected by a change in Air Force policy, e.g., combat aviators, female aviators, and non-aviators. Altogether, 1,700 commanders and 6,500 Air Force members received surveys. Response rates of 70% for commanders and 52% for Air Force members ensured a diversity of opinions were captured. Results obtained from this survey were presented to the Presidential Commission on the Assignment of Women in the Armed Forces on 13 July 1992.

Operation DESERT SHIELD/DESERT STORM brought the largest number of women closer to actual combat than at any other time in the history of the Air Force. Involvement of over 8,000 active duty women from the Air Force alone, rekindled the highly emotional debate over allowing women in combat. With only a few of the NATO countries (e.g., Canada, Netherlands) currently having all combat positions open to women, most studies have been relegated to public opinion. Matthews, Melton & Weaver (1988) and Matthews and Weaver (1992), using data from the 1982 General Social Survey conducted by the National Opinion Research Center (NORC), first identified race and gender differences and then later educational, financial standing and age differences in Americans' attitudes toward womens' roles in the military. Women were more supportive of non-traditional military roles. Blacks, regardless of gender, were least supportive of women in the military. Level of education and financial standing were directly related to whether someone would support women in combat roles while age was inversely related. Not surprisingly, the roles Americans supported least were hand-to-hand combat soldier and crew member on a combat ship. A nurse in a combat zone and typist in the Pentagon were roles most supported. A majority supported the role of fighter pilot for women, a role ranked midway between the two extremes.

Evidence that women are more than willing to take on such non-traditional roles can be found with Major Peterson's (1988) survey of all female Air Force pilots. With a response rate of 70% from 322 women, 100% agreed they were capable of flying combat aircraft. Ninety-three percent agreed that Air Force combat aircraft and missions should be open to women aviators; 81% personally wanted to fly combat aircraft.

Since the 1982 study, public opinion appears to have become more receptive to women in combat roles. A recent study conducted by the Roper Organization polled 1,700 adults via telephone. The results showed "clear majorities favor assignments of women to combat aircraft, combat ships, and most ground combat" (Air Force Times, September 21, 1992). Results from a recent Roper paper and pencil survey, conducted for the Presidential Commission, based on returns from 4,442 Armed Service members, reported similar views of women military roles as the 1982 NORC poll (Air Force Times, October 12, 1992, pg. 6). Sixty-two percent of the service members agreed women should be pilots or other crew members on a bomber or a fighter aircraft fighting the enemy from the air. However, women as marines landing on shore to attack or as infantry soldiers fighting hand-to-hand received only 27% and 25% respectively of service members' support.

Perhaps, as stated so succinctly by David Hackworth (Washington Post, October 4, 1991, pg. 25), the proper research question for the upcoming commission to answer should be "What do the soldiers who will be most affected by the proposed changes think?"

Method

With the 1991 repeal of combat exclusions laws and the establishment of a Presidential Commission on the Assignment of Women in the Armed Services, the Air Force initiated several studies to assist in defining a position on the issue of women in combat. This study examined the impact of assigning women to combat positions on unit morale, cohesion, readiness and performance. A sampling strategy was developed to capture the attitudes and opinions from all Air Force members with specific emphasis on those who would be directly affected by a change in combat exclusion policies. Air Force personnel were divided into two groups: commanders, those individuals responsible for implementing potential policy changes, and Air Force service members. Substrata for each of these populations were chosen based on the following categories: officer/enlisted, gender, aviator/support personnel and participation in DESERT SHIELD/DESERT STORM. For each stratification category, members were randomly selected from the Air Force-wide population to ensure representation at a 95% confidence level with a 5% margin of error. The resulting sample sizes were 1,756 for commanders and 6,421 for Air Force service members.

Results/Discussion

By the end of April 1992, survey returns had been received from 3,254 Air Force members: 1,678 men (639 enlisted and 1,039 officers of which 603 were aviators) and 1576 women (624 enlisted and 952 officers of which 121 were aviators). The returns for each of the categories were evenly split between deployed and non-deployed personnel. Because sampling was representational, all data were weighted to population proportions.

One quarter of Air Force commanders and all the wing commanders were sampled. Responses were received from 1,229 commanders (officers), 1,079 men and 143 women. There were 465 aviators of which 153 were currently commanding flying operations units. Senior Air Force leadership was well represented by 177 commanders of units above the squadron level, 107 of whom were wing commanders.

Perceptions of Women in Air Combat Roles

Overall, a majority of Air Force members (56%) and commanders (62%) agreed female aircrew members should be assigned to combat aircraft. Non-aviators who would not be directly affected positively influenced the overall results with 60% of the men and 75% of the women agreeing to a change in policy. When the responses are examined by combat aviators (men only) and female aviators, substantial differences appear. Of the men who fly combat aircraft (e.g., fighter, bomber, special operations), 27% agreed and 66% disagreed with assigning female crew members to combat aircraft. Of female aviators, 95% agreed and only 2% disagreed with changing the exclusion policy. Of all personnel groups surveyed, combat aviators and female aviators were the most different in their responses on all women in combat issues. Of some interest, aviators of non-combat aircraft, which are open to female aviators, were more supportive (47% agree, 41% disagree) than the combat aviators who had never flown with women.

Commanders were generally more supportive of assigning women in combat roles than were the service members; however, differences between combat and non-combat units remained. Commanders of combat aircraft units were less supportive of opening combat exclusions for air combat (41% agree, 46% disagree) than non-combat aircraft units (69% agree, 20% disagree).

Over half of the responses were accompanied with written comments. Two comments that reflect the opposing positions of male and female aviators are:

"Take time to incorporate [women] into Services and educate those who have not yet had the chance to work with competent women." Captain, C-9 Pilot, Deployed, Female

"I believe women could do their job in combat. It's their effect on everyone else that should exclude them from combat duty in all-male units." Captain, F-16 Pilot, Deployed, Male

Bonding/Cohesion

One of the issues in the charter was to examine the impact of women on unit cohesion. From the literature review, the importance of bonding among male pilots was put forth as a crucial argument for excluding women. The presence of women would undermine that trust and implicit understanding which had been developed through male camaraderie, resulting in decreased performance in battle. Within the survey, bonding was defined as "The psychological process that occurs within a group that has shared experiences that set it apart from other groups. Bonding allows an individual to feel comfortable within a group and to set aside his or her personal needs for the good of the entire group."

When asked what the impact of the presence of women would be on work group bonding, 24% of all respondents reported a negative impact; 26% a positive impact and 50% said no impact. However, most combat aviators (47%) reported a negative impact; 11% said there would be a positive impact, while 42% said there would be no impact. Female aviators were the mirror image: 6% reporting a negative impact, 40% a positive impact and 54% saying no impact.

The following comments from fighter pilots best reflect their attitude on bonding:

"It is imperative that fighter pilots feel at ease and feel a bond with their squadron mates, a feeling that will not be present if women are allowed to tear apart cur squadron." Captain, F-15 Pilot, Deployed, Male

"Under no circumstances should women fly fighter aircraft in combat or peacetime. The effect of a cultural change of this nature would devastate the traditions, esprit, and ultimately the combat capability of the fighter force." Lieutenant Colonel, F-16 Squadron Commander, Deployed, Male

For cohesion, analysis was focused on those Air Force members who could best report what effect women might have: those deployed to the Persian Gulf. Overall, 24% of deployed Air Force members stated that the effect of the presence of women in the unit was negative, about a quarter said positive but half reported cohesion was not affected. Combat aviators reported a less positive effect (10%) but the majority (54%) reported no effect with 36% reporting a negative effect. Female aviators were most positive with 67% stating women would have a positive effect on cohesion, 31% declaring no effect. Only two percent cited a negative effect.

Commanders who had been deployed to the Persian Gulf, regardless of what type of flying unit they were in charge of, overwhelmingly (70%) felt there was no effect on cohesion from the presence of women in their units.

Differences in how the male aviators perceive cohesion versus all other service members is reflected in the following comments.

"I believe women in a fighter unit would have strong detrimental effects on the combat effectiveness of that unit. More so than any military group, fighter squadrons have historically formed a "fraternal" bonding that goes to the heart and soul of that unit. The difference between a "good" or "bad" fighter squadron can most likely be determined by this one point." Captain, F-15 Pilot, Deployed, Male

"I was assigned to two different units in Saudi Arabia...I found that women and men worked together cohesively. Everyone for the most part did his or her job. I think having women in the unit helped reduce stress and made the situation seem more reasonable." Sergeant, Secure Communications Specialist, Deployed, Female

Morale

When asked about the effect of the presence of women on unit morale, we saw slightly more positive responses from deployed personnel than with the issue of cohesion. Overall, 31% of Air Force members cited a positive effect on morale, 44% said there was no effect and 25% said women had a negative effect on morale. Again combat aviators reported less positive effects (17%), with 45% claiming no effect and 38% stating a negative effect. Female aviators were consistent in their opposing view, 63% relating a positive effect, 33% no effect and 4% negative effect.

Commanders were also more positive, especially the ones responsible for non-combat units. Only 14% of the combat unit commanders cited a positive effect from the presence of women, although 66% felt there had been no effect on morale and 20% said there had been a negative effect. Thirty percent of commanders of non-combat flying units reported a positive effect, 59% no effect and only 11% cited a negative effect. Non-flying unit commanders were most positive (39%) while 44% reported no effect and 17% stated there had been a negative effect.

Impact of Women on Air Combat Performance

As we have just shown, the majority of Air Force members and their commanders felt the presence of women would not negatively affect cohesion and morale of the unit. However, when it's time to face the enemy, what did our aviators think about how women would effect performance of flight crews during combat? This question was worded so that the respondent was to consider both single crew aircraft in formation as well as multi-member aircraft. Overall, the majority of deployed members (55%) felt that performance would not be affected; 35% felt performance would be diminished and 10% felt performance would be enhanced.

However, male aviators expressed less confidence in the performance of female aviators. Thirty percent of the combat aviators felt performance levels would stay the same and only 3% reported an enhancement. Sixty-seven percent felt combat performance would be diminished. Female aviators had considerably more confidence in their abilities, with 31% reporting performance would be enhanced and 64% stating performance would stay the same. Five percent reported diminished performance.

The majority of commanders of combat units (50%) felt performance of mixed gender aircrews would remain the same, while 47% felt that performance would be diminished. For all other unit commanders, 70% felt that performance would stay the same; 26% said performance would be diminished.

Two comments that capture the dichotomy between the commanders and the combat pilots follow:

"With our smaller, more global force, the luxury of having any noncombatant airmen is gone. Everyone must be prepared to arm and fight responsibly. It's the humane and right way to run a global Air Force." Lieutenant Colonel, Operations Management Staff Officer, Deployed, Female

"I think it is a mistake to assign women to combat positions. If I had to be under enemy fire again, I would choose to fly with a man. If ordered, I would salute smartly and fly with a woman, but I wouldn't trust her to carry her load and I would expect her to fold when the bullets and missiles started flying." Captain, B-52 Navigator, Deployed, Male

Readiness

Throughout the literature, the issue of how pregnancy impacts readiness is a recurring theme. We approached the topic of readiness through three items: ability to respond immediately for deployment, reasons for not deploying, reasons for returning early from deployment. Response choices were presented in a "laundry list" fashion. The top three reasons were identified based upon the largest percentage of respondents. For inability to deploy, both men and women cited having to make arrangements for dependent care, off-duty education and transportation problems as the top reasons. Pregnancy or spouses' pregnancy was cited by only 10% of the women and 5% of the men as the primary reason for not responding to a deployment.

The top reasons why Air Force members did not actually deploy were their skill was not needed or others vclunteered. On pregnancy, 3% of the women reported they did not deploy and 1% of the men. Top reasons for returning early from deployment from the gulf were skills no longer needed and family health (excluding pregnancy). Again, a small percent of both men and women were returned early; 3% of the women returned and 4% of the men returned because of pregnancy or spouse's pregnancy. Even though this data was captured through self-reported, these results clearly show that pregnancy does not have a major impact on readiness and that the readiness for both men and women is about equal.

Perceptions on Potential Mixed Aircrew Missions

In order to gain a perspective on Air Force service members' willingness to fight side-by-side, we asked three questions. For male aircrew members flying combat support missions with women, nine out of ten would do so again. Eight of ten female aircrew members would willingly accept a chance to fly combat missions. Six of ten male aircrew members who had never flown with women would willingly fly combat missions with women. Interestingly enough, if the male aircrew members are fighter pilots, only four out of ten would willingly fly/fight on combat missions with women.

Here we see an underlying thread running through the service members' responses. If men are currently working or have worked with women in their units, they tend to respond more positively to accepting women in combat roles.

Summary

When we asked airmen what they think about women in combat we found a large disparity in opinions between the all-male combat aviators and the female aviators across the issues of cohesion/bonding, morale and performance. Combat aviators, the group most affected by a change in policy, were unmistakably opposed to opening combat cockpits to women. In sharp contrast to the combat aviators, female aviators presented a cohesive group ready to accept the challenge of combat roles. Overall, excluding male combat aviators, Air Force personnel appear receptive to the concept of women in air combat. In fact, the presence of women is perceived by the majority to have no effect (either positive or negative) on cohesion/bonding, morale, performance or readiness.

References

Air Force Times (September 21, 1992). Poll finds more Americans support women in combat jobs. 53rd YEAR, No. 7, p. 6

Air Force Times (October 12, 1992). Women at War. 53rd YEAR, No. 10, p. 6

Matthews, M.D., Melton, E.C., & Weaver, C. N. (1988). Attitudes towards women's roles in the military as a function of gender and race. Proceedings of Psychology in the Department of Defense, Eleventh Symposium. Colorado Springs, Colorado: United States Air Force Academy.

Matthews, M.D., & Weaver, C. N. (1992). Demographic and Attitudinal Correlates of Women's role in the Military. Proceedings of Psychology in the Department of Defense, Twelfth Symposium. Colorado Springs, Colorado: United States Air Force Academy.

Peterson, T. M. (1988). USAF Women Pilots - The Combat Issue. (Air Combat & Staff College Report #88-2110). Maxwell AFB, Alabama: Air University.

Washington Post (October 4, 1991). Women Warriors. p. 25

Evaluating Changes in Test Content: The ASVAB Review Technical Committee

Bruce Bloxom and Lauress Wise
Defense Manpower Data Center

The ASVAB Review Technical Committee

Two problems have to be addressed in considering ways of improving the Armed Services Vocational Aptitude Battery (ASVAB). One problem is the very large number of ideas about ways in which the battery could potentially be improved. The other problem is the very large number of ways in which an idea must be evaluated to provide convincing evidence that implementing it would be an improvement. The first problem is being addressed, in part, by using policy input to constrain the number of ideas being seriously considered. Major constraining factors include projections of operational and financial perturbations which would be created by implementing each idea, as well as immediate and long-term perceived benefits which would result from the implementation. This is the subject of another paper in this session.

The second problem -- the very large number of ways in which each idea must be evaluated -- has been handled in typical governmental fashion by referring it to a committee. This Committee -- the ASVAB Review Technical (ART) Committee -- has held four meetings to define the range of technical information needed to support implementing revisions in the ASVAB and to set priorities on what information is required at various stages in the process of deciding about revisions.[1] Table 1 shows the topical organization of the information under discussion for each of three types of changes in the ASVAB. Note that the same broad categories are used to organize the collection of information about each type of potential change. However, the particulars and priorities for information gathering differ across the three types.

Table 2 shows an example of the specific kinds of information being sought regarding one type of change in the ASVAB -- the use of new and/or alternative subtests. Note that underlined topics indicate types of information needed prior to recommending the inclusion of a new test in the battery. Other topics indicate types of information needed to provide an implementable form of the new test in the context of the entire operational battery. Where the first kind of information is needed, proposals have been written to indicate how the information can be provided. Summaries of these proposals -- as of June 1992 -- are available in ASVAB Review Technical Information Summary, a living document available on request from Defense Manpower Data Center[2].

[1] Defining the range of technical information relied heavily on previous work by Malcolm Ree, when he was chairing the technical subgroup of the Joint-Services Selection and Classification Working Group, and by the Technical Advisory Selection Panel, consisting of technical representatives of Defense Manpower Data Center and the Military Services.

[2] Address requests to Dr. Bruce Bloxom, Defense Manpower Data Center, 99 Pacific Street, Suite 155a, Monterey, CA 93940.

As discussion of the alternative changes to the ASVAB have progressed, a clearer separation between operational and technical issues has emerged. Changes in the mode or location of testing have significant impact on testing operations. Cost-benefit analyses to evaluate these changes are described in another paper in this session (McBride & Hogan, 1992). Changes in the content and length of the battery involve very significant technical issues and attention to these issues has become a major focus of the ART Committee. The remainder of this paper will describe technical issues in identifying and evaluating changes to the contents of the ASVAB and will outline proposed approaches.

Changes in ASVAB Content and Length

Potential Changes. Three types of changes are being considered including: (1) lengthening the entire battery -- adding new spatial, psychomotor, or memory tests that are included in a current validation effort; (2) holding test administration time constant by replacing some existing tests with new tests; and (3) shortening the battery by combining or deleting tests. The impact of these changes depends on whether a CAT version of the ASVAB will be used or whether we will continue with paper-and-pencil administration procedures. With a CAT version, new tests can be added while total administration time is still reduced (to about two hours) in comparison with the current three hour battery. With computer administration, potential changes in battery length will add or subtract 15 to 20 minutes and will not have a major operational impact. In the paper-and-pencil mode, the first option of adding tests could add as much as 30 minutes to testing time while the third option of combining or deleting tests might reduce testing time by as much as an hour.

General Evaluation Criteria. With ten subtests in the current ASVAB and another nine subtests under evaluation, there are more than half a million possible combinations of tests that might be considered as candidates for the new ASVAB. Several general criteria, in addition to testing time, are being used in evaluating all of these different combinations, including: (1) average increments in predictive validity for all jobs; (2) improvement in classification efficiency (improved predictive validity for some jobs but not others); (3) reduction in adverse impact for minorities and women; and (4) resistance to practice and coaching effects that might degrade test validity and/or distort the score scale.

The general strategy that we are following in evaluating test content changes is to sort all possible combinations of tests by total administration time in paper-and-pencil mode focussing on the two hour, three hour, and three-and-a-half hour levels. For each of the three administration time levels, we will find the combination of tests that optimizes the above evaluation criteria. We will then compare the benefits derived from the longer batteries with the costs of the additional testing time.

Issues. Several issues must be addressed in carrying out the proposed evaluation of alternative test battery content. First and foremost is limitations on the jobs and samples for which test scores and outcomes are available. Three data sets are being analyzed: (1) the ECAT Validity Study, (2) the Navy Validity Study, and (3) the Army's Project A. Table 3 shows the number of different jobs, range of sample sizes, and type of predictor and criterion measures associated with each of these data sets. In the aggregate, the total number of jobs is limited compared to the hundreds of jobs for which selection and classification decisions must be made. In addition, with limited sample sizes and fallible criteria it is not possible to identify exactly the optimal weighting of subtests for each job nor estimate precisely the absolute and differential validity of a given battery, even for these jobs for which some data are available.

Another issue that limits the evaluation of alternative battery concerns constraints on the ways in which test scores are used operationally. We can model or simulate gains in expected performance associated with more accurate predictor information, but operational decisions tend to be suboptimal because of (1) the need for sequential decision-making as opposed to selection and assignment from an annual pool, (2) the fact that the different jobs are in comp϶tition with each other for the same pool of applicants, and (3) the need to accommodate applicant choices and preferences.

A third issue that must be addressed in evaluating alternative test batteries is the difficulty of quantifying benefits from improved decision-making (or cost of reduction in job match quality). There is some hope of achieving a consensus regarding the cost of additional testing time, but the savings from improved prediction of training success and job performance is a much tougher issue. Can we really trade training costs or manpower levels for testing time?

Preliminary results. Some preliminary analyses were reported by the Services at the APA Convention in August, 1992. Analyses of the Marine Corps data on helicopter and ground mechanics indicated that one of the spatial tests, Assembling Objects, did significantly improve the prediction of job performance criteria; it also had minimal practice effects (Carey, 1992). Spatial tests were also found to add significantly to predictive validity in analyses of Navy validity data (Wolfe & Alderton, 1992; Alderton & Larson, 1992). The Army reported results from a special study of coaching and practice effects on three of the spatial tests and concluded one of the three tests (not Assembling Objects) was significantly coachable (Busciglio & Palmer, 1992).

The Army is supporting additional analyses of the data collected from Project A, including scores on seven of the new predictors for roughly 40,000 new recruits (Peterson, Oppler, & Rosse, 1992). Preliminary findings indicate the importance of a spatial measure in predicting hands-on job performance test scores across a wide range of jobs and the utility of psychomotor measures in increasing classification efficiency (by allowing for capitalization on differences in psychomotor skill in assigning applicants at a given level of cognitive ability). Their results also indicate that gains associated with longer testing times are quite minimal and may be totally negligible without better data for identifying optimal uses of new test information.

The Navy has collected both complete predictor data and extensive training performance measures in the Enhanced Computer Administered Test (ECAT) Validity Study. Preliminary analyses of these data also show a number of areas where the use of new tests can significantly enhance the prediction of training outcomes. Results from this effort are expected in the next few months.

Remaining steps. Several steps remain in completing the evaluation of changes to the content of the ASVAB battery. First, results must be combined across the different studies that have been completed or are now in progress. The Navy is designing a meta-analytic approach to assessing incremental validity associated with new tests in the ECAT Validity databa͵e and in the other data sets as well. Some combination across criteria, as well as across data sets, will be required.

To date, relatively little attention has been paid to adverse impact. Some of the new tests do lead to reductions in adverse impact for minorities. Women do score somewhat lower on the spatial tests in comparison to men, but the differences are generally smaller than gender differences on technical subtests in the current battery. It is difficult to evaluate adverse impact without careful consideration of how each new test would be used. More attention to this issue will be required before final recommendations can be made.

Finally, a dollar value must be placed on improvements in selection and classification decisions. Several approaches are under review including: (1) a traditional Cronbach/Gleser utility assessment where improvements in job performance are estimated and valued, (2) estimation of cost savings associated with reductions in training failures (assuming a fixed standard for success), and (3) estimated savings in recruiting costs associated with making better use of more readily available applicants.

In the end, the data and analyses supporting recommendations about the length and contents of the battery will necessarily be incomplete. We are hopeful that the results of our efforts will be clear and compelling. Even if they are not, we will have succeeded in mounting a coordinated effort to do the best evaluation possible under time and resource constraints, and we should set important precedents for evaluation of future changes to the battery.

References

Alderton, D.L. & Larson, G.E. (1992). ECAT Battery: Descriptions, constructs, and factor structure. Paper presented at the centenniel convention of the American Psychological Association, Washington, DC.

Busciglio, H.H. & Palmer, D.R. (1992). An empirical assessment of coaching and practice effects on three Army tests of spatial aptitude. Paper presented at the centennial convention of the American Psychological Association, Washington, DC.

Carey, N.B. (1992). New predictors of mechanic's job performance: Marine Corps findings. Paper presented at the centennial convention of the American Psychological Association, Washington, DC.

Cronbach, L.J. & Gleser, G.C. (1957). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Curran L.A. & McBride J. R. (1992). The development of alternative operational concepts. Paper presented at the annual convention of the Military Testing Association, San Diego, CA.

McBride, J.R. & Hogan, P. (1992). Evaluation of alternative operational concepts. Paper presented at the annual convention of the Military Testing Association, San Diego, CA.

Peterson, N.G., Oppler, S.H., & Rosse, R. (1992). Optimal battery analyses of Project A/Career Force data. Presentation to the ASVAB Review Technical Committee, San Antonio, TX.

Wolfe, J.H. & Alderton, D.L. (1992). Navy technical school validity of the new predictor battery. Paper presented at the centennial convention of the American Psychological Association, Washington, DC.

Table 1
Types of Information Required for Technical Review of ASVAB

I. Information required to evaluate new subtests
   a. Studies of quality of decisions in selection & classification using ASVAB with new tests
   b. Studies to define operational procedures for use of new tests
   c. Studies to define procedures for developing new forms of new tests
   d. Studies of costs of implementing, using and maintaining new tests

II. Information required to evaluate computer-based testing
   a. Studies of quality of decisions in selection and classification using ASVAB with computer-based tests
   b. Studies to define operational procedures for use of computer-based tests
   c. Studies to define procedures for developing new forms of computer-based tests
   d. Studies of costs of implementing, using and maintaining computer-based tests

III. Information required to evaluate multi-stage testing
   a. Studies of quality of decisions in selection & classification using ASVAB with multi-stage testing
   b. Studies to define operational procedures for use of multi-stage testing
   c. Studies to define procedures for developing new forms in multi-stage testing
   d. Studies of costs of implementing, using and maintaining multi-stage testing

Table 2
Studies of Quality of Decisions in Selection and Classification Using ASVAB with New Tests[3]

1. Constructs being assessed
   - definition of construct and its facets
   - proneness to sensitiviey and DIF problems

2. Time required for sufficient measurement precision
   - reliability and/or score-conditional precision estimate as a function of test length
   - precision estimate(s) for composite(s) including new test

3. Construct and predictive validity
   - factorial validity, DIF, adverse impact, unique variance, predictive validity, incremental validity, substitutional validity, predictive fairness
   - how to synthesize results of validity studies

4. Norm requirements and availability
   - reference population requirements
   - definition of score scale for continuous scored test
   - procedures for obtaining/approximating norms

5. Accuracy and stability of score scale in operational use
   - retest effect size as function of time, practice, fatigue, and coaching
   - development of instructions and practice items to stabilize score scale

[3] Information needed by March of 1993 is indicated by underlining.

Table 3
Summary of Available Data

Data Set	Predictors	Criteria	# of Job Samples	Sample Ns	
				Min	Max
ECAT Validitation Study	**ASVAB Tests** -*Verbal* --Word Knowledge(WK) --Paragraph Comp.(PC) -*Quantitative* --Math Knowledge(MK) --Arith. Reasoning(AR) -*Science & Technical* --General Science(GS) --Mechanical Comp.(MC) --Auto&Shop Inf.(AS) --Electronics Inf.(EI) -*Clerical Skills* --Coding Speed(CS) --Numerical Operations(NO) **New Tests** -*Spatial Ability* --Integrating Details (ID)* --Assembling Objects(AO) --Spatial Orientation(SO) -*Non-Verbal Reasoning* --Mental Counters(ME)* --Sequential Memory(SM)* --Figural Reasoning(FR) -*Psychomotor Skill* --One-Hand Tracking(T1)** --Two-Hand Tracking(T2)** -*Perceptual Speed* --Target Identification(TI)**	**Training Outcomes** -Final School Grades -Performance Tests -Written Tests	20	111	1508
Army's Project A	**ASVAB Tests** (same as above) **New Tests** -- AO, SO, FR, TI, T1, T2 (as above) --Short-Term Memory(ST)	**Training Outcomes** -Written Tests -Perf. Ratings **Job Performance** -Hands-on Tests -Written Tests -Perf. Ratings	9	206	696
Navy Validity Study	**ASVAB Tests** (same as above) **New Tests** -- ID, ME,SM, FR (as above) --Space Perception(SP) --Perceptual Speed(PS)	**Training Outcomes** -Final School Grades -Performance Tests	9	122	1169

*-predictors requiring computer administration; **-predictors requiring special response pedestal

Status of Computer Administration of the ASVAB*

David L. Alderton
Navy Personnel Research and Development Center
San Diego, CA 92152-6800

Paper presented at the
Annual meeting of the Military Testing Association

27-29 October 1992
San Diego, CA

Introduction and Purpose

The Armed Services Vocational Aptitude Battery, or ASVAB, is the sole selection and classification test battery for military enlisted personnel. The ASVAB is a traditional multiple-aptitude, paper-and-pencil test battery and has been used by all Services since 1976. Today, however, we are on the verge of major and perhaps fundamental changes in the ASVAB. There are two lines of military aptitude research that will be given particular consideration in the process of revamping the ASVAB. One is the Computerized Adaptive Testing version of the ASVAB, or CAT-ASVAB. The second is the Enhanced Computer Administered Test Battery, or ECAT. In June 1992 these two independent yet symbiotic strains of research were united in an Operational Test and Evaluation (OT&E). The purpose of this paper is to provide a thumbnail sketch of the history of research that produced CAT-ASVAB, ECAT, and their unification, and thereby foretell some of the changes we could soon see in the ASVAB.

History

In 1973, following the end of the Vietnam War the draft was terminated and the Services reverted to an All-Volunteer Force and adopted a uniform selection and classification test battery, the ASVAB. Over the next decade the quality of service applicants declined severely (Eitelberg, Laurence, Waters, & Perelman, 1984), as did the quantity (Ramsberoer & Means, 1987). Further complicating the grim manpower outlook was the military's tremendous technological modernization, a development placing even greater intellectual demands on the average enlistee. Yet even as the military's need for talented people grew more acute, our tools for identifying talent, i.e., aptitude tests, came under growing attack from critics. In 1978, the *Uniform Guidelines on Employee Selection Procedures* (EEOC, 1978) were adopted by the Federal Government — an action that led,

* The opinions expressed or implied in this paper are the author's, are not official and do not necessarily reflect the views of the Navy Department or the US Government.

Portions of this paper were presented at the August 1992 annual meeting of the American Psychological Association and the May 1992 Army Research Institute's Conference on Selection and Classification.

in part, to the 1981 Congressional directive requiring the Services to better document the relationship between education, test scores, and job performance. Collectively, these forces fostered a modern, resurgent interest in military manpower testing. There were two main *ASVAB improvement* themes, as mentioned earlier. The first was an effort to reformat the ASVAB, while the second theme was to add new aptitude constructs and thus broaden the ASVAB.

CAT-ASVAB

The ASVAB reformatting project was CAT-ASVAB which formally began in 1979 when the Navy Personnel Research and Development Center was designated as Lead Laboratory for Computerized Testing. The program was initiated because of the need to improve military selection and classification. The program's approach was shaped by two technological advances, the availability of powerful microcomputers and developments in statistical theories of test scores. Blending these two advances produced the concept for CAT-ASVAB which was to develop a computer administered version of the ASVAB redesigned using modern psychometric theory, referred to as Item Response Theory. By combining Item Response Theory with computer administration, the ASVAB's power tests could be made adaptive in that test items could be specifically selected for an examinee based on his or her previous responses. Adaptive test administration could reduce test length by as much as one-half while improving reliability, particularly in the extremes of score distributions where applicant discrimination was poor. Moreover, administratively, computer-based testing could among other things, improve test security, reduce scoring errors, and provide immediate feedback to examinees and their recruiters.

Although adaptive tests had been theoretical possibility for a number of years, no one had ever successfully produced an adaptive, multiple aptitude test battery intended for large-scale use. As such, hundreds of difficult, pragmatic, and unanticipated problems had to be solved in the development of a working CAT-ASVAB system. Problems such as how the system should be organized, what fail-safe and failure-recovery procedures should be included, what hardware and networking system should be chosen, how items should be protected, and how the frequency of item use should be controlled. These and many more problems were solved to produce a functional delivery system.

More importantly though, is that there were several critical research questions that had to be answered as a prelude to operational use. One issue was a concern that the medium of administration alone, i.e., paper-and-pencil versus computer, would produce important differences in test items and scores. A large scale 1987 study explicitly addressed this concern and found it to be generally unwarranted (Hetter, 1992). A second concern was that test score intercorrelations, within and across mediums of administration, would differ markedly. This issue was addressed in an important 1988 study (Moreno & Segall, 1992) and the results clearly demonstrated that there were no substantial differences among the intercorrelation matrices of ASVAB tests, either within or across test mediums.

Having solved the practical testing issues and assuaged the concerns of many psychometricans and policy makers, a final step was required before CAT-ASVAB could actually be used. Specifically, conversion or equating tables were required that would allow CAT-ASVAB and paper-and-pencil ASVAB scores to be used interchangeably. In 1988, an elaborately designed study was conducted, requiring data collection in several sites across the country. The data were used to develop preliminary tables equating CAT and ASVAB test scores. However, since the original equating tables were based on individuals who were required to take several non-operational versions of the ASVAB, the validity of the equating tables had to be verified in one final study. This study was initiated, and as a result, in September 1990, CAT-ASVAB was operationally used for the first time. CAT-ASVAB has become the first operational, computer administered, adaptive selection and classification battery in use.

ECAT

While efforts to reformat the ASVAB were focused and localized, attempts to broaden the abiiities measured by the ASVAB were dispersed, with each of the Services conducting research. In 1981 the Army's Project A was commissioned with a very broad charter and sweeping objectives. In the same time frame, the Air Force's Learning Abilities Measurement Project (LAMP) began with the goal of developing new predictors of learning. Smaller testing programs were also started in the Navy.

Several common contextual stimulants independently shaped the Services' attempts to broaden the ASVAB, producing similarities in their research programs. For example, the availability of inexpensive microcomputers and the momentum behind the computerization of the ASVAB, lead the Services to develop new tests that were primarily computer-based. Moreover, the cognitive *zeitgeist* in American psychology during the mid-70s and 80s strongly influenced the programs. For example, all of the Services investigated the use of reaction time measures; the Air Forces' program was built around a cognitive model; and, the Navy's research was driven by cognitive theories of aptitude, working memory, and mental imagery.

Though there were commonalities across Services, there was little collaboration. However, as work on CAT-ASVAB progressed and a national renorming of ASVAB was anticipated, additional impetus was provided to the possibility of adding new aptitude dimensions to the ASVAB. In December of 1988, the Office of the Assistant Secretary of Defense (Force Management and Personnel) (OASD/FM&P) redirected the CAT-ASVAB program to "include a Joint-Service validation of the Services' new computerized cognitive and psychomotor tests" (Sellman, 14 Dec 1988). This directive was in recognition of two facts. First, an early cost-benefit study suggested that fielding a computer version of the ASVAB may not be cost effective relative to the paper-and-pencil version (CACI, 1988). Second, other research indicated that broadening the ASVAB's ability measures could result in large improvements in productivity per accession (Schmidt, Hunter, & Dunn, 1987). Combining these findings, a new computer-based ASVAB augmented with new ability measures could produce a better and cost effective selection and classification system. Just as importantly though, the directive was a realization that if decisions were to be made about the usefulness of new ability measures, they needed to be evaluated in a single study using the most probable delivery system for a computerized ASVAB, the CAT-ASVAB system. This formally integrated the two research strains to improve the ASVAB.

In response to OASD's redirection, the Technical Advisory Selection Panel (TASP) was established in January of 1989 to evaluate and select tests for the Joint-Service validation battery. The Panel's charter was to select the best tests in terms of their psychometric properties and theoretical justifications within the constraint that the battery could not exceed three hours. Across Services, hundreds of pages of documentation were submitted supporting the use of dozens of new aptitude measures. Nine tests were chosen and combined into a battery named ECAT. A research design was approved, the necessary software and hardware were developed and/or acquired, and in February 1990 the study began. Twenty-one months and 16,000 examinees later, testing ended. The sample included enlisted personnel in the Army, Navy, Air Force, and Marine Corps, representing 19 Military Occupational Specialties (MOS). Criteria data collection has been completed and analyses will be complete in the winter; with a full evaiuation of the study taking place in early 1993. This undoubtedly represents the largest single validation of a computerized test battery ever undertaken.

The Face of a New ASVAB

Thus far I've discussed the process of broadening the ASVAB but have said nothing of the content of either test battery. Before moving on, a brief description of the ASVAB is necessary. Table 1 shows the eight power and two speeded tests that comprise the ASVAB. The paper-and-pencil test requires three hours for administration. These tests define four ability or achievement factors: Verbal Ability, Mathematical Ability, Technical Knowledge, and Perceptual Speed. In selecting the ECAT tests, the assumption was made that the scope of human intellectual and non-intellectual skills was much greater than that represented by the ASVAB, and that capturing this breadth held the greatest promise for improving personnel selection and/or classification. Table 2 shows the nine tests that comprise the ECAT battery including a brief description of each of the tests. The battery requires a maximum of three hours with most individuals finishing in just under two hours. The specifics of these tests are unimportant for this discussion, suffice it to say that the battery measures: Spatial Ability, Non-Verbal Reasoning, Perceptual Speed, and Psychomotor Skill. The next Table combines ASVAB tests/constructs and the supplemental ECAT tests/constructs. One interesting point is that, even though the Services independently chose new aptitude constructs for research, there was a large measure of agreement as to which ability dimensions were the best candidates to improve the ASVAB. This consensus was clearly reflected in the proposals to the ECAT test selection panel.

For example, of the intellectual abilities identified in the Army's Project A review and not measured by the ASVAB, Spatial Ability is the most evident shortcoming given that a spatial factor is the most frequently reported psychometric construct, next to verbal ability. Independently, the Navy arrived at the same conclusion: the ASVAB should include spatial ability tests. Both Army and Navy spatial tests are now part of ECAT. The Services also independently concluded that information processing measures were good candidates for supplementing the ASVAB. Information processing measures are broadly defined here to cover the whole array of test types that emerged from cognitive and experimental psychology beginning in the mid-1970s; including, simple and choice reaction time tests, reaction time based reasoning tasks, divided and selective attention measures, and power-oriented working memory tests. Literally, hundreds of such tests have been developed and field tested.

If we look at the ASVAB and ECAT tests as two groups, one thing becomes immediately apparent — the military's collective efforts to broaden the dimensionality of the ASVAB have produced almost a classic verbal versus performance or crystallized versus fluid intelligence dichotomy, with the addition of a psychomotor dimension which is indicated in the Table.

This, then, is the outline of the future ASVAB. An ASVAB that will be computer administered. An ASVAB that will be less achievement and knowledge based and provide a better sampling of the breadth of human intellectual performance. The specific features of the new face of the ASVAB will be known soon, based largely on the results of the ECAT validation study.

Current Status

In June 1992, based on the success of CAT-ASVAB and promising early results from the ECAT study (and its precursor), we began an Operational Test and Evaluation (OT&E) of CAT-ASVAB combined with tests from the ECAT battery. CAT-ASVAB scores will be operational while the ECAT scores will be used only for research purposes. The OT&E will use several sites across the country and continue for two years. For now, the CAT-ECAT OT&E battery is our best guess at what the new face of the ASVAB will look like in the future.

References

CACI, Inc. & Automated Sciences Group (1988, March). *CAT-ASVAB Program: Concept of operation and cost/benefit analysis.* Contractor report for Department of the Navy, Office of the Chief of Naval Personnel, Military Personnel Policy Division (Contract No. DE-AC05-860R21642).

Eitelberg, M. J., Laurence, J. H., Waters, B. K., & Perelman, L. S. (1984, September). *Screening for Service: Aptitude and Education Criteria for Military Entry.* Washington, DC: Office of Assistant Secretary of Defense (Manpower, Installations, and Logistics).

Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice (1978, August 25). *Uniform Guidelines on Employee Selection Procedures.* 43 Fed. Reg. 166, 38290-38309.

Hetter, R. D. (1992). *Item Calibration Medium Effect on CAT Scores.* Paper presented at the annual meeting of the Military Testing Association, October, San Diego, CA.

Moreno, K. E. & Segall, D. O. (1992). *CAT-ASVAB Precision.* Paper presented at the annual meeting of the Military Testing Association, October, San Diego, CA.

Ramsberber, P. & Means, B. (1987). *Military performance of low-aptitude recruits: A re-examination of data from Project 100,000 and the ASVAB misnorming period* (Final Report 87-31). Alexandria, VA: Human Resources Resources Research Organization.

Schmidt, F. L., Hunter, J. E., & Dunn, W. L. (1987, November). Potential Utility Increases from Adding New Tests to the Armed Services Vocationa Aptitude Battery (ASVAB). Contractor report to the Navy Personnel Research and Development Center, San Diego, CA (Contract No. TCN-86-698-DO-0053).

The Development of Alternative Operational Concepts

James R. McBride
Human Resources Research Organization (HumRRO)

Linda T. Curran
Defense Manpower Data Center

The Department of Defense is preparing for a major revision of the Armed Services Vocational Aptitude Battery (ASVAB), to take place in the latter part of this decade. In the earliest stages of revision planning, a number of alternatives are being considered. These alternatives can be grouped into four major dimensions:

1. WHERE ASVAB is given -- changes in the site of ASVAB administration;
2. WHEN ASVAB is given -- changes in the schedule of administering different parts of the battery ;
3. HOW ASVAB is given -- changes in the medium of administration; and
4. WHAT ASVAB consists of -- changes in the content of the battery.

Clearly, these categories of possible change go beyond mere revision of the ASVAB tests themselves; in the aggregate, they constitute revisions to the entire concept of operations for administering ASVAB. This paper summarizes the development of alternative concepts of operations that are being considered in the planning for ASVAB revision.

Where, when, how, and what? There are a number of alternative answers to each of those four questions -- so many that at the outset of the project, we were faced with at least 540 different possible combinations, each of which represented a plausible operational concept. Presently, we are considering the following:

- three different classes of test administration sites: MEPS, MET sites, and Contract Testing Centers;
- two alternative classes of test administration stages: single-stage and two-stage testing;
- four alternatives as to test administration media: print, computer, digital keypad, and a special-purpose response pedestal;
- three alternative content configurations: the current one, a shortened battery, and a battery with expanded content.

Most of the alternative concepts consist of innovations added to the current operational concept. At this writing, the field has been narrowed to just seven innovations. Each of these innovations is discussed below, under the heading of the major dimension of change it pertains to.

Test Administration Sites

The ASVAB Enlistment Testing Program is conducted primarily by administering the tests in Military Entrance Processing Stations (MEPS) and in their associated Mobile Examining Team (MET) sites. At this writing there are almost 70 MEPS, and over 800 MET sites, so the operational ASVAB is given in about 900 locations throughout the U.S. The MEPS' mission is evaluation and processing of candidates for entrance into the Armed Services; the MET sites' mission is to make it more convenient for applicants to take -- and less costly for the Government to administer -- the ASVAB battery by reducing the distance applicants must travel to take the tests which are prerequisites to enlistment. The MET sites make it easier for applicants to take the tests; in doing so, they are thought to facilitate recruiting by lessening somewhat the burden of applying for enlistment.

We are considering a new form of MET site: contract testing centers (CTCs). CTCs would be dispersed sites, operated by contractors, at which ASVAB can be administered by appointment. The CTC concept differs from current MET sites in several respects: (a) CTCs would supplement MET sites, making ASVAB available at even more locations than it is now; (b) ASVAB would essentially be individually administered in CTCs, as opposed to the small group administration typical in MET sites; and (c) appointments would be available on short notice -- typically one day in advance -- as a convenience to applicants and recruiters.

Test Administration Stages

Recruiting personnel have expressed concern that the length of the ASVAB may hinder recruiting by creating a psychological obstacle between a momentarily motivated prospect and the evaluation and processing that are prerequisite to consummating an enlistment contract. The recruiter's interest, once he or she has identified a promising prospect, is to complete the process while prospect is interested and motivated. The concept of two-stage administration of the battery is intended to address these concerns.

The idea of two-stage administration of the ASVAB is to reduce the duration of the initial test administration by breaking up the battery into two sessions that take place at different times. The first session would consist of just those tests that are used to determine enlistment eligibility -- essentially, the Armed Forces Qualification Test (AFQT) subtests. The second session would contain the remainder of the battery -- the additional tests whose scores are needed to determine applicants' eligibility for military job specialty training. The first stage would be much shorter than the current battery -- and hence would be less of an obstacle to recruiting. The second stage would be administered only to applicants wno attained qualifying scores in the first stage; knowing they are at least minimally qualified for enlistment these applicants would be expected to be motivated to continue the testing process.

Test Administration Media

Three new means of administering ASVAB remain under consideration: (a) computerized adaptive testing (CAT-ASVAB), (b) a special-purpose response pedestal capable of administering certain psychomotor ability measures (in addition to CAT-ASVAB), and (c) a digital keypad that would replace printed answer sheets with an electronic input device to facilitate test administration and scoring.

CAT-ASVAB has essentially completed its research phase, and has been successfully equated to the printed ASVAB. Another computerized testing program, ECAT, is evaluating the incremental usefulness of nine innovative tests, six of which can only be administered by computer. Three of the ECAT tests measure psychomotor abilities, and require the use of a special-purpose response pedestal in addition to the CAT-ASVAB computer. An operational test and evaluation of CAT-ASVAB began in selected MEPS in late FY92; as part of that OT&E, CAT-ASVAB will be augmented by two or more ECAT tests.

The digital keypad is about the size and weight of an electronic calculator. Using it instead of a scannable answer sheet offers some of the administrative advantages of computerized testing at lower cost. It is also smaller and lighter than a computer, and thus may be more suitable for use in MET sites where equipment portability is a concern. Digital response keypads are used in a number of personnel testing programs outside the Department of Defense.

Test Battery Content

Two alternatives to the current battery content are being considered: (a) a shortened battery, formed by consolidating some of the current 10 subtests, and (b) an expanded battery, consisting of some or all of the current content plus one or more new tests drawn from the battery of 9 ECAT measures.

The rationale for a shortened battery is to eliminate redundancy by consolidating ASVAB tests with similar content, and in the process reduce the overall length of the battery. The most obvious candidates for consolidation are the ASVAB subtests measuring quantitative and verbal abilities. Arithmetic Reasoning, Math Knowledge, and Numerical Operations might be consolidated into a single test of quantitative ability. Word Knowledge and Paragraph Comprehension might similarly be consolidated into a single measure of verbal ability. The most extreme form of consolidation would be to reduce the battery to just four subtests, one measuring each of the four ability factors that are thought to constitute ASVAB -- Verbal, Quantitative, Technical, and Speed. A consolidation of any kind might significantly reduce the time it takes to administer ASVAB; perhaps an hour or more of testing time could be eliminated in this way.

The rationale for expanding the battery is to take advantage of years of research into new measures conducted jointly and separately by the Services. This research has culminated in trials of the ECAT battery: nine tests measuring dimensions not represented in the ASVAB, including spatial abilities, memory, and psychomotor skills. The hope is that expanding the scope of human abilities measured in the ASVAB will improve the prediction of training and job performance.

Choosing Among Alternative Concepts

Earlier we mentioned that a slate of 540 different concepts had been reduced to 50. How is the Department of Defense going about the process of choosing alternative concepts for possible operational implementation in the revision of ASVAB? Ultimately, the decision will hinge on a comparison of the costs and benefits of the current system with those of candidate alternatives; to this end, DMDC has commissioned a contractor to develop a cost-benefit evaluation process to be applied to both the current system and the most attractive alternatives. Before that process is applied, however, a few among the many alternative concepts must be selected; that selection is being done through a process intended to develop a consensus as to which of the many alternatives are the "most attractive" ones. This part of the paper presents a summary of that process.

Clearly, the current operational concept for ASVAB administration should not be replaced without compelling reasons. This implies that appreciable benefits must accrue, or there is no reason to change. One of the first steps in the process was to study the current system, identify its strength, weaknesses, and who might be affected, and in what ways. This led immediately to enumeration of a number of issues that need to be considered, and different stakeholders whose interests need to be considered.

Issues. Changes to the concept of operations for administering ASVAB may have an impact on organizations and systems that are interrelated with ASVAB. This section identifies a number of issues that should be considered in evaluating alternatives to the present ASVAB and its operational concepts.

Impact on Recruiting. Any change in the operational concepts for ASVAB administration and use may have an impact, favorable or unfavorable, on recruiting; such impacts will have to be assessed and evaluated for any operational concept change that is seriously considered. There are at least three different areas in which the recruiting impact might be felt: 1) recruiter productivity, 2) enlistment propensity of prospects, and 3) resources involved in recruiting.

Impact on Military Personnel Accession Systems. Some changes in operational concepts might also have impact on military personnel accessioning systems. Any operational concept that is considered will need to be evaluated in terms of its impact on such matters as the following: 1) statutory DoD-wide standards for enlistment qualification; 2) standards set by the individual Armed Services for both personnel selection and classification; 3) practices of each of the Services pertaining to personnel classification and training assignment; 4) DoD and Service formal policies and automated systems that support recruitment and accessions processing. Each of these is discussed in somewhat more detail below.

The impact of ASVAB changes on policy directives and automated accessioning, assignment, and information systems must be carefully assessed. DoD and each of the Services maintains formal directives that specify policy for personnel accessions as well as training and job assignment. Additionally, there are a number of automated information systems that use or record

ASVAB data; examples include the MEPS Reporting System (MEPRS), the DMDC enlisted personnel data files, and the four Services' separate automated accessions systems (REQUEST, PRIDE, PROMIS, and ARMS). Changes in ASVAB content, score metrics, norms, or operational concepts could require substantial changes to existing systems of policy directives and automated data management.

Impact on Military Personnel Training. Changes in ASVAB or its operational concepts may have a substantial post-enlistment impact, as well. Changes in classification and assignment standards or practices may affect performance in military job specialty training, and may entail a requirement to conduct an extensive body of new predictive validation research. Additionally, to the extent that existing selection composites are eliminated by ASVAB changes, the existing validity research base may be rendered meaningless.

Costs. Changes to the present concepts of operations inevitably entail changes in the cost of those operations, so cost is an issue that must be considered in evaluating alternative concepts. There are a number of cost components that may be affected by operational concept changes. Included among these are capital investment costs, operating and support costs, personnel costs, recruiting costs, and job training costs, to name but a few.

Implementation. The actual implementation of any change in the concept of operations is an issue unto itself, due to the magnitude of the system and to the number of organizations that will necessarily be involved. This is because of the large scale of ASVAB administration, and because operational changes will require implementation measures not only by U.S. MEPCOM, but also by four Armed Services and the Coast Guard. Within each of the affected Services, there may need to be implementation measures taken by the respective recruiting arms, by accession policy organizations, and by the training establishments. The scale of implementation activities may be quite large; the number of different organizations that will have to participate makes implementation also potentially complex.

Acceptance. Acceptance of operational concept changes is also an issue, and one with many facets. Any operational concept change will have to be accepted by the test operator (U.S. MEPCOM), by the Armed Services recruiting commands, and by manpower policy officials in each of the Services. Additionally, if the change in operational concept is a highly visible one, there may be questions of its acceptability outside the Department of Defense, where there may be questions of social and political acceptance by the public at large and by special interest groups. This kind of issue may be more prominent in the case of introduction of a change to computerized test administration, or to adaptive testing.

Test Security. Procedures for safeguarding the security of ASVAB materials are well established for the current concept of operations. Operational changes in medium of administration, administration sites, or authorized test administration personnel may create requirements for different or additional test security precautions, to the extent that current security procedures do not apply to the new operational concepts. This is another issue that must be addressed in considering any alternatives to the present operational concepts.

Test Validity. The prospect of changing the ASVAB battery itself, the medium of administration, and operational conditions such as testing sites and test administration personnel -- all of these things raise the issue of what impact they may have on test validity. In this context, the term "test validity" is interpreted in its broadest sense, as the extent to which decisions and inferences based on test scores are justified scientifically and supported by appropriate evidence. In this sense, test validity includes not only predictive validity, but also construct validity, test equity and fairness, and the fundamental measurement properties of the tests. All of these things may be affected by changes in test content, battery composition, conditions of administration, and other aspects of new operational concepts.

Stakeholders. The interested stakeholders in ASVAB revision are a number of diverse groups and organizations, including the recruiting services, accession policy offices of Defense and the four Services, the Military Entrance Processing Command (MEPCOM), as well as the Services' personnel research and training communities. Representatives of each of these groups were invited to participate in three different forums: a project advisory panel, a personel research technical committee, and a series of Joint Services workshops. The study of the current system, and the forums for addressing possible changes, are described below.

Analysis of current operations. The contractor conducted a study of the current operational concept, and prepared a report (Hogan and Mullin, 1992) that describes current recruiting, applicant processing, and ASVAB testing operations. They found the current mode of ASVAB operations to be generally effective and efficient, but they were able to identify several areas for potential improvement that might be addressed in revising the ASVAB concept of operations. Included among these were the following:

1. Information about available military jobs or enlistment incentives for which an applicant is qualified is not available to the applicant until he or she has invested a good deal of time and effort. This may discourage or frustrate some applicants; providing such information earlier in the process may increase enlistment rates among qualified applicants.
2. MET site testing costs are relatively high because there is a fixed cost per test session, and the number of examinees per session is frequently small -- less than 7.
3. Official AFQT scores, and scores of other ASVAB tests administered at MET sites are not available until the answer sheets have been scored at the host MEPS -- a delay of one or more days. This interrupts the flow of the recruiter's effort to convert a prospect to an enlistment, and may effect some applicants' decisions not to enlist.
4. Aptitude information contained in ASVAB scores is not fully exploited in matching recruits to job assignments, before or after enlistment.
5. Two-day processing of applicants is the norm at the MEPS, so many applicants must remain overnight, at government expense and some personal inconvenience. Significant, direct cost savings would result if most MEPS processing could be accomplished in a single day. Indirect savings, in the form of increased recruiter productivity, are likely if the incidence of two-day MEPS processing were reduced.
6. Recruiters frequently escort prospects to MET sites or MEPS for processing. Because these test administration sites are often located far from the applicant's home and the recruiter's office, a good deal of recruiter time is taken up by this practice. If ASVAB could be administered in more convenient locations, recruiter productivity might benefit substantially.

The Concept of Operations Planning and Evaluation (COPE) Advisory Panel. This is a panel, composed of delegates from OASD, USMEPCOM and the four Services' manpower policy and recruiting commands, formed to advise DMDC in the development and evaluation of alternative concepts for ASVAB. The panel first convened in July 1991, and has convened on four subsequent occasions to date. At its September 1991 meeting, a number of goals and objectives for revised ASVAB operations were enumerated (COPE, 1991). These included the following:

1. Improvement of pre-screening tests (EST and CAST);
2. Development of a shorter qualification test (1 to 2 hours) with immediate scoring.
3. Reduction of the number of MET sites with low ratios of examinees to test administrators.
4. Re-instituting one-day processing of applicants at MEPS.
5. Increasing high school participation in the ASVAB School Testing Program.
6. Using computerized technology to increase flexibility of testing-related operations.

Several of these objectives are directly reflective of recruiter concerns.

An ASVAB Revision Workshop. In March, 1992, DMDC convened a Joint Services workshop to identify and discuss issues related to the revision of ASVAB and its concepts of operations. An important function of the workshop was to bring together representatives of all the stakeholders in the ASVAB, to give them an opportunity to present points of view, to hear those of others, and to identify issues and considerations that were of critical importance. The workshop included an overview of alternative operational concepts, a presentation of significant issues that must be considered in evaluating such concepts, and introductions to several specific alternative concepts that were used as illustrative examples.

The ASVAB Review Technical (ART) Committee. This committee is an ad hoc subcommittee of the DOD Manpower Accession Policy Working Group, formed for the express purpose of identifying and resolving psychometric and technical issues associated with the changes that are now being contemplated. While the ART Committee is especially interested in questions that pertain to possible changes in ASVAB content and psychometric characteristics, they are also watchful for the psychometric implications of any other changes in the concept of operations. The work of the ART Committee was described in the paper by Bloxom and Wise, elsewhere in this symposium.

Based on input from the COPE Advisory Panel, and the Joint Services ASVAB Revision Workshop, we have recently narrowed the list of alternatives from 540 to 50, and selected the best combinations of those 50 alternatives for further study, including a detailed evaluation of cost-effectiveness. The evaluation process, and today's list of the most promising operational concepts, are described by McBride and Hogan (1992) in their paper in this symposium.

Evaluation of Alternative Concepts

James R. McBride
Human Resources Research Organization

Paul F. Hcgan
Decision Science Consortium, Inc.

Earlier in this symposium, McBride and Curran indicated that a large number of candidate alternative operational concepts were identified. Over the course of the project their number has been reduced from over 500 to 50; although that is a significant reduction, 50 discrete operational concepts is still too many to deal with in detail. Ultimately, a decision must be made between a small number of the best alternatives--possibly including the status quo. This paper summarizes an evaluation process that will facilitate movement toward that decision. The evaluation process can be thought of as consisting of two stages. In the first of these, qualitative criteria have been applied, subjectively, leading to narrowing the field to a small number of alternative concepts to be evaluated further. In the second, quantitative criteria will be applied objectively to compare those few alternative concepts against each other and the status quo; this will lead toward a recommendation regarding the ultimate revision of ASVAB itself.

Narrowing the Field of Alternative Operational Concepts

In March 1992, a Joint Services ASVAB Revision Workshop was conducted; all of the major stakeholders in ASVAB revision participated. The consideration of alternative concepts of operations was a major agenda item. The participants responded to a variety of broad proposals, rejected some, and endorsed others for further study. One outcome of the workshop was the reduction of the number of possible alternatives from 540 to 72. Each of those 72 alternative concepts consisted of a unique combination of the remaining alternatives for ASVAB testing sites, medium of administration, content, and administration schedule (single- vs. two-stage). Figure 1 depicts these 72 alternatives in matrix form.

As indicated in Figure 1, some 22 of these alternatives either represent the status quo, or are not considered viable; that leaves just 50 alternatives for serious consideration. To reduce the number further, we subjected these 50 candidates to a formal evaluation. A qualitative rating process was developed, and a panel of raters convened to conduct the evaluation. The rating task was simple but laborious: Each rater was asked to judge whether a candidate alternative would improve on the current system, make things worse, or make no difference. Every rater had to make such a judgment with respect to some 39 separate items spanning 6 dimensions: impacts on recruiting, MEPCOM operations, enlistment processing costs, post-enlistment training, research and development requirements, and military personnel accession policies and systems. Scores for each alternative were computed by averaging the ratings across judges.

Separate scores were computed for each of the six dimensions; the overall score was the sum of the six dimension scores. For the purposes of this paper, we will limit the discussion to the overall ratings for each alternative.

Figure 1 lists the overall scores for each alternative. To interpret these scores, keep in mind that a score of 0 means an average judgment of no difference between the current ASVAB operational concept and the alternative in question. A positive score means that more things were judged to be improved than made worse; a negative score means more things were thought to be made worse than improved. With that in mind, take a holistic look at the pattern of average ratings shown in Figure 1. Note that there is a consistent pattern of negative scores for all of the alternatives that involve expanded ASVAB content. Another consistent pattern occurs on the right half of the figure: the ratings for the alternatives involving two-stage testing were consistently low or negative.

Now look at the ratings for the other alternatives -- those involving one-stage testing with either the current battery or a shorter one. Our interpretation of the pattern of these scores goes like this:

(a) the concept of Contract Testing Centers (CTCs) as alternative sites for administering ASVAB got high marks;

(b) both alternative administration media -- Digital Response Pads (DRP) and computers -- were judged to improve on the paper-and-pencil system, overall;

(c) computer-administered testing was judged somewhat higher than the Digital Response Pad concept for testing in MEPS and Contract Testing Centers, but the Digital pads were judged to be somewhat better for MET site testing;

(d) a shortened ASVAB battery was judged to be an overall improvcment over the current content in the case of the paper-and-pencil medium, but not in the case of automated testing using either computerized testing or the Digital Response Pad.

These evaluation results were used in an advisory capacity to reduce the scope of alternative concepts to be considered further. Essentially, the two-stage and expanded battery concepts were eliminated, leaving just 16 alternatives. Each of these 16 is a concept for testing in one kind of site or another. A complete ASVAB operational concept would consist of two or more of these in combination, because there may be different approaches to testing in the MEPS, MET sites, and CTCs (if they are implemented). If we consider all possible combinations of these 16 alternatives, we again have a large problem. At this point, however, we made some practical judgments, and identified a small set of rules for eliminating some combinations. Space does not permit going into the detail of this part of the process, so we will move directly to the bottom line: in conjunction with the COPE Advisory Panel, we identified 8 combinations to be evaluated further, including the current operational concept. These combinations are identified in Figure 2, which lists the medium of administration and battery content of each, by type of testing site.

Two things about Figure 2 should be pointed out. The first is that combination A is the current operational concept; it will provide a baseline against which all other operational concepts can be compared. The second is that the presence of combination H may come as a surprise, since it involves a two-stage ASVAB process employing the ECAT response pedestal to administer expanded content in the second stage. It is listed among these "semi-finalist" concepts out of two concerns: (a) After the research investment that has been made in developing and evaluating the 9 ECAT tests, it would be imprudent to eliminate all of them from consideration before the results of a Joint Services validity study are published; (b) based on earlier analyses, there is a possibility that computerized testing may not be cost-effective unless the scope of ASVAB battery content is expanded.

Quantitative Evaluation of Candidate Operational Concepts

Each of the eight combinations of alternative concepts, listed in Figure 2, will be the subject of a cost evaluation. These evaluations will use a common methodological framework designed to capture key costs and cost savings associated with each of the alternatives. The methodology employed will be broad enough to capture the highly diverse scope of costs and benefits associated with the alternatives; at the same time, it will be consistent with established economic analysis principles, as well as relevant Department of Defense guidance for conducting and reporting such analyses, such as the CIM (Corporate Information Management) format. This section of the paper summarizes the approach that will be employed.

The cost evaluation will employ a cost-effectiveness framework; that is, it will compare the costs of the various alternatives, holding constant the output of the system. For most -- but not all -- of the analysis, this "output" is the number of enlisted accessions. Within this framework, the emphasis will be on costs that are likely to vary among alternatives; in other words, the focus is to look for differences in the costs of the alternatives, and not to attempt a complete cost accounting. The analysis will be forward-looking; it will extend over a period of time long enough to encompass the time path of the output.

Earlier in the project, Hogan and Mullin conducted a study of the current recruiting and enlistment processing system. They developed a description of the system in terms of several sequential stages of processing. The cost-effectiveness analysis will use what is known about those stages, and develop parametric stage-of-processing "spreadsheet" models of costs and benefits of each alternative operational concept.

Separate but interrelated spreadsheet models will be developed for each of three levels; the three levels will differ in terms of the costs they encompass, as well as in complexity and uncertainty.

Level 1 will entail developing models of direct processing costs, including such items as travel costs, test administration costs, meals, lodging, and so forth. This level has the most readily measurable costs, and has the added advantages of being the one most closely tied to operating budgets, and for which the most objective data exist. At level 1, for one operational concept to be deemed more cost-effective than the current system, it will have to produce the same number of accessions at less cost.

Level 2 will entail models of recruiting costs and applicant propensities. Level 2 models will only be relevant if an operational concept in some way affects either (a) the amount of recruiter time required in enlistment processing (recruiter efficiency) or (b) the propensity for an applicant to enlist or at least continue to a subsequent stage of processing (and thus affects recruiter productivity). In contrast to Level 1, where all cost elements have to do with processing efficiency, at Level 2 we include some cost elements of that are linked to behavior, as we attempt to measure recruiter productivity and changes in the propensity of applicants to enlist.

Level 3 will entail models of the "quality of the job match", and will address the cost-benefit of improvements in predicting training and job performance. This level of model will only apply to operational concepts that provide improved aptitude information, and hence improved selection or classification. It will have a fundamentally different output measure than Level 1 and Level 2: The output measure at the first two levels is enlisted accessions; at Level 3, it is "success" as indicated by (a) an enlistee's completion of initial skill training or (b) satisfactory performance on the job. For the latter criterion -- job performance -- the approach will be to model benefits as reductions in recruiting and training costs resulting from improved force performance.

An example of what these spreadsheet models will look like is given in Figure 5. This is a draft spreadsheet encompassing both level 1 and level 2 data. It allows the analyst to specify the number of annual enlistments or accessions, then compare an operational concept against a baseline in terms of the total costs of recruiting and processing enough applicants to yield that number of enlistments. The column labeled "Default Value" contains actual FY91 data; the data in that column constitute the baseline. The column labeled "Desired Value" is the spreadsheet model; by varying the input values associated with travel costs, processing costs, flow rates and propensities (continuation rates), we can assess the impact of these variables on total costs.

The spreadsheet in Figure 5 is just an example to illustrate the framework and the approach to cost analysis. The actual spreadsheet models are still under development. Baseline data are still being collected; additional variables are still being added to the model; and interrelationships among the variables are still under study. In the final analysis, we will need to be able to specify the impact of each alternative operational concept on the spreadsheet variables that make a substantial difference in total cost.

Figure 1. A Schema for the Alternative Concepts, and their Overall Ratings

	1-Stage			Multi-Stage		
	MEPS	MET Sites	CTCs	MEPS	MET Sites	CTCs
Current Battery						
Paper-Pencil	0*	0*	1.4	-6.7	-4.3	2.3
Digital Pad	4.9	8.2	8.8	-3.1	.7	1.8
Computer	9.0	7.7	15.0	-1.8	2.7	4.1
ECAT Pedestal	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
Shortened Battery						
Paper-Pencil	2.0	3.4	5.7	-3.6	-1.1	-1.3
Digital Pad	3.3	5.0	7.1	-2.9	1.2	1.7
Computer	4.4	3.0	8.4	-2.3	.9	1.9
ECAT Pedestal	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx	xxxxx
Expanded Battery						
Paper-Pencil	-10.8	-13.7	-13.4	-14.0	xxxxx	xxxxx
Digital Pad	-8.2	-4.1	-2.3	-12.1	xxxxx	xxxxx
Computer	-2.9	-1.0	2.0	-8.8	xxxxx	xxxxx
ECAT Pedestal	-10.8	-3.6	1.4	-7.2	xxxxx	xxxxx

Notes: 1) * indicates current operational concept.
2) xxxxx indicates an invalid operational concept.
3) N = 9 raters.

Figure 2. Alternative Concepts Nominated for Further Study

		SITE Type		CONTENT	Comments
ID	MEPS	MET Sites	CTCs		
A	Paper	Paper	None	Current	Current operational concept.
B	Digital	Digital	None	Current	Automates all testing.
C	CAT	Digital	None	Short	Avoids CAT portability issue.
D'	Paper	CAT	None	Current	Highest investment cost.
E'	Paper	Digital	CAT	Short	
F'	CAT	Digital	CAT	Current	
G	CAT	CAT	CAT	Expanded	All CAT.
H	ECAT	Digital	None	Current	Two-stage ASVAB. Stage 2 CAT/ECAT given to qualified ETP and STP applicants.

Figure 3. A 3-level Framework for Modeling Costs of Operational Concepts

Level	Primary Purpose	Output Measure
Processing Costs	Capture enlistment processing costs associated with a specific concept of ASVAB operations.	Entrance processing costs of achieving a given number of accessions.
Recruiting Costs	Capture the effect of an operational concept on recruiter productivity/applicant propensity.	Recruiting costs of achieving a specified number of accessions.
Job Match: A	Measure the value of improved selection and classification in terms of initial skill training.	Costs of achieving specified numbers of "successful" recruits, i.e. recruits completing initial skill training.
Job Match: B	Measure the value of improved selection and classification in terms of first-term job performance.	Recruiting and training costs of achieving first-term performance goals.

Considerations for the Development of New ASVAB Norms

Dr. Linda T. Curran
Defense Manpower Data Center

Background

The Armed Services Vocational Aptitude Battery (ASVAB) is administered to students in Grades 10, 11, 12, and postsecondary school in the Department of Defense (DoD) Student Testing Program (STP) and to military applicants in the Enlistment Testing Program (ETP). ASVAB norms are used in both programs to show each examinee's standing relative to a reference population to determine enlistment eligibility qualifications and to classify individuals into occupations. The STP also uses separate norms by grade and gender to aid career counseling. STP materials help students link their interests, aptitudes from the ASVAB, and personal preferences to civilian and military occupations.

Current ASVAB Norms

Current ASVAB norms are from an administration of the ASVAB in 1980 to about 12,000 American youth about 16-23 years of age (DoD, 1982). This project was part of a joint effort between DoD and the Department of Labor (DoL). The ASVAB was administered in conjunction with the DoL's National Longitudinal Survey of Youth Labor Force Behavior to a sample that was nationally representative in terms of gender, ethnicity/race, geographic region, urban/rural area, and socioeconomic status. An oversampling of Hispanics, Blacks, economically disadvantaged Whites, and youth in the military was obtained to allow for more precise analyses of these groups. Norms were developed by weighting the sample to represent the national population distributions for all groups. For the ETP, norms were developed from scores of 18-23 year olds in this sample, and for the STP, norms were developed separately for Grades 11 and 12 and postsecondary school students.

Need for New Norms

The DoD is exploring alternatives to the content, testing mode, and administration of the ASVAB as described in previous papers presented in this symposium. Normative information will be required before new or significantly changed tests can be implemented. Regardless of DoD's decisions on ASVAB content and mode of administration, it is time to plan for new ASVAB norms as the 1980 norms are becoming dated. This is evidenced by the changing demographic distributions of the nation as demonstrated by the 1990 census and the changing distribution of abilities as illustrated by test scores on nationally administered tests.

Considerations for Future ASVAB Norms

The previous ASVAB norming study was very costly (approximately $4 million for DoD's share of the effort). To conduct a normative study today, that is as extensive as the effort in 1980, would require monies that are probably not available in today's budget environment. However, the need for new norms exists and the challenge is to design a norming study that will be psychometrically sound yet cost effective. The ASVAB misnorming in 1976-1980 which led to the erroneous enlistment of about 400,000 military applicants at a significant cost is a reminder that DoD should not sacrifice psychometric accuracy for cost.

What follows are descriptions of issues that need to be considered when planning the development of norms for use in the selection and classification of military and for the secondary purpose of career counseling in the STP. Included in the descriptions is the identification of issues that are common to norming efforts regardless of test content or administration media and those that are peculiar to testing via computer or electronic answer sheet (digital response pad), options being considered in the ASVAB Review and Concept of Operation projects.

Objectives of Norms. The first step in a normative effort is to identify the norming objectives. One major objective of ASVAB norms is to fulfill the legal requirement which prohibits the enlistment of individuals below the tenth percentile on the Armed Forces Qualification Test (AFQT), a measure of verbal and math abilities. To meet this legal requirement and for annual reporting purposes to Congress on the quality of recruits, each of the Services use the AFQT to determine whether an applicant meets minimum qualification standards. Also, each of the Services uses the AFQT to determine whether applicants qualify as "quality" recruits (high school graduates who score at or above the 50th percentile). Some of the Services provide enlistment bonuses to quality recruits who enlist in certain occupations. Therefore, the accuracy of new ASVAB norms is imperative and could effect costs associated with bonuses, attrition, and training and job performance.

A second objective is to provide aptitude information for use in the classification of military applicants to specific occupational specialties. Based on Service-specific studies analyzing the utility of the ASVAB to predict training and on-the-job performance, each of the Services define composites using different ASVAB subtests. The Services determine the cut scores on the composites above which an applicant must score to be considered for a particular occupation. For the first two objectives (to meet AFQT legal requirements and provide classification information), then, would require both accurate and up-to-date normative information on American youth that meet general enlistment eligibility standards based on age.

A final objective for ASVAB norms only applies to the STP. In addition to meeting the objectives discussed previously, STP ASVAB norms are used to allow students to compare their aptitudes with those of their peers for the purposes of facilitating students' career explorations. ASVAB content changes under consideration include changing the content of the STP ASVAB to decrease its length. Military recruiters argue that the test used in the STP should be shortened to increase participation in the program which is now perceived as too long. One recommendation is to keep the AFQT subtests in the STP battery and to add other tests with content useful for career guidance. The objectives of the STP norms and consequently the design of a norming study which applies to the STP will depend on DoD's decision regarding this recommendation.

Target Normative Populations. After norming objectives are identified, the next step is to define the target normative population based on the objectives. For the selection and classification objectives of ASVAB norms for the ETP, the appropriate normative population is enlistment eligible American youth. This population is broadly defined as U.S. citizens or legal aliens between the ages of 17 and 35 who are not physically handicapped or have court convictions. There are other general and Service-specific standards; however, in a norming effort it would be more difficult to identify individuals meeting these other criteria.

The current STP objectives are to provide enlistment selection and classification and career guidance information. All three of these objectives can be met with a target normative population defined as American youth in Grades 10-12 and postsecondary schools who meet eligibility standards except age.

Sampling Plan. To develop accurate norms, the sample must be representative of the population and contain a sufficient number of examinees in each demographic cell. The number of examinees and other demographic characteristics of the sample will depend largely on the availability of funds, availability of examinees, and the requirements of a joint sponsor (if one is found). To provide cell sizes needed to estimate score distributions with accuracy, oversampling can be accomplished, as in the ASVAB 1980 norming effort. With oversampling, population weights need to be developed and applied to individual cases in the sample to make it representative of the entire population. ETP and STP ASVAB norms can be collected in one data collection effort as in 1980; however, the sample must be defined in terms that will meet the objectives of both programs.

To provide an acceptable sample of the population, one must balance the sample in terms of demographic characteristics which are known, or suspected, to effect distributions of aptitude test scores and consequently can effect the accuracy of the norms. Stratification of various characteristics of the population will be accomplished using the most recent census data available; the 1990 national census contains the most recent information on American youth. Key demographic variables include age, gender, level of education, socioeconomic status, ethnicity/race, geographic region, and an urban/suburban/rural indicator. Based on the proportion of the national population in each cell of the sampling matrix a sampling plan calling for target national proportions in various demographic conditions can be developed.

If DoD decides to develop a test battery for the STP which does not contain the same content as the battery administered in the ETP, then DoD may use a different sampling strategy for test content which will be used solely for career guidance purposes. Some would argue that for career exploration purposes, the accuracy of norms is not as much of an issue as it would be for a test which is being used to make personnel decisions. Therefore, it may be possible to develop norms for some portions of a new STP battery using a sample of schools that participate in the program. To generalize information collected on a sample of schools to either the population of all students in Grades 10, 11, and 12, and postsecondary schools or the population of students who participate in the STP, weights must be developed to make the sample representative of the intended population. Demographic characteristics of both students and schools should be considered in determining the weights and these include: for students, gender, age, ethnicity, race, socioeconomic status of parents, and intended field of study; and for schools, size of school, type of support (public, religious, or private), pupil-teacher ratio, per pupil expenditure, curricular emphasis, and proportion of students who are college bound (Angoff, 1984).

If a sample of students/schools participating in the STP are used for the development of norms, it will be necessary to examine the demographic characteristics of the institutions and students currently participating in the STP to determine how representative they are of the national population of high schools and postsecondary schools. Using such a sample would likely introduce systematic bias because the students and, to some extent, the schools who participate are self-selected and do not represent the student population at large. The vast majority of all schools participate in the program, so self-selection of students is the primary problem. It will be necessary to identify a representative sample of schools for which representative samples of students are identified and encouraged to participate. Information on non-participants from these samples would be used to reweight the sample to reduce nonresponse bias.

Administration Plans. Test administration plans will differ depending on whether the ASVAB will be normed in a paper-and-pencil, computerized, digital response pad, or some combination of media format. The norming of paper-and-pencil tests and tests which use digital response pads is more straightforward and would require the administration of the entire battery to all examinees.

The development of norms for a computerized test is not as straightforward and would require the administration of some portion of test items from a larger item bank to each examinee. Computerized item banks are large and it would not be feasible to administer every item to each examinee. Ignoring costs, issues to consider in the norming of a computerized test include: 1) inclusion of common items to place entire bank on the same scale, 2) balance of test content administered to each examineee (unbalanced content could effect the item response theory estimation of item and ability parameters), and 3) the logisitcs of getting computers to examinees or vice versa.

The Navy developed a computerized version of the ASVAB called the Computerized Adaptive Test-ASVAB (CAT-ASVAB) which was equated to the paper-and-pencil battery. If the content of the ASVAB does not require administration solely by computer, another norming alternative would be to administer a comparable paper-and-pencil version of the battery and equate the computerized version to the paper-and-pencil form. However, if DoD decides to use computerized testing in operations, it would be preferable to norm the ASVAB using computers since it would involve norming the test under conditions very similar to which the test would be used operationally.

Analysis Plans. The easiest part of the analysis will be to compute the percentile rankings and standard scores for each subtest for each normative sample (enlistment eligible youth, Grades 10, 11, and 12, and postsecondary school students). To get to the point where percentiles can be computed will require plans to 1) conduct data editing, 2) perform tests of sampling assumptions, 3) apply missing data algorithms where necessary, and 4) develop weights that will make the sample representative of the population. For normative data collected via computers, the analysis plans need to include specifications as to the models that will be used to estimate item and ability parameters and methods for linking them onto a common scale. Then the representative (weighted) distribution of thetas can be used to compute percentile rankings.

Evaluation of Costs and Psychometric Feasibility. DoD plans are to develop three alternative designs for developing norms for computerized and paper-and-pencil tests. DoD will evaluate each design in terms of cost and psychometric feasibility.

Cost factors include the development of sampling design specifications and the identification of individuals to test (in 1980, elaborate methods were used to identify households and eventually individuals to test; these methods were costly and perhaps more convenient samples, i.e., students participating in the STP, might be psychometrically acceptable and less costly); contact of individuals to test; printing of booklets/answer sheets/administration manuals and other materials; renting of test facilities; test administrator/proctor salaries; scanning/scoring; data editing and analysis; payments of incentives to examinees; and the generation of reports. Costs associated with computerized testing include the development of software for test administration; renting test sites (sufficient size, storage, electrical outlets); procurement of computer hardware; printing of test manuals, and other materials; salaries of test administrators/proctors; specialized training of test administrators; examinee incentives; transport of data (i.e., disks, modems); logistical concerns (movement of computers to examinee location); data editing and analyses; and the generation of reports.

Obtaining examinee participation will be problematic. One option, as was done in the 1980 norming, would be to collect the normative data in combination with another national testing project (or survey). DoD has initiated contacts with other government agencies to explore the possibility of a joint effort. Additionally, the payment of incentives to examinees may enhance participation.

Next, each of the alternative plans will be analyzed to determine their psychometric and operational feasibility, especially from a technological and logistical perspective. If necessary, policy and procedural changes necessary to evaluate the alternatives shall be identified and their potential impact assessed.

REFERENCES

Angoff, W. H. (1984). **Scales, Norms, and Equivalent Scores**. Princeton, NJ: Educational Testing Service.

U.S. Department of Defense. (1982). **Profile of American Youth: 1980 nationwide administration of the Armed Services Vocational Aptitude Battery**. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs and Logistics).

Discussant Remarks
ASVAB Contents and Operations:
Preparing for Changes

by

W. S. Sellman
Office of the Assistant Secretary of Defense
(Force Management and Personnel)

The military enlistment testing system has remained virtually unchanged since the beginning of the all volunteer force. Applicants take the paper-and-pencil ASVAB at some 68 Military Entrance Processing Stations or about 850 satellite testing locations across the country. Enlistment eligibility is quickly determined, and applicants negotiate with Service counselors about occupations and enlistment options and benefits before going to basic training. In the 1970s, it was a buyers' market so the Services had to tell young people about jobs and options before closing the sale. Today, we still administer ASVAB to all applicants, but the Air Force and Marine Corps generally do their specific job placement once their recruits arrive at basic training.

In the past 20 years, there has been tremendous progress made in the area of computerized testing. Its progress is both with adaptive testing and the use of computers to administer new tests not possible with paper-and-pencil formats such as cognitive processing, long and short-term memory, perceptual speed, and psychomotor skills. We also know that computerized tests can be equated to paper-and-pencil tests so when implemented we can retain our AFQT scale, and we can compare aptitude levels of recruits across Services and across time.

One area of computerized testing that has proved almost intractable is hardware development and selection. We started out many years ago to build our own systems; these would be computerized testing systems designed to administer our test with our hardware in our testing environment. Then, we moved toward use of off-the-shelf equipment. In 1984, at the MTA conference in Munich, Lt. General E. A. Chavarrie, my boss, was delivering the keynote speech, and the speech contained a major section about CAT-ASVAB. General Chavarrie had just visited some German enlistment stations where he saw computerized tests. Oh course, they were not adaptive, but General Chavarrie did not know that. Wnat he did know was that he saw the Germans giving tests on computers, and his speech said it would be another three or four years before United States implementation. By the end of the speech, General Chavarrie had accelerated the program. Thus, we abandoned the notion of developing specific hardware and moved to off-the-shelf equipment. It now has been eight years since that meeting, and we still have not implemented CAT-ASVAB.

Computer technology is evolving so rapidly that a system we bought today for CAT-ASVAB might be obsolete before it could be delivered. Buying enough computers to test over 750,000 people a year also would be expensive. In the budgetarily constrained world of the 1990s, the only way we will be able to obtain funds for procuring computers for testing is to demonstrate that such technology will pay for itself by improving recruiting, personnel selection, and job classification. Several years ago, we did a cost-benefit analysis of CAT-ASVAB, and it was disappointing--modest increases in validity did not lead to selection and classification utility sufficient to amortize the costs of procuring the hardware.

This analysis led to research on the validity of the ECAT, and we are very close now to seeing its results. Maybe both CAT and ECAT together will increase differential validity and make computerized testing worthwhile. But, just in case, the Defense Manpower Data Center initiated the concepts of operation research about which you just heard. Every aspect of enlistment testing has been reviewed. I am particularly impressed by the contract testing center concept. It would solve two of our most difficult problems. The contractor would always use state-of-the-art technology, and we would not have to buy computers that would soon be obsolete.

Finally, let me make some observations about Linda Curran's paper on norming. Linda's discussion about the problems facing us in developing new enlistment test norms is very complete. Developing norms would be complicated if we already knew what our testing system of the future will be; unfortunately, we don't. Add to that the cost of the norming study. It cost about $10 million in 1980 to do the Profile of American Youth Study in conjunction with the Department of Labor. I do not know what replication of that study would cost today. It was 35 years between our last norming efforts--in 1944 and when we renormed the ASVAB in 1980. The use of test scores for military personnel management is too important for us to wait another 35 years, to 2015, to renorm.

The way the youth population is changing there is no doubt that the 1980 norms soon will have no relevance whatsoever in terms of comparing the youth population and the quality of new recruits. So planning for a norming study in conjunction with the concepts of operation project is particularly appropriate. As is true in 1992 for all issues pertaining to Defense manpower, in the future there will not be much money available. I only hope we can get the funds to do these very important projects that we discussed here today.

Recognition-Primed Decisions: Implications for the Military

Thomas E. Miller
Gary Klein

Klein Associates Inc.
Fairborn, OH 45324

The past several years have seen the development of a new approach to understanding how people make decisions in real-world settings (Klein, 1989). Naturalistic decision making is an attempt to understand how humans actually make decisions in complex real-world settings. This work has focused on situations marked by a few key features: dynamic and continually changing conditions, real-time reactions to these changes, ill-defined tasks, time pressure, significant personal consequences for mistakes, and experienced people making the decisions. These task conditions are found often in military operational environments, so it is essential that we determine how people typically handle these conditions.

Previous models of decision making have been limited in their ability to encompass these operational features. Classical approaches to decision making such as multi-attribute utility analysis and decision analysis prescribe very analytical and systematic methods for weighing evidence and selecting an optimal course of action. Multi-attribute utility analysis decision makers are encouraged to generate a wide range of options, identify criteria for evaluating these options, assign weights to the evaluation criteria, rate each option on each criterion, and tabulate the scores to find the best option. Decision analysis is a technique for constructing various branches of responses and counter-responses and postulating the probability and utility of each possible future state in order to calculate maximum and minimum outcomes.

On the surface these strategies may seem adequate, yet they fail to consider some important factors inherent in real-world decisions. Classical strategies deteriorate when confronted with time pressure. They simply take too long. Under low time pressure, they still require an extensive amount of work and they lack flexibility for handling rapidly changing conditions. It is difficult to factor in ambiguity, vagueness, and inaccuracies when applying analytical methods. Another problem is that the classical methods have primarily been developed and evaluated using inexperienced subjects, typically college students.

There is a group of decision researchers who are trying to derive models that describe how experienced decision makers actually function. Rasmussen (1985) used protocols and critical incident interviews to study nuclear power plant operators. He has a three-stage typology of skills (sensori-motor, rule-based, and knowledge-based) which highlights how differential expertise creates differences in decision strategy. Hammond, Hamm, Grassia, and Pearson (1987) studied highway engineers and found that intuitive decision strategies were more effective for tasks such as judging aesthetic qualities of a road; while analytical strategies were more valuable for other tasks, such as estimating amount of traffic.

Pennington and Hastie (1987) studied jury deliberation as a complex decision task and found that the jurors attempted to fit all of the evidence into a coherent account of the incident. Their assessment was then based on this account or story, rather than on likelihood judgments of the evidence introduced. The jurors focused on whether the prosecution's or defense's story fit more easily into a major activity. The work of Noble (in press) with Naval Command-and-Control officers and Lipshitz (in press) with infantry soldiers has generated the same conclusions--under operational conditions, decision makers rarely use analytical methods, and nonanalytical methods can be identified that are flexible, efficient, and effective.

Our own work shows how people can make effective decisions without performing analyses. For the past several years, we have been studying command-and-control decision making and have generated a recognitional model of naturalistic decision making. We began by observing and obtaining protocols from urban fireground commanders (FGCs) who are in charge of allocating resources and directing personnel. We studied the decisions they made in handling non-routine incidents during emergency events. Some examples of the types of decisions these commanders had to make included whether to initiate search and rescue, whether to initiate an offensive attack or concentrate on defensive precautions, and where to allocate resources.

The FGCs' accounts of their decision making did not fit into a decision-tree framework. The FGCs argued that they were not "making choices," "considering alternatives," or "assessing probabilities." They saw themselves as acting and reacting on the basis of prior experience; they were generating, monitoring, and modifying plans to meet the needs of the situations. We found no evidence for extensive option generation. Rarely were even two options concurrently evaluated. We could see no way in which the concept of optimal choice might be applied. Moreover, it appeared that a search for an optimal choice could stall them long enough to lose control of the operation altogether. The FGCs were more interested in finding an action that was "workable," "timely," and "cost effective."

Nonetheless, the FGCs were clearly encountering choice points during each incident. They were aware that alternative courses of action were possible, but insisted that they rarely deliberated about the advantages and disadvantages of the different options. Instead, the FGCs relied on their ability to recognize and appropriately classify a situation. Once they knew it was "that" type of case, they usually also knew the typical way of reacting to it. Imagery might be used to "watch" the option being implemented, to search for flaws, and to discover what might go wrong. If problems were foreseen, then the option might be modified or rejected altogether and the next most typical reaction explored. This mental search would continue until a workable solution had been identified.

We have described these strategies as a Recognition-Primed Decision (RPD) model. For this fireground task environment, a recognitional strategy appears to be highly efficient. The proficient FGCs we studied used their experience to generate a workable option as the first one considered. If they had tried to generate a large set of options, and then systematically evaluated these, it is likely that the fires would have gotten out of control before they could make any decisions.

The RPD model is presented in Figure 1. The simplest case is one in which the situation is recognized and the obvious reaction is implemented. A somewhat more complex case is one in which the decision maker performs some conscious evaluation of the reaction, typically using imagery to uncover problems prior to carrying it out. The most complex case is one in which the evaluation reveals flaws requiring modification, or the option is judged inadequate and rejected in favor of the next most typical reaction.

The model is characterized by the following features:

- Situational recognition allows the decision maker to classify the task as familiar or prototypical.

- The recognition as familiar carries with it recognition of the following types of information: plausible goals, cues to monitor, expectancies about the unfolding of the situation, and typical reactions.

- Options are generated serially, with a very typical course of action as the first one considered.

- Option evaluation is also performed serially to test the adequacy of the option, identify weaknesses and find ways to overcome them.

- The RPD model includes aspects of problem solving and judgment along with decision making.

- Experienced decision makers are able to respond quickly, by using experience to identify a plausible course of action as the first one considered rather than having to generate and evaluate a large set of options.

- Under time pressure, the decision maker is poised to act while evaluating a promising course of action, rather than paralyzed while waiting to complete an evaluation of different options. The focus is on acting rather than analyzing.

We do not propose the RPD model as an alternative to analytic approaches. Rather, we postulate that recognitional and analytical decision strategies occupy opposite ends of a decision continuum similar to the cognitive continuum described by Hammond et al. (1987). At one extreme are the conscious, deliberated, highly analytic strategies such as multi-attribute utility analysis and decision analysis. Slightly less analytic are noncompensatory strategies such as elimination-by-aspects. At the alternate end of the continuum are Recognition-Primed Decisions (RPD), which involve non-optimizing and non-compensatory strategies and require little conscious deliberation. The RPDs are marked by an absence of comparison among various options. They are induced by a starting point that involves recognitional matches that in turn evoke generation of the most likely action in the situation.

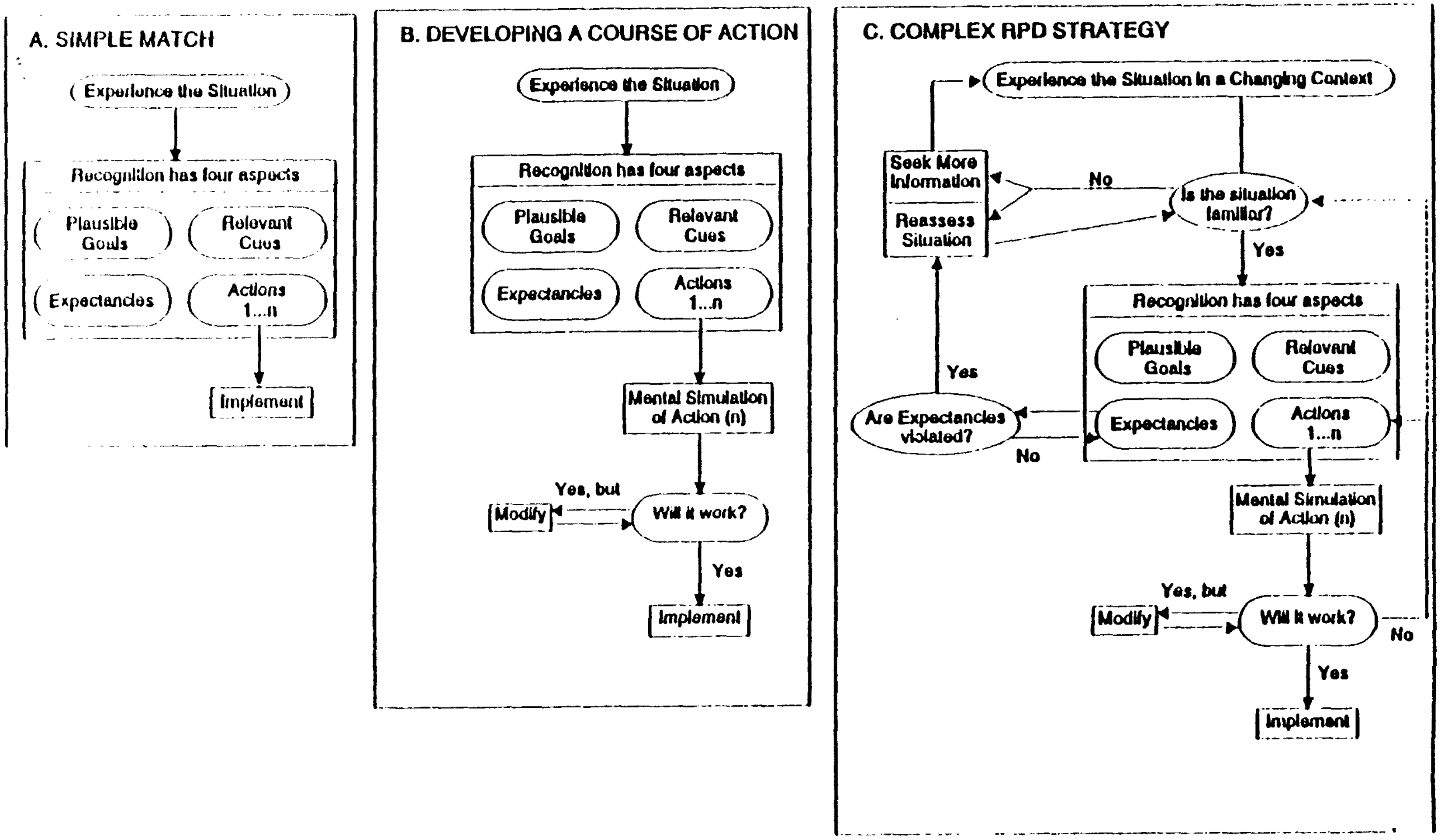


Figure 1. Recognition Decision Making model.

KLEIN Associates Inc.

We have tested applications of the model in a variety of tasks and domains, including fireground command, battle planning, critical care nursing, corporate information management, and chess tournament play. These studies have shown good support for the validity and utility of the model presented in Figure 1 as it applies to individual decision makers. Our coding was evaluated as having 87% to 94% inter-rater reliability.

What are the implications of the naturalistic decision-making approach? A workshop was held in Dayton, Ohio, in Fall 1989, to take stock of the current state of knowledge and to explore implications and future research directions. The workshop brought together researchers who have been active in naturalistic decision making. In attendance were 31 professionals who represented decision research being conducted by the military, NASA, private firms, and academic institutions. The domains studied spanned tactical operations, medical decision making, weather forecasting, nuclear power plant control, and executive planning, among others. This workshop was sponsored by the Army Research Institute (ARI) which began a research program in 1985 on Planning, Problem Solving, and Decision Making. The goal of this program is to make decision research more relevant to the needs of the applied community.

The Dayton workshop enabled researchers working with different domains and paradigms to find commonalities and to identify remaining questions. The participants have also contributed to a book "Decision Making in Action: Models and Methods" edited by Gary Klein, Judith Orasanu, Roberta Calderwood, and Caroline Zsambok (in press). A number of collaborative efforts were initiated, particularly directed at using our current models for guidance in developing training programs and decision support systems.

We are also exploring the implications of naturalistic decision making as it applies to military human-computer interface (HCI) design. For example, in the wake of the catastrophic shootdown of an Iranian airliner by the U.S. Vincennes, we have been studying the naturalistic decision-making requirements of the Combat Information Center (CIC) in the AEGIS Combat System (ACS). To develop effective decision aids and human-computer interfaces in such environments, we must understand the decisions that the intended operators will have to make, and the various pieces of information they will need to make those decisions.

We have developed an approach to Cognitive Task Analysis (CTA) that is designed to identify the naturalistic decision requirements of the tasks in a CIC. Decision requirements include the decisions that system users must make, the strategies they invoke to make these decisions, and the cues essential for making these decisions. The resulting understanding of cognitive performance affects how computer interfaces augment and support the decision making of the user.

Our approach to CTA is to collect critical incidents from people highly experienced with the AEGIS CIC. The critical incidents are analyzed with the objective of identifying the decisions made by the system operators in each critical incident, how CIC personnel made these decisions, the cues and factors that affected the decisions and any relationships between cues that were important. The analysis of the critical incidents is important to the design of the HCI. For example, if relationships between cues are important to decision making, the

HCI should enable the operator to perceive the relationship easily or display the relationship directly. In the CIC example, the relationship between vertical airspeed and horizontal airspeed is important. An unidentified aircraft that is descending and accelerating is often considered a hostile threat. It is, therefore, important to communicate this relationship to the operator in a non intrusive, yet visually salient way in order to support naturalistic decision making.

The intent of this paper has been to describe a decision process used frequently in naturalistic settings. The paper also presents data about the conditions favoring Recognition-Primed Decisions, and implications for applied decision research.

References

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. IEEE Transactions on Systems, Man, and Cybernetics, SMC-17(5), 753-770.

Klein, G. A. (1989). Recognition-primed decisions. In W. B. Rouse (Ed.), Advances in Man-Machine System Research, 5, 47-92. Greenwich, CT: JAI Press, Inc.

Klein, G. A., Orasanu, J., Calderwood, R., & Zsambok, C. E. (Eds.) (in press). Decision making in action: Models and methods. Norwood, NJ: Ablex Publishing Corporation.

Lipshitz, R. (in press). Converging themes in the study of decision making in realistic settings. In G. A. Klein, J. Orasanu, R. Calderwood, and C. E. Zsambok (Eds.), Decision making in action: Models and methods. Norwood, NJ: Ablex Publishing Corporation.

Noble, D. (in press). A model to support development of situation assessment aids. In G. A. Klein, J. Orasanu, R. Calderwood, and C. E. Zsambok (Eds.), Decision making in action: Models and methods. Norwood, NJ: Ablex Publishing Corporation.

Pennington, N., & Hastie, R. (1987). Explaining the evidence: Further tests of the Story Model for juror reasoning. Unpublished manuscript, University of Colorado, Boulder, CO.

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decision making and system management. IEEE Transactions on Systems, Man and Cybernetics, SMC-15(2), 234-243.

Evaluating Decisions

Dr. Lawrence J. Fogel
ORINCON Corporation

Success of the military depends on effective decision making. But what does "effective" mean? How can we measure the worth of individual decisions, regardless of the subject matter?

A decision occurs whenever one of a number of options is selected in order to achieve some purpose. The worth of that decision is then the degree to which that purpose is achieved, for means have value *only* in terms of ends. The evaluation of any decision must therefore begin by asking the decision maker to indicate his specific intent at the time the choice was made.

This is not a simple matter, for all too often, we simply focus on what must be achieved. But what if the most desired outcome isn't realized? What is the worth of lesser degrees of achievement. And what about that future that is to be avoided at all cost? In other words, purpose becomes well defined only when there is a statement of each of the significantly different futures and their relative worth. But these futures may be difficult to discern, no less distinguish in relative importance. Rather than search for such alternative entireties it is far more convenient to express purpose in the form of a Valuated State Space and normalizing function. The dimensions of the Valuated State Space are the parameters of concern. Each is made measurable in terms of class intervals that designate the significantly different degrees of achievement. Each class interval is then attributed some value for that degree of achievement. A normalizing function indicates the manner in which contributions on the various parameters can be combined into a single overall worth.

To illustrate, suppose three dimensions are of concern: X, Y, and Z. These dimensions have relative importance, say, six, nine, and two, respectively. As depicted in Table 1, these are the coefficients. The horizontal lines following each dimension are partitioned to indicate the class intervals that correspond with the significantly different degrees of achievement. The thresholds identify the limits of these mutually exclusive class intervals. Usually, the number of class intervals on a parameter reflects its relative importance; that is to say, the more important the dimension, the greater the specificity of its measure. A parameter of little concern may be adequately specified in a binary sense. The value attributed to being in each class interval is indicated on a ratio or magnitude scale. For convenience, the class intervals are arranged from left to right in order of decreasing worth.

Table 1. A Valuated State Space, the Check Marks Indicating the Current Situation

6	X	3	1 √	0		
9	Y	10	9 √	2	1	0
2	Z	10	0 √			

The number of possible states is then the product of the number of class intervals on each of the dimensions, in this case, 30. Achieving the most valuable class interval on each of the dimensions corresponds with being in the state of highest overall worth; thus, a measure of 1.0 (or 10 on a 10 scale, 100 on a percent scale). Achieving no success on any dimension corresponds with an overall worth of zero.[1] Any intermediate state has some worth depending upon the particular combinatory function. For example, using the weighted arithmetic mean, the situation indicated by the profile of check marks shown in Table 1 has a worth of

[1] In situations where a negative worth is ordinarily associated with some class intervals, the scale can be linearly transformed so that the worst possible degree of achievement has zero value.

$$W_o = \left(\frac{1}{3}\right)\left(\frac{6}{17}\right) + \left(\frac{9}{10}\right)\left(\frac{9}{17}\right) + \left(\frac{0}{10}\right)\left(\frac{2}{17}\right) = 0.59 \text{ or } 59\% \qquad (1)$$

Suppose the testing of some weapon system concerns Reliability, Maintainability, Ease of Operation and Resistance to Enemy Countermeasures, these having relative importance weights of ten, eight, three and six, respectively. Table 2 indicates the class intervals that correspond with the significantly different degrees of achievement. The check marks indicate a current (or future) state.

Table 2. The Valuated State Space for a Weapon System

10		Reliability:	
		10	≥ 99%
		09	≥ 95% but > 99%
		08	≥ 90% but < 95%
	√	06	≥ 80% but < 90%
		04	≥ 60% but < 80%
		01	≥ 40% but < 60%
		00	< 40%
08		Maintainability:	
		10	Easily maintained at field level by operational personnel
		07	Maintainable at field level by unit military technicians
	√	05	Maintainable at field level with the assistance of specialized military technicians
		03	Skilled civilian field technicians required for all but minor maintenance procedures
		01	Must be returned to factory or depot for all but minor, routine maintenance procedures
		00	Non-repairable
03		Ease of Operation:	
		10	Routinely operated by regular troops
		08	Operated by regular troops directed by specially trained supervisor
	√	04	Operated by specially trained crew
		01	Operated only by highly technical and highly skilled personnel
		00	Inoperable
06		Resistance to Enemy Countermeasures	
		10	Not susceptible to countermeasure
		08	Contains self-contained capability to defeat enemy countermeasures
	√	05	Requires special support forces to suppress enemy countermeasures
		03	Highly degraded by sophisticated countermeasures
		00	Easily defeated by simple countermeasures

Taking the weighted arithmetic mean, the overall worth of this situation is,

$$W_o = \left(\frac{6}{10}\right)\left(\frac{10}{27}\right) + \left(\frac{5}{10}\right)\left(\frac{8}{27}\right) + \left(\frac{4}{10}\right)\left(\frac{3}{27}\right) + \left(\frac{5}{10}\right)\left(\frac{6}{27}\right) = 0.\ 5259 \text{ or } 52.\ 6\% \qquad (2)$$

Now improving the Reliability from, say, the 4th to the 5th class interval (i.e., from between 80% and 90% to between 90% and 95%), increases the overall worth to:

$$W_o = \left(\frac{8}{10}\right)\left(\frac{10}{27}\right) + \left(\frac{5}{10}\right)\left(\frac{8}{27}\right) + \left(\frac{4}{10}\right)\left(\frac{3}{27}\right) + \left(\frac{5}{10}\right)\left(\frac{6}{27}\right) = 0.\ 6000 \text{ or } 60\% \qquad (3)$$

It is often useful to explicate each of the dimensions in terms of lower level indicators, and then in terms of still lower level measures. Class intervals and attributed values are only assigned to the lowest level dimensions of such a hierarchic Valuated State Space.

For any given situation, the remaining deficiencies (the problems still to be addressed) are evident and clearly prioritized[2]. The overall worth of resolving these to any specific degree can be readily measured.

In many situations, achievement on any dimension contributes to the overall worth. In others, all the parameters are critical; that is, failure in any single regard corresponds with no overall worth whatsoever. In this case, it is appropriate to use the weighted geometric mean. Other more complex normalizing functions may be required to reflect the manner in which the parameters contribute to the overall worth. In any case, the Valuated State Space and normalizing function is a convenient, multiattribute utility function that defines the significantly different situations and corresponding overall measures of worth.

But why bother with the Valuated State Space when the testing requirements have already been established? The stated requirements are certainly worthy of reference, but it is dangerous to accept these without additional understanding. Are the stated requirements a complete set, or might additional requirements be added at a later date? Are all the testing requirements of equal importance? If some are more important than others, which ones are these? How much more important are they?

Quite separately, are any or all of these requirements critical; that is, does a lack of achievement in any specific regard nullify the worth of the system? If only certain parameters are critical, which ones are these? Are some partially critical? If such a parameter is not met, how bad is the effect? What about failing to meet two such requirements, and so forth?

Requirements are usually stated in binary form. Either they are met or not met. What if a requirement is exceeded? Is the greater achievement worth something, nothing, or was the extra effort devoted to attaining that additional achievement a waste of effort at some cost?

Do the requirements consistently indicate thresholds that define utopia, or do these specify the very lowest level of acceptability? Are some requirements thresholds of utopia while others define some lower level of acceptability? Why not have two or more requirements on the same parameter, each corresponding to different levels of required achievement? Surely, such greater specificity would make the stated purpose more complete and meaningful. Indeed, this is realized by defining the thresholds and thus the class intervals on the parameters of the Valuated State Space.

All too often, weapon system requirements have been technology-driven rather than threat-driven. That is, they were generated primarily to ensure technological progress rather than to meet a specific operational need. If the stated requirements are threat-driven, what are the specific anticipated missions? For what time frames? How important are these time frames? What is the relative importance and likelihood of each of the missions? How are these missions

[2] Future situations are uncertain. Each class interval has some estimated probability. The expected value is then the contribution of that parameter.

combined into a single composite intent? What is the presumed mission and level of sophistication of the enemy? Might he be intelligently interactive during the mission? Are there other enemies to be countered? The stated requirements can be properly interpreted and integrated into the Valuated State Space in a meaningful manner if, and only if, these and the related affordability issues have been properly framed.

In terms of testing, do the requirements reflect a real need, or were they established to yield some particular level of product acceptance? If the former be true, how were the requirements generated? If the latter, why that particular level of acceptance? And certain other questions remain to be answered. In point of fact, it would be particularly useful to frame the testing requirements in the form of a Valuated State Space and normalizing function.

This discussion has not presumed any particular mental model or mode of decision making. It does, however, suggest that thoughtful decisions begin with an understanding of the intent to be realized, then proceed to reference the alternative options (combinations of the allocable resources), projecting these into future situations that can be measured in worth by reference to the purpose. The best is then selected for implementation. Errors in judgment can arise in each of these regards. Valuable options may have been overlooked, the forecasting may be in error, the outcomes may not have been properly scored. The best option may not have been selected. But most important of all, the purpose may have been ill-defined or, worse yet, left undefined.

The decision process can be improved by adopting this kind of intellectual rigor. Testing decisions deserve such a top-down approach.

Author Index