



**ZeE – Publikationen**

Reihe empirische  
Evaluationsmethoden

Band 23

---

# **Empirische Evaluationsmethoden Band 23**

## **Workshop 2018**

Herausgegeben vom Zentrum für empirische Evaluationsmethoden e. V.  
Berlin 2019

Die Reihe empirische Evaluationsmethoden ist eine Dokumentation von Beiträgen der jährlichen Workshops des Zentrums für empirische Evaluationsmethoden e.V.. Ziele sind die gemeinsame Diskussion von Problemen der Evaluationsforschung und ihrer Anwendungen, die Entwicklung und Verbreitung neuer methodischer Zugänge und die Weiterbildung in diesen Bereichen. Die Beiträge sind ausschließlich für diese Zielstellung bestimmt, jede Art der Vervielfältigung ist untersagt. Die Reihe empirische Evaluationsmethoden kann von Interessenten beim ZeE gegen eine Gebühr von 5.00 € je Band erworben werden.

2019. ZeE Verlag: Berlin

1. Auflage

Zentrum für empirische Evaluationsmethoden e.V.

Homepage: <http://www.z-e-e.de>

E-Mail: [vorstand@z-e-e.de](mailto:vorstand@z-e-e.de)

ISSN: 1618-5552



**ZeE – Publikationen**

Reihe empirische  
Evaluationsmethoden

Band 23

---

**Empirische Evaluationsmethoden  
Band 23**

**Workshop 2018**

Herausgeber:

Reinhard Beyer, Bodo Krause, Jens Nachtwei



# Empirische Evaluationsmethoden

## Band 23

### Inhaltsverzeichnis

		Seite
Gerlach, R., Beyer, R., van der Meer, E., Nimke-Sliwinski, B., Foerges, R.	Evaluation von zielgruppenspezifischen Antilittering-Maßnahmen im Feld mit Hilfe des Einsatzes von Nudging	5
Hochstrate, M	Schlüssel-Metriken im Personalmarketing	23
Fahrner, M., Wolf, B.	Untersuchung des Einflusses von Mathematikkompetenzen auf Studienerfolg in den Wirtschaftswissenschaften	31
Beyer, M L., Großmann, I., Schüttpelz-Brauns, K.	Varianzanalytische Auswirkungen faktorenanalytischer Skalenkonstruktionen	41
Krause, B.	Zur Stabilität von Zuverlässigkeitsmaßen	49
Bieniok, M., Reich, A., Hesse, P. U.	Kann die Blickrichtung die emotionale Qualität der Erinnerung von belastenden oder traumatischen Ereignissen verändern?	57
	Anschriften der Autoren	79
	bereits erschienen	81



## **Evaluation von zielgruppenspezifischen Antilittering-Maßnahmen im Feld mit Hilfe des Einsatzes von Nudging**

Rebekka Gerlach, Elke van der Meer, Birgit Nimke-Sliwinski, Rainer Foerges & Reinhard Beyer

### **Zusammenfassung**

Ein neuer Akzent der letzten eigenen Studien lag in der besonderen Beachtung der Meinung jüngerer Stadtbewohner. Wie sich zeigte, nehmen Kinder und Jugendliche Sauberkeitsmängel im öffentlichen Raum sensibler wahr als Erwachsene, gleichzeitig sind sie erzieherischen (Antilittering-)Maßnahmen gegenüber eher skeptisch. Deshalb wurde in der aktuellen Evaluationsstudie ein besonderes Augenmerk darauf gelegt, die zu evaluierenden Maßnahmen in Form und Inhalt modern und explizit reaktanzvermeidend zu gestalten, um auf diese Weise auch Kinder und Jugendliche anzusprechen. Ziel der vorliegenden Littering-Feldstudie ist die gezielte Überprüfung einer Antilittering-Maßnahme und einer Maßnahmenkombination. Ausgewählt wurde eine Maßnahme zur Aufmerksamkeitsfokussierung (grüne Fußspuren Richtung Abfallbehälter, sog. Nudging) und die Begleitung dieser Maßnahme durch eine Plakatkampagne, die über eine moderne Ansprache insbesondere zur Sensibilisierung von Jugendlichen und jungen Erwachsenen beitragen soll, also einer spezifischen relevanten Zielgruppe im Kontext von Littering. Die Studie wurde in drei Untersuchungsgebieten in Köln durchgeführt. Erhoben wurden die subjektive Einschätzung der Sauberkeit der Straßenzüge sowie das objektive Littering-Aufkommen in diesen Gebieten zu mehreren Messzeitpunkten. Im Ergebnis konnte unter Verwendung dieser subjektiven und objektiven Kennwerte ein positiver Effekt der Maßnahmen nachgewiesen werden, insbesondere dann, wenn eine Kombination beider Maßnahmen zum Einsatz kam.

### **Schlüsselwörter**

Littering, Sauberkeit, Prävention, Intervention, Evaluation

### **1. Ausgangspunkt und Zielstellung**

In Anlehnung an die 2009 realisierte Feldstudie wurden unter Einbeziehung neuer Erkenntnisse der bisherigen Replikationsstudien (bspw. Spezifik der Ansprache von Jugendlichen und jungen Erwachsenen) zwei Antilittering-Maßnahmen für die Evaluation in einer Feldstudie ausgewählt:

1. Einsatz von Nudges (grüne Fußspuren auf dem Gehweg in Richtung Abfallbehälter) und
2. die Begleitung dieser Maßnahme durch eine Plakatkampagne.

Kriterien für die Auswahl der beiden Maßnahmen waren neben ihrer Relevanz in den bisher durchgeführten Befragungen die Möglichkeit einer kostengünstigen und zeitnahen Durchführung sowie die Angemessenheit der konkreten Maßnahmen in einer der beteiligten Städte. Auf dieser Grundlage wurde als Untersuchungsort Köln bzw. der Verantwortungsbereich der AWB Köln festgelegt.

Ziel der Studie ist eine differenzierte Bewertung der Maßnahmen „Einsatz von Nudges“ (Fußspuren in Richtung Abfallbehälter) und „Plakatkampagne“ in einer Feldstudie in Köln. Ein Effekt der Einzelmaßnahmen und/ oder der Maßnahmenkombination sollte sich in einer verbesserten subjektiven Beurteilung des Sauberkeitszustandes der betreffenden Straßenzüge durch die Bewohner äußern sowie in einer Verminderung des objektiven Litteraufkommens.

## 2. Untersuchungsplanung

Im Jahr 2009 fand im Rahmen des Projektes „Littering – Merkmale, Ursachen & Prävention“ die erste Maßnahmenevaluation statt (siehe van der Meer et al. 2013, VKS Information 80). Hier konnte gezeigt werden, dass eine auffälligere Gestaltung von Abfallbehältern sowohl zu einer Reduktion des Litteringverhaltens als auch zu einem verbesserten subjektiven Sauberkeitseindruck führt. Beide Effekte verstärkten sich durch die Kombination mit einer Plakatkampagne, die gezielt auf die neugestalteten Abfallbehälter fokussierte und die Mitverantwortung der Bürger für die Sauberkeit der Stadt ansprach.

Die aktuelle Feldstudie sollte unter vergleichbaren Bedingungen wie im Jahr 2009 stattfinden. Entsprechend dieser Maßgabe erfolgten die Auswahl der Untersuchungsorte (dieselben drei Straßenzüge in Köln wie 2009) und die Auswahl der zu evaluierenden Maßnahmen (aufmerksamkeitsfokussierende Maßnahme in Kombination mit sensibilisierender Plakatkampagne).

Die ausgewählten Antilittering-Maßnahmen und Untersuchungsstandorte sowie das methodische Vorgehen werden im Folgenden beschrieben und erläutert.

### 2.1 Auswahl der Maßnahmen

Wie oben bereits erwähnt, wurden zwei Antilittering-Maßnahmen für die Evaluation in dieser Feldstudie ausgewählt: Einsatz von Nudges (grüne Fußspuren auf dem Gehweg in Richtung Abfallbehälter) und die Begleitung dieser Maßnahme durch eine für das Thema „Littering“ sensibilisierende Plakatkampagne.

**Einsatz von Nudges (grüne Fußspuren Richtung Abfallbehälter):** Den Einsatz sog. Nudges ("Anstuber") fanden viele Experten der Reinigungsunternehmen hilfreich, um Passanten an korrektes Abfallentsorgungsverhalten zu erinnern „Nudges“ lenken das Verhalten von Menschen nicht direkt. Sie machen also keine Verhaltensvorschläge (oder verbieten ein Verhalten). Sie lenken lediglich die Aufmerksamkeit auf das erwünschte Verhalten, machen es salient (es tritt ins Bewusstsein)(Thaler & Sunstein 2009).

Nudges sind beispielsweise auffällig farbige Fußspuren auf dem Gehweg, die zum Abfallbehälter führen. Diese erleichtern zum einen das korrekte Abfallentsorgungsverhalten, indem sie konkret auf die Existenz eines Abfallbehälters hinweisen (der sog. "making things easy" part of the nudging-doctrine; siehe Jespersen 2012; Hansen & Jespersen 2013). Andererseits - und dies bezeichnen Hansen & Jespersen als entscheidend für die Initiierung des gewünschten Verhaltens - aktivieren sie die soziale Norm des korrekten Abfallentsorgungsverhaltens.

In Absprache mit der AWB wurde beschlossen, etwa die Hälfte der Abfallbehälter in den Experimentalgebieten jeweils mit oben beschriebenen Nudges auszustatten. Die entsprechenden Abfallbehälter wurden mit Fußspuren auf dem Gehweg in



auffälligem Grün in Richtung Abfallbehälter gekennzeichnet. Abbildung 1 zeigt ein Beispiel dieser Nudges.



**Abb. 1:** Einsatz von Nudges (Fußspuren Richtung Abfallbehälter) in Köln in den Experimentalgebieten 1 und 2

Der Einsatz von Nudges dieser Art (grüne Fußspuren auf dem Gehweg Richtung Abfallbehälter) besitzt den Vorteil, dass er bei Erfolg als Antilittering-Maßnahme mit relativ wenig Aufwand von anderen Städten nachgenutzt werden kann.

**Gestaltung der Plakate für eine Kampagne:** Plakatkampagnen wurden in allen bisherigen Studien als sehr wirksam eingeschätzt, insbesondere dann, wenn kurz- und mittelfristige Effekte erzielt werden sollen. Deshalb fiel die Entscheidung bei der Entwicklung einer zweiten Maßnahme zugunsten dieser Maßnahmenvariante. Diese zweite Maßnahme sollte zudem eine inhaltlich sinnvolle Ergänzung der ersten Maßnahme ermöglichen (aber prinzipiell auch unabhängig einsetzbar sein). Folglich wurden die Nudges (Fußspuren Richtung Abfallbehälter) als zentraler Bildgegenstand gewählt, die Plakate sollten direkt auf die eingeführte Neuerung hinweisen. Hinzu kam eine kurze verbale humorakzentuierte Ansprache der Passanten. Studien zeigen (siehe Bode 2016; Kaiser 2017), dass knapp formulierte, humorvolle Botschaften die Wirksamkeit von Plakatkampagnen positiv beeinflussen. Zudem werden solche Botschaften besser gespeichert als sachlich-neutrale.

Hinzu kommt, dass der Slogan „Beifall für Abfall“ ein wortspielerisches Spannungselement setzt, das im Folgenden durch den Nachsatz „Klar, wenn er reinfällt.“ aufgelöst wird. Solche Spannungselemente (siehe Jarren & Röttger 2015; Bode 2016) fördern die Bereitschaft, sich auf ein Thema einzulassen (in diesem Fall korrektes Abfallentsorgungsverhalten) und können somit Einstellungs- und

Verhaltensänderungen anstoßen. Durch diese Art der humorvollen Werbung für korrektes Abfallentsorgungsverhalten sollen vor allem Jugendliche und junge Erwachsene (die am häufigsten litternde Personengruppe; siehe Litteringstudie 2016) angesprochen werden. In der Litteringstudie 2016 zeigte sich, dass insbesondere Jugendliche zwar besonders sensibel auf Verschmutzungen reagieren, aber durch sensibilisierende bzw. erzieherische Maßnahmen schwerer zu erreichen sind. Die aktuelle Plakatkampagne ist daher explizit reaktanzvermeidend (jegliche Belehrungen werden vermieden) konzipiert. Stattdessen wird das korrekte Abfallentsorgungsverhalten auf humorvolle Weise beworben. In Abbildung 2 ist die Plakatvariante dargestellt, die in der Feldstudie zum Einsatz kam.



**Abb. 2:** Plakat mit Hinweis auf die Fußspuren Richtung Abfallbehälter und humorakzentuiertem (wortspielerischem) Slogan mit Hinweis auf korrektes Abfallentsorgungsverhalten

## 2.2 Auswahl der Untersuchungsstandorte

Um die Wirkungsmessung der zu evaluierenden Maßnahmen unter vergleichbaren Bedingungen wie 2009 durchführen zu können, wurden dieselben Straßenabschnitte gewählt wie 2009:

- Kalker Hauptstraße (Viektorstraße bis Breuerstraße, 492 Meter),
- Venloer Straße (Ehrenfeldtgürtel bis Gutenbergstraße, 568 Meter),
- Neusser Straße (Kempener Straße bis Florastraße, 497 Meter).

Grundsätzlich ist für die Auswahl der Untersuchungsgebiete von Bedeutung, dass diese in möglichst vieler Hinsicht ähnlich sind (z.B. Sozialstruktur, Frequentierung, Art des Littering, Verschmutzungsgrad, Personengruppe der Litterer). Schon 2009 zeigte sich, dass es nicht möglich war, alle Kriterien gleichermaßen zu berücksichtigen. Letztlich fiel die Entscheidung auf vergleichbar große Abschnitte in den oben genannten Straßenzügen, da sich diese Gebiete die in Hinblick auf Frequentierung, Art des Littering (Einkaufsstraßen mit Verweilgelegenheiten) und Verschmutzungsgrad ähnlich sind. Einen wichtigen Aspekt der Vergleichbarkeit stellt die Anzahl der Abfallbehälter in den Testgebieten dar. Die Ausstattung mit Abfallbehältern in den drei Untersuchungsgebieten fällt sehr ähnlich aus: In der Venloer Straße sind es 21 Doppelpapierkörbe, in der Neusser Straße 25 Doppelpapierkörbe und in der Kalker Hauptstraße 21 Doppelpapierkörbe.

Die Neusser Straße wurde als Kontrollgebiet definiert, das heißt, hier wurden über die verschiedenen Untersuchungszeitpunkte (Studie 1 bis 3) keine Veränderungen vorgenommen. Die Venloer Straße stellte das Experimentalgebiet 1 dar. Hier wurden vor dem zweiten Untersuchungszeitpunkt grüne Fußspuren vor den Abfallbehältern schabloniert (Maßnahme 1). Als Experimentalgebiet 2 diente die Kalker Hauptstraße. Hier wurden ebenfalls vor dem zweiten Untersuchungszeitpunkt grüne Fußspuren vor den Abfallbehältern schabloniert (Maßnahme 1) und vor dem dritten Untersuchungszeitpunkt zusätzlich die oben beschriebene Plakataktion (Maßnahme 2) durchgeführt.

### **2.3 Auswahl der Methoden zur Messung der Sauberkeit und des Litteraufkommens**

Zur Prüfung der Effekte der Maßnahmen wurde die subjektive Beurteilung der Sauberkeit erfasst und die objektive Sauberkeit sowie das Litteraufkommen gemessen. Zur Erfassung des subjektiven Sauberkeitseindrucks wurden Fragebogentechniken eingesetzt (Sauberkeitseindruck, Wahrnehmung von Mängeln, Störungsempfinden, Beurteilung der Maßnahmen und wahrgenommene Verbesserungen). Zur Erhebung des objektiven Sauberkeitszustandes wurden u.a. DSQS-Analysen (computergestützte Erfassung des Sauberkeitszustandes) und die Messung des Litteraufkommens (korrekt entsorgte Menge des Abfalls in Abfallbehältern im Verhältnis zur gelitterten Menge (Straßenkehricht)) realisiert. Die Befragungen bzw. Messungen erfolgten zu Beginn der Untersuchung (Ausgangswert, Prä-Untersuchung), 1 Monat nach dem Setzen der Nudges (Post-Untersuchung 1) und einen Monat nach dem Start der Plakatkampagne (Post-Untersuchung 2).

Insgesamt nahmen ca. 1000 Probanden an den Befragungen teil. Pro Untersuchungsgebiet und Untersuchungszeitpunkt wurden ca. 110 Personen befragt. Das Durchschnittsalter der Probanden lag bei 33,8 Jahren. Der jüngste Proband zum Zeitpunkt der Erhebung war 10 Jahre alt, der älteste 82.

### **2.4 Ablauf der Untersuchung**

Aus den bisherigen Erläuterungen ist die Grundstruktur der Untersuchung bereits deutlich geworden:

- **Auswahl von drei Untersuchungsgebieten:** Neusser Straße (Kontrollgebiet ohne Maßnahme, Venloer Straße (Experimentalgebiet 1: Nudges/ grüne Fußspuren Richtung Abfallbehälter) und Kalker Hauptstraße

(Experimentalgebiet 2: Nudges/ grüne Fußspuren Richtung Abfallbehälter und Einführung einer Plakatkampagne).

- **Auswahl von zwei Maßnahmen gegen das Littering:** Einsatz von Nudges (grüne Fußspuren Richtung Abfallbehälter), Plakatkampagne.

Aus dieser Konstellation ergaben sich für den Ablauf der Studie drei Untersuchungsabschnitte:

- **Prä-Untersuchung:** Erfassung des Ausgangszustandes bezüglich der Sauberkeit in allen drei Untersuchungsgebieten (subjektive Bewertung, DSQS, Erfassung Litteraufkommen).
- **Post-Untersuchung 1:** Erfassung des Sauberkeitszustandes in allen drei Untersuchungsgebieten nach dem Einsatz von Nudges (grüne Fußspuren) in Experimentalgebiet 1 und 2 (subjektive Bewertung, DSQS, Erfassung Litteraufkommen).
- **Post-Untersuchung 2:** Erfassung des Sauberkeitszustandes in allen drei Untersuchungsgebieten nach zusätzlicher Einführung der Plakatkampagne in Experimentalgebiet 2 (subjektive Bewertung, DSQS, Erfassung Litteraufkommen).

### 3. Hypothesen

Aus dem Untersuchungsdesign ergeben sich zwei generelle Auswertungsvarianten:

- Verlauf des Sauberkeitszustandes innerhalb eines Untersuchungsgebietes,
- Sauberkeitszustand im Experimentalgebiet jeweils im Vergleich zum Kontrollgebiet.

Bei einem Vergleich innerhalb des Gebietes würde man beispielsweise beim Kontrollgebiet keine Veränderung erwarten, während die stärkste Veränderung bei Experimentalgebiet 2 auftreten sollte (2 Maßnahmen). Beim Vergleich zwischen den Gebieten sollten die Experimentalgebiete ab Untersuchungszeitpunkt 2 gegenüber dem Kontrollgebiet in den subjektiven und objektiven Parametern besser abschneiden. Letzteres setzt allerdings eine Gleichwertigkeit der drei Untersuchungsgebiete bei der Ausgangswertmessung voraus. Da diese nicht in ausreichendem Maße gegeben ist, konzentriert sich der Schwerpunkt der Auswertung auf den Vergleich der Ergebnisse zu den verschiedenen Untersuchungszeitpunkten und richtet das Augenmerk damit auf Veränderungen innerhalb der jeweiligen Gebiete.

### 4. Ergebnisse

Die Ergebnisse werden in drei Abschnitten dargestellt: 1. Ergebnisse zur subjektiven Sauberkeitsbewertung, 2. Ergebnisse zur objektiven Sauberkeitsbewertung und 3. Ergebnisse zur Einschätzung der Arbeit des Kölner Reinigungsunternehmens AWB.

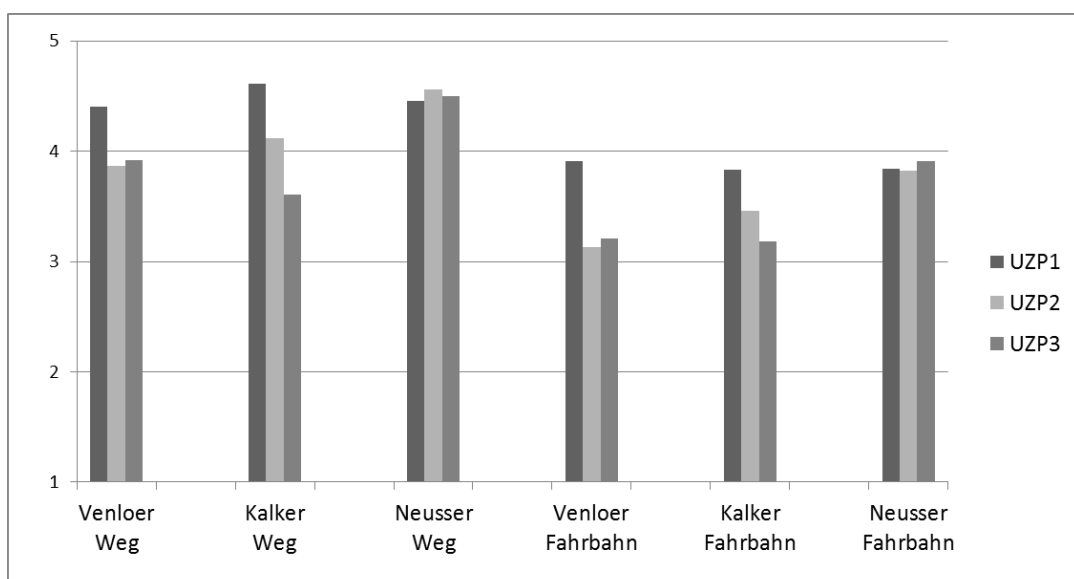
#### 4.1 Ergebnisse zur subjektiven Sauberkeitsbewertung

Von besonderem Interesse war, wie die Bewohner die Sauberkeit in den untersuchten Stadtgebieten einschätzen und ob die Maßnahmen zu einer Verbesserung des Sauberkeitseindrucks geführt haben. In Tabelle 1 werden die mittleren Sauberkeitsurteile für die untersuchten drei Straßen zusammengestellt. Es zeigt sich, dass bereits durch die erste Maßnahme, das Anbringen von Nudges, eine Verbesserung der Einschätzung der Sauberkeit in den beiden Experimentalgebieten erzielt werden konnte, während sich im Kontrollgebiet keine Veränderung zeigt. Durch die Ergänzung der Plakatkampagne in Experimentalgebiet 2 kann sogar eine weitere Verbesserung des subjektiven Sauberkeitseindrucks erreicht werden.

**Tab 1:** Mittlere Urteile der Anwohner zur Sauberkeit für die drei Untersuchungsgebiete (Experimentalgebiet 1 (EG1): nur Nudges, Experimentalgebiet 2 (EG2): Nudges und später Plakatkampagne und Kontrollgebiet (KG: keine Maßnahmen, jeweils für die drei Untersuchungszeitpunkte Prä, Post 1 und Post 2 (1= stimmt nicht, 7= stimmt))

		Prä	Post 1	Post2
Straßenzug	EG1	3,75	4,38	4,00
	EG2	3,59	3,97	4,07
	KG	4,20	4,23	3,91

Dieser Befund kann durch eine andere Bewertungsvariante noch unterstützt werden. Wenn die Probanden direkt nach wahrgenommenen Mängeln auf Gehwegen und Fahrbahnen in den drei Untersuchungsgebieten gefragt werden, zeigt sich nur in den beiden Experimentalgebieten eine Verbesserung, im Experimentalgebiet 2 (Kalker Hauptstraße) ist auch hier ein zusätzlicher positiver Effekt nach Einführung der Plakatkampagne nachweisbar. Dieser Befund wird in Abbildung 3 sehr gut sichtbar.



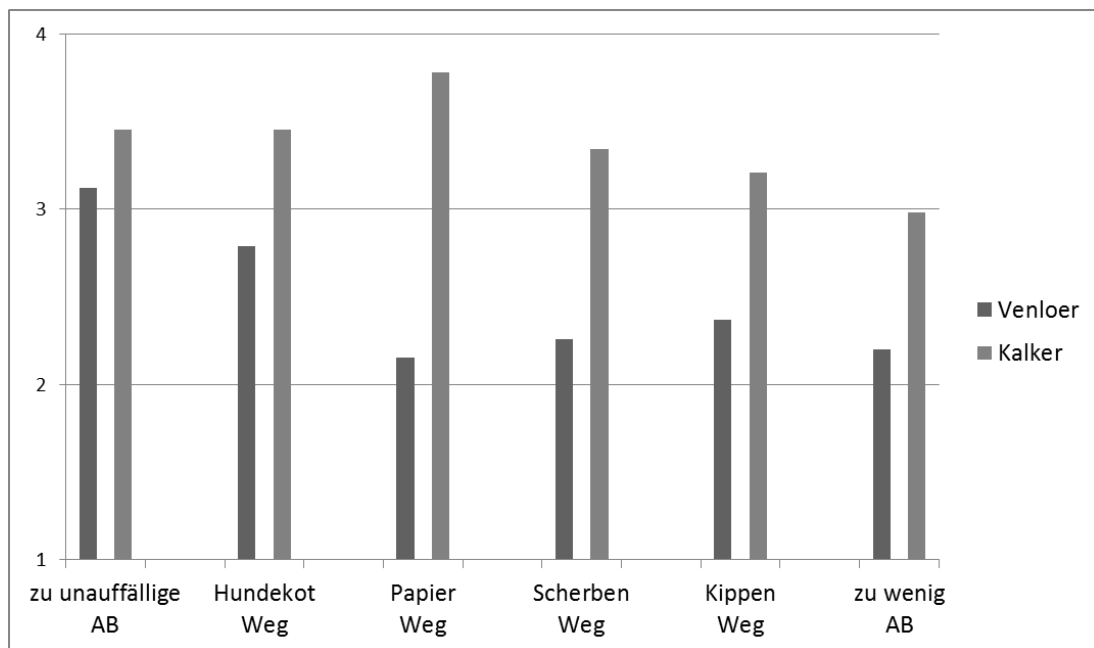
**Abb.3:** Mittlere Urteile zu wahrgenommenen Mängeln auf Gehwegen und Fahrbahnen in den drei Untersuchungsgebieten Venloer Straße (Experimentalgebiet 1), Kalker Hauptstraße (Experimentalgebiet 2) und Neusser Straße (Kontrollgebiet); 1= Mangel nicht vorhanden, 7= Mangel liegt stark vor; UZP= Untersuchungszeitpunkt

Da die Abfallbehälter direkt von den Maßnahmen betroffen waren, und Abfallbehälter bei den Stadtbewohnern in allen bisherigen Studien ein zentrales Kriterium für die Sauberkeitsbeurteilung waren, haben wir Merkmale von Abfallbehältern gezielt und separat bewerten lassen. In Tabelle 2 ist das Ergebnis für die zwei wichtigsten Merkmale „überfüllter Abfallbehälter“ und „zu wenig Abfallbehälter“ dargestellt. Es zeigt sich, dass sich durch die beiden Maßnahmen der Sauberkeitseindruck bei den Bewohnern deutlich verbessert. Bemerkenswert ist vor allem, dass die Einschätzung zur Anzahl der vorhandenen Abfallbehälter insbesondere nach Einführung der Plakatkampagne signifikant positiver bewertet wird. Und dies obwohl sich die Anzahl der Abfallbehälter nicht verändert hat. Offensichtlich wurde die Aufmerksamkeit der Bewohner durch die Maßnahmen wirksamer auf die Abfallbehälter gelenkt, wodurch sie besser wahrgenommen und vermutlich auch intensiver genutzt werden können.

**Tab. 2:** Mittlere Urteile zu wahrgenommenen Mängeln im Bereich Abfallbehälter (AB) für die beiden Experimentalgebiete (Experimentalgebiet 1 (EG1): nur Nudges, Experimentalgebiet 2 (EG2): Nudges und später Plakatkampagne), jeweils für die drei Untersuchungszeitpunkte Prä, Post 1 und Post 2 (1= Mangel nicht vorhanden, 7= Mangel liegt stark vor)

subjektiv Mängel AB		Prä	Post 1	Post 2
überfüllte AB	EG1	4,17	3,56	3,58
	EG2	4,79	4,56	4,20
zu wenige AB	EG1	3,46	3,31	3,39
	EG2	4,78	4,48	3,07

Um die Aussagen zur Verbesserung des Sauberkeitszustandes aus subjektiven Sicht noch sicherer zu machen, wurde eine weitere Frageperspektive gewählt: Die Probanden sollten am Ende des Untersuchungszeitraumes beurteilen, ob sie Verbesserungen bei verschiedenen Sauberkeitsmerkmalen erkannt haben.



**Abb.4:** Mittlere Urteile über wahrgenommene Verbesserungen in beiden Experimentalgebieten Venloer Straße (Experimentalgebiet 1, nur Nudges), Kalker Hauptstraße (Experimentalgebiet 2, Nudges und später Plakatkampagne) (1=Verbesserung nicht sichtbar bis 7=Verbesserung sehr stark sichtbar; AB=Abfallbehälter)

In Abbildung 4 sind die Ergebnisse zu den Merkmalen „Auffälligkeit und Anzahl Abfallbehälter“ sowie „Hundekot, Papier, Scherben und Kippen auf dem Weg“ zusammengestellt. Bei der Interpretation der Werte ist zu beachten, dass bereits ab einen Skalenwert von 2 Verbesserungen zumindest im Ansatz subjektiv sichtbar geworden sind. In allen oben genannten Kategorien sind in diesem Sinne Verbesserungen im Erscheinungsbild der Stadt bezüglich der genannten Merkmale aus Sicht der Bevölkerung erkennbar. Eine signifikante Steigerung des Effektes ergibt sich nach Einführung der Plakatkampagne in Experimentalgebiet 2 (Kalker Hauptstraße).

Außerdem war von Interesse, ob die Bewohner von Köln die eingesetzten Maßnahmen generell als sinnvoll empfunden haben und ob sie glauben, dass dadurch das eigene Litteringverhalten tatsächlich beeinflusst werden kann. In Tabelle 3 sind die entsprechenden mittleren Urteile eingetragen.

**Tab. 3:** Mittlere Urteile zu Sinn und prinzipieller Wirksamkeit der Maßnahmen in beiden Experimentalgebieten (EG 1, EG 2) zu den beiden Untersuchungszeitpunkten Post 1 und Post 2 (1= nicht wirksam, 7= sehr wirksam)

Maßnahme		Meinung zum Sinn der Kampagne*		Wirkung auf das eigene Litteringverhalten**	
		Post 1	Post 2	Post 1	Post 2
Nudges	EG1	<b>4,46</b>	<b>5,07</b>	2,49	2,87
	EG2	5,05	5,11	<b>2,57</b>	<b>3,20</b>
Plakat-kampagne	EG2		5,02		2,84
Kombi-nation			<b>5,38</b>		<b>3,33</b>

Es wird deutlich, dass beide Maßnahmen von den Bürgern als sehr sinnvoll eingeschätzt werden und sich dieser Eindruck mit zunehmender Präsenz der Maßnahmen verstärkt. Als besonders sinnvoll wird die Kombination beider Maßnahmen im Experimentalgebiet 2 eingeschätzt. Deutlich geringer fallen die Zustimmungswerte bei der Frage der tatsächlichen Verhaltenswirksamkeit aus. Dies ist übrigens typisch bei derartigen Untersuchungen. Dennoch zeigt sich bei einem Vergleich der Werte in beiden Experimentalgebieten zwischen Messung Post 1 und Post 2 eine höhere geschätzte Wirksamkeit der Nudges auf das Littering-Verhalten. Bei einer Kombination der Maßnahmen kann eine weitere, numerisch zwar nur kleine, aber statistisch bedeutsame weitere Steigerung der Wirksamkeit auf das Verhalten belegt werden. Das Ergebnis spricht dafür, dass sich durch eine längere Präsenz und vor allem durch die Kombination der Maßnahmen eine signifikante Steigerung der Zustimmungswerte erzielen lässt. Wie oben erwähnt, wurden in Feldstudie 3 Kinder und Jugendliche im Alter zwischen 10 und 17 Jahren gezielt in die Befragungen einbezogen. Insbesondere bei dem in Tabelle 3 dargestellten Befund zum Sinn der Maßnahmen und zur Verhaltenswirksamkeit zeigten sich einige Besonderheiten beim Vergleich der Altersgruppen. So nahmen Kinder und Jugendliche die „Nudges“ (grüne Fußspuren Richtung Abfallbehälter) in den Experimentalgebieten vor allem zum Untersuchungszeitpunkt Post 1 deutlich stärker wahr und fanden diese auch sinnvoller zur Verbesserung der Sauberkeit im Gebiet als Erwachsene. Den Einfluss der Maßnahme auf ihr eigenes Litteringverhalten beurteilten sie tendenziell höher als Erwachsene. Zum zweiten Untersuchungszeitpunkt Post 2 glichen sich die Ergebnisse zwischen Kindern/Jugendlichen und Erwachsenen wieder an. Das spricht für zwei Erklärungen: 1. Kinder/Jugendliche fühlten sich durch die Maßnahmen besser angesprochen, was auch ein wichtiges Ziel von Feldstudie 3 war. 2. Erwachsene wurden bei längerfristiger Präsenz der Maßnahme Nudges und deren Ergänzung durch die Plakatkampagne auch gut erreicht. Bei den anderen zuvor diskutierten Themen und Befunden ergaben sich keine wesentlichen Unterschiede zwischen Kindern/Jugendlichen und Erwachsenen.



## 4.2 Ergebnisse zur objektiven Sauberkeitsbewertung

Zur Einschätzung der objektiven Sauberkeit wurde durch die Mitarbeiter der AWB an zwei Tagen je Woche der Sauberkeitszustand der drei Straßenzüge mit dem softwaregestützten Qualitätssicherungssystem „DSQS“ (INFA) erfasst. Außerdem wurde an fünf Tagen in der jeweiligen Untersuchungswoche am Ende der Maßnahmen-Einführung die Menge des Abfalls (Tagesmenge) in den Abfallbehältern und des gelitterten Abfalls (Straßenkehricht Tagesmenge) verwogen. Daraus kann ein Verhältniswert zwischen korrekt entsorgter Menge des Abfalls in Abfallbehältern und gelittertem Abfall (Straßenkehricht) bestimmt werden. In Tabelle 4 sind die Ergebnisse der DSQS-Messungen für die drei Untersuchungszeitpunkt und die drei Untersuchungsgebiete dargestellt (niedrige DSQS-Werte sprechen für mehr Sauberkeit). Es zeigt sich, dass sich nach Einführung der Nudges in Experimentalgebiet 1 und nach der Ergänzung der Maßnahme Nudges durch eine Plakatkampagne in Experimentalgebiet 2 signifikante Verbesserungen nachweisen lassen, während im Kontrollgebiet über den gesamte Untersuchungszeitraum keine wesentlichen Veränderungen bzw. Verbesserungen erkennbar sind.

**Tab. 4:** Sauberkeitsbewertung nach DSQS zu den drei Untersuchungszeitpunkten Prä, Post 1 und Post 2 für die drei Untersuchungsgebiete (fett markiert: signifikante Unterschiede zwischen Untersuchungszeitpunkten; niedrige DSQS-Werte sprechen für mehr Sauberkeit)

	Prä	Post 1	Post 2
<b>EG1</b>	<b>11,47</b>	<b>10,18</b>	10,93
<b>EG2</b>	11,06	<b>11,73</b>	<b>10,45</b>
<b>KG</b>	<b>9,04</b>	<b>10,19</b>	9,92

Die Ergebnisse der Abfallverwiegung werden als prozentualer Anteil am Gesamt-Abfallaufkommen (jeweils verwogener Anteil des Abfalls in Abfallbehältern und Straßenkehricht) im jeweiligen Untersuchungsgebiet ausgewiesen. Zu wünschen wäre natürlich, dass sich das Verhältnis dieser beiden Werte zugunsten des Abfalls in den Abfallbehältern verändert. Tabelle 5 gibt einen Überblick über die Entwicklung im gesamten Untersuchungszeitraum in allen drei Untersuchungsgebieten. Für beide Experimentalgebiete ergaben sich zumindest tendenziell Verbesserungen im Verhältnis von korrekt entsorgtem zu gelittertem Abfall: Tendenzierter Anstieg Abfallmenge in Abfallbehältern und gleichzeitig tendenzielles Absinken Abfallmenge auf der Straße in den Experimentalgebieten, während sich im Kontrollgebiet keinerlei Veränderungen zeigten.

Bei der Interpretation der nicht so deutlichen Verbesserungen in den objektiven Parametern ist zu beachten, dass die dafür wesentlichen Untersuchungszeitpunkte Post 1 und Post 2 jahreszeitlich in den Sommer- und Spätsommerzeitraum fielen. In dieser Zeit muss generell mit mehr Abfallaufkommen durch die sehr viel stärkere Frequentierung der öffentlichen Räume gerechnet werden. Insofern ist es umso bemerkenswerter, dass insbesondere nach DSQS immer noch Verbesserungen in der absoluten Sauberkeit erreicht und empirisch belegt werden konnten.

**Tab. 5:** Abfallverwiegung durch die AWB: prozentualer Anteil der verwogenen Mengen Abfall (in kg) in Abfallbehältern (AB) und Straßenkehricht (Littering) zu den drei Untersuchungszeitpunkten Prä, Post 1 und Post 2 für die drei Untersuchungsgebiete (Experimentalgebiet 1 (EG1): nur Nudges, Experimentalgebiet 2 (EG2): Nudges und später Plakatkampagne und Kontrollgebiet (KG: keine Maßnahmen))

	Prä		Post 1		Post 2	
	AB	Littering	AB	Littering	AB	Littering
<b>EG1</b>	82,6	17,3	83,0	16,9	84,9	15,0
<b>EG2</b>	88,3	11,67	86,9	13,1	87,3	12,6
<b>KG</b>	84,0	15,91	84,3	15,7	83,0	16,9

#### 4.3 Ergebnisse zur Einschätzung der Arbeit des Reinigungsunternehmens ABW Köln

In einer weiteren Fragestellung interessierte uns, wie die Arbeit des zuständigen Reinigungsunternehmens, in diesem Fall der AWB Köln, durch die Bürger in einigen Reinigungskennwerten beurteilt wird. Zwei Befunde seien exemplarisch genannt und dargestellt. In Tabelle 6 sind die Ergebnisse zur Zufriedenheit der Bürger in den Experimentalgebieten mit drei Arbeitsaufgaben der AWB für die drei Erhebungszeitpunkte dargestellt. Mit den Maßnahmen konnten die jeweiligen Zufriedenheitswerte signifikant gesteigert werden und diese Werte liegen nach der Intervention ausnahmslos im positiven Bereich der Antwortskala.

**Tab. 6:** Beurteilung der Zufriedenheit mit ausgewählten Tätigkeitsbereichen der AWB zu den drei Untersuchungszeitpunkten Prä, Post 1 und Post 2 für die drei Untersuchungsgebiete (Experimentalgebiet 1 (EG1): nur Nudges, Experimentalgebiet 2 (EG2): Nudges und später Plakatkampagne und Kontrollgebiet (KG: keine Maßnahmen) (AB= Abfallbehälter, RU= Reinigungsunternehmen) 1= sehr unzufrieden, 7= sehr zufrieden)

Zufriedenheit mit		Prä	Post 1	Post 2
<b>Gründlichkeit Gehwegreinigung</b>	EG1	4,45	4,88	4,55
	EG2	5,29	4,96	4,98
<b>Leerungshäufigkeit der AB</b>	EG1	4,28	4,90	4,64
	EG2	3,30	3,37	4,31
<b>Öffentlichkeitsarbeit des RU</b>	EG1	4,30	4,55	4,48
	EG2	3,90	3,83	4,65

Im Sinne eines langfristigen Vergleichs interessierte uns, ob es bei der Zufriedenheit der Bürger mit der AWB Köln 2017 Veränderungen gegenüber der Untersuchung von 2009 gab. Um einen Effekt zu messen, der unabhängig von den aktuellen Maßnahmen gilt, haben wir jeweils die Zufriedenheitswerte der Bürger zum Zeitpunkt jeweils vor der Einführung der Maßnahmen festgestellt. Dabei sollte die Zufriedenheit mit der AWB generell und mit einigen spezielleren Aufgaben beurteilt werden. In Tabelle 7 ist ein Ausschnitt der Ergebnisse dargestellt. Die relativ gute generelle Bewertung der AWB Köln von 2009 kann 2017 noch einmal signifikant gesteigert werden und erreicht mit mehr als 5 Punkten einen Wert im oberen Bereich der Bewertungsskala. Bei den eher spezielleren Aufgabenbereichen „Gründlichkeit Gehwegreinigung“ und „Häufigkeit Gehwegreinigung“ sind ebenfalls positive Veränderungen zu verzeichnen. Keine Veränderung wurde bei der Zufriedenheit mit der Öffentlichkeitsarbeit festgestellt. Eine etwas schlechtere Bewertung zeichnet sich bei der Bewertung der Leerungshäufigkeit der Abfallbehälter ab, sie bleibt jedoch noch im mittleren Bereich der Bewertungsskala. Dies gilt, wie erwähnt, vor der Maßnahmen-Einführung. Freie Kommentare der Bewohner nach der Applikation der Maßnahmen signalisieren auch in diesen beiden Aspekten der Arbeit der AWB Köln eine noch positivere Einschätzung.

**Tab. 7:** Beurteilung der Zufriedenheit mit ausgewählten Tätigkeitsbereichen der AWB Köln - Vergleich 2009/ (1= sehr unzufrieden bis 7= sehr zufrieden; fett markiert: signifikante Unterschiede 2009/ 2017; AB= Abfallbehälter, RU= Reinigungsunternehmen )

Zufriedenheit mit	2009	2017
<b>Arbeit der AWB gesamt</b>	<b>4,65</b>	<b>5,13</b>
<b>Gründlichkeit Gehwegreinigung</b>	<b>4,09</b>	<b>4,66</b>
<b>Häufigkeit Gehwegreinigung</b>	<b>3,67</b>	<b>4,23</b>
<b>Leerungshäufigkeit AB</b>	<b>4,54</b>	<b>4,01</b>
<b>Öffentlichkeitsarbeit des RU</b>	4,29	4,24

## 5. Zusammenfassung der Ergebnisse und Schlussfolgerungen

Ziel der Untersuchung war die exemplarische Prüfung der Effektivität von Maßnahmen gegen das Litteringverhalten von Bewohnern einer Großstadt mittels einer Feldstudie in Köln. Auf der Basis der Befunde der Feldstudie im Jahr 2009 sowie der Replikationsstudien „Wahrnehmung von Sauberkeit im öffentlichen Raum“ und „Littering – Merkmale, Ursachen, Prävention“ wurden zwei Maßnahmen für diese Evaluation ausgewählt:

1. eine Maßnahme zur Aufmerksamkeitsfokussierung (Nudging) auf die Abfallbehälter (grüne Fußspuren Richtung Abfallbehälter)

2. die Begleitung dieser Maßnahme durch eine inhaltlich ergänzende Plakatkampagne, die über eine moderne Ansprache insbesondere zur Sensibilisierung von Jugendlichen und jungen Erwachsenen beitragen soll, also einer spezifischen relevanten Zielgruppe im Kontext von Littering.

Die Effekte der Maßnahmen sollten sich sowohl in der Beurteilung der Sauberkeit in diesen Straßen durch die Anwohner (subjektive Kriterien) als auch im festgestellten Litteraufkommen (objektive Kriterien) äußern.

Die wichtigsten Ergebnisse der aktuellen Feldstudie beziehen sich einerseits auf die Wirkung der „Nudges“ (grüne Fußspuren auf dem Gehweg in Richtung Abfallbehälter), zum anderen auf die Wirkung der Maßnahmenkombination „Nudges“ plus Plakatkampagne:

### **Wirkung der „Nudges“ (grüne Fußspuren auf dem Gehweg in Richtung Abfallbehälter)**

Bereits die Umsetzung der Maßnahme „Nudges“ führt zu signifikanten Verbesserungen in der Sauberkeit, die sich anhand verschiedener subjektiver und auch objektiver Kriterien belegen lässt.

Folgende Ergebnisse sind dabei hervorzuheben:

- Reduktion wahrgenommener Sauberkeitsmängel (z.B. positivere Mangelbeurteilungen für Zigarettenkippen und Hundekot auf Gehwegen in beiden Experimentalgebieten),
- z.T. stärker realitätsbezogene und handlungsorientierte Mangelwahrnehmung bezüglich der Abfallbehälter (z.B. positivere Beurteilung des Mangels „überfüllte Abfallbehälter“)
- Wahrnehmung von Verbesserungen z.B. hinsichtlich „zu unauffälliger Abfallbehälter“ und „Papierabfall auf Gehwegen“,
- stärkere Gewichtung der Verantwortung der Bürger für die Sauberhaltung des öffentlichen Raums,
- Verbesserung objektiver Sauberheitskriterien: Verbesserte DSQS-Beurteilung im Experimentalgebiet 1.

Zudem konnte die Wirkung der Maßnahme „grüne Fußspuren Richtung Abfallbehälter“ in Hinblick auf verschiedene Personengruppen differenziert werden, z.B.:

Kinder und Jugendliche

- nahmen die „Nudges“ (grüne Fußspuren Richtung Abfallbehälter) in den Experimentalgebieten deutlich stärker wahr als Erwachsene,
- gaben ausschließlich positive Rückmeldungen zu den grünen Fußspuren,
- fanden die „Nudges“ sinnvoller zur Verbesserung der Sauberkeit im Gebiet als Erwachsene,
- beurteilten den Einfluss der Maßnahme auf ihr eigenes Litteringverhalten tendenziell höher als Erwachsene.

## **Wirkung der Maßnahmenkombination**

Im Rahmen der Studie sollte zum einen die längerfristige Wirkung der Maßnahme 1 (grüne Fußspuren auf dem Gehweg in Richtung Abfallbehälter) untersucht werden. Vor allem aber sollte die Wirksamkeit der Kombination von aufeinander abgestimmten Antilittering-Maßnahmen (grüne Fußspuren auf dem Gehweg in Richtung Abfallbehälter plus inhaltlich ergänzende Plakatkampagne im Experimentalgebiet 2) überprüft werden.

Die Befunde belegen einerseits die anhaltende Wirkung der Antilittering-Maßnahme 1 (grüne Fußspuren) auch über den Untersuchungszeitraum 3 hinweg. Zum anderen verweisen die Ergebnisse klar auf eine zusätzliche Wirkung der Plakatkampagne vor allem auf subjektive, aber auch auf objektive Parameter der Sauberkeitsbewertung.

Hervorzuheben ist außerdem die positive Resonanz auf die initiierten Maßnahmen (bzw. die Maßnahmenkombination), insbesondere auch von Kindern und Jugendlichen. Damit übertrifft die Zustimmung für diese Maßnahme in den Experimentalgebieten die Zustimmung im Jahr 2009 für die Maßnahme „auffälligere Gestaltung der Abfallbehälter“ (reflektierenden Markierungen am Korpus und der Einwurföffnung der Abfallbehälter).

Besonders positiv wurde – auch von Kindern und Jugendlichen - gewertet, dass durch die grünen Fußspuren die eher unauffälligen Abfallbehälter der AWB stärker ins Auge fallen. Sehr positiv wurde auch hervorgehoben, dass auf diese Weise Bürger für das korrekte Entsorgen der Abfälle ohne erhobenen Zeigefinger sensibilisiert würden. Gelobt wurde auch das mit der Maßnahme verbundene Engagement der AWB.

Als generelle Schlussfolgerung kann damit festgehalten werden, dass mit zwei relativ wenig aufwendigen Maßnahmen signifikante Effekte bei der Wahrnehmung der Sauberkeit und bei der objektiven Verbesserung der Sauberkeit in einem Teil einer Großstadt erzielt werden können.

## **Ansatzpunkte für weiterführende Antilittering-Maßnahmen und deren Evaluation**

Aufmerksamkeitsfokussierende Maßnahmen wie die in der aktuellen Feldstudie evaluierten „grünen Fußspuren Richtung Abfallbehälter“ und Plakatkampagnen gehören zu den Antilittering-Maßnahmen mit kurz- bis mittelfristiger Wirkung. Ansatzpunkte für eine weiterführende Feldstudie ergeben sich daher beispielsweise aus der Frage nach dem Ausmaß und der Nachhaltigkeit der festgestellten Effekte.

## **Differenzierte Prüfung der Nachhaltigkeit der Ergebnisse**

Relativ einfach ließe sich ein langfristiger Erfolg der bisher realisierten Maßnahmen durch weitere Nachuntersuchungen prüfen. Da sich die Kombination von subjektiven und objektiven Bewertungstechniken als sehr sinnvoll erwiesen hat, dürfte eine Erhebung dieser Daten zu weiteren Zeitpunkten sehr informativ in Hinblick auf die Dauer der Wirkung nach Beendigung der Antilittering-Maßnahme sein. Daraus kann geschlussfolgert werden, in welchen Zeitabständen ähnliche Maßnahmen erforderlich sind, um die Bürger für die städtische Sauberkeit und ihre Verantwortung dafür zu sensibilisieren. Insbesondere die längerfristige Wirkung auf Kinder und Jugendliche ist von großem Interesse, da die Vermutung nahe liegt, dass Gewöhnungseffekte bei dieser Altersgruppe eher eintreten als bei Erwachsenen.

## **Evaluation einer Antilittering-Maßnahme mit stärker sensibilisierendem Charakter und längerfristiger Wirkung**

In Betracht zu ziehen ist auch die Evaluation einer Antilittering-Maßnahme, die stärker auf die nachhaltige Sensibilisierung von Menschen für die Sauberkeit ausgerichtet ist. Im Jahr 2011 wurde mit diesem Ziel eine zweite Feldstudie initiiert. Ausgewählt wurde das Projekt der „Kümmerer“/ Koordinatoren in Hamburg. Es handelt sich hier um Mitarbeiter der SR Hamburg, die u.a. in ihrer Rolle als Multiplikatoren Werte und Normen vermitteln und damit aktiv die soziale Kontrolle in den Stadtgebieten fördern, in denen sie eingesetzt sind. Die Ergebnisse damals zeigten, dass mit dem Einsatz von Koordinatoren deutlich positive Effekte in subjektiven und objektiven Sauberkeitskennwerte erzielt werden konnten. Eine Replikation und Erweiterung dieser Feldstudie wäre insofern sinnvoll, als bspw. spezifische Effekte auf bestimmte Zielgruppen, z.B. Kinder und Jugendliche, untersucht werden könnten.

## **Evaluation einer zielgruppenspezifischen Kampagne**

Die Ergebnisse der aktuellen Maßnahmenevaluation und der Litteringstudie 2016 legen nahe, das Potenzial von Antilittering-Kampagnen verstärkt zu nutzen, indem sie beispielsweise auf spezifische Litterergruppen (wie Raucher oder Hundehalter) zugeschnitten werden oder sich auf besonders sensible Bereiche der Sauberkeitswahrnehmung oder spezifische Orte (z.B. Spielplätze) konzentrieren.

## **6. Literaturverzeichnis**

Bode, C. (2017). Psychologische Strategien zur Optimierung von Kampagnen (im Kontext von Litteringverhalten). Bachelorarbeit (unveröffentlicht). Humboldt-Universität zu Berlin.

Cialdini, R., Demaine, L., Sagarin, B., Barrett, D., Rhoads, K. & Winter, P. (2006). Managing social norms for persuasive impact. *Psychological Press, Social Influence*, 1 (1), 3-15.

Fehr, G., Veit, M., Kamm, A., Geisseler, L. (2014). Littering in der Schweiz- Studie zur Wirksamkeit von Massnahmen unter Berücksichtigung verhaltensökonomischer Erkenntnisse. Zürich: FehrAdvice & Partners AG.

Foerges, R. (2009). Sauberkeitsstudie: Sauberkeit als Standortfaktor. In (Hrsg. VKS im VKU): *Stadt- und Straßenreinigung, Reihe Kommunalhandbuch*. S. 22-36. Lehrte: Beckmann Verlag.

Hansen, P. G. (2016). The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove?. *European Journal of Risk Regulation*, 7(1), 155-174.

Hansen, P. G., & Jespersen, A. M. (2013). Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 4(1), 3–28.

Jarren, O., & Röttger, U. (2015). Public Relations aus kommunikationswissenschaftlicher Sicht (pp. 29-46). Springer Fachmedien Wiesbaden.

Jespersen, S.M. (2012). Green nudge: Nudging into the litter bin. [<http://www.inudgeyou.com>] [letzter Zugriff: 12.11.2017]

Kaiser, N. (2017). Einfluss sozialer Normen in Anti-Littering-Botschaften auf Littering-Verhalten. Masterarbeit (unveröffentlicht). Humboldt-Universität zu Berlin.

Thaler R.H. und Sunstein C.R. (2009): Nudge – Wie man kluge Entscheidungen anstösst, Ullstein Buchverlage GmbH. Econ Verlag Berlin.

Thorun, C., Diels, J., Vetter, M., Reisch, L., Bernauer, M., Micklitz, H. W., ... & Sunstein, C. R. (2016). Nudge-Ansätze beim nachhaltigen Konsum: Ermittlung und Entwicklung von Maßnahmen zum „Anstoßen“ nachhaltiger Konsummuster.

Van der Meer, E., Beyer, R., Gerlach, R. (2010). Littering – Merkmale, Ursachen, Prävention. Köln: *VKS Information 76*.

Van der Meer, E., Beyer, R., Gerlach, R. (2013). Massnahmen gegen Littering. Überprüfung der Wirksamkeit im Feldexperiment. Köln: *VKS Information 80*.





# Schlüssel-Metriken im Personalmarketing

Marc Hochstrate

## Zusammenfassung

Zunächst werden einige auf Personalmarketing wirkende Einflussfaktoren, das grundlegende Anliegen des Controllings und die Bedeutung von Schlüssel-Metriken kurz angesprochen. Im Anschluss daran wird ein einfacher theoretischer Rahmen gespannt, anhand dessen durch Kennzahlen zu kennzeichnende kritische Prozesse im Personalmarketing identifiziert werden. Zum Abschluss werden einige Kennzahlen für diese Prozesse vorgeschlagen und in ihrer Erhebung und Interpretation skizziert.

## Schlüsselwörter

Personalmarketing, Metrik, KPI (key performance indicator), KMU(kleine und mittlere Unternehmen), HR (human resource)

## 1. Einleitung und Zielstellung

Erfolgreiches Bestehen in einer sich verändernden Umwelt erfordert die Fähigkeit zur Anpassung. Dies gilt für Organisationen gleichermaßen wie für Individuen. Ein wichtiger Teil der Unternehmensumwelt ist der Arbeitsmarkt, dessen Wandel viel diskutiert und systematisch untersucht wird. In der Folge entstehen in Unternehmen Handlungsstrategien, die ihrerseits einen Wandel in der Arbeitsumwelt derer, die im Personalwesen tätig sind, nach sich ziehen. Die praktische Konsequenz dieses Wandels sind meist neue und zusätzliche Aufgaben, die es zu erfüllen gilt. Während in großen Unternehmen schnell neue Teams aufgebaut sind, die sich mit neuen Fragestellungen befassen, sind Mitarbeiter kleinerer Unternehmen oft gefordert, solche Arbeiten zusätzlich in Angriff zu nehmen.

Dieser Artikel will jenen eine Hilfestellung bieten, die sich möglicherweise erstmalig mit der Auswahl, Durchführung und Beurteilung von Maßnahmen im Personalmarketing beschäftigen müssen. Der gedankliche Ausgangspunkt ist ein imaginärer Vorgesetzter, der nach einem Konferenzbesuch von seinen Mitarbeitern im Personalbereich möglichst schnell ein Konzept zur Umstellung auf ein kennzahlengetriebenes Vorgehen haben möchte.

## 2. Inhaltlicher Wandel der Personalarbeit

In (Petry & Vaßen, 2017, Seite 311) wird festgestellt, dass „[b]ei einer Betrachtung der Einflusskräfte auf das Personalmarketing [...] aktuell zwei Megatrends im Fokus [...] stehen]. Dies ist zum einen die demographische Entwicklung, die u. a. zu einer sinkenden Anzahl an Fach- und Führungskräften führt [...] und z]um anderen wirkt die zunehmende Social-Media-Nutzung auf das Personalmarketing ein.“ Zu einer ähnlichen Einschätzung gelangt (Krüger, 2018, Seiten 12-20). Sie fasst den zweiten Wirkfaktor jedoch deutlich weiter und betrachtet den allgemeinen Wertewandel, der mit dem Wechsel der Generationen (Generation Y, Generation Z) von Arbeitnehmern einhergeht. Auch (Esch, Baumgartl, Stahl, & Jaeger, 2016, Seite 2) benennt diese Faktoren und konstatiert: „Die Machtverhältnisse auf dem Arbeit[s]markt verschieben sich somit zunehmend. Im Wettbewerb stehen nicht länger die Arbeitnehmer, die um

eine bestimmte Position konkurrieren, sondern immer häufiger die Arbeitgeber, welche um die Gunst der Right Potentials buhlen. [...] Entsprechend gewinnt das Personalmarketing an strategischer Wichtigkeit.“

Dabei wird (Esch, Baumgartl, Stahl & Jaeger, 2016, Seite 2 f.) erläutert: „Das Personalmarketing ist mit der marktorientierten Ausrichtung des Human Resource Management befasst.“ Es wird zwischen internen (Profilierung und Motivation) und externen (Akquisition) Funktionen unterschieden, wobei „[d]as nach außen orientierte Personalmarketing [...] zum einen das Recruiting als eher operative Funktion und zum anderen das strategisch ausgerichtete Employer Branding [umfasst].“ Petry & Vaßen stellen fest, dass der Begriff Personalmarketing in der Literatur zum Teil sehr eng und auf die kurzfristige Personalbeschaffung fokussiert definiert wird. Mit dem „Ziel [...] der] Schaffung von Voraussetzungen zur langfristigen Sicherung der Versorgung des Unternehmens mit qualifizierten und motivierten Mitarbeitern“ (Petry & Vaßen, 2017, Seite 312) sei jedoch ein weiter gefasstes Verständnis zu bevorzugen:

„Personalmarketing i. w. S. umfasst alle Maßnahmen zur strategischen Positionierung des Unternehmens auf dem Arbeitsmarkt, mit dem Ziel sich extern und intern als attraktiver Arbeitgeber darzustellen und so gegenwärtige Mitarbeiter zu halten und zu motivieren und neue Mitarbeiter zu gewinnen.“ (Petry & Vaßen, 2017, Seite 313) Drastischer wird die Erweiterung des Inhaltsbereichs durch Tometschek vertreten: „Employer Branding wird in vielen Unternehmen leider immer noch als neuer Modebegriff für das gute alte Personalmarketing gesehen. Halten Sie es etwa auch noch so?“ (Tometschek, 2013, Seite 77)

Die genannten Autoren und viele andere analysieren im Weiteren Konsequenzen sowie Maßnahmen und Maßnahmenbündel, die geeignet sind, um den erweiterten inhaltlichen Anforderungen Rechnung zu tragen. So diskutiert Krüger zum Beispiel die diesbezüglichen Eigenschaften verschiedener Personalbeschaffungswege und unterscheidet unter anderem zwischen traditionellen Wegen (BA, Printmedien, Mitarbeiterempfehlung, (Hochschul-) Messen, ...), klassischen Internetansätzen (Unternehmenshomepage, Internetstellenbörsen, ...) und Social Media bzw. Sozialen Netzwerken. Von letzteren werden Xing und Facebook detailliert betrachtet. (Krüger, 2018, Kapitel 4)

### **3. Kennzahlgetriebenes Controlling**

Neue Anforderungen in der Personalarbeit betreffen aber nicht nur den Inhaltsbereich, sondern auch die Organisation der Arbeitsprozesse selbst. Unter anderem befeuert durch die Möglichkeiten der computergestützten Datenverarbeitung (Business Intelligence), Forderungen nach mehr Transparenz und der Notwendigkeit in sich schneller verändernden Märkten mit mehr und neuen Teilnehmern agieren zu müssen, sind Kennzahlen (Geschäftsmetriken) zunehmend in den Fokus des Managements geraten (siehe z.B. Ossola-Haring, Schlageter & Schöning, 2019, oder Posselt, 2018).

Eine Geschäftsmetrik ist ganz allgemein eine quantitative (numerische) Maßzahl, die mit einem Geschäftsprozess in Zusammenhang steht und diesen in charakteristischer Weise kennzeichnet. Geschäftsmetriken dienen der Informationsgewinnung und Informationsverdichtung und werden zur Darstellung

wichtiger unternehmerischer Sachverhalte herangezogen. Im Rahmen des Controllings (betriebliche Planung und Überwachung) werden sie häufig zur Definition von Zielvorgaben sowie zur Ermittlung von Abweichungen (Ist-Soll-Vergleich) eingesetzt, um möglichst frühzeitig auf Probleme und Vorteile im betrieblichen Geschehen aufmerksam zu werden.

Der (gerade bei Verfügbarkeit von entsprechenden Datenverarbeitungssystemen bestehenden) Tendenz jeden Geschäftsprozess möglichst kleinteilig mit Metriken zu untermauern, steht die Notwendigkeit gegenüber, die entstehenden Kennzahlen interpretieren zu können, um sie in einem sinnvollen Zusammenhangsgefüge als Entscheidungs- und Handlungsgrundlage zu nutzen. Praktische Erfahrungen mit den Problemen einer „Kennzahlenflut“ haben die Hinwendung zu Schlüssel-Metriken (key performance indicators) bewirkt. Diese zeichnen sich dadurch aus, dass sie einerseits genau die zentralen für den Unternehmenserfolg kritischen Geschäftsprozesse (bzw. spezifische strategische Ziele) in ihren wesentlichen Eigenschaften kennzeichnen und andererseits durch das unternehmerische Handeln (in ihren Wertausprägungen) beeinflussbar also kontrollierbar sind.

Im Zusammenspiel mit den dargelegten inhaltlichen Veränderungen müssen Mitarbeiter im Personalmanagement nicht nur mehr und neue Maßnahmen ersinnen und koordinieren, sondern auch hinsichtlich ihrer Zielstellung, Wirksamkeit und Angemessenheit beurteilen bzw. darstellen und vertreten können.

#### **4. Theoretischer Rahmen**

Der theoretische Rahmen soll hauptsächlich als Gerüst dienen, anhand dessen Kennzahlen ausgewählt und begründet werden. Eine zweite wichtige Funktion ist die Schaffung von Anknüpfungspunkten für das Management, um Akzeptanz und Unterstützung sicher zu stellen. Da die neuartigen Inhaltsbereiche im Personalmarketing dem Marketing nahestehen, ist es ein plausibler Schritt, sich dort nach geeigneten Konzepten umzusehen. Ein im Marketing sehr bekanntes Schema ist das auf Stimulus-Organismus-Reaktion-Theorien beruhende AIDA-Modell. Der Name leitet sich von den ersten Buchstaben der im Modell dargestellten Phasen ab (Krings, 2017, Seite 33):

- Attention (Aufmerksamkeit erzeugen)
- Interest (Interesse wecken)
- Desire (Wunsch auslösen)
- Action (Handlung herbeiführen)

Bezogen auf das Personalmarketing besagt dieses Modell, dass (potentielle) Mitarbeiter das Unternehmen zunächst kennen und als interessant wahrnehmen müssen. Aus dem Interesse soll der Wunsch, bei diesem Unternehmen zu arbeiten, erwachsen und schließlich in die Handlung einer Bewerbung münden.

Eine andere beliebte Art der Modellierung sind Trichtermodelle, die auf Prozesse ausgerichtet sind, bei denen in einer mehrstufigen Abfolge ein immer feineres und enger umrissenes Ergebnis entsteht. Angewandt auf das Personalmarketing entsteht ein Trichter, an dessen oberen breiten Ende alle potentiellen Arbeitnehmer sind, die beim Unternehmen arbeiten könnten. Der Trichter verengt sich dann zunächst auf den Kreis derer, die beim Unternehmen arbeiten möchten und weiter auf die, die sich

auch bewerben. Die abschließende Verengung reduziert das Ergebnis auf die Arbeitnehmer, die zum Unternehmen und zu den verfügbaren Arbeitsangeboten passen. Am unteren Austrittsende des Trichters stehen schließlich genau die Arbeitnehmer, mit denen ein (dann möglichst langfristiges) Arbeitsverhältnis zustande kommt.

Unabhängig von den zahlreichen Details und Spezifika der verschiedenen Branchen und Unternehmungen sind dies offensichtliche Kernprozesse, die überall dort ablaufen müssen, wo es zu Anstellungen und Arbeitsverhältnissen kommen soll. Grundsätzlich interessante Kennzahlen betreffen damit die Ausstrahlung (Bekanntheit) und die Anziehung (Attraktivität) des Unternehmens sowie die Passung und die Bindung von Arbeitnehmern. Die Tätigkeiten im Personalmarketing zielen prinzipiell darauf ab, dass das Unternehmen bekannt und beliebt ist und in der Konsequenz viele Bewerbungen erhält.

Für die abschließende Betrachtung muss noch bedacht werden, dass die Bearbeitung von Bewerbungen sowie die Durchführung von Maßnahmen Aufwand und Kosten verursacht und am tatsächlichen Bedarf an neuen Angestellten zu orientieren ist. Dies betrifft den Aspekt der Steuerung. Maßnahmen sollten möglichst gezielt und nicht überflüssig durchgeführt werden. Bei der Kennzahlbildung müssen ineinander verzahnte punktuelle und kontinuierlich ablaufende Geschäftsprozesse Berücksichtigung finden.

## **5. Kennzahlen**

Für die Bildung von Kennzahlen stehen zwei grundsätzliche Strategien zur Verfügung. Man kann Daten nutzen, die im Geschäftsprozess ohnehin anfallen, oder zusätzliche Daten erheben, die ausschließlich der Kennzahlbildung dienen. Während der erste Weg meist weniger Aufwand und Kosten verursacht, produziert der zweite Weg potentiell mehr und wertvollere Informationen. Gerade bei dem Versuch Maßnahmen in ihren Effekten zu kennzeichnen, ist die Erhebung zusätzlicher Informationen während der Durchführung häufig unabdingbar.

Es fällt auf, dass die verschiedenen Maßnahmen (siehe z.B. Krüger, 2018, Kapitel 4), sich in einigen Eigenschaften prinzipiell unterscheiden, die Einfluss auf die Möglichkeiten der Kennzahlenbildung und Erhebung zusätzlicher Daten haben. Die wichtigsten Eigenschaften sind (Vor-)Selektion und Verfügbarkeit. Selektion betrifft die Frage, welche Gruppe von Personen überhaupt in Kontakt mit einer gegebenen Maßnahme kommt und wie sich Eigenschaften dieser Gruppe auf die Kennzahlen auswirken können. Verfügbarkeit betrifft die Frage, inwiefern und wie lange während einer Maßnahme Zugriff auf die von der Maßnahme betroffenen Personen besteht. So zeichnen sich Maßnahmen im Internet durch geringe Selektion und geringe Verfügbarkeit aus, Events haben oft mittlere Selektion und Verfügbarkeit während Programme (z.B. Praktika, Trainee-Wochen) hohe Selektion und Verfügbarkeit aufweisen.

### **Ausstrahlung**

Die Ausstrahlung des Unternehmens ist wesentlich durch Faktoren bestimmt (z.B. Unternehmensgröße), die außerhalb des Einflussbereichs des Personalmarketings liegen. Insofern ist dieser Bereich zur Ableitung von allgemeinen Schlüsselmetriken

schlecht geeignet, da es an der Kontrollierbarkeit mangelt und die Erhebung die Abteilungsgrenzen überschreitet.

Zur Beurteilung von Maßnahmen kann er sehr wohl herangezogen werden. Gerade bei kleinen und mittleren Unternehmen ist dabei zu bedenken, dass das Ziel meist nicht in globaler Bekanntheit, sondern in Bekanntheit bei einer bestimmten Gruppe besteht. Eine Vorbetrachtung hinsichtlich der Relevanz der erreichten Gruppe muss also im Kontext des Unternehmens erfolgen. Maßnahmen sind dann sinnvoll durch die Anzahl der erreichten Personen (Teilnehmerzahlen, Besucherzahlen, Anzahl von Kontaktaufnahmen) gekennzeichnet und können hinsichtlich dieser verglichen werden.

### Anziehung

Wie die Ausstrahlung wird auch die Anziehung durch Faktoren (z.B. allgemeines Ansehen der Branche) beeinflusst, die den Wirkungsbereich des Personalmarketings überschreiten. Die Schaffung einer starken anziehenden Employer Brand ist jedoch eines der zentralen Kernthemen und damit von hoher Relevanz. Die konkrete inhaltliche Ausgestaltung der Arbeitgebermarke muss wiederum im Kontext des Unternehmens erfolgen, kann sich aber an allgemeinen Untersuchungen zur Arbeitgeberattraktivität orientieren.

Im Anschluss kann die Erfüllung der gesetzten Ziele am einfachsten durch Befragungen geprüft werden. Als Startpunkt eignen sich dabei sogenannte Single-Item-Skalen, die aus einer einzelnen möglichst präzise formulierten Frage bestehen, die mit ja oder nein beantwortet werden kann. Die Möglichkeit, keine Antwort zu geben, sollte vorgesehen werden.

Zur Erhebung wird einer Menge von Personen befragt. Als Wert ergibt sich die Anzahl der positiven Antworten verringert um die Anzahl der negativen Antworten bezogen auf die Gesamtanzahl der gefragten Personen. Jede gestellte Frage resultiert in einer Kennzahl, die eine Werteausprägung zwischen -1 und 1 annehmen kann.

Eine mögliche Frage, um die Anziehung zu kennzeichnen, ist zum Beispiel:

„Ich würde einem Freund empfehlen, ein passendes Anstellungsangebot von (dem Unternehmen) anzunehmen.“ (Ja/Vielleicht/Nein)

Die resultierenden Kennzahlen können sowohl für die punktuellen als auch für die kontinuierlichen Prozessaspekte zum Einsatz kommen. Werden die betreffenden Fragen allen Teilnehmern einer Maßnahme gestellt, ergibt sich eine Kennzeichnung der einzelnen Maßnahme, die den Vergleich von Maßnahmen gestattet. So sind zum Beispiel bei der Erhöhung der Ausstrahlung die Maßnahmen besonders sinnvoll, bei denen von Vorneherein eine hohe Anziehung gegeben ist. Bei Maßnahmen mit hoher Verfügbarkeit und dem Ziel bestimmte Inhalte der Arbeitgebermarke zu schärfen, ist es auch denkbar die Kennzahlen am Beginn und Ende der Maßnahme zu erheben und die direkte Veränderung zur Maßnahmenkennzeichnung heranzuziehen.

Zur kontinuierlichen Kennzeichnung sollten Referenzzeiträume (Monate, Quartale, Jahre) definiert werden, innerhalb derer Personen befragt werden. Für jeden Referenzzeitraum wird dann ein Wert für jede Kennzahl ermittelt und die Veränderung zwischen den Referenzzeiträumen betrachtet. Dabei ist auf mögliche Effekte der Selektion der befragten Personen zu achten. Dieses Vorgehen bietet sich insbesondere im Rahmen wiederholter Mitarbeiterbefragungen an, um die Wahrnehmung der angestrebten Markeninhalte und Erfüllung der angestrebten Markenversprechen im Inneren zu prüfen.

### Bewerberpassung

Bei der Passung ist zwischen der Passung (der Person) zur Organisation (P-O-Passung) und der Passung zur konkreten Arbeitsaufgabe, dem Job, (P-J-Passung) zu unterscheiden (Esch, Baumgartl, Stahl & Jaeger, 2016, Seite 9), die Einfluss auf den Schritt zur Bewerbung haben. Die Erzeugung der Wahrnehmung einer hohen Passung auf der Seite des Bewerbers ist im gewählten theoretischen Rahmen im Bereich der Anziehung verortet. Passung bezieht sich hier auf die Beurteilung der eingehenden Bewerbungen hinsichtlich der prinzipiellen Eignung aus der Sicht des Unternehmens. Im Vordergrund steht dabei Effizienz bzw. das Bestreben die durch Bewerbungen entstehenden Kosten zu berücksichtigen. Je niedriger die Bewerberpassung ist, desto mehr Bewerbungen müssen generiert werden, um eine geeignete Kandidatin zu finden.

Da die eingehenden Bewerbungen ohnehin gesichtet und klassifiziert werden müssen, können sie ganz allgemein dahingehend beurteilt werden, ob sie prinzipiell „interessant“ oder „passend“ waren, z.B. durch die Frage:

„Ich würde mir wünschen, dass mehr derartige Bewerbungen eingehen.“  
(Ja/(Egal)/Nein)

Die Kennzahl der Passung wird entweder wie im Abschnitt zuvor beschrieben ermittelt oder noch einfacher als Verhältnis von passenden zu eingehenden Bewerbungen mit einem Wertebereich zwischen 0 und 1. Analog zum vorherigen Abschnitt ist der Einsatz für kontinuierliche Prozessaspekte (durch Kennzahlermittlung in Referenzzeiträumen) und für die punktuelle Kennzeichnung (durch Bezugnahme auf alle im Rahmen einer bestimmten Maßnahme eingegangenen Bewerbungen) denkbar.

In diesem Zusammenhang spricht man statt von Maßnahmen auch häufig von Kanälen und betont dabei die Wiederholung von gleichartigen Maßnahmen unter bestimmten Rahmenbedingungen (z.B. Kanal „Karrieremesse HU“ oder „Internetportal X23“). Eine derartige Konstellation ist auch geeignet, beide Betrachtungsweisen zu kombinieren, indem Kennzahlen pro Referenzzeitraum für alle (interessierenden) Kanäle ermittelt werden.

### Bedarfspassung

Ein anderer Blickwinkel stellt die betriebswirtschaftlichen Aspekte in den Vordergrund und betrifft die Passung der Bewerbungsmenge zum Bedarf des Unternehmens an Bewerbungen. Der bestmögliche Fall tritt ein, wenn in dem Moment, in dem eine Stelle besetzt werden muss, direkt ein geeigneter Bewerber zur Verfügung steht, der

eingestellt werden kann. Im Normalfall wird mit zum Teil sehr langen Verzögerungen zu rechnen sein und eine Stelle mehreren Kandidaten parallel offeriert, um die Wahrscheinlichkeit der tatsächlichen Besetzung zu erhöhen. Zur Kennzeichnung dieses Sachverhalts wird vorgeschlagen, die Anzahl offener Bewerbungsverfahren (vom Eingang bis zum positiven oder negativen Abschluss) auf die Summe aus offenen Stellen und laufenden Bewerbungsverfahren zu beziehen. Die resultierende Kennzahl bewegt sich zwischen 0 und 1 und zeigt ein etwas schwieriger zu interpretierendes Verhalten.

Sind keine offenen Stellen zu besetzen, so ist ihr Wert 1. Gibt es offene Stellen, jedoch keine Bewerbungen, so nimmt sie den Wert 0 an. Wenn auf jede offene Stelle genau ein Bewerber fällt, wird der Wert 0,5 angenommen. Dieses Verhältnis ist minimal zu erfüllen, um alle offenen Stellen besetzen zu können, jedoch bestehen dann keinerlei Wahlmöglichkeiten (Ablehnung ungeeigneter Bewerber). Ein günstiger Bereich liegt bei 0,65 (zwei Bewerber pro Stelle) bis 0,85 (5-6 Bewerber pro Stelle), muss jedoch wiederum im Kontext des Unternehmens – insbesondere hinsichtlich der mittleren Bewerberpassung – festgelegt werden.

Die Bedarfspassung ist insbesondere auf den kontinuierlichen Prozessaspekt ausgelegt, wobei im besten Fall eine tagesaktuelle Ist-Zahl vorliegt, die bei jedem Bewerbungseingang und jeder Stellenbesetzung aktualisiert wird. Alternativ kann in regelmäßigen Abständen der aktuelle Stand erhoben und verzeichnet werden. Ab einer gewissen Unternehmensgröße kann die Aufspaltung dieser Kennzahl (zum Beispiel in administrative und fachliche Stellen) sinnvoll sein.

Es ist möglich, diese Kennzahl auch zur Bewertung von Maßnahmen (Kanälen) heranzuziehen, indem man anstatt aller offenen Bewerbungen nur die über den betrachteten Kanal eingegangenen einbezieht. Das nicht lineare Verhalten der Kennzahl erschwert jedoch den direkten Vergleich der Kanäle auf dieser Basis, so dass in diesem Fall besser und einfacher mit der absoluten Anzahl der (im Referenzzeitraum) auf dem Kanal eingegangenen Bewerbungen arbeitet. Noch besser aber aufwändiger ist es, zusätzlich pro Kanal die auf diesem aktuell verfügbaren (veröffentlichten) Stellenangebote zur erfassen und somit die Bedarfspassung pro Kanal zu bestimmen, die man dann direkt miteinander vergleichen kann.

## **6. Prozessorganisation**

Die tatsächliche Prozessorganisation hängt stark vom Unternehmen ab und ist permanenter Entwicklung unterworfen. Aus diesem Grund sollen zum Abschluss nur zwei prinzipielle Aspekte kurz angesprochen werden.

Um auf der Grundlage von Kennzahlen steuern zu können, ist es notwendig, den erwarteten oder erwünschten Verlauf zu antizipieren und auf sich ergebende Abweichungen (korrigierend) zu reagieren. Das betrifft den in Abschnitt 3 angesprochenen Aspekt der Kontrollierbarkeit von Schlüssel-Metriken. Zentral ist es also zunächst den erwarteten Bedarf an zu besetzenden Stellen möglichst langfristig vorherzusagen. Mittels der bisherigen Bedarfs- und Bewerberpassung kann dann die Entwicklung der Kenngröße unter dieser Annahme bestimmt werden (unter der Annahme der Konstanz bisheriger Bewerberzahlen und Bewerberqualität). Anhand dieser Prognose können spezifischer Handlungsbedarf und die ihm entsprechende

Maßnahmen abgeleitet werden. Durch die Ermittlung von Kennzahlen für die verschiedenen Maßnahmen entsteht nach und nach die Basis, zwischen geeigneten und nicht geeigneten Maßnahmen zu unterscheiden, bzw. Maßnahmen zielgerichtet zu verändern.

Generell bietet es sich schon früh an, offene Stellen mit einer eindeutigen Identifikationsnummer zu versehen und diese im Bewerbungsprozess (z.B. bei Anzeigen) zu kommunizieren. Zusätzlich ist es sinnvoll, eine Identifikationsnummer für Maßnahmen und/oder Kanäle einzuführen, so dass die Grundlage für die Erhebung und Zuordnung der dargelegten Kennzahlen geschaffen wird. Die Identifikation eines konkreten Stellenangebotes folgt dann dem Muster „Stelle-Kanal-Maßnahme“ (z.B. „345-SZ-191011“ = Stelle 345 inseriert in der Südlichen Zeitung am 11. Oktober 2019) Dabei kann es sinnvoll sein, zusätzliche Kanäle wie z.B. Initiativbewerbungen zu berücksichtigen, die nicht aktiv mit Stellenangeboten versehen werden.

## 7. Schlussbemerkungen

Sowohl beim Ausformen der konkreten Employer-Branding-Strategie als auch bei der Erstellung von Schlüssel-Metriken ist die individuelle (strategische) Situation des Unternehmens mit all seinen Spezifika, Stärken und Schwächen von zentraler Bedeutung. Insofern ist ein allgemeiner Artikel zu diesem Thema ein Widerspruch in sich und müsste mit einer reichlichen Anzahl an „wenn“ und „falls“ dekoriert sein. Tatsächlich finden sich im vorliegenden Text entsprechend viele unscharfe Formulierungen. Sie sollten den Leser mahnen, dass hier nur eine Hilfestellung und ein möglicher Ausgangspunkt, jedoch kein vollständiges Rezept zum Vorgehen angeboten wurde, und zahlreiche relevante Fragen nicht einmal angesprochen werden konnten.

## 8. Literaturverzeichnis

Esch, F.-R., Baumgartl, C., Stahl, M. & Jaeger, A. (2016). Employer Branding: Right Potentials gewinnen. Esch, F.-R. (eds). *Handbuch Markenführung*. Springer Reference Wirtschaft. Springer Gabler, Wiesbaden. 1-21.

Krings, T. (2017). *Erfolgsfaktoren effektiver Personalauswahl*. Springer Gabler, Wiesbaden.

Krüger, K. (2018). *Herausforderung Fachkräftemangel, Erfahrungen, Diagnosen und Vorschläge für die effektive Personalrekrutierung*. Springer Gabler, Wiesbaden.

Ossola-Haring, C., Schlageter, A. & Schöning, S. (2019). *11 Irrtümer über Kennzahlen, Mit den richtigen Erkenntnissen führen*. Springer Fachmedien Wiesbaden.

Petry, T. & Vaßen, M. (2017). Personalmarketing 2.0 – State of the Art, Praxisbeispiele und Erfolgsfaktoren. Rosenberger, B. (eds). *Modernes Personalmanagement*. Springer Gabler, Wiesbaden. 311-324.

Posselt, G. (2018). *Renditeorientierte Führungsstrategien, Kennzahlen motivierend einsetzen*. Springer Fachmedien Wiesbaden.

Tometschek, R. (2013). Employer Branding: Innen beginnen. Buckmann, J. (eds). *Einstellungssache: Personalgewinnung mit Frechmut und Können*. Springer Fachmedien Wiesbaden. 77-90



# **Untersuchung des Einflusses von Mathematikkompetenzen auf Studienerfolg in den Wirtschaftswissenschaften**

Marius Fahrner & Birgit Wolf

## **Zusammenfassung**

Der vorliegende Artikel untersucht den Einfluss mathematischer Kompetenz auf Studienerfolg in einem wirtschaftswissenschaftlichen Studiengang. Hierzu wurden Regressionsanalysen durchgeführt, bei denen die Studienabschlussnote mit der Durchschnittsnote in Mathematik und einigen Kontrollvariablen wie beispielsweise Abiturnote und Studiendauer in Beziehung gesetzt wurde. Die Analysen zeigen, dass die Durchschnittsnote Mathematik als Indikator für Studienerfolg in den Wirtschaftswissenschaften angesehen werden kann.

## **Schlüsselwörter**

Studienerfolg, Mathematikkompetenzen, Regressionsanalysen, Abiturnote

## **1. Einleitung**

Die in diesem Artikel verwendeten Daten sind einer Kohorte von Studierenden entnommen, die ein wirtschaftswissenschaftliches Bachelorstudium an einer Hochschule in Berlin erfolgreich absolviert haben. Die Unterrichtssprache ist Englisch; es kann ein deutscher und ein amerikanischer Studienabschluss erworben werden. Wie in vorangegangenen Untersuchungen gezeigt wurde (Fahrner/Wolf, 2018, S.79), existiert eine äußerst hohe positive Korrelation (über +0,7) zwischen der Durchschnittsnote Mathematik während des Studiums und der gewichteten Studienabschlussnote (Grade Point Average, GPA). Nun ließe sich aus der hohen positiven Korrelation einerseits folgern, dass sich die Mathematikkompetenzen unmittelbar auf den Studienerfolg auswirken; andererseits wäre es aber auch möglich, dass andere Faktoren originär entscheidend für den Studienerfolg sind und die hohe Korrelation lediglich darauf zurückzuführen ist, dass diese Faktoren sowohl die Durchschnittsnote Mathematik als auch den GPA maßgeblich beeinflussen und somit kein unmittelbarer Zusammenhang zwischen der Durchschnittsnote Mathematik und dem GPA besteht. Um die Hypothese zu testen, dass kein unmittelbarer Zusammenhang zwischen der Durchschnittsnote Mathematik und dem GPA besteht, wurden Regressionsanalysen mit zahlreichen Kontrollvariablen wie Alter, Geschlecht, Studiendauer durchgeführt. Sollte die Hypothese, basierend auf den durchgeführten Regressionsanalysen, verworfen werden, könnte eine Optimierung der Mathematikmodule angestrebt werden, um auf diese Weise eine Verbesserung der Mathematikkompetenzen der Studierenden und somit eine Verbesserung der Studienabschlussnote zu erzielen.

Der vorliegende Artikel untersucht die Leistung von Studierenden, die ihren Bachelorabschluss in dem Studienfach Management im Zeitraum von 2005 bis 2016 erworben haben. In Abschnitt 2 werden die Kontrollvariablen beschrieben, die bei den Regressionsanalysen verwendet wurden. In Abschnitt 3 werden die Ergebnisse der Regressionsanalysen mit den in Abschnitt 2 beschriebenen Kontrollvariablen dargestellt. In Abschnitt 4 werden die Ergebnisse aus den Regressionen analysiert und interpretiert. Schließlich wird in Abschnitt 5 auf Einschränkungen sowie auf mögliche Auswirkungen der Untersuchung hingewiesen.

## 2. Kontrollvariablen

Um die Wahl der Kontrollvariablen zu rechtfertigen, wurde für die metrischen Variablen die Korrelation zwischen dem GPA und der jeweiligen (metrischen) Variable berechnet; eine deutlich von null verschiedene Korrelation gibt Anlass zu der Vermutung, dass sich die entsprechende metrische Variable signifikant auf die Studienabschlussnote auswirkt und somit alternativ zu der Durchschnittsnote Mathematik ein Indikator für Studienerfolg sein könnte. Für die nominalskalierten Variablen wurde für jede Kategorie der Mittelwert des GPAs berechnet; diese Mittelwerte wurden miteinander verglichen. Bei deutlich voneinander abweichenden Mittelwerten liegt die Vermutung nahe, dass die Zugehörigkeit zu einer Kategorie bereits Aufschluss über den späteren Studienerfolg gibt. Neben diesen rein quantitativen Argumenten wird auch auf Argumente aus der Literatur verwiesen.

### 2.1 Geschlecht

Einer Studie der Ökonomin *Lin* zufolge erzielen weibliche Studierende der Wirtschaftswissenschaften im Durchschnitt eine bessere Studienabschlussnote als männliche Studierende; zudem erhalten männliche Studierende während ihres Studiums seltener Stipendien als weibliche Studierende, was ebenfalls eine bessere Studienleistung weiblicher Studierender im Vergleich zu männlichen Studierenden nahelegt (Lin, 2017, S.128). Auch die Berechnungen aus den für den vorliegenden Artikel verwendeten Daten ergeben eine im Durchschnitt deutlich bessere Studienabschlussnote weiblicher Studierender gegenüber männlichen Studierenden (weiblich: 1,68, männlich: 1,97), sodass insgesamt die Annahme, dass sich das Geschlecht auf die Studienabschlussnote auswirkt, plausibel erscheint.

### 2.2 Ort des Schulabschlusses

Neben dem Geschlecht könnte sich auch der Ort des Schulabschlusses (im Inland (Deutschland) oder im Ausland) auf den Studienerfolg auswirken. Zwar konnte, ausgehend von den vorliegenden Daten, hinsichtlich der Studienabschlussnote kein entscheidender Unterschied zwischen Studierenden, die ihren Schulabschluss in Deutschland erworben haben, und Studierenden, die ihren Schulabschluss nicht in Deutschland erworben haben, festgestellt werden (Schulabschluss in Deutschland: 1,86, Schulabschluss im Ausland: 1,87); dennoch erscheint im Kontext der äußerst hohen Diversität der Studierendenschaft an der untersuchten Hochschule in Bezug auf das Land des Schulabschlusses und in Anbetracht der Verschiedenheit der jeweiligen Bildungssysteme die Wahl der Kontrollvariable Ort des Schulabschlusses als sinnvoll.

### 2.3 Art des Schulabschlusses

Die Diversität der Schulabschlüsse der Studierenden erstreckt sich nicht nur auf den Ort, sondern auch auf die Art der jeweiligen Abschlüsse. So erwarben einige der untersuchten Studierenden als Hochschulzugangsberechtigung neben dem Abitur bzw. dem High School-Abschluss auch das International Baccalaureate (IB) sowie eine Vielzahl anderer Schulabschlüsse, die aufgrund ihres seltenen Auftretens unter sonstige Abschlüsse zusammengefasst werden. Eine Analyse der GPA-Mittelwerte hinsichtlich der jeweiligen Arten der Schulabschlüsse ergibt erhebliche Unterschiede

bezüglich der Studienabschlussnote je nachdem, mit welchem Schulabschluss sich die Studierenden an der Hochschule bewarben (Abitur: 1,79, High School: 1,87, IB: 1,98, sonstiger Abschluss: 1,93). Daher scheint die Vermutung, dass sich die Art des Schulabschlusses auf die Studienabschlussnote auswirkt, naheliegend.

## 2.4 Abiturnote

Wie soeben dargelegt, erzielen Studierende mit Abitur im Durchschnitt den größten Studienerfolg unter allen Studierenden. Aus dieser Beobachtung ergibt sich die Frage, inwiefern bereits die Abiturnote Aufschluss über den späteren Studienerfolg geben könnte. Eine gute Abiturnote ist im Allgemeinen mit hoher Lernmotivation und Disziplin assoziiert, was sich häufig auch auf den Studienerfolg positiv auswirkt (Rotter, 2011, S.9). Die Bildungswissenschaftlerin *Rotter* verweist zudem auf Studien, bei denen hohe positive Korrelationen zwischen der Abiturnote und der Studienabschlussnote in verschiedenen Studienfächern nachgewiesen werden konnten – und die Abiturnote sogar als „der beste Einzelprädiktor für Studienerfolg im Sinne von Studiennoten“ (Rotter, 2011, S.11) ausgewiesen wurde. Diese Ergebnisse lassen sich auch mit der für diesen Artikel durchgeführten Analyse, bei der sich eine Korrelation von 0,56 zwischen Abiturnote und GPA ergab, in Einklang bringen, sodass die Wahl der Kontrollvariable Abiturnote gerechtfertigt erscheint.

## 2.5 Bachelorarbeit

Studierende, die einen Bachelorabschluss an der untersuchten Hochschule im amerikanischen System erlangen wollen, müssen hierzu nicht notwendigerweise eine Bachelorarbeit verfassen. Das Verfassen einer Bachelorarbeit stellt für diese Studierenden also einen zusätzlichen, nicht vorausgesetzten Aufwand dar. Daher liegt es nahe, dass neben einer guten Abiturnote auch die Bereitschaft, eine Bachelorarbeit zu verfassen, auf eine hohe Lernmotivation und eine ausgeprägte Disziplin verweist, was sich wiederum positiv auf die Studienabschlussnote auswirken könnte. Diese Vermutung wird auch durch die vorliegenden Daten gestützt, auf deren Basis berechnet werden konnte, dass Studierende, die (freiwillig) eine Bachelorarbeit verfasst hatten, im Durchschnitt einen GPA von 1,73 erzielten, während Studierende, die keine Bachelorarbeit verfasst hatten, im Durchschnitt einen GPA von 1,94 erzielten. Insgesamt liegt es demnach nahe, die Bereitschaft, eine Bachelorarbeit zu verfassen, als Kontrollvariable in die Untersuchung zu integrieren.

## 2.6 Studieneintrittsalter und Studiendauer

Neben den bereits genannten Variablen sollten auch solche Variablen Eingang in die Untersuchung finden, die eine zeitliche Dimension während oder vor dem Studium abbilden. Hierzu wurden das Studieneintrittsalter sowie die Studiendauer als potentielle Kontrollvariablen ausgewählt. Hinsichtlich des Studieneintrittsalters wurde vermutet, dass ältere Studierende aufgrund einer größeren Lebenserfahrung und umfangreicherer beruflicher Vorkenntnisse bessere Studienleistungen erzielen als jüngere Studierende. Ein Anstieg der Studienleistung mit zunehmendem Alter konnte auch von *Owen* aufgezeigt werden (Owen, 2003, S.183). Dabei handelt es sich um ein Ergebnis, das sich ebenfalls in Studien nachweisen ließ, die webbasierte Kurse zum Gegenstand hatten (Vella/Turesky/Hebert, 2016, S.593), und das in Einklang mit der Korrelationsberechnung unter Verwendung der für diesen Artikel relevanten

Daten zu bringen ist. Es ergab sich eine Korrelation von  $-0,32$  zwischen Alter und dem GPA, sodass ältere Studierende im Durchschnitt einen niedrigeren GPA hatten, was äquivalent zu einer besseren Studienleistung ist. Insgesamt ist es daher naheliegend, das Alter zu Beginn des Bachelor-Studiums als Kontrollvariable zu wählen. Schließlich konnte eine auffallend hohe Korrelation von  $0,45$  zwischen Studiendauer und Studienabschlussnote berechnet werden; Studierende, die sehr lange für ihren Abschluss brauchten, schnitten demnach in der Regel deutlich schlechter ab als Studierende, die für ihren Abschluss nicht (erheblich) länger als die Regelstudienzeit benötigten. Daher wurde auch die Studiendauer als Kontrollvariable in die Untersuchung eingeführt.

### 3. Analyse

In allen Fällen wurden Regressionsanalysen durchgeführt. Hierbei war stets die abgängige Variable die gewichtete Studiendurchschnittsnote (Grade Point Averages) und stets eine unabhängige Variable die Durchschnittsnote Mathematik. Sowohl der GPA als auch die Mathematiknoten lagen zwischen  $1,00$  und  $4,00$ , wobei eine kleinere Zahl eine bessere Note angab. Für jede berücksichtigte metrische Variable konnte gezeigt werden, dass ein nahezu linearer Zusammenhang zwischen dieser und dem GPA besteht. Außerdem lag der Determinationskoeffizient  $R^2$  in allen Fällen bei über  $0,5$ , sodass eine hinreichend gute Anpassungsgüte für jedes Modell angenommen werden konnte. Des Weiteren waren alle Residuen nahezu normalverteilt; es herrschte Homoskedastizität unter den Residuen und der Durbin-Watson-Test ergab in allen Fällen einen Wert zwischen  $1,7$  und  $2,1$ , sodass bei den durchgeführten Regressionsanalysen keine signifikanten Autokorrelationen nachgewiesen werden konnten. Schließlich wurden die unabhängigen Variablen auf Multikollinearität getestet; auch Multikollinearität konnte in allen Fällen ausgeschlossen werden, da der Variance Inflation Factor (VIF) stets kleiner als  $1,7$  war. Basierend auf den soeben genannten Resultaten konnten in allen Fällen lineare Regressionen durchgeführt werden.

#### 3.1 Mathematik und Geschlecht

Zunächst wurde eine lineare Regression mit  $N=203$  Studierenden durchgeführt (116 männlich, 87 weiblich). Neben der abhängigen Variable GPA sowie der unabhängigen Variable Durchschnittsnote Mathematik wurde die unabhängige Variable Geschlecht (männlich vs. weiblich) verwendet. Die Variable Geschlecht wurde als Dummy-Variable kodiert, wobei weiblich als Referenzvariable verwendet wurde. Für männlich wurde ein Regressionskoeffizient von  $0,81$  berechnet, der nicht signifikant war, da der p-Wert signifikant größer als  $0,11$  war. Für die Durchschnittsnote Mathematik wurde ein Regressionskoeffizient von  $0,481$  berechnet, der signifikant war, da der p-Wert bei unter  $0,01$  lag. Daher ergab sich ein signifikanter Einfluss der Durchschnittsnote Mathematik auf den GPA, während kein signifikanter Einfluss des Geschlechts auf Studienerfolg nachgewiesen werden konnte. Allerdings ergab die Analyse, dass weibliche Studierende im Durchschnitt etwas besser abschnitten als männliche (der Regressionskoeffizient von  $0,81$  bei der Variable männlich weist darauf hin, dass männliche Studierende im Durchschnitt einen höheren GPA als weibliche Studierende erzielten und somit schlechter abschnitten).

### 3.2 Mathematik und Ort des Schulabschlusses

In einem nächsten Schritt wurde eine lineare Regression mit  $N=203$  Studierenden (130 mit deutschem Schulabschluss, 73 mit Schulabschluss mit im Ausland erworbenem Schulabschluss) und der Kontrollvariable Ort des Schulabschlusses Inland (Deutschland) vs. Ausland durchgeführt. Die Variable Ort des Schulabschlusses wurde als Dummy-Variable kodiert, wobei Ausland als Referenzvariable verwendet wurde. Für Ausland ergab sich ein Regressionskoeffizient von 0,001, der nicht signifikant war, da der p-Wert bei über 0,97 lag. Für die Durchschnittsnote Mathematik wurde ein Regressionskoeffizient von 0,494 berechnet, der signifikant war, da der p-Wert [bei] unter 0,01 lag. Es ergab sich also ein signifikanter Einfluss der Durchschnittsnote Mathematik auf den Studienerfolg – während Studierende, die ihren Schulabschluss in Deutschland erworben hatten, und Studierende, die ihren Schulabschluss im Ausland erworben hatten, im Durchschnitt fast den gleichen GPA erzielten.

### 3.3 Mathematik und Art des Schulabschlusses

Nachdem nachgewiesen werden konnte, dass der Ort des Schulabschlusses keinen signifikanten Einfluss auf den Studienerfolg hat, ergab sich die Frage, ob hingegen die Art des Schulabschlusses (Abitur, High School, International Baccalaureate (IB), sonstiger Abschluss) sich womöglich entscheidend auf den GPA auswirkt. Um dieser Frage nachzugehen, wurde eine lineare Regression mit  $N=202$  Studierenden (85 mit Abitur, 61 mit High School-Abschluss, 10 mit IB und 46 mit sonstigem Abschluss) und der Kontrollvariable Art des Schulabschlusses durchgeführt. Art des Schulabschlusses wurde hierzu als Dummy-Variable kodiert; Abitur wurde als Referenzvariable verwendet. Für High School wurde ein Regressionskoeffizient von 0,006 berechnet, der nicht signifikant war, da der p-Wert bei über 0,92 lag; für IB und sonstiger Abschluss wurden Regressionskoeffizienten von 0,72 sowie 0,17 berechnet, die jeweils nicht signifikant waren, da die p-Werte bei über 0,55 (IB) sowie über 0,79 (sonstiger Abschluss) lagen. Der Regressionskoeffizient der Variable Durchschnittsnote Mathematik lag bei 0,493; dieser war signifikant, da der p-Wert (erneut) kleiner als 0,01 war. Somit ergab sich ein signifikanter Einfluss der Durchschnittsnote Mathematik auf den Studienerfolg, während keine signifikante Auswirkung von der Art des Schulabschlusses auf den Studienerfolg nachgewiesen werden konnte. Nichtsdestotrotz konnte gezeigt werden, dass Studierende mit Abitur im Mittel besser abschnitten als Studierende mit anderen Schulabschlüssen.

### 3.4 Mathematik und Abiturnote

In einem weiteren Schritt wurde untersucht, inwiefern die Abiturnote den Studienerfolg voraussagt. Die Abiturnote liegt zwischen 1,00 und 4,00 (eine kleinere Zahl spiegelt eine bessere Note wider). Um den Einfluss der Abiturnote unter Berücksichtigung der Durchschnittsnote Mathematik auf den GPA zu testen, wurde eine lineare Regression mit  $N=71$  Studierenden durchgeführt; hierbei wurde die unabhängige metrische Kontrollvariable Abiturnote genutzt. Die Berechnungen ergaben einen Regressionskoeffizienten von 0,286 für die Abiturnote, der signifikant war, da der p-Wert bei ca. 0,02 lag. Für die Durchschnittsnote Mathematik erhielt man einen Regressionskoeffizienten von 0,392, der aufgrund eines p-Wertes von unter 0,01 ebenfalls signifikant war. Zusammenfassend konnte sowohl ein signifikanter Einfluss der Abiturnote als auch ein signifikanter Einfluss der

Durchschnittsnote Mathematik auf den GPA nachgewiesen werden, wobei die Durchschnittsnote Mathematik im Vergleich mit der Abiturnote einen stärkeren Einfluss auf den GPA ( $0,392 > 0,286$ ) hatte.

### **3.5 Matematik und Bachelorarbeit**

Wie in Abschnitt 2 beschrieben, liegt die Vermutung nahe, dass der Bereitschaft, eine Bachelorarbeit zu verfassen, eine höhere intrinsische Motivation der Studierenden zugrunde liegt, welche sich wiederum positiv auf den Studienerfolg auswirken könnte. Daher wurde eine lineare Regression mit  $N=203$  Studierenden (85 schrieben eine Bachelorarbeit, 118 schrieben keine) und der Kontrollvariable Bachelorarbeit (ja vs. nein) durchgeführt. Die Variable Bachelorarbeit wurde als Dummy-Variable kodiert und nein als Referenzvariable genutzt. Für ja ergab sich ein Regressionskoeffizient von  $-0,19$ , der als signifikant eingestuft werden konnte, da der p-Wert bei unter  $0,01$  lag. Für die Durchschnittsnote Mathematik wurde ein Regressionskoeffizient von  $0,483$  berechnet, der ebenfalls bei einem p-Wert von unter  $0,01$  signifikant war. Dadurch ergab sich, dass einerseits die Durchschnittsnote Mathematik einen signifikanten Einfluss auf Studienerfolg hat und andererseits, dass auch die Bereitschaft, eine Bachelorarbeit zu verfassen, sich signifikant positiv auf den Studienerfolg auswirkt (ein negativer Regressionskoeffizient von  $-0,19$  bedeutet einen niedrigeren GPA bei Studierenden, die eine Bachelorarbeit verfassten, und daher eine bessere Gesamtleistung dieser Studierenden).

### **3.6 Mathematik und Studieneintrittsalter**

Nachfolgend wurde die Beziehung zwischen Studienerfolg und Studieneintrittsalter untersucht. Hierzu wurde eine lineare Regression mit  $N=203$  Studierenden und der Kontrollvariable Studieneintrittsalter durchgeführt. Für diese Variable ergab sich ein Regressionskoeffizient von  $0,006$ , der bei einem p-Wert von über  $0,12$  nicht als signifikant eingestuft werden konnte. Für die Durchschnittsnote Mathematik hingegen ergab sich ein signifikanter Regressionskoeffizient (p-Wert  $< 0,01$ ) von  $0,493$ . Daraus resultierte erneut ein signifikanter Einfluss der Durchschnittsnote Mathematik auf Studienerfolg, während für das Studieneintrittsalter eine nur sehr geringe Bedeutung für den GPA nachweisbar war.

### **3.7 Mathematik und Studiendauer**

Als letzte Kontrollvariable wurde die Studiendauer getestet. Hierzu wurde eine lineare Regression mit  $N=203$  Studierenden vorgenommen. Für Studiendauer ergab sich ein Regressionskoeffizient von  $0,01$ , der nicht signifikant war, da der p-Wert bei über  $0,07$  lag. Für die Durchschnittsnote Mathematik konnte ein Regressionskoeffizient von  $0,458$  berechnet werden, der signifikant war, da der p-Wert bei unter  $0,01$  lag. Wie zuvor konnte ein signifikanter Einfluss der Durchschnittsnote Mathematik auf den GPA nachgewiesen werden, während dies in Bezug auf die Studiendauer nicht der Fall war.

### **3.8 Regression mit allen Variablen**

Schließlich wurde noch eine lineare Regression mit der abhängigen Variable GPA, der unabhängigen Variable Durchschnittsnote Mathematik sowie allen bislang getesteten Kontrollvariablen durchgeführt ( $N=71$ ). Hierbei ergab sich lediglich für die

Durchschnittsnote Mathematik (Regressionskoeffizient 0,368 bei p-Wert  $< 0,01$ ) und die Abiturnote (Regressionskoeffizient 0,251 bei p-Wert von ca. 0,04) ein signifikanter Einfluss auf den GPA. Da der Regressionskoeffizient der Durchschnittsnote Mathematik höher war als derjenige der Abiturnote ( $0,368 > 0,251$ ), lässt sich ein stärkerer Einfluss der Durchschnittsnote Mathematik im Vergleich zu der Abiturnote auf den GPA aufzeigen; dieses Ergebnis lässt sich mit der vierten Regressionsanalyse in diesem Abschnitt in Einklang bringen.

### 3.9 Interaktionseffekt

Das Resultat der letzten Regressionsanalyse führte zu der Frage, ob die Abiturnote sich signifikant auf die Durchschnittsnote Mathematik auswirkt, und ob sie daher als ursprünglich verantwortlich für Studienerfolg anzusehen ist. Zur Klärung dieser Frage wurde der Regressionskoeffizient des Interaktionsterms zwischen Abiturnote und der Durchschnittsnote Mathematik berechnet. Die Berechnung ergab einen Regressionskoeffizienten von 0,172, der nicht signifikant war, da der p-Wert größer als 0,12 war. Dadurch konnte ein signifikanter Interaktionseffekt ausgeschlossen werden.

Abschließend sollte bemerkt werden, dass alle soeben beschriebenen Regressionsanalysen auch mit einer leicht modifizierten abhängigen Variablen durchgeführt wurden. Anstelle des GPA wurde der GPA ohne die Mathematiknoten verwendet. Es ergaben sich dabei keine signifikanten Unterschiede zu den genannten Ergebnissen.

## 4. Fazit

Es lassen sich mehrere Schlüsse aus der Analyse ziehen. Der Wichtigste ist der, dass alle Regressionsanalysen auf einen signifikanten Einfluss der Durchschnittsnote Mathematik auf den Studienerfolg hinweisen (bei allen Regressionen lag der p-Wert [bei] unter 0,01). Zudem hatte die Durchschnittsnote Mathematik den stärksten Einfluss auf die Studienabschlussnote. Neben der Durchschnittsnote Mathematik scheint auch die Abiturnote Studienerfolg vorherzusagen; da allerdings nachgewiesen werden konnte, dass die Abiturnote keinen signifikanten Einfluss auf die Durchschnittsnote Mathematik hat, ist davon auszugehen, dass es sich bei der Durchschnittsnote Mathematik tatsächlich um einen unmittelbaren Indikator für Studienerfolg handelt. Daher kann, basierend auf den durchgeführten Analysen, die in Abschnitt 1 formulierte Hypothese verworfen werden. Dies legt nahe, die Mathematikmodule konzeptionell zu überarbeiten. Neben der Durchschnittsnote Mathematik und der Abiturnote hatte auch die Bereitschaft, eine Bachelorarbeit zu verfassen, einen signifikant positiven Einfluss auf die Studienabschlussnote. Dies lässt vermuten, dass Studierende, die eine Bachelorarbeit verfassen, eine höhere (intrinsische) Motivation für ihr Studium aufbringen, was sich wiederum positiv auf den Studienerfolg auswirken könnte [diese Hypothese könnte Anlass für eine weitere Untersuchung sein]. Für alle weiteren getesteten unabhängigen Variablen konnte kein signifikanter Einfluss auf die Studienabschlussnote festgestellt werden. Allerdings konnte gezeigt werden, dass weibliche Studierenden etwas besser abschnitten als männliche, jüngere Studierende etwas besser als ältere und dass eine längere Studiendauer sich im Durchschnitt leicht negativ auf die Studienabschlussnote auswirkt. Interessanterweise konnte in Bezug auf Studienerfolg (fast) kein Unterschied

zwischen Studierenden mit deutschem Schulabschluss und Studierenden mit im Ausland erworbenem Schulabschluss ermittelt werden.

## 5. Zusammenfassung und Ausblick

Einschränkend muss erwähnt werden, dass für die Regressionsanalysen lediglich eine kleine Auswahl an Kontrollvariablen genutzt wurde. Es wäre denkbar, dass unter Verwendung weiterer Kontrollvariablen die Ergebnisse der Regressionen den Schluss zuließen, dass die Durchschnittsnote Mathematik keinen direkten Einfluss auf Studienerfolg hat. Um die Hypothese, dass die Durchschnittsnote Mathematik keinen direkten Einfluss auf Studienerfolg hat, noch eingehender zu untersuchen, wäre es daher lohnenswert, Regressionsanalysen mit weiteren Kontrollvariablen durchzuführen. Eine weitere Einschränkung betrifft die Auswahl der statistischen Verfahren. So wurden im Rahmen der vorliegenden Untersuchung neben Korrelationen lediglich lineare Regressionen durchgeführt. In einer weiteren Untersuchung mit neuen Kontrollvariablen ließe sich die Auswahl an statistischen Verfahren etwa durch die Verwendung von Strukturgleichungsmodellen erweitern.

Obwohl aufgrund der limitierten Anzahl an Kontrollvariablen Zweifel daran erlaubt sind, dass die Durchschnittsnote Mathematik einen direkten Einfluss auf Studienerfolg hat, deuten doch alle in diesem Artikel dargestellten Analysen sowie zahlreiche andere Studien auf einen unmittelbaren Einfluss der Durchschnittsnote Mathematik auf die Studienabschlussnote hin (Große/Geller-Urban, 2017, S.22; Parker, 2005, S.30). Dies legt die Vermutung nahe, dass der Erwerb mathematischer Kompetenzen während des Studiums einen signifikant positiven Effekt auf die Gesamtleistung der Studierenden hat. Daher scheint es sinnvoll, in den Mathematikmodulen neue Lehrmethoden anzuwenden, die es den Studierenden ermöglichen, komplizierte mathematische Sachverhalte grundlegend besser zu verstehen und dadurch ihre Mathematikkompetenzen sowie ihre Studienleistungen im Allgemeinen zu verbessern.

Geplant ist unter anderem die Einführung alternativer didaktischer Methoden (z.B. der sokratischen Methode) in den verschiedenen Modulen des Faches Mathematik. Sollte sich die Einführung der sokratischen Methode als erfolgreich erweisen, ließe sich auch über deren Anwendung in weiteren Modulen nachdenken (Kanakis, 1997, S.234 ff.). Darüber hinaus könnte ein modifiziertes Angebot – etwa in Form zusätzlicher (unbenoteter) Module oder eines zusätzlichen Helpdesk Mathematik – langfristig zu einer Verbesserung der Mathematikkompetenzen der Studierenden und somit der Studienleistung insgesamt führen (Fahrner/Wolf, 2018, S.81).

## 6. Literaturverzeichnis

Fahrner, M. & Wolf, B. (2018). Analysis of Skills at Entry and Overall Academic Performance in Undergraduate Management Studies from 2005/06 to 2015/16. *ZeE-Publikationen, Reihe empirische Evaluationsmethoden*, 22, 75-82.

Große, A. & Geller-Urban, K. (2017). Monitoring der Studieneingangsphase im Fach Mathematik an der Ernst-Abbe-Hochschule Jena / Fachbereich Grundlagenwissenschaften. *Die neue Hochschule 1* (2017), 22-25.

Kanakis, I. (1997). Die sokratische Lehrstrategie und ihre Relevanz für die heutige Didaktik. *International Review of Education* 43, no. 2-3, 225-240.



Lin, S. (2017). Women win through averages, Men win through extremes: An empirical examination of Business Student Gpa Data. *Chinese Sociological Dialogue* 2, no. 3-4, 120-135.

Owen, T. R. (2003). Retention implications of a relationship between Age and Gpa. *College Student Journal* 37, no. 2, 181-190.

Parker, M. (2005). Placement, Retention, and Success: A longitudinal study of Mathematics and Retention. *The Journal of General Education* 54, no. 1, 22-40.

Rotter, C., (2011). Vertrauen ist gut, Kontrolle Ist besser?! Hochschulische Auswahlverfahren von Studierenden im internationalen Vergleich. *Tertium Comparationis* 17, no. 1, 3-18.

Vella, E. J., Turesky, E.F. & Hebert, J. (2016). Predictors of academic success in Web-based courses: Age, Gpa, and instruction mode. *Quality Assurance in Education* 24, no. 4, 586-600.



## **Varianzanalytische Auswirkungen faktorenanalytischer Skalenkonstruktionen**

Monika Luzi Beyer, Katrin Schüttpelz-Brauns & Inga Großmann

### **Zusammenfassung**

In der Psychologie ist es üblich, mit Hilfe von Faktoranalysen Skalen für Fragebögen zu bilden und hinter diesen Skalen ein Merkmalskonstrukt zu postulieren. Doch wie zuverlässig und replizierbar sind Faktoranalysen? Um diese Frage zu beantworten, werden Persönlichkeitsdaten aus verschiedenen Subgruppen einer Gesamtstichprobe von  $N=3.986$  zur faktorenanalytischen Skalenkonstruktion herangezogen und anschließend verglichen. Die Ergebnisse zeigen, dass die Zuverlässigkeit der Faktorenlösung nicht nur maßgeblich mit der Größe der gezogenen Stichprobe zusammenhängt, sondern zusätzlich auch mit deren Homogenität in den gemessenen Faktorwerten. Für die Praxis kann dies bedeuten, dass Skalenkonstruktionen immer anhand von Stichproben mit  $n > 1.000$  durchgeführt werden müssen. Nur so sind die gemessenen Personenwerte, die in die Faktorenlösung mit einfließen, für die Gesamtpopulation repräsentativ und weisen somit repräsentative Varianzen und Mittelwerte auf.

### **Schlüsselwörter**

Faktorenanalyse, Stabilität, Testkonstruktion, Skalenkonstruktion, Reliabilität, Fragebogen

### **Abstract**

In psychology it is common to use factor analysis to form scales for questionnaires and to postulate a characteristic construct behind these scales. But how reliable and replicable are factor analyses? To answer this question, personality data from different subgroups of a total sample of  $N=3,986$  are used for factor-analytical scale construction and subsequently compared. The results show that the reliability of the factor solution is not only significantly related to the size of the sample, but also to its homogeneity in the measured factor values. In practice, this can mean that scale constructions must always be carried out on the basis of samples with  $n > 1,000$ . Only in this way, the measured personal values included in the factor solution are representative for the total population and thus exhibit representative variances and mean values.

### **Keywords**

factor analysis, stability, test construction, scale construction, reliability, questionnaire

## **1. Stand der Forschung zu Robustheit von Faktoranalysen**

Skalenkonstruktionen in der Psychologie und in den Erziehungswissenschaften erfolgen mitunter über explorative Faktoranalysen. Allein für das Jahr 2018 erhält man mit dem Suchalgorithmus „factor analysis“ and „scale development“ in ERIC insgesamt 24 Treffer. Von diesen Treffern gibt es in 15 Abstracts Hinweise auf die Anzahl der Items und Versuchspersonen mit denen explorativen Faktoranalysen

zur Konstruktion von Skalen berechnet wurden. Die Anzahl der Items beträgt über alle 15 Abstracts hinweg im Mittel  $M=25$  ( $SD=8,28$ ; Range: 13-44; Median:  $Med=23,5$ ) und die Anzahl der Versuchspersonen  $M=910$  ( $SD=1942,99$ ; Range: 106-8132; Median:  $Med=320,5$ ). Betrachtet man das Verhältnis von Versuchspersonen pro Item finden sich im Mittel  $M=764,97$  ( $SD=131,62$ ; Range: 6,5-542,13; Median:  $Med=12,8$ ). Wie man sieht, gibt es einen sehr großen Range in der Anzahl der Items, Versuchspersonen und dem Verhältnis von Versuchspersonen pro Item.

Die häufigsten Empfehlungen in der Literatur beziehen sich auf die absoluten Stichprobengrößen (N Anzahl der Personen). Typische Empfehlungen für Stichprobengrößen N sind hier als Übersicht zusammengestellt und variieren von mindestens 100 (beispielsweise Gorsuch, 1983 oder Kline, 1979) über mindestens 200 (als bekanntester Vertreter hier Guilford, 1954) bis hin zu mindestens 250 (z.B. Cattell, 1978).

Andere Autoren differenzieren bei ihren Handlungsempfehlungen nach verschiedenen Kriterien. Beispielsweise kann basierend auf der Stichprobengröße eine Qualitätseinschätzung vorgenommen werden: Comrey and Lee (1992) bezeichnen Stichprobengrößen ab 100 als „poor“, ab 200 als „fair“, ab 300 als „good“, ab 500 als „very good“ und ab 1000 als „excellent“. MacCallum, Widaman, Zhang & Hong (1999) postulieren, dass zusätzlich zur Stichprobengröße auch die Korreliertheit der Variablen Einfluss auf die Stabilität der Ergebnisse habe.

Noch differenzierter argumentieren Autoren, die ein bestimmtes Verhältnis aus Stichprobengröße N und Anzahl der zu faktorierenden Variablen p beschreiben. Nach Cattell (1978) soll das Verhältnis N:p beispielsweise mindestens 3 und höchstens 6 betragen, was bedeutet, es soll mindestens 3 bis 6 mal so viele Fälle(N) wie Variablen (p) geben. Bei einem  $N=100$  dürfte es daher zwischen 17 und 33 zu faktorierenden Variablen geben. Gorsuch (1983) spricht von einem Verhältnis von mindestens 5, Everitt (1975) von mindestens 10. Bei einem  $N=100$  entspräche dies 10 bis 20 Variablen.

Mögliche weitere Einflüsse auf die Stabilität der Faktorenlösungen neben N und p, sind bisher weitestgehend ungeklärt. Beispielsweise wurde ein möglicher Effekt der Homogenität der gezogenen Stichprobe bisher nicht untersucht, wenn auch diskutiert (Harrison & List, 2004).

Ziel der vorliegenden Studie ist es daher, verschiedene Einflüsse auf die Zuverlässigkeit von faktoranalytischer Skalenkonstruktion zu untersuchen.

## 2. Methoden

### Stichprobe

Als Grundlage diente ein Beispieldatensatz, veröffentlicht im „Statistik mit SPSS“ (Diehl & Staufenberg, 2007), mit einer Stichprobengröße von  $N=3.986$  (Frauen  $n=2.868$ , Männer  $n=1.045$ ). Es handelt sich hierbei um Studierende, die 9 Rechenaufgaben mit dichotomem Ergebnis (0 und 1 – richtig und falsch) gelöst haben.

Im nächsten Schritt wurden homogene und heterogene Subgruppen erzeugt. Homogene Subgruppen wurden zum einen über das Geschlecht eingeteilt (= 2 Untergruppen) und zum anderen über den Fähigkeitswert des Hauptfaktors (12 Untergruppen). Die Einteilung nach dem Fähigkeitswert soll die Homogenität der Stichproben variieren und damit ein typisches Problem aus der Praxis zeigen.

Im Allgemeinen werden in der Praxis keine Zufallsstichproben aus Grundgesamtheiten gezogen, sondern anfallende Stichproben zur Grundlage

genommen (zum Beispiel Studierende, Mitarbeiter eines Standortes, dieselbe Gemeinschaft usw.). Dies gilt als unproblematisch, „wenn sie [die Teilerhebung] in der Verteilung aller interessierenden Merkmale der Gesamtmasse entspricht, d.h. ein zwar verkleinertes, aber sonst wirklichkeitsgetreues Abbild der Gesamtheit darstellt.“ (Berekoven u.a. 1999, S. 50). Das „interessierende Merkmal“ hat jedoch im Allgemeinen eine unbekannte Verteilungsfunktion. Genau das ist meist der Grund der Untersuchung. So schreibt beispielsweise Rogge (1981) „Das Ziehen einer Stichprobe aus den Studenten einer Hochschule für die Durchführung eines Produkttests einer Joghurtmarke ist ein relativ schnelles Stichprobenverfahren. Das Geschmacksempfinden [...] ist aber wegen des geringen Einkommens [Unterstreichungen von den Verfassern] von Studenten keineswegs repräsentativ für die mögliche Gesamtkäuferschaft“ (Rogge 1981, S. 168). Diese anfallenden Stichproben haben oft eine geringere Varianz und repräsentieren zusätzlich meist nicht den Mittelwert der Grundgesamtheit, beispielsweise zeigen Mitarbeiter eines Standortes oder Studierende in bestimmten Bereichen einheitlichere Einstellungs-, Fähigkeits- und Interessenwerte als im Vergleich zur Gesamtpopulation. Auch Harrison und List (2004) beschreiben das methodische Problem, welches entsteht, wenn Aussagen von Studierenden als übertragbar auf das Feld interpretiert werden. Gleiches kann von allen selektiven Stichproben angenommen werden, wenn deren Verteilung des interessierenden Merkmals nicht der der Grundgesamtheit entspricht.

Tabelle 1 beinhaltet die deskriptive Statistik der Subgruppen für den Fähigkeitswert des Hauptfaktors „mathematische Fähigkeit“ aus der Gesamtstichprobe.

**Tab. 1:** Deskriptive Statistiken der Subgruppen „Fähigkeitswert“ und „Geschlecht“

	N	Mittelwert	Std.-Abweichung
1	344	-.781	.088
2	381	-.560	.030
3	169	-.474	.015
4	100	-.422	.014
5	267	-.322	.035
7	879	-.228	.039
8	328	-.043	.041
9	179	.076	.025
10	194	.218	.043
11	380	.463	.120
12	129	.826	.160
Gesamt	3350	-.169	.408
Frauen	2868	.025	1.025
Männer	1045	-.070	.925

*Anmerkung:* Stichprobengröße, mittlerer Fähigkeitswert (Wertebereich -1 bis 1) und Standardabweichung der Stichproben.

Für die Simulation heterogener Subgruppen wurden drei Zufallsstichproben unterschiedlicher Größe erzeugt, entsprechend der Empfehlungen  $n=150$  mit  $n:p=16,7$ ,  $n=200$   $n:p=22,2$  und  $n=300$   $n:p=33,3$ .

### 3. Statistische Analyse

Es wurden zunächst Hauptkomponentenanalysen mit Varimaxrotation sowohl für die Gesamtstichprobe als auch für die homogenen und heterogenen Subgruppen berechnet. Dann wurden der Einfluss von Geschlecht, Stichprobengröße, Varianz und Ausprägung der einschränkung der Varianz der Fähigkeitswerte auf die Reliabilität der Faktorenlösung evaluiert. Die Reliabilität der Subgruppen bzgl. der Gesamtstichprobe wurde sowohl auf Personen- als auch auf Itemebene überprüft. Dazu wurden einerseits die Faktorwerte und andererseits die Personenkenwerte der Subgruppe mit denen der Gesamtstichprobe korreliert.

Die Zuverlässigkeit wurde also aus zwei Perspektiven bestimmt:

1. kann in den Faktorenlösungen der Substichproben eine ähnliche Struktur der einzelnen Variablen erzeugt werden (Ladung der Items) und
2. werden die Personenwerte in den Substichproben ähnlich geschätzt wie in der Faktorenlösung der Gesamtstichprobe?

Im weiteren Verlauf wurde deshalb geprüft, ob (1) der Fähigkeitswert, den Personen auf dem ersten Faktor haben, auch in den Subgruppen repliziert werden kann (Korrelation der Personenkenwerte mit dem ersten und dem zweiten Faktor = Zuverlässigkeit der Vorhersage der Personenparameter) und ob (2) die Ladungen der Items auf den Faktoren zu vergleichbaren Lösungen kommen. Korreliert werden also die Itemladungen der ersten zwei Faktoren der Subgruppen mit den Ladungen der Gesamtstichprobe (Korrelation der Itemstrukturen = Zuverlässigkeit der Bestimmung der Itemordnung).

Von einer ähnlichen Faktorenlösung kann ausgegangen werden, wenn der Fähigkeitswert der Subgruppe auf Faktor 1 mit dem der Gesamtstichprobe auf Faktor 1 hoch korreliert ( $r_{11}$ ) und der Fähigkeitswert der Subgruppe auf Faktor 2 mit dem der Gesamtstichprobe auf Faktor 2 ( $r_{22}$ ), während Nullkorrelationen für die Kombinationen von Faktor 1 und 2 ( $r_{21}$  und  $r_{12}$ ) vorliegen. Analog gilt dies für die Reliabilitätsbetrachtung der Itemladungen.

### 4. Ergebnisse

Reliabilitätsbetrachtung der Fähigkeitswerte

Tabelle 2 zeigt, dass sowohl die Einteilung nach Geschlecht als auch die Zufallsstichproben den Fähigkeitswert der Personen beim ersten Faktor mit  $r > .9$  sehr zuverlässig replizieren konnten. Dies erfolgt unabhängig von der Stichprobengröße bereits bei einem  $n$  von 145. Der Fähigkeitswert des zweiten Faktors kann bei einigen dieser Substichproben ebenfalls repliziert werden. In den Subgruppen, welche die Personen mit unterschiedlichen Fähigkeitswerten enthalten, ist dies nicht der Fall. Auffällig ist hier aber, dass in der Zufallsstichprobe 150 (bei der mit listenweisem Fallausschluss 145 Personen in die Berechnung aufgenommen wurden) die Varianzaufklärung des zweifaktoriellen Modells überschätzt wird, mit 63,5% erklärter Varianz anstatt 44%.

**Tab. 2:** Ähnlichkeiten der Personenstrukturen: Reliabilität der Fähigkeitswerte.

Subgruppe	n	Aufgeklärte Varianz [%]	$r_{11}$ <b>Fähigkeits- wert</b>	$r_{21}$ Fähigkeits- wert	$r_{12}$ Fähigkeits- wert	$r_{22}$ <b>Fähigkeits- wert</b>
0	3.913	44.0				
w	2.868	43.3	<b>1.00</b>	.059	-.024	<b>.997</b>
m	1.045	46.8	<b>.996</b>	-.191	.082	<b>.951</b>
Zufall300	293	43.4	<b>.993</b>	.034	-.011	.575
Zufall200	196	44.1	<b>.991</b>	.033	.095	<b>-.853</b>
Zufall150	145	63.5	<b>.999</b>	-.392	-.007	.486
1	563	61.5	<b>.995</b>	-.021	.084	<b>-.989</b>
2	344	33.3	.058	.496	-.170	-.606
3	381	37.9	-.411	-.159	-.252	-.030
4	169	38.8	-.331	.466	-.441	.360
5	100	38.1	.498	-.431	-.087	.332
6	267	31.1	.162	-.836	-.090	-.318
7	879	34.1	.540	-.502	-.123	-.768
8	328	31.3	-.267	<b>.947</b>	-.154	-.168
9	179	33.0	-.138	.824	.240	-.478
10	194	29.6	.127	-.503	-.063	.037
11	380	30.5	-.059	.725	.001	.375
12	129	32.7	.116	.576	.071	.306

*Anmerkungen.* Korrelationen Gesamtlösung und Subgruppenlösung. Die zwei Fähigkeitswerte der Personen in der Stichprobenlösung werden mit denen der Gesamtlösung korreliert (Spalten 4 bis 7). Wobei die Indices der Korrelation ( $r$ ) den Faktor der Subgruppe (1 oder 2) und Gesamtstichprobe (1 oder 2) zeigen. Gezeigt wird außerdem, wie viel Varianz die zweifaktorielle Lösung der Stichproben erklärt (Spalte 3).

#### Reliabilitätsbetrachtung der Itemkennwerte

Bei der Prüfung der Itemstruktur der Faktoren zeigen sich insgesamt keine reliablen Ergebnisse. Lediglich bei den sehr großen Stichproben (Männer und Frauen  $n > 1000$ ) haben die Items vergleichbare Ladungen auf den Faktoren. Die gefundenen Strukturen in den kleineren Substichproben kommen nicht zu ähnlichen Lösungen wie es bei der Analyse der Personenwerte der Fall war.

**Tab. 3:** Ähnlichkeiten der Faktorenstrukturen: Reliabilität der Itemladungen.

	$r_{11}$ Itemladung	$r_{21}$ Itemladung	$r_{12}$ Itemladung	$r_{22}$ Itemladung
w	<b>.949</b>	-.326	-.578	<b>.997</b>
m	<b>.864</b>	<b>-.892</b>	-.637	<b>.951</b>
Zufall293	.604	-.071	-.385	.533
Zufall196	.479	-.044	.376	<b>-.800</b>
Zufall145	.599	<b>-.922</b>	-.374	.476
1	.291	.429	.450	<b>-.966</b>
2	.017	.124	<b>.968</b>	-.633
3	.301	-.136	-.101	.103
4	.503	.275	-.034	.181
5	.129	-.405	.000	.005
6	.602	-.791	.042	-.218
7	.636	-.474	.457	-.754
8	-.662	<b>.935</b>	-.587	-.137
9	-.440	.787	.619	-.526
10	-.362	-.369	.297	-.035
11	-.592	.686	-.561	.357
12	-.656	.552	-.273	.338

*Anmerkungen.* Korrelationen Gesamtlösung und Subgruppenlösung. Die zwei Itemladungen der Substichproben werden mit denen der Gesamtstichprobe korreliert (Spalten 4 bis 7). Wobei die Indices der Korrelation ( $r$ ) den Faktor der Subgruppe (1 oder 2) und Gesamtstichprobe (1 oder 2) zeigen. Gezeigt wird außerdem, wie viel Varianz die zweifaktorielle Lösung der Stichproben erklärt (Spalte 3).

## 5. Interpretation und Ergebnisdiskussion

Ist das Ziel der Faktorenanalyse lediglich die korrekte Bestimmung von Fähigkeitswerten von Personen, so scheint dies bereits bei kleinen Zufallsstichproben möglich zu sein. In den meisten Anwendungsfällen, in denen faktorielle Strukturen konstruiert werden, ist dies jedoch nicht das Ziel. Vielmehr wird die Faktorenstruktur inhaltlich meist als die zugrundeliegende tatsächliche Struktur unabhängiger Eigenschaftsaspekte gewertet.

Sollen die erzielten faktoriellen Strukturen als Grundlage für solcherlei inhaltliche Interpretationen herangezogen werden, so sehen wir im vorgestellten Fall mit nur neun Items und einer N:p-Ratio  $\geq 111$  erst bei Stichproben mit einem N größer 1.000 verbindliche Strukturreplikationen.

Homogene Stichproben, bezogen auf die Variablen, die in die Faktorenanalyse eingehen, sind völlig ungeeignet, um Personen- und Itemparameter zuverlässig zu bestimmen. Hierfür dürfen nur heterogene (Zufalls-)Stichproben verwendet werden.

## 6. Literaturverzeichnis

Berekoven, L., Eckert, W. & Ellenrieder, P. (1999). Marktforschung: Methodische Grundlagen und praktische Anwendung, 8. Auflage, Wiesbaden

Cattell, R. B. (1978). The scientific use of factor analysis. New York: Plenum.



- Comrey, A. L., & Lee, H. B. (1992). *A first course in factoranalysis*. Hillsdale, NJ: Erlbaum.
- Everitt, R. S. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, 126, 257-240.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, 68(3), 532-560.
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
- Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, 19, 149-162.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic literature*, 42(4), 1009-1055.
- Kline, P. (1979). *Psychometrics and psychology*. London: Academic Press.
- Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best practices in exploratory factor analysis. *Best practices in quantitative methods*, 86-99.
- Rogge, H.-J. (1981). *Marktforschung: Elemente und Methoden betrieblicher Informationsgewinnung*, Hanser, München/Wien



## Zur Stabilität von Zuverlässigkeitsmaßen

Bodo Krause

### Zusammenfassung

In den letzten 20 Jahren gab es verstärkt Studien, die die Stabilität klassischer Zuverlässigkeitsaussagen zum Gegenstand hatten. Die Reliabilitätstheorie ist ein allgemeiner Zugang zur Kennzeichnung der Zuverlässigkeit empirischer, insbesondere testtheoretischer Befunde.

Für metrische Beobachtungsdaten sind das Kennzahlen, die auf dem Ansatz der klassischen Testtheorie basieren, die als Modell den Beobachtungswert  $X$  einer Person als Summe eines wahren Werts  $t$  und eines davon unabhängigen Fehlerwerts  $E$  annimmt. Als Modell für eine Grundgesamtheit ergibt sich

$$X = T + E.$$

Die Variabilität der Daten in einen wahren und einen Fehleranteil aufspalten, wobei deren Verhältnis die Zuverlässigkeit kennzeichnet:

$$\text{Rel}(X) = \text{Var } T / \text{Var } X = 1 - \text{Var } E / \text{Var } X.$$

Die Aussagekraft dieses Maßes wird kritisiert, da in diesem Ansatz einerseits der Fehler, genauer der unaufgeklärte Rest, nicht differenziert wird und andererseits zusätzlich eine Populationsabhängigkeit der Aussagen besteht. Neben einer metaanalytischen begründeten verallgemeinerten Reliabilität entwirft Bentler (2016) durch die Einbeziehung von Kovariablen einen differenzierenden Zugang.

Für nichtmetrische, vor allem dichotome Beobachtungsdaten wird die Zuverlässigkeit wesentlich auf der Grundlage der Formeln von Kuder-Richardson (1937), vor allem KR 20 und KR21 gekennzeichnet. Daraus abgeleitet die Formeln von Guttman (1945) und der gebräuchliche Koeffizient  $\alpha$  von Cronbach (1951). Zur Beurteilung der Stabilität dieser Maßzahl entwickelten Pauly und Umlauf (2018) einen Zugang, der auf dem Resampling basiert.

Insgesamt verbleibt die Frage, ob und in welchem Umfang diese Zuverlässigkeitsmaße der „technischen Güte“ stabil gegenüber einer Variation der Messobjekte sind. Die Konsequenzen aus den beiden Zugängen werden dargestellt und diskutiert.

### Keywords

reliability, reliability generalization, stability, permutation, bootstrapping

### 1. Zum Gegenstand

Die Reliabilitätstheorie ist ein allgemeiner Zugang zur Kennzeichnung der (sogenannten „technischen“) Zuverlässigkeit empirischer, insbesondere testtheoretischer Befunde.

Für metrische Beobachtungsdaten sind das Kennzahlen, die auf dem Ansatz der klassischen Testtheorie basieren. Diese postulieren einen Modellansatz, in dem die Beobachtungswerte (Messwerte)  $X$  der Personen als Summe ihrer wahren Werte  $T$  und davon unabhängigen Fehlerwerte  $E$  annimmt. Die Fehlerwerte werden zusätzlich

als voneinander unabhängig und normalverteilt vorausgesetzt. Für eine Grundgesamtheit (Population) ergibt sich der Modellansatz der klassischen Testtheorie

$$X = T + E.$$

Aus diesem Modellansatz ergibt sich, dass die Variabilität der Daten in einen wahren und einen davon unabhängigen Fehleranteil aufgespalten wird:

$$\text{Var}(X) = \text{Var}(T) + \text{Var}(E).$$

Das Verhältnis dieser Varianzanteile kennzeichnet dann die Zuverlässigkeit der Beobachtungsmethode (Reliabilität):

$$\text{Rel}(X) = \rho^2(X, T) = \frac{\text{Var}(T)}{\text{Var}(X)} = 1 - \frac{\text{Var}(E)}{\text{Var}(X)}$$

Die Aussagekraft dieses Maßes wird problematisiert, da in diesem Ansatz

a) der Fehler, genauer der unaufgeklärte Rest, und die Modellannahme der Unabhängigkeit nicht differenzierter betrachtet werden, und

b) zusätzlich eine Populationsabhängigkeit des Reliabilitätskoeffizienten und damit der Zuverlässigkeitsaussage bestehen kann.

Schon 1999 betonen Wilkinson & APA Task Force on Statistical Inference in ihrem Beitrag „Statistical methods in psychological journals: Guidelines and explanations“, dass “a test is not reliable ore unreliable. Reliability is a property of the score on a test for particular population of examinees”.

Damit wird eine Unterscheidung von Beobachtungswert (Messwert, Score) und dem Beobachtungsverfahren (Messverfahren, Messinstrument) betont, die auch eine differenziertere Betrachtung der Gütwerte bedingt.

Bezogen auf die Zuverlässigkeitsaussage ergibt sich daraus, dass diese sowohl bzgl. des Beobachtungswerts (Score-Reliabilität) als auch bzgl. des Beobachtungsverfahrens (verallgemeinerte Reliabilität, Reliability generalization RG) erfolgen kann. (Das alles gilt analog auch bzgl. der Gültigkeitsaussage (Validität), vgl. Schmidt & Hunter, 1977).

Für die Beurteilung der verallgemeinerten Reliabilität (RG) einer Beobachtungsmethode ist ein methodischer Grundzugang durch die Metaanalyse gegeben und auch durch Vacha-Haase (1998) begründet worden. Er setzt allerdings, wie bei Metaanalysen erforderlich, eine hinreichende Anzahl unterschiedlicher Studien mit dieser Beobachtungsmethode voraus.

Als eine Alternative zur Metaanalyse für die Beurteilung der verallgemeinerten Reliabilität entwirft Bentler (2016) durch die Einbeziehung und Berücksichtigung von Kovariablen einen differenzierenden Zugang, mit dem der Einfluss dieser Kovariablen und damit deren Wirkung (und auch eine Populationsabhängigkeit) geprüft werden kann. Als Analysemethoden können sowohl die Kovarianzanalyse als auch Strukturgleichungsmodelle genutzt werden.

Für nichtmetrische, vor allem dichotome Beobachtungsdaten wird die Zuverlässigkeit wesentlich auf der Grundlage der Formeln von Kuder-Richardson (1937), vor allem KR 20 und KR21 gekennzeichnet. Daraus abgeleitet sind die Kennwerte  $\lambda_3$  und  $\lambda_6$  von Guttman (1945) und der gebräuchliche Koeffizient  $\alpha$  von Cronbach (1951). Die

Stabilität dieser Maßzahl kann ebenfalls durch die Einbeziehung von Kovariablen differenzierter beurteilt werden.

Einen weiteren alternativen Zugang für eine solche Beurteilung entwickelten Pauly und Umlauf (2018) auf der Basis des Vergleichs von Stichprobenverteilungen, die durch ein Resampling aus einer Grundstichprobe erzeugt werden. Hierbei wird die unbekannte Verteilungsfunktion in der Grundgesamtheit durch die empirische Verteilungsfunktion aus der Stichprobe ersetzt und die aus der Literatur bekannte Prozeduren der Permutation (vgl. Fisher, 1935; Good, 1995, 2000) und des Bootstrapping (begründet von Efron, 1985; Efron & Tibshirani, 1993) genutzt. Die statistische Beurteilung erfolgt dann über die Nullhypothese, dass alle diese Stichproben hinsichtlich ihrer Verteilung und/oder von Verteilungsparametern homogen sind.

Wir wollen nachfolgend die Grundgedanken dieser Zugänge darstellen und ihre Konsequenzen für die Stabilitätsaussage von Zuverlässigkeitsmaßen diskutieren.

## 2. Der kovarianzanalytisch begründete Zugang von Bentler (2016)

### 2.1 für metrische Daten

Ausgangspunkt für Bentler sind u.a. Darstellungen von Raykov und Marcoulides (2011, 2013) über die bekannte Populationsabhängigkeit von Gütekriterien im Rahmen des Modellansatzes der Klassischen Testtheorie. Insbesondere betont werden dabei metaanalytische Aussagen über die Stabilität der Reliabilität und eine damit begründete generalisierte Reliabilitätsaussage. Dafür können auch Strukturgleichungsmodelle eingesetzt werden.

Alle diese generalisierenden Zugänge stehen vor dem Problem, hinreichend viele Studien zu einem Parameter bei variierenden Populationen zu erhalten. Dies begründet die Frage nach alternativen Möglichkeiten zur Beurteilung der Stabilität. Einen alternativen Zugang zur Beurteilung der Stabilität entwickelt Bentler, indem er einen möglichen Einfluss von Kovariablen in die Beurteilung der Aussagekraft explizit aufnimmt.

Genauer: Bentler geht davon aus, dass die Wirkung einer Menge  $Z$  von Kovariablen dazu führen sollte, dass sich der wahre Testpunktwert  $T$  in zwei voneinander unabhängige Komponenten aufspalten lässt (Regression von  $T$  bzgl.  $Z$ ) :

$$T = T^{(Z)} + T^{(\sim Z)}$$

Dabei kennzeichnet  $T^{(Z)} = T(Z)$  den von den Kovariablen abhängigen Anteil von  $T$  und  $T^{(\sim Z)} = T - T(Z)$  den davon unabhängigen Anteil im True Score. Damit kann dann auch die Reliabilität in zwei unabhängige Anteile aufgespalten werden:

$$\text{Rel}(X) = \rho^2(X, T) = \rho^2(X, T^{(Z)}) + \rho^2(X, T^{(\sim Z)}).$$

Entscheidend wird dann, ob der Anteil  $T^{(\sim Z)}$  eine Veränderung in der Größe der Reliabilität bewirkt oder diesbezüglich unabhängig ist. Ist  $\rho^2(X, T^{(\sim Z)})$  groß im Vergleich zu  $\rho^2(X, T)$ , dann gibt es kaum eine Populationsabhängigkeit des Reliabilitätskoeffizienten, er ist weitgehend generalisierbar. Ist dagegen  $\rho^2(X, T^{(Z)})$  groß im Vergleich zu  $\rho^2(X, T)$ , dann ist eine Populationsabhängigkeit anzunehmen. Wichtig in diesem Zusammenhang ist, dass für diesen Zugang die Kovariablen bekannt und erfasst sein müssen. Thompson (1994) verweist darauf, dass u.a. Alter,

Geschlecht und Bildungsgrad als wirksame Kovariable in der Literatur ausgewiesen sind.

## 2.2 für nichtmetrische Daten

Für nichtmetrische, insbesondere dichotome Daten wird die Zuverlässigkeit über die Konsistenz von Skalen bestimmt, genauer über das Verhältnis der Variabilität der Items  $Y_i$  zur Variabilität des Gesamtscores  $X$ , also  $X = \sum_i Y_i$ .

$$\lambda_3 = \alpha = \frac{N}{N-1} \left( 1 - \frac{\sum_i \text{Var}(Y_i)}{\text{Var}(X)} \right).$$

Alternativ kann dieses Maß der internen Konsistenz auch über die durchschnittliche Korrelation  $\bar{r}$  Items bestimmt werden:

$$\lambda_3 = \alpha = \frac{N \cdot \bar{r}}{1 + (N-1) \cdot \bar{r}}.$$

Entscheidend nun, dass bei Kenntnis einer Menge  $Z$  von Kovariablen aus den Werten der Kovarianzmatrix heraus auch dieser Koeffizient in zwei Anteile aufgespalten werden kann:

$$\alpha = \alpha^{(\sim Z)} + \alpha^{(Z)}$$

und damit, vergleichbar zum Reliabilitätskoeffizienten, Aussagen über den Einfluss der Kovariablen und damit zur Populationsabhängigkeit der Zuverlässigkeitsaussage möglich werden.

## 3. Resampling als weiterer alternativer Zugang (Pauly und Umlauf, 2018)

Resampling allgemein bedeutet eine Prozedur, bei der aus einer vorliegenden Stichprobe unabhängige Stichproben generiert werden. Dies dient allgemein der Simulation der Variabilität der Daten und hat damit direkten Bezug zu den Zuverlässigkeitsmaßen, die durch die Varianzanteile in den Daten bestimmt sind. Durch diese Vorgehensweise können die Eigenschaften der Verteilungsfunktion, speziell von Schätzern und Testgrößen, untersucht werden. Als Zugänge zum Resampling werden verschiedene Prozeduren vorgeschlagen, u.a. die Permutation und das Bootstrapping-Verfahren (neben der Jackknife-Methode und der Kreuzvalidierung).

Wir wollen die statistischen Aussageformen aus solchen Resamplingprozeduren zunächst am Beispiel von zwei Stichproben darstellen. Die dabei abgeleiteten Testprozeduren lassen sich dann direkt auf das allgemeine Mehrstichprobenproblem übertragen.

### 3.1 Die Grundidee am Beispiel von zwei Stichproben

Ausgangspunkt sind im einfachsten Fall zwei unabhängige Stichproben  $X_1, \dots, X_{n1}$  und  $X_{n1+1}, \dots, X_N$ . Dabei bezeichnet der Gesamtscore  $X$  immer den Summenwert der zugrundeliegenden Itemwerte  $Y_i$ , also  $X = \sum_i Y_i$ . Unter diesen Voraussetzungen sollen aus jeder Stichprobe die  $\alpha$ -Koeffizienten bestimmt (abgeschätzt) werden, also die Parameter  $\alpha_1$  und  $\alpha_2$ . Ihre Beurteilung erfolgt über die Nullhypothese

$H_0: \alpha_1 = \alpha_2$  bei ein- oder zweiseitiger Fragestellung.

Diese Nullhypothese entspricht der Annahme keiner Stichprobenunterschiede und spricht damit für eine Populationsunabhängigkeit der  $\alpha$ -Koeffizienten.

Die statistische Beurteilung kann über die Testgröße

$$M_n = \sqrt{\frac{n_1 \cdot n_2}{N}} \cdot (\hat{\alpha}_1 - \hat{\alpha}_2)$$

erfolgen, die asymptotisch standard-normalverteilt ist (Maydeu-Olivares et al. 2007).

Mit dieser Teststatistik wird daher die Aussage möglich, ob die Variation der Stichproben einen Einfluss auf die Konsistenzaussage hat, beide Stichproben also aus der gleichen Grundgesamtheit entstammen oder nicht. Dies entspricht der Ablehnung der Nullhypothese.

Die Prüfung kann auch über die studentisierte Teststatistik

$$T_n = \frac{M_n}{\hat{\sigma}}$$

erfolgen.

### 3.2 Der allgemeinere Permutationszugang

Die Permutation begründet ein mögliches Vorgehen zum Generieren von unabhängigen Stichproben aus einer Grundstichprobe (Resampling). Entscheidend für die Permutation ist das zufällige Ziehen von Stichprobenelementen ohne Zurücklegen. Jede so gewonnene Stichprobe ist dann eine Permutation der Grundstichprobe.

Als Verallgemeinerung des allgemeinen Zweistichprobenzugangs können dabei mehrere unabhängige Stichproben auftreten. Es entsteht also ein Mehrstichprobenproblem, das durch Verallgemeinerung des obigen Zugangs gelöst wird. Die verallgemeinerte Grundfrage ist, ob alle diese Stichproben (Permutationen) der gleichen Grundgesamtheit entstammen, wobei die Nichtablehnung dieser Nullhypothese wieder für eine Populationsunabhängigkeit spricht.

Formalisiert bedeutet dies:

$X$  bezeichne alle  $N$  Datenvektoren  $X = (X_1, \dots, X_N)$ . Durch die zufällige Stichprobenziehung ohne Zurücklegen entsteht dann eine Permutation  $\pi$  dieser Datenvektoren  $(X_{\pi(1)}, \dots, X_{\pi(N)}) = \pi(X)$ . Unter der Nullhypothese keiner Verteilungsunterschiede sind diese Permutationen genauso verteilt wie die Permutationen der Zahlen von 1 bis  $N$ . Dies entspricht der Permutationsverteilung und gibt die Möglichkeit, unter diesen Bedingungen die Wahrscheinlichkeiten des Auftretens einer jeden Permutation zu bestimmen und damit die Zufälligkeit über einer vorgegebenen Irrtumswahrscheinlichkeit abzuschätzen. Als Testgröße dient in Verallgemeinerung des obigen Zugangs hier:

$$T_n(\pi(X)) = \frac{M_n \pi(X)}{\hat{\sigma}(\pi(X))}$$

### 3.3 Das parametrische Bootstrapping als alternativer Zugang

Das Resampling nach der Bootstrap-Prozedur basiert auf dem Ziehen von Zufallsstichproben aus der Grundstichprobe mit Zurücklegen. Macht man keine Verteilungsannahmen über die unbekannte Verteilung in der Grundstichprobe, dann spricht man von einem nichtparametrischen Bootstrap. Kann man jedoch Annahmen begründen, dann wird diese zusätzliche Information einen parametrischen Zugang zum Bootstrap ermöglichen.

Pauly u.a. gehen nun davon aus, dass es sich um normalverteilte Verteilungsfunktionen mit dem Erwartungswert 0 und den zugehörigen Kovarianzmatrizen  $\Sigma_i$  handelt. Daraus wird eine vergleichbare Teststatistik begründet und wie die obige studentisierte Testgröße beurteilt.

### 4. Fazit

a) Mit der Unterscheidung von Score-Reliability und Reliability Generalization (RG) wird eine Präzisierung der Zuverlässigkeitsaussage begründet. Im Rückblick auf 12 Jahre RG-Studien betonen Vacha-Haase & Thompson (2011) „ Score reliability characterizes the degree to which scores measure “something” as opposed to “nothing” (e.g. are completely random)“. Letzteres ist nach wie vor die essentielle Grundaussage für eine substantielle Betrachtung und Interpretation des Scores.

b) Für die Beurteilung der Stabilität der Reliabilität (Reliability Generalization) und damit der Beobachtungsmethode sind (unabhängig vom Datenniveau) drei Zugangsformen ausgewiesen:

- die Metaanalyse als klassischer Zugang unter der Voraussetzung hinreichend vieler Studien,
- die Kovarianzanalyse als Zugang, bei dem der Einfluss von bekannten Kovariablen auf die Stabilität der Reliabilität geprüft wird und
- die Methodik des Resampling, bei der unabhängige Stichproben aus einer Basisstichprobe gezogen werden und daraus die Verteilungsparameter abgeschätzt und beurteilt werden.

Ziel aller vorgestellten Verfahren ist die Prüfung, ob die Variabilität der Daten und Kennwerte konsistent und damit populationsunabhängig ist oder nicht. Der wesentliche Unterschied der Zugänge liegt in den oben skizzierten Voraussetzungen.

c) Als entscheidende Konsequenz ergeben sich unterschiedliche Bedeutungen der Zuverlässigkeitsaussage:

- Score-Reliabilität kennzeichnet die Zuverlässigkeit des Beobachtungsergebnisses (Messwertes) und kann traditionell über die Split-half- oder Retest-Methode geschätzt werden.
- Reliability generalization kennzeichnet die Zuverlässigkeit der Beobachtungsmethode (Messinstrument). Urbina (2004) verdeutlicht diesen Bezug auf das Messinstrument „If a test is described as reliable, the implication is, that its reliability has been established permanently, in all respects for all uses, and all users“ (S. 120).



Die Variabilität der Reliabilitätskoeffizienten, genauer ihre Stabilität, begründet deren Stabilitätsaussage.

d) Als prototypisches Beispiel für die Bedeutung der verallgemeinerten Reliabilität (RG) kann die Metaanalyse von Viswesvaran & Ones (2000) gelten, die das Persönlichkeitsinventar der Big-Five-Faktoren untersuchten. Aus den Befunden von 848 Stabilitätskoeffizienten und 1359 Reliabilitätsaussagen schlussfolgern sie „both types of reliabilities were found to be reasonable high, at least on the average, for all dimensions of personality examined here“. Also sind die Scores zuverlässige Schätzer der Eigenschaftsdimensionen und dies konsistent gegenüber variierender Populationen.

## 5. Literaturverzeichnis

Bentler, P.M. (2016). Covariate-free and covariate-dependent reliability. *Psychometrika*, 81(4), 907-920.

Efron, B. (1985). *The jackknife, the bootstrap and other resampling plans*. Chapman & Hall: London.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, London: Oliver and Boyd.

Good, P. (1995). *Permutation test: a practical guide to resampling methods for testing hypotheses*. New York: Springer.

Good, P. (2000). *Permutation tests*. Heidelberg: Springer.

Kuder, G.F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151-160.

Maydeu-Olivares, A., Coffman, D. L. & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods* 12(2), 157.

Pauly, M. & Umlauf, M. (2018). Resampling-based inference methods for comparing two coefficients alpha. *Psychometrika*, 83(1), 203-222.

Raykov, T. Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Raykov, T. Marcoulides, G. A. (2013). Meta-analysis of scale reliability using latent variable modeling. *Structural Equation Modeling*, 20, 338-353.

Schmidt, F. L. & Hunter, J. K. (1977). Development of a general solution to the problem of generalization. *Journal of Applied Psychology*, 62, 529-540.

- Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurements, 54*, 837-847.
- Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ: John Wiley.
- Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20.
- Vacha-Haase, T. & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*(3), 159-168.
- Viswesvaran, C. & Ones, D. (2000). Measurement error in "Big Five Factors" personality assessment: reliability generalization across studies and measures. *Educational and Psychological Measurement, 60*(2), 224-235.
- Wilkinson, L. & APA Task Force on Statistical Inference. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

## **Kann die Blickrichtung die emotionale Qualität der Erinnerung von belastenden oder traumatischen Ereignissen verändern?**

Majken Bieniok, Alexander Reich & Peter Uwe Hesse

### **Zusammenfassung**

Bis dato stellt sich die grundlegende Frage der Traumatherapie, wie mit der Konfrontation von emotional herausfordernden und belastenden Erinnerungen (Exposition) möglichst ressourcenorientiert und nicht retraumatisierend (Überaktivierung) umgegangen werden kann, sodass die Orientierung in der Gegenwart bestehen bleibt, eine kognitive (Neu-) Verarbeitung möglich wird, und negative Erinnerungen ein explizites Narrativ erhalten können. Die Anwendung von Augenbewegungen und die Traumverarbeitung auf spezifischen Blickrichtungen scheinen einen vielversprechenden Ansatz zur Regulation von emotionaler Erregung während der Exposition darzustellen.

In diesem Kontext führten obige Autoren eine explorative Studie (N=29) zu belastenden Erinnerungen durch. Per Fragebogen wurde unter anderem die Intensität und Qualität der begleitenden Emotionen unter zwei Blickrichtungsbedingungen während einer 5-minütigen Exposition (Erinnerung an ein belastendes Ereignis sowie Erfassung der Blickrichtung der höchsten emotionalen Belastung, die im Körper spürbar ist, und Erfassung der Blickrichtung der geringsten spürbaren Belastung bei der Erinnerung an das identische Ereignis) erfasst. Die Ergebnisse weisen auf signifikant unterschiedliche Erregungsniveaus und emotionale Qualitäten in Abhängigkeit der Blickrichtung während der Exposition hin. Es wurden keine bedeutsamen Gruppenunterschiede bezüglich des Belastungsgrades der Versuchspersonen (vorhandene Hinweise auf Posttraumatische bzw. Dissoziative Störungsbilder - PC-PTSD-5 / Prins et al, 2015 und SDQ 5 / Nijenhuis, 2010) auf die emotionale Reaktion gefunden. Die Ergebnisse werden diskutiert und in den gegenwärtigen Stand der Erkenntnisse eingeordnet.

### **Schlüsselwörter**

Trauma, PTBS, Emotionen, Augenbewegung, Blickrichtung

### **1. Einleitung**

Die vorliegende explorative Pilot-Studie erkundet die Bedeutung von Blickrichtungen (in verschiedenen Methodenbeschreibungen auch als Blick- oder Augenpositionen, visuelle Spots, Brainspots oder visuelle Foci benannt) für die Veränderung der emotionalen Qualität von belastenden oder traumatischen Ereignissen, in der Hoffnung, dass sich daraus in der Folge Hinweise für die effektive Behandlung von belasteten Patient\*innen und/oder Traumapatient\*innen mittels verschiedener Arbeitsmodi generieren lassen, welche das Visusfeld des Patienten nutzen. So werden im Bereich der Traumatherapie beispielsweise die Wirkweisen und wirksamen Komponenten der Augenbewegung bzw. Blickrichtungen im Visusfeld von Techniken wie EMDR mit schnellen sakkadischen Augenbewegungen (Sack et al, 2016), EMI mit langsamen oder auch smooth-pursuit Augenbewegungen (Beaulieu, 2005) und Brainspotting mit einem stabilen Fokus (Corrigan & Grand, 2013) diskutiert. Die praktische Erfahrung zeigt, dass bei der Anwendung von bewegten bilateralen Augenarbeitstechniken die Augenbewegungen von Klienten\*innen

stocken oder gebremst werden und dadurch relevante Blickrichtungen für spezifische Themen und Erinnerungen für Therapeuten\*innen beobachtbar werden. Scheinbar spielen eben diese fokussierten Spots eine wichtige Rolle in dem mentalen Prozess der Traumaverarbeitung. Die mit solchen Spots assoziierten und für Klienten\*innen erlebbaren emotionalen Qualitäten gilt es näher zu erforschen.

Neben einfachen, belastenden Ereignissen (z.B.: Ärger und Streit mit Familienangehörigen oder Arbeitskolleg\*innen), kann ein mit dem Visusfeld arbeitender Therapieansatz auch bei traumatisierter Klientel eingesetzt werden. Wir greifen dabei zur Definition der Posttraumatischen Belastungsstörung (PTBS) mit dem folgenden Wortlaut für eine Traumasituation, wie sie demnächst mit Publikation des ICD-11 genutzt werden wird: „Ein extrem bedrohliches oder entsetzliches Ereignis oder eine Reihe von Ereignissen“ (Gysi, 2018, S.1). Im Fall der komplexen PTBS: „Ein extrem bedrohliches oder entsetzliches Ereignis oder eine Reihe von Ereignissen, meistens längerdauernde oder wiederholte Ereignisse, bei denen Flucht schwierig oder unmöglich war (z.B. Folter, Sklaverei, Genozidversuche, längerdauernde häusliche Gewalt, wiederholter sexueller oder körperlicher Kindesmissbrauch)“ (ebd., S.1). In der Traumatherapie wird zwischen Typ I-Trauma (Monotraumata, einmalig, eher kurzfristige und plötzliche traumatisierende Ereignisse) und Typ II-Trauma (Serielle Traumata, sich wiederholende oder länger andauernde Stressoren, welche zu einer erwartbaren Belastung werden) unterschieden. Dieser Unterteilung und den unterschiedlichen Folgen der beiden Traumtypen trägt das ICD-11 Rechnung. Maercker (2013) unterscheidet diese beiden Grundtypen jeweils weiter anhand der drei Aspekte akzidentiell (in Form eines Unfalls) und interpersonell bzw. man made.

Die möglichen ultimativen Reaktionen auf eine sehr bedrohlich erlebte Situation lassen sich mit den englischsprachigen Begriffen fight (Kampf/Angriff), flight (Flucht), und freeze (Totstellen/Immobilisation) beschreiben und sind mit unterschiedlichen Aktivitäten des Nervensystems assoziiert. Vereinfacht dargestellt, werden bei Kampf und Flucht charakteristischerweise insbesondere über die sympathischen adrenergen Nervenfasern Informationen an die periphere Muskulatur gesendet, um diese motorisch aktive Reaktion hervorrufen zu können (Levine, 2010; Porges, 2001). Erscheint diese Reaktion aussichtslos oder führte bisher nicht zur Rettung, wird der Vagusnerv und das dorsale vagale System aktiviert, welches die inneren Organe im Bauch-, Brust- und Halsbereich versorgt. Es kommt dann zu einer Erstarrung oder sogar Erschlaffung des Körpers und zu einer Immobilisierung (Totstellen). Bei Überleben und Nachlassen der Gefahr kann dann wiederum das sympathische Nervensystem übernehmen und entsprechende rettende Aktivitäten (z.B. Rennen, Schlagen ...) einleiten. Ist die Umgebung dann wieder sicher, kann das myelinisierte ventrale System seine kommunikativen, sozialen, interaktiven und emotionalen Ruhefunktionen übernehmen, welche ansonsten in der Notsituation automatisch gehemmt werden (Levine, 2010; Porges, 2001).

Als Traumafolgestörungen imponieren depressive- und/oder Angststörungen, Zwangsstörungen, dissoziative Störungen, PTBS, Essstörungen, Suchterkrankungen, psychosomatische Störungen, Persönlichkeitsstörungen u.a. ICD Diagnosen.

Erlebte Traumatisierungen können im Organismus zu einer Vielfalt von Symptomen führen. Beobachtbar werden beispielsweise akute Belastungsreaktionen ebenso wie

eine verzögerte, protrahierte Symptombildung mit intrusivem Erleben (z.B. visuelle oder Körperflashbacks), Vermeidung von Situationen/Emotionen oder Dissoziation, Hypervigilanz (wie erhöhter Schreck- und Reizbarkeit, Schlafstörungen u.a.). Herman (2010) beschreibt in diesem Kontext Störungen in der Regulation von Affekten und Impulsen; Störungen der Wahrnehmung oder des Bewusstseins; gestörte Selbstwahrnehmung; gestörte Wahrnehmung des Täters/der Täter, Beziehungsprobleme und Veränderungen des Wertesystems.

Zu den psychophysiologischen Effekten, die eine erlebte Traumatisierung bewirken, kann zählen: Reizüberempfindlichkeit, Hyperarousal/Veränderungen in energetischen Hirnmustern, fehlende Habituation an Schreckerfahrungen, hinzu kommen neurohormonelle Effekte, neuroanatomische Veränderungen sowie immunologische Effekte (Levine, 2010; van der Kolk, 2014).

## **2. Traumaarbeit im Visusfeld**

Im Folgenden beziehen wir uns exemplarisch auf verschiedene Verfahren, wie Eye Movement Desensitization and Reprocessing (EMDR), Eye Movement Integration (EMI) sowie Brainspotting, welche die Arbeit im Visusfeld nutzen um traumatische Erlebnisse zu verarbeiten. Uns ist es nicht wichtig, diese Verfahren umfassend als Methoden der Traumapsychotherapie vorzustellen. Es wird vielmehr auf die unterschiedliche Nutzung von Blickrichtungen bzw. Spots fokussiert. Wichtige weitere Vertreter werden an dieser Stelle auch nur kurz benannt. Es fehlen beispielsweise bestimmte Methoden der Hypnotherapie, des Eyes Closure Eyes Movement (ECEM, nach Hollander & Bender, 2001 und Hollander, 2009), der Observed Experiential Integration (OEI, nach Bradshaw, Cook, & McDonald, 2011) und weitere, die an dieser Stelle nur kurz benannt werden.

Mit der Entwicklung von traumaverarbeitenden Methoden wie der Eye Movement Integration (EMI) sowie des Eye Movement Desensitization and Reprocessing (EMDR) aus dem Neurolinguistischen Programmieren (NLP) heraus, standen erstmals gut anzuwendende Methoden der Traumaarbeit neben der Hypnose/Hypnotherapie zur Verfügung. Interessanterweise war das NLP in den 70er Jahren des letzten Jahrhunderts dabei nicht als therapeutisches Verfahren entwickelt worden, sondern stellte im Kern ein Kommunikationsmodell dar, welches durch Modellieren von bekannten Therapeuten wie Fritz Perls (Gestalttherapie), Virginia Satir (Familientherapie) und Milton Erickson (Hypnotherapie) sowie durch Beobachtung des Kommunikationsstiles Gregory Batesons entstanden war (Hesse, 2017, S. 47). Die Hypnose und Hypnotherapie mit ihren verschiedenen Nutzungen des Augenfokus zur Induktion von Trancezuständen (Blickfixation u.a. Techniken) stand insbesondere vertreten durch Milton Erickson am Beginn der Entwicklung der modernen Visusfeldarbeit (ebd.).

Mit dem Augenbewegungsmodell von Robert Dilts, welches er im Rahmen des NLP entwickelte, entstand ein Grundmodell, welches dazu einlud, zu erforschen, ob und wie sich von außen angeleitete Augenbewegungen therapeutisch nutzen ließen (ebd.). In den von ihm initiierten Arbeitsgruppen forschten Francine Shapiro, welche späterhin das EMDR gründete, sowie Connirae und Steven Andreas, welche das EMI entwickelten. EMDR als Methode verbreitete sich von der Westküste der USA aus über die USA und in der Folge weltweit, das EMI verbreitete sich von der Westküste vor allem zuerst nach Kanada. Erst die kanadische Psychotherapeutin

Danie Beaulieu führte es in das europäische psychotherapeutische Methodenarsenal ein und verbreitete es dann ebenfalls weltweit. Beide Verfahren entwickelten sich zeitgleich (ebd.). Griff Shapiro zur Verarbeitung traumatischen Erlebens auf horizontale, bilaterale, sakkadische Augenbewegungen zurück und fügte ihrem Repertoire später bilaterales Tapping (kinästhetisches Repräsentationssystem) und Snipping (akustisches Repräsentationssystem) hinzu, so begriffen Connirae und Steven Andreas das Visusfeld einer Person als Ort, von dem aus man über bestimmte Blickpositionen und Spots Kontakt zu abgelaufenen positiven oder negativen Erlebnissen aufnehmen kann. Sie entwickelten ein System von 22 Basisbewegungen in Form langsamer, wenig oft wiederholter Augenbewegungen, welches geeignet war, das Visusfeld abzudecken (Beaulieu, 2005).

Durch die Augenbewegungen wird eine Pendelbewegung zwischen negativ-traumatisch besetzten Spots und positiv besetzten Augenpositionen erzeugt, wodurch die negative Qualität nach einigen Wiederholungen nachlässt. Postulierte Shapiro einen Mechanismus einer beschleunigten Reizverarbeitung, welcher durch die schnellen Sakkaden abgedeckt würde, so waren Connirae und Steven Andreas in der Annahme eines generellen Wirkmechanismus zurückhaltender. Neuere Ideen zur Wirkungsweise vermuten bereits in der Pendelbewegung zwischen traumatisch besetztem Spot/Trauma und angenehm besetztem Spot/Ressource einen zentralen Wirkmechanismus zur Auflösung impliziter traumatischer Erinnerungen, ähnlich der Pendelbewegung/Pendulation im kinästhetischen Sinneskanal zwischen Kontraktion und Expansion in Peter Levines Somatic Experiencing (Levine, 2015, S.96).

Beide Arbeitsmodelle lassen sich gut in ihrer Wirkung durch das Grundmodell der Gedächtnisrekonsolidierung (Ecker, Ticic & Hulley, 2016, S. 100) erklären, wonach ein Verfahren, welches Wirkung im Bereich der Aufarbeitung negativer Erlebnisse/Gedächtnisinhalte zeigt, folgende Bearbeitungsabfolge beinhaltet: 1) Beschreibung des Symptoms/Problems in konkreten, erfahrungsbezogenen Details, 2) Gewahrsein des darin enthaltenen implizit-emotional Gelernten, 3) Kontakt zu emotional echtem, widerlegendem Wissen (Ressource), 4) Pendelbewegung zwischen symptom erzeugendem Gelernten und widerlegendem Wissen. Dieses Nebeneinanderstellen und Pendeln zwischen diesen beiden Erfahrungen eröffnet die Rekonsolidierung und Überschreibung des negativ Gelernten durch die Ressource. Diese Schrittfolge ist von einer Verifizierung gefolgt, sodass das Symptom/Problem tatsächlich weggefallen ist (Schritt 5). Ähnlich wie Levine, fokussieren auch Ecker, Ticic und Hulley (2016) auf Pendelbewegungen als zentrale Technik in der Verarbeitung belastender und/oder traumatischer Erinnerungen.

Matthijssen (2018) und Sack et al. (2016) legten Studien vor, die aufzeigten, dass ein fixierter Blickpunkt wie in der Glitch Massaging oder dem Brainspotting letztlich gleich effektiv in der Arbeit mit traumatischen Erinnerungen ist, wie die bilaterale Stimulation des EMDR oder EMI. Aus der Beobachtung des traumabesetzten Spots heraus, die sich im EMDR dadurch zeigen, dass sich Störungen und Unregelmäßigkeiten in der sakkadischen Augenbewegung ergaben, wenn ein traumabesetzter Spot berührt wurde (ein Defokussieren des Blickes, Blinzeln, Verlieren des zu verfolgenden bewegten Fingers, Muskelbewegungen u.a.) entwickelten sich wiederum Verfahren, welche diesen Spot nutzten, der sich so in der Visusarbeit zeigte.

Die Technik Glitch Massaging, eine der drei basalen Techniken des OEI - neben Switching und Sweeping - entwickelt von Audrey Cooks (Bradshaw, 2011), war scheinbar die erste Technik, welche die therapeutische Nutzung des Spots wiederentdeckte. OEI vereinigt Ideen des Focusing ebenso wie des Somatic Experiencing (Levine, 1997) und der Sensorimotor Psychotherapy (Ogden, Pain, & Fisher, 2006), die ähnliche Annahmen und Techniken verwenden. Ein weiteres Verfahren, welches eine größere Verbreitung erfuhr, wurde von David Grand entwickelt und als Brainspotting bekannt (Grand, 2013). Alle diese therapeutischen Methoden befinden sich in stetiger Weiterentwicklung.

Mit der Rückkehr zum Fokus oder Spot war der Kreis der Methodenentwicklungen geschlossen, der mit der Nutzung von Blickfixationen oder Spots durch die Hypnose begann. Die Hypnotherapie postuliert, dass eine Fokussierung von Aufmerksamkeit das Eintreten von Tranceprozessen bewirkt, und so ergeben sich natürlich auch Tranceprozesse und damit veränderte Bewusstseinszustände bei der Fokussierung der Aufmerksamkeit auf einen bestimmten Blickspot, was wiederum zur Verarbeitung von Störungsmustern führt, welche auf diesem Spot in den Kontakt genommen werden.

Grands (2013) Brainspotting macht erlebbar, dass es für jeden im Körper fühlbaren Gedächtnisinhalt auch eine Blickrichtung gibt, bei welcher diese Empfindung besonders intensiv wahrgenommen wird. Er nutzt Zeigestäbe/Pointer, um mit dem Patienten zusammen den für ein Symptom oder Problem relevanten Brainspot zu finden. Die Prozessierung des Symptoms/Problems erfolgt dann durch Halten des Blickes auf diesem Brainspot. Die Blickfokussierung in der entsprechenden Richtung der maximalen Aktivierung der körperlichen Empfindung scheint die Verbindung mit impliziten Gedächtnisinhalten herzustellen und dabei zugleich eine innere Distanzierung zum Erlebten (durch das Schauen auf den Stab) zu ermöglichen. Werden Patient\*innen dazu aufgefordert, über ihr Problem zu berichten, so neigen sie bei Abrufen der Gedächtnisinhalte dazu, den Blick in bestimmten Richtungen zu fixieren (engl. to gaze). Folglich können auch über Gazing (Gazespotting) relevante Spots gefunden werden. Micic, Ehrlichman und Chen (2010), haben den Zusammenhang zwischen der Blickrichtung und dem Abrufen von Inhalten aus dem Langzeitgedächtnis bereits aufgezeigt.

Allen beschriebenen Ansätzen ist gemein, dass ergänzend zu den gewählten Blickrichtungen, auch die gleichzeitig empfundenen Emotionen und körperlichen Sensationen bzw. deren Veränderung im Laufe der Arbeit im Visusfeld einfließen.

Da bisher wenig Forschung in diesem Feld stattfand, interessierte uns in einem ersten Schritt, ob die Blickfixierung auf einen solchen mit emotionaler Belastung besetzten Spot/Blickposition tatsächlich in der Lage ist, die emotionale Qualität des belastenden Ereignisses für den Klienten zu verändern.

### **3. Methodische Vorgehensweise**

In der vorliegenden Pilot-Studie wurde ein Fragebogen, bestehend aus 2 Teilen verwendet. Der erste Teil (Prä-Testung) wurde vor einer Konfrontation mit emotional herausfordernden und belastenden Erinnerungen (Exposition) ausgehändigt. Der zweite Teil wurde nach Beendigung ausgehändigt (Posttestung). Der gesamte Ablauf und das Fragebogeninstrument werden nachfolgend dargestellt.

Nach einer kurzen Aufklärung über den Ablauf und die Ziele der Pilotstudie „Emotionalität in Abhängigkeit der Blickrichtung“, der Freiwilligkeit und des Haftungsausschlusses, sowie möglicher emotionaler Belastung während und nach Teilnahme, wurden per Papier-und-Bleistift-Fragebogen (1. Teil bzw. Prä-Testung) anonymisiert, soziodemographische Daten (Alter, Geschlecht, Bildungsstand ...) sowie weitere Daten, unter anderem zum aktuellen und generellen täglichen Stresslevel, zum allgemeinen Gesundheitszustand und zum aktuellen Glücksempfinden mit unterschiedlichen Frageformaten erfasst. Um unterschiedliche Stresslevel zu erfassen, wurde ein Item „Bitte beschreiben Sie Ihr Stresslevel: 1) allgemeines tägliches Stresslevel 2) aktuelles Stresslevel in diesem Moment“ mit einer 5-stufigen Likert-Skala (hoch:1 bis niedrig:5) verwendet. Das aktuelle Glücksempfinden wurde mit einem Happy-Single-Item „Wie glücklich und zufrieden fühlen sie sich jetzt gerade?“ (11-stufige Likertskala / nicht glücklich:0 – sehr glücklich:10) in Anlehnung an Marselle et al. (2016) erfasst. Um die individuelle traumatische Belastung der Probanden besser einschätzen zu können, wurden zudem etablierte Screening-Items bezüglich des Vorliegens einer PTBS (PC-PTSD-5 / Beispielitem: „Habe Sie in Ihrem Leben jemals eine Erfahrung gemacht, die so beängstigend, schrecklich oder erschütternd war, dass Sie in der Zeit danach Alpträume davon hatten oder daran gedacht haben, wenn Sie es nicht wollten?“ mit binärem Antwortformat: „Ja“ vs. „Nein“ / Prins et al., 2015), sowie bezüglich des Vorliegens von Dissoziationsmechanismen (SDQ-5 / Beispielitem: „Wie stark oder häufig erleben Sie das folgende Symptom im letzten Jahr? ... ist mein Körper oder ein Teil davon schmerzempfindlich gewesen.“ mit einer 5-stufigen Likert-Skala (überhaupt nicht:1 bis sehr stark:5) / Nijenhuis, 2010) verwendet. Des Weiteren wurden zwei offene Fragen, in Anlehnung an gängige Diagnoseinstrumente, wie z.B. dem DSM-5 (PC-PTSD-5 oder CLINICIAN-ADMINISTERED PTSD SCALE FOR DSM-5), zu traumatischen bzw. belastenden Lebensereignissen gestellt („Haben Sie in Ihrem Leben schon einmal eine Situation erlebt, die Sie als traumatisierend bezeichnen würden z.B. Naturkatastrophen, Verkehrsunfälle, plötzlicher Tod eines geliebten Menschen, Raubüberfälle, körperliche Angriffe, schwere körperliche Bestrafung/Gewalt in der Kindheit, sexuelle Missbrauchserfahrungen ... oder waren Zeuge(in) eines solchen Ereignisses?“ bzw. „Haben Sie in Ihrem Leben schon einmal eine Situation erlebt, die Sie als (sehr) belastend, aber nicht traumatisierend, bezeichnen würden z.B.: Zielscheibe von verbalen Wutausbrüchen, Mobbing, oder Beschämung oder Erleben von Hilflosigkeit, Angst, Ausgeliefertsein ...)“ mit den Antwortmöglichkeiten: Nein vs. Ja – bitte benennen Sie das Ereignis in kurzen Worten: ...).

Der emotionale Zustand wurde bezugnehmend auf zwei Zeiträumen/-punkte a) im Allgemeinen in den letzten Tagen sowie b) jetzt gerade, mit der- Brief Measures of Positive & Negative Affect Scale (Watson et al., 1988), mit den Subskalen positive Emotionen (10 Items zu den Affekten: aktiv, interessiert, freudig erregt, stark, angeregt, stolz, begeistert, wach, entschlossen, aufmerksam; 5-stufige Likert-Skala: sehr wenig oder gar nicht:1 – äußerst/sehr stark:5) und negative Emotionen (10 Items zu den Affekten: bekümmert, verärgert, schuldig, feindselig, gereizt, beschämt, nervös, durcheinander, ängstlich; 5-stufige Likert-Skala: sehr wenig oder gar nicht:1 – äußerst/sehr stark:5) erfasst.

Es folgte die Präsentation eines meditativen Films für die Dauer von 8 Minuten um eine emotional wenig anspruchsvolle und kontrollierte Ausgangssituation für alle Probanden zu gestalten. Danach wurden die Probanden aufgefordert, sich an ein



belastendes Ereignis zu erinnern, welches eine mittlere bis hohe emotionale Aktivierung erzeugt (Exposition, SUD von 5-8), alle Körpersensationen wahrzunehmen und – falls vorhanden - eine Blickrichtung zu finden, bei der diese Sensationen am intensivsten zu spüren sind (höchste Aktivierungsblickrichtung). Der dem Probanden gegenüber sitzende Versuchsleiter unterstützte die Suche der Blickrichtung mit einem Teleskop-Zeige-Stab, indem alle Quadranten des Gesichtsfeldes langsam und systematisch abgesucht wurden. Um die gefundene Blickrichtung mit der stärksten Aktivierung aufrecht zu erhalten und die Körperwahrnehmung, Emotionen und Kognitionen intensiver zu können, wurde die Blickrichtung mit der Stabspitze nach Angaben des Probanden „im dreidimensionalen Raum fixiert“, indem der Versuchsleiter den Teleskop-Zeige-Stab für 5 Minuten entsprechend festhielt. Danach wurde, immer noch unter Erinnerung des ausgewählten belastenden Ereignisses und unter Anwendung der identischen Prozedur eine Blickrichtung gesucht, bei der die Körpersensationen am geringsten ausfallen (geringste Aktivierungsblickrichtung). Wurde eine solche Blickrichtung gefunden, so wurde auch diese für 5 Minuten unter zur Hilfenahme der Stabspitze aufrechterhalten und alle Körperwahrnehmung und Emotionen wahrgenommen.

Es folgte die Erhebung der Post-Daten mit dem 2. Teil des Papier-und-Bleistift-Fragebogens. Die Probanden wurden zuerst aufgefordert das ausgewählte traumatische oder belastende Ereignis in kurzen Worten zu benennen. Es folgten offene verbale Fragen (bzw. eine Anweisung zum Einzeichnen) bezüglich der höchsten Aktivierungsblickrichtung, d.h. bezüglich der genauen Richtung, in die geschaut wurde, der Körpergefühle und Lokalisation dieser Empfindungen. Zudem wurde die subjektive Belastung auf einer 10-stufigen Likert-Skala erfasst (Subjective Units of Distress Scale - SUD: sehr schwach:1 bis sehr stark:10 / in Anlehnung an Wolpe, 1969). Des Weiteren wurde erneut die Brief Measures of Positive & Negative Affect Scale (Watson et al., 1988) eingesetzt, um die emotionale Qualität zu bestimmen. Identische Items wurden auch für die geringste Aktivierungsblickrichtung erhoben. Der Fragebogen schloss mit den schon beschriebenen Items aus der Prä-Testung zum aktuellen Stresslevel und zum aktuellen Glücks-/Zufriedenheitsempfinden sowie einer offenen Frage zu den gemachten Erfahrungen ab (Welche Erfahrung haben Sie gerade gemacht? Möchten Sie uns noch etwas mitteilen?).

### **3.1 Stichprobenbeschreibung**

Potentielle Probanden wurden über den bekannten Therapeutenkreis und über den Pool von eigenen Klienten\*innen kontaktiert. Die Teilnahme war freiwillig und wurde nicht entlohnt. Des Weiteren konnte die Teilnahme jeder Zeit abgebrochen und widerrufen werden. Diese Möglichkeit wurde jedoch von keiner Person genutzt. Die anfallende und selbstselektierte Stichprobe bestand aus  $N=29$  deutschsprachigen Personen (weiblich  $n=20$ ; männlich  $n=9$ , Altersdurchschnitt  $M=44$  Jahre;  $Std=9,87$ ) davon waren  $n=19$  ledig,  $n=2$  geschieden und  $n=8$  verheiratet. Allerdings lebten  $n=11$  alleine und  $n=18$  zusammen mit Partner\*in und/oder den Kindern. Das allgemeine Bildungsniveau der Stichprobe ist als hoch einzuschätzen ( $n=1$  Person mit Realschulabschluss,  $n=12$  Personen mit Abitur und  $n=16$  Personen mit Fachhochschul- bzw. Hochschulabschluss). Die Stichprobe enthielt  $n=5$  Personen, die eine ihnen bekannte psychiatrische Diagnose in eigenen Worten folgendermaßen angaben: Generalisierte Angststörung, Depression, Agoraphobie mit Depression, Interaktionsstörung und Angstattacken. Das erlebte allgemeine tägliche Stresslevel

wurde als etwas niedriger  $M=2,91$  ( $Std=0,88$ ) angegeben als das aktuelle Stresslevel in diesem Moment vor der Exposition (Prä-Stress)  $M=3,82$  ( $Std=1,05$ ) auf einer 5-stufigen Likert-Skala (hoch: 1; niedrig: 5). Bezüglich des allgemeinen Gesundheitszustandes „Bitte beschreiben Sie Ihren allgemeinen Gesundheitszustand“ auf einer 5-stufigen Likert-Skala (ausgezeichnet:1 bis sehr schlecht:5) gaben die Personen im Durchschnitt einen guten Wert von  $M=2,52$  ( $Std=0,78$ ) an. Die Probanden fühlten sich im Schnitt glücklich und zufrieden (Prä-Glück:  $M=6,52$ ;  $Std=1,57$ ).

Die Gesamtstichprobe lässt sich in Abhängigkeit der berichteten traumatischen Symptome in zwei Unterstichproben einteilen: 1) wenig belastete Probanden und 2) höher belastete Probanden. Die verwendenden Cut-Off-Werte, um in die höherbelastete Gruppe an Probanden mit Verdacht auf PTSD und/oder Dissoziation eingeordnet zu werden, lagen für das Screening bezüglich einer PTBS bei  $cut-off=4$  (PC-PTSD-5 / Prins et al, 2015) und bezüglich des Vorhandenseins von Dissoziation bei  $cut-off \geq 8$  (SDQ-5 / Nijenhuis, 2010). Resultierend lassen sich  $n=13$  Personen der höherbelasteten Gruppe und  $n=16$  in die weniger belastete Gruppe einteilen. Alle Personen, die eine psychiatrische Diagnose angaben, sind anhand der verwendeten Cut-Off-Werte automatisch der höherbelasteten Gruppe zugeordnet worden.

Sortiert man die erlebten und beschriebenen Trauma-Typen nach dem von Maercker (2013) vorgeschlagenen Schema (Typ I: einmalig, kurzfristig; Typ II: mehrfach, langfristig; medizinisch bedingt; jeweils mit der Unterscheidung in akzidentiell vs. interpersonell / man made) ein, so ergibt sich folgendes Bild (siehe Tabelle 1): Die Gruppe, die in dem Screening höher punktet und dementsprechend belasteter ist, beschreibt augenscheinlich auch mehr traumatische Ereignisse, insbesondere vom interpersonellen Typ II. Bei der Beschreibung von interpersonellen Belastungen unterscheiden sich die Gruppen augenscheinlich etwas weniger.

Neben dem unterschiedlichen psychischen Belastungsgrad unterscheiden sich die beiden Gruppen an Probanden in ihren sozio-demographischen Variablen (Alter, Bildungsstand ...), den allgemeinen sowie dem aktuellen (Prä-Stress und Post-Stress) Stresslevel, dem allgemeinen Gesundheitszustand und dem aktuellen Glücksempfinden (Prä-Glück und Post-Glück) nicht signifikant voneinander (Mann und Whitney U-Test für unabhängige Stichproben).

**Tab. 1:** Übersicht über die berichteten Lebenserfahrungen der Probanden (in Anlehnung an Maercker, 2013), die entweder unter der Frage nach traumatischen Erfahrungen (Trauma) oder nach Belastungssituationen (Belastung) geantwortet wurden (Mehrfachnennung waren möglich; unterschiedliche Ereignisse, die von einer Person berichtet wurden, wurden auch getrennt codiert -einfach- und nicht als mehrfach/langfristiges Ereignis).

Trauma Typ nach Maercker	weniger belastete Gruppe (n=16): Häufigkeiten		höher belastete Gruppe (n=13): Häufigkeiten		gesamte Stichprobe (n=29): Häufigkeiten	
	Trauma	Belastung	Trauma	Belastung	Trauma	Belastung
<b>Typ I: einmalig, kurzfristig</b>						
akzidentiell	6	1	9	4	15	5
interpersonell	6	5	4	4	10	8
<b>Typ II: mehrfach, langfristig</b>						
akzidentiell	-	-	-	-	-	-
interpersonell	3	16	16	15	19	31
<b>medizinisch bedingt</b>						
akzidentiell	-	12	2	2	2	3
Interpersonell	1	-	3	-	4	-
<b>Summe</b>	16	23	34	24	50	47
<b>keine Antwort bzw. keine entsprechende Erfahrung</b>	7	-	-	-	7	-

## 4. Ergebnisse

Nachfolgend werden zuerst Ergebnisse der deskriptiven und qualitativen Datenanalyse zu den relevanten Blickrichtungen, Körperempfindungen und Kognitionen, d.h. den gewählten belastenden oder traumatischen Erinnerung dargestellt. Anschließend werden die Ergebnisse qualitativer Datenanalysen und Gruppenvergleiche erörtert.

### 4.1 Deskriptive und qualitative Ergebnisse

Die von den Probanden ausgewählten Erinnerungen, welche eine mittlere bis hohe emotionale Aktivierung erzeugten, bezogen sich auf unterschiedliche interpersonelle Situationen im Erwachsenen- bzw. Kindes-/Jugendalter und den privaten Kontext der Familie (Eltern, Geschwister, Partner und eigene Kinder) sowie auf einen nicht-privaten Kontext wie z.B. den Arbeitsplatz. Die von  $n=8$  Personen gewählten belastenden oder traumatisierenden interpersonellen Erinnerungen des Kindes- und

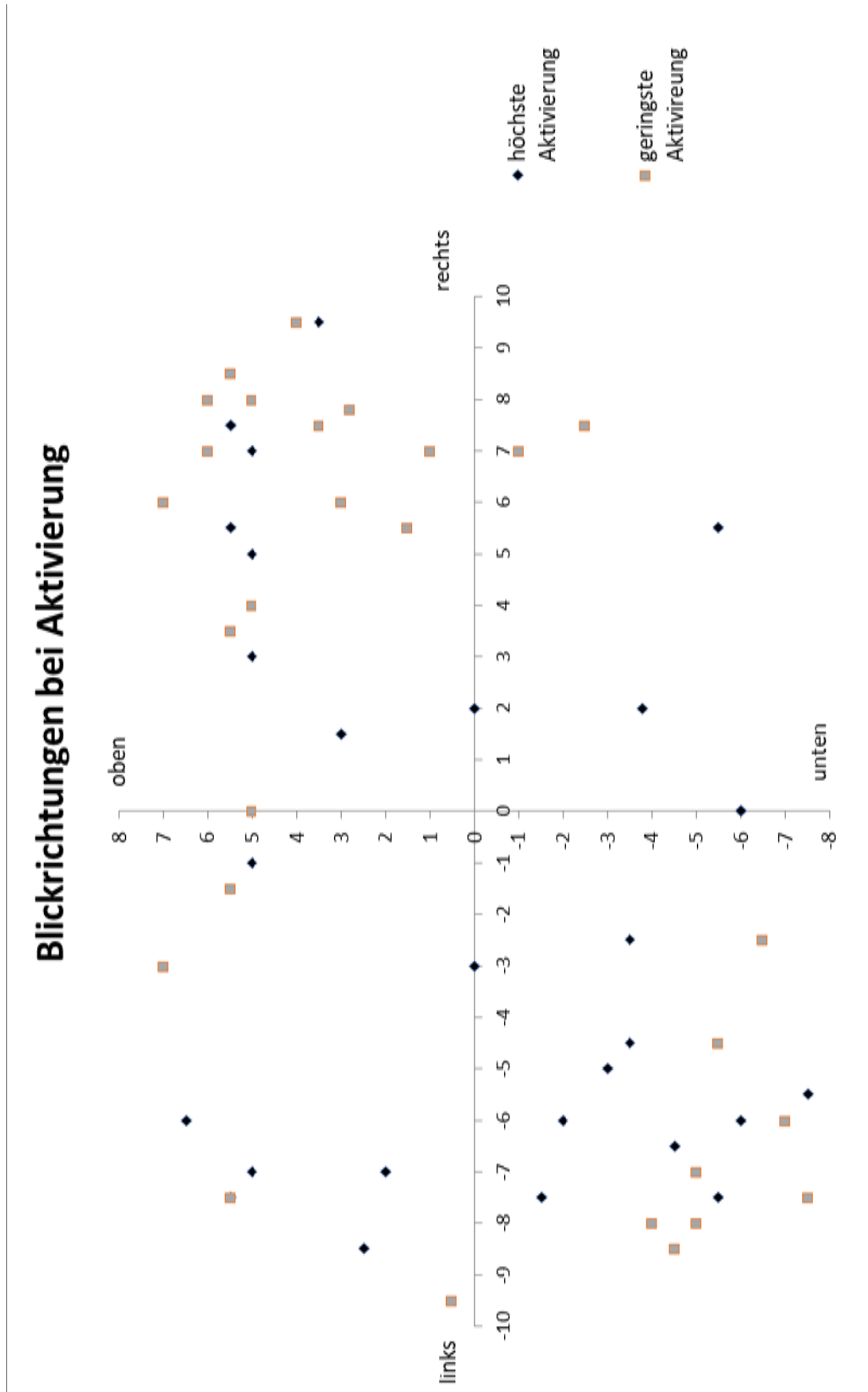
Jugendalters bezogen sich ausschließlich auf den familiären Kontext und insbesondere auf die Interaktion mit Bindungspersonen (z.B. Verlassenheitserlebnisse, verbale Abwertung oder Gewaltanwendung durch die Elternteile). Ein weiterer Themenbereich bezog sich auch auf die interpersonelle Interaktion im privaten Umfeld im Erwachsenenalter (z.B.: Streit und Enttäuschungen oder Krankheit und Verlusterlebnisse von nahstehenden Menschen wie Eltern, Partner und Kinder) und wurde von  $n=10$  Personen ausgewählt. Ausgewählte interpersonelle Erinnerungen in Bezug auf das Erwachsenenalter decken zudem auch einen nicht-privaten Bereich ab (z.B.: Stress und Belastung am Arbeitsplatz oder in halb- und öffentlichen Räumen, mit Kollegen, Vorgesetzten oder Fremden sowie verbunden mit Gefühlen von Abwertung, Beschämung, Ängsten, Kontrollverlust und Unterlegenheit) und wurden von  $n=11$  Personen gewählt.

Allen Probanden war es möglich zu diesen belastenden oder auch traumatisierenden Ereignissen während ihrer Erinnerung (Exposition) eine entsprechende spezifische Blickrichtung zu finden, bei der sie eine höchste Aktivierung verspüren und ebenso eine spezifische Blickrichtung, bei der die emotionale Aktivierung am geringsten ausfällt.

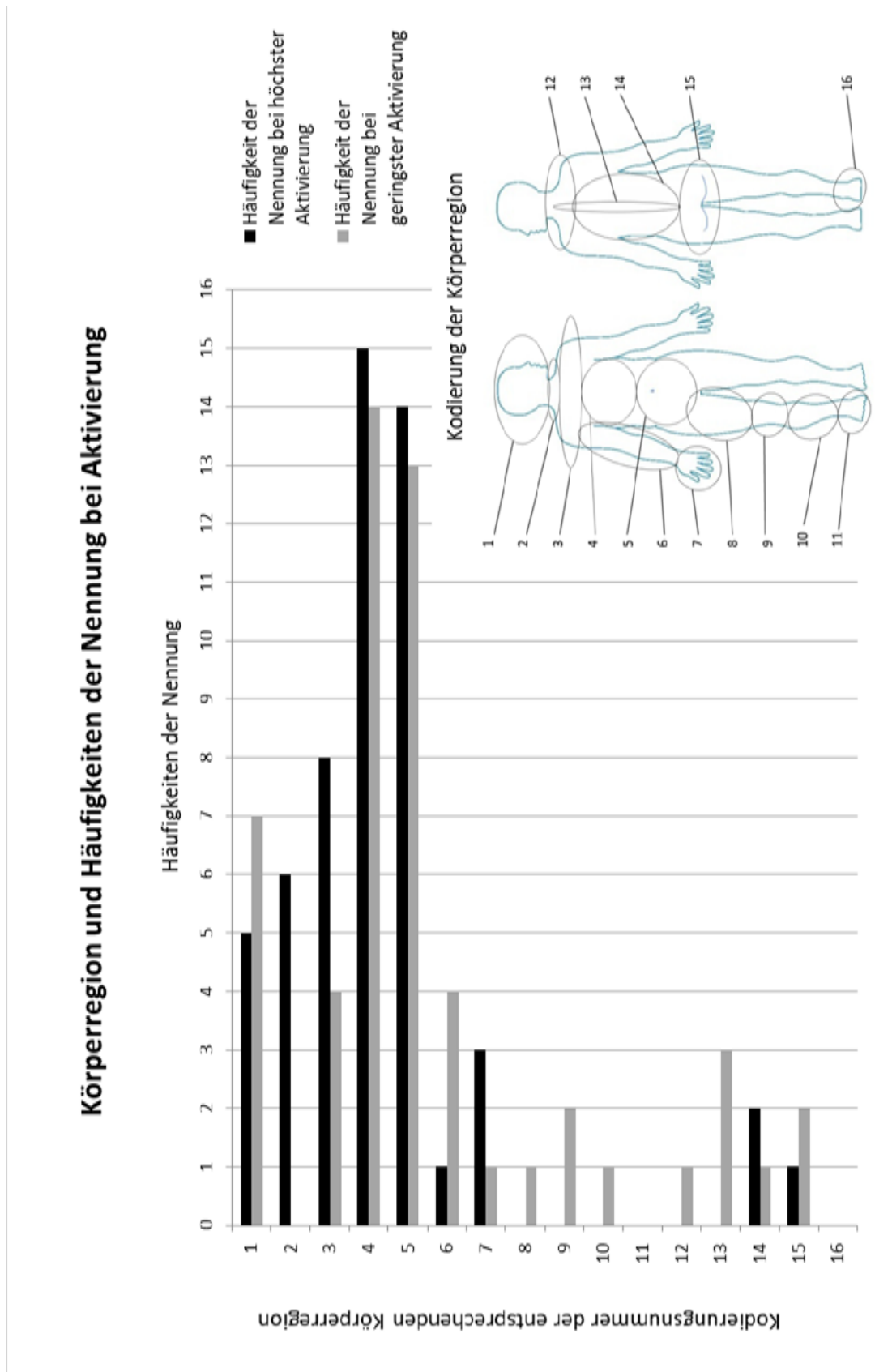
Die gefundenen Blickrichtungen verteilen sich für beide Bedingungen (höchste und geringste Aktivierung) über das gesamte Visusfeld (Abb.1), wobei häufig für die unterschiedliche emotionale Intensität der Aktivierung die jeweils diagonal gegenüberliegende Blickrichtung ( $n=19$ ) gewählt wurde oder zumindest eine deutliche Änderung der Ausrichtung auf einer Achse (oben/unten mit  $n=6$  oder links/rechts mit  $n=4$ ).

Die Probanden fühlten, während sie in die Richtung der höchsten Aktivierung schauten, ein weites Spektrum an negativer Körpersensationen, die wir an dieser Stelle nur exemplarisch andeuten möchten: Zittern, Schweiß, Verkrampfungen der Muskeln, Druck, Enge, Tränen, Hitze, Taubheit, Übelkeit, Kälte, Müdigkeit, Atemnot, Schwäche, Magen- und Halsprobleme, Herzrasen ... Während der Blick in die Richtung der geringsten Aktivierung ausgerichtet war, verspürten die Probanden vorwiegend Sensationen wie beispielsweise Leichtigkeit, Entlastung, Ruhe, Weite, Entspannung, Schwere, Müdigkeit, Wärme, Kribbeln, Beweglichkeit, Frische, tiefes/freies Atmen und Lächeln.

Die unterschiedlichen körperlichen Sensationen waren augenscheinlich bei der höchsten sowie bei der geringsten Aktivierung tendenziell in ventralen, zentralen und kranialen Regionen d.h. im Bereich des Kopfes, Halses, der Schultern, des Brustkorbes und der Bauchregion verortet, wohingegen distale und dorsale Körperregionen tendenziell eher bei einer geringeren Aktivierung zu spüren waren. Eine Analyse ergab keine signifikanten Unterschiede in den Häufigkeiten der Körperregionsnennung bei der höchsten und geringsten Aktivierung (T-Test bei gepaarten Stichproben:  $M=0,06$ ;  $Std=2,43$ ;  $T(15)=-1,03$ ;  $n.s.$ ). Die Abbildung 2 illustriert die Häufigkeiten der Nennungen von Körperregionen.



**Abb. 1:** Blickrichtungen der Probanden bei höchster und geringster Aktivierung (Die Zahlenwerte der Achsen dienen nur der Ausrichtung der Blickrichtung in der Darstellung und die Bezeichnungen rechts/links sind aus Sicht der Probanden zu verstehen.)



**Abb. 2:** Lokalisation der Sensation in Körperregion mit Kodierung bei höchster und geringster Aktivierung (Eine Mehrfachnennung war möglich.)

## 4.2 Quantitative Ergebnisse

Es erfolgt die Darstellung der quantitativen Ergebnisse. Bei dem Vergleich der empfundenen positiven und negativen Emotionen sowie des SUD (Subjective Units of Distress Scale) während der unterschiedlichen Blickrichtungen lassen sich signifikante Unterschiede mit hohen Effektstärken  $r < .5$  (Field, 2009) feststellen (Tabelle 2). Die Blickrichtung der geringsten Aktivierung ist im Vergleich zur höchsten Aktivierungsblickrichtung signifikant mit mehr positiven und weniger negativen Emotionen sowie mit einem generell niedrigeren Belastungsniveau verbunden. Betrachtet man die emotionalen Qualitäten differenzierter, so wird deutlich, dass die Intensitäten aller Emotionen (positive und negative) mit Ausnahme von den Emotionen a) aktiv, b) wach und c) aufmerksam zu sein, signifikante Unterschiede in Abhängigkeit der Blickrichtung aufweisen. Die Effektstärken liegen im mittleren ( $r > .3$ ) und hohen ( $r > .5$ ) Bereich.

Abbildung 3 zeigt detailliert die Intensität aller Emotionen in unterschiedlichen Situationen, im Alltag, in der aktuellen Situation vor der Exposition (Prä-Test) und während der Exposition a) bei der höchsten Aktivierungsblickrichtung und b) bei der geringsten Aktivierungsblickrichtung. Das Emotionsprofil der Situation vor der Exposition (Prä-Test) und während der Exposition bei der geringsten Aktivierungsblickrichtung erscheinen sehr ähnlich. Das alltägliche Emotionsprofil weist mehr Abweichungen und insgesamt höhere Intensitäten sowohl bei positiven als auch bei negativen Emotionen auf. Die Intensität der Emotionen bei der höchsten Aktivierungsblickrichtung ist, im Vergleich zu allen anderen Situationen, durch wenig intensive Positivität und eine intensivere Negativität gekennzeichnet. Allein die Aufmerksamkeit wird als immer identisch berichtet.

Bei Teilung der Stichprobe in höher und weniger belastete Probanden zeigen die Ergebnisse einer einfaktoriellen Anova einen signifikanten Unterschied mit mittlerer Effektstärke ( $r > .3$ ) bezüglich der positiven Emotionen während der höchsten Aktivierungsblickrichtung. Weitere signifikante Unterschiede wurden nicht gefunden (Tabelle 3). Vertrauensvoll auf die Robustheit auch bei Nichterfüllen der relevanten Voraussetzungen dieser Analyse und bei annähernd gleich großen Gruppengrößen wurde auf die Testung der Normalverteilung und weiterer Voraussetzungen (Field, 2009; S.347ff) verzichtet.

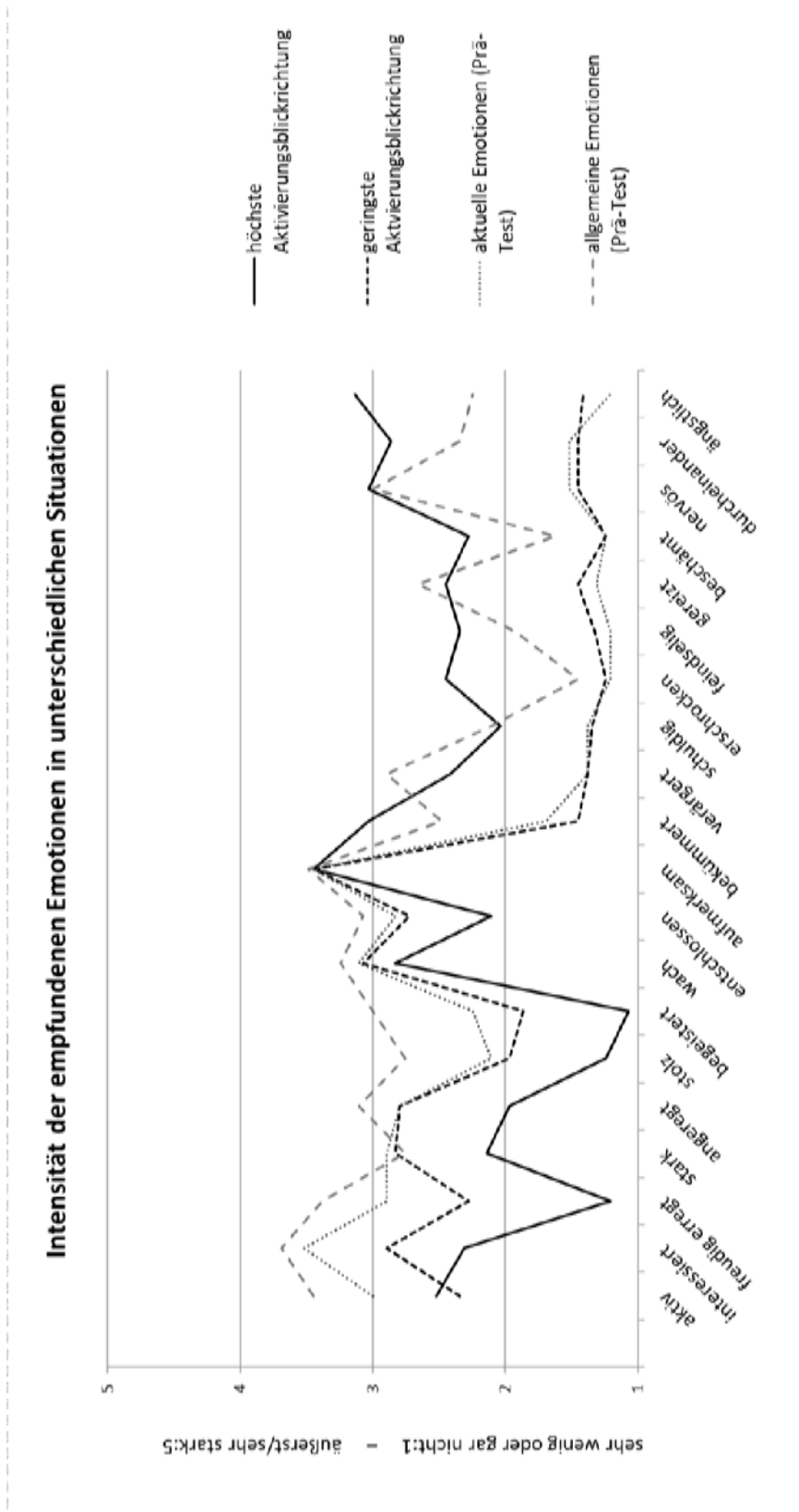
Weder das Stresslevel noch das empfundene Glück haben sich über den Zeitraum der Testung (Prä-Post-Vergleich aktuelles Stresslevel:  $M_{prä}=3,82$  mit  $Std=1,05$ ;  $M_{post}=3,36$  mit  $Std=1,32$ ;  $z=-1,35$ ; n.s. / Prä-Post-Vergleich Glück:  $M_{prä}=6,52$  mit  $Std=1,67$ ;  $M_{post}=6,65$  mit  $Std=1,70$ ;  $z=0,76$ ; n.s.) signifikant verändert, obwohl augenscheinlich eine Tendenz zur Verbesserung in beiden Parametern zu erkennen ist.

**Tab. 2:** Ergebnisse für den Vergleich der Emotionen bei unterschiedlichen Blickrichtungen (Wilcoxon-Test für verbundene Stichproben)

	<i>höchste Aktivierungs- blickrichtung M=</i>	<i>geringste Aktivierungs- blickrichtung M=</i>	<i>Testwert z=</i>	<i>Signifikanz p=</i>	<i>Effektstärke r=</i>
<b>Positive Emotionen (gesamt)</b>	2,08 (Std=0,53)	2,62 (Std=0,79)	3,16	.00**	.59
<b>Negative Emotionen (gesamt)</b>	2,60 (Std=0,74)	1,37 (Std=0,50)	-4,35	.00**	-.81
<b>SUD</b>	6,51 (Std=1,57)	3,17 (Std=1,66)	4,55	.00**	.84
aktiv	2,52	2,34	-0,31	.75 (n.s.)	-
interessiert	2,31	2,90	2,12	.03*	.39
freudig erregt	1,21	2,28	3,57	.00**	.66
stark	2,14	2,83	2,20	.03*	.41
angeregt	1,97	2,79	3,24	.00**	.60
stolz	1,24	1,97	3,16	.00**	.59
begeistert	1,07	1,86	3,23	.00**	.60
wach	2,83	3,07	0,86	.39 (n.s.)	-
entschlossen	2,10	2,72	2,25	.02*	.42
aufmerksam	3,45	3,45	-0,08	.93 (n.s.)	-
bekümmert	3,03	1,45	-4,10	.00**	-.76
verärgert	2,41	1,38	-3,16	.00**	-.59
schuldig	2,03	1,34	-2,39	.02*	-.44
erschrocken	2,45	1,24	-3,69	.00**	-.68
feindselig	2,34	1,33	-3,32	.00**	-.62
gereizt	2,45	1,45	-2,93	.00**	-.54
beschämt	2,28	1,24	-3,09	.00**	-.57
nervös	3,03	1,45	-4,00	.00**	-.74
durcheinander	2,86	1,45	-3,92	.00**	-.73
ängstlich	3,14	1,41	-4,14	.00**	-.77

\*signifikant auf p&lt;.05 Level \*\* signifikant auf p&lt;.01 Level





**Abb. 3:** Intensität der empfundenen Emotionen in unterschiedlichen Situationen (höchste und geringste Aktivierungsblickrichtung, aktuelle Intensität von Emotionen vor Exposition und allgemeine Intensität von Emotionen)

**Tab. 3:** Ergebnisse des Gruppenvergleichs von unterschiedlich stark belasteten Probanden

<b>positive Emotionen höchste Aktivierung</b>	geringer belastete Probanden	16	1,91	0,51	Zwischen den Gruppen	4,44	0,04*	.37
	höher belastete Probanden	13	2,30	0,49				
	Gesamt	29	2,08	0,53				
<b>negative Emotionen höchste Aktivierung</b>	geringer belastete Probanden	16	2,56	0,77	Zwischen den Gruppen	0,10	0,75 (n.s.)	
	höher belastete Probanden	13	2,65	0,74				
	Gesamt	29	2,60	0,74				
<b>positive Emotionen geringste Aktivierung</b>	geringer belastete Probanden	16	2,59	0,85	Zwischen den Gruppen	0,06	0,81 (n.s.)	
	höher belastete Probanden	13	2,66	0,74				
	Gesamt	29	2,62	0,79				
<b>negative Emotionen geringste Aktivierung</b>	geringer belastete Probanden	16	1,38	0,39	Zwischen den Gruppen	0,00	0,99 (n.s.)	
	höher belastete Probanden	13	1,37	0,63				
	Gesamt	29	1,37	0,50				
<b>SUD höchste Aktivierung</b>	geringer belastete Probanden	16	7	2	Zwischen den Gruppen	0,40	0,53 (n.s.)	
	höher belastete Probanden	13	6	2				
	Gesamt	29	7	2				
<b>SUD geringste Aktivierung</b>	geringer belastete Probanden	16	3	2	Zwischen den Gruppen	0,38	0,54 (n.s.)	
	höher belastete Probanden	13	3	2				
	Gesamt	29	3	2				

\*signifikant auf  $p < .05$  Level

## 5. Diskussion

Wie im vorhergehenden Abschnitt beschrieben, konnten alle Versuchsteilnehmer\*innen jeweils eine spezifische Blickrichtung im Visusfeld einnehmen, bei der sich die unterschiedlichen Qualitäten einer höchsten Aktivierung und eine andere Blickrichtung mit der geringsten Aktivierung, während der Erinnerung an das gewählte belastende bzw. traumatische Ereignis, erleben ließen. Dies war zudem für alle Personen unabhängig von der psychischen Vorbelastung und dem ausgewählten belastenden oder traumatischen Ereignis möglich.

Kohärent mit den Annahmen erfahrener Trauma-Therapeut\*innen und Forscher\*innen (siehe Levine, 2010; Porges, 2001) beschrieben die Personen bei der Blickrichtung mit der höchsten Aktivierung bzw. empfundenen Belastung entsprechende vom Vagusnerv und dem dorsal vagalen Nervensystem innervierte und jetzt deutlich spürbare Körperpartien, die anatomisch mit den Bereichen der inneren Organe (Bauch-, Brustraum und Halsregionen) verbunden sind. Es wird davon ausgegangen, dass die Aktivierung dieses nervalen Teilsystems für Reaktionen der Immobilisation und des Erstarrens zuständig ist (ebd.). Bei Einnahme der wenig belastenden Blickrichtung beschreiben die Personen augenscheinlich mehr periphere und distale Körperpartien, die nun unterschiedlichste Sensationen empfinden. Die Sensationen in Beinen und Armen könnten auf eine damalige oder aktuelle Aktivierung des sympathischen Systems und die damit assoziierte Kampf- oder Fluchtreaktion hinweisen.

Eine naheliegende Spekulation wäre, dass unsere Versuchspersonen aktuell körperlich eine geringere Bedrohung erleben und sich auf aktive Reaktionen einstellen, während sie diese weniger belastende Blickrichtung einnehmen. Die höhere Belastung scheint eher mit der erlebten oder nun wieder wahrgenommenen Immobilitätsreaktion assoziiert zu sein. Überprüfen können wir das mit den vorliegenden Ergebnissen und dem gewählten Ansatz dieser Pilotstudie leider nicht. Hierzu hätte man die Probanden noch genauer befragen, bzw. physiologische Parameter erheben müssen. Dennoch diskutieren wir nachfolgend weitere stützende als auch kritisch zu bedenkende Hinweise und Interpretationsmöglichkeiten der vorliegenden Befunde dieser Pilotstudie. So wäre es auch denkbar, dass allein die suggestive Aufforderung, eine belastende und eine möglichst wenig belastende Blickrichtung zu wählen, sowie das gewählte Forschungssetting, diese und die weiteren, aus unserer Sicht, erfreulichen Ergebnisse im Sinne einer Fremd- oder Autosuggestion hervorruft. Wir können dies an dieser Stelle nicht ausschließen.

Die berichteten Ergebnisse zeigen zum einen, dass die gewählte Blickrichtung die emotionale Intensität, und zum anderen aber auch die empfundene emotionale Qualität von erinnerten belastenden/traumatischen Ereignissen signifikant verändert. Die signifikanten Unterschiede der erlebten Belastung (SUD) in Abhängigkeit der gewählten Blickrichtung könnten ebenso von suggestiv induzierten Erwartungen betroffen sein. Das signifikante Absinken der wahrgenommenen Belastung könnte zudem mit dem gewählten Versuchsaufbau zusammenhängen, da erst die Blickrichtung der höchsten Aktivierung (5min lang) eingenommen wurde, es ggf. schon hier zu einer Habituation gekommen ist, und erst danach die Blickrichtung der geringsten Aktivierung gesucht und eingenommen wurde. Allerdings spricht die Bestimmung und das Aufsuchen der initial gewählten Blickrichtung (höchste Aktivierung) durch den Probanden selbst logischerweise gegen diese Annahme.

Alleinig die Existenz einer Blickrichtung mit höchster Aktivierung bedingt an sich schon die anderen schwächer aktivierenden Blickrichtungen, unabhängig von möglichen Habituationsprozessen.

Einhergehend mit der reduzierten Belastung, in Abhängigkeit der Blickrichtung, werden auch signifikant mehr positiv konnotierte Emotionen und signifikant weniger negativ konnotierte Emotionen berichtet. Dies deutet auf eine weniger bedrohlich empfundene Interpretation des gewählten Ereignisses in der Erinnerung hin, wobei nochmals zu betonen ist, dass es sich um die identische traumatische/belastende Situation in der Erinnerung handelt. Es wurde ausschließlich die Blickrichtung im Visusfeld gewechselt.

Keine signifikanten Unterschiede finden sich bezüglich der einzelnen emotionalen Aspekte aktiv, wach und aufmerksam, welche auf eine allgemeine Wachsamkeit, während des gesamten Erinnerungsprozesses unabhängig von der Blickrichtung schließen lassen. Dieser Befund scheint naheliegend, da die Personen aufgefordert waren, Ereignisse zu erinnern, die auch ein gewisses Erregungsniveau hervorrufen können. Zudem unterstützen die Befunde der Emotionsqualitäten, die Annahme einer Wirkung der Blickrichtung unabhängig von möglichen Einflüssen der Suggestion, da diese emotionalen Qualitäten nicht Inhalt der Versuchsinstruktionen waren.

Warum stärker belastete Probanden signifikant stärker empfundene positive Emotionen während der Exposition (höchste Aktivierungsblickrichtung) im Vergleich zu weniger belasteten Probanden aufweisen, vermögen wir an dieser Stelle nicht zu sagen. Es wäre möglich, dass sie aktiv ein höheres Leiden während der Exposition eingehen und beispielsweise mehr negative Gedächtnisinhalte zulassen. Sie könnten auch andere Coping-Stile verwenden und weniger kognitive Vermeidungsstrategien oder beispielsweise auch weniger dissoziative Phänomene nutzen. Dennoch ist diesbezüglich keine abschließende Interpretation möglich, da die Probanden bezüglich ihrer kognitiven Strategien und Erlebnisse während der Studie nicht weiter befragt wurden. Weitere signifikante Gruppenunterschiede wurden nicht gefunden. Es bleibt festzuhalten, dass die Pilotstudie auch keine Belastung für die Probanden darstellte. Stresslevel und empfundenes Glück haben sich über den Zeitraum der Testung nicht signifikant zum Schlechteren verändert, ganz im Gegenteil. Augenscheinlich ist eine Tendenz zur Verbesserung in beiden Parametern zu erkennen.

Um methodischer Kritik vorzubeugen, wäre es bei einer Nachfolgestudie sinnvoll die Hälfte der Probanden zuerst mit der Blickrichtung der höchsten Aktivierung und die andere Hälfte der Probanden mit der Blickrichtung der geringsten Aktivierung zu testen.

Für zukünftige Forschung wäre es zudem notwendig, die Stichprobengröße zu erhöhen, um zu empirisch belastbareren Aussagen zu gelangen. Des Weiteren wäre es wünschenswert, die Daten mit einerseits mehr subjektiv erlebbaren Parametern und andererseits auch mit physiologischen Parametern (Hautleitwert, u.a.) zu unterfüttern. Interessant wäre auch die Betrachtung von kompletten Therapiesitzungen und Behandlungsverläufen. Der Bedarf an qualitativ hochwertiger Forschung in diesem Bereich bleibt wie bisher bestehen.

Trotz der beschriebenen, kritisch zu bedenkenden Aspekte und nicht zu unterbewertenden Limitationen dieser kleinen Pilotstudie, freuen wir uns, dass folgende vielversprechende Ergebnisse zu berichten sind:

- Alle Probanden fanden zu einem persönlichen, spezifischen, belastenden/traumatisierenden Ereignis eine entsprechende spezifische Blickrichtung, die mit dem Teleskop-Zeigestab im Raum gesucht wurde, bei der sie eine höchste Aktivierung verspüren sowie eine weitere, bei der die emotionale Aktivierung am geringsten ausfällt.
- Die gewählte Blickrichtung im Visusfeld kann die empfundene emotionale Qualität sowie emotionale Intensität von erinnerten belastenden/traumatischen Ereignissen signifikant verändern.
- Dies ist nicht wesentlich abhängig von der Stärke der Vorbelastung (Traumatisierung), allerdings berichten höher belastete Probanden signifikant stärker empfundene positive Emotionen während der Exposition.

Abschließend möchten wir allen Unterstützer\*innen dieser Pilotstudie für ihr Engagement vielmals danken.

## 6. Literaturverzeichnis

Beaulieu, D. (2005). *Eye Movement Integration Therapy (EMI): The Comprehensive Clinical Guide*. Wales UK: Crown House Publishing.

Bradshaw, R. A., Cook, A., & McDonald, M. J. (2011). Observed & experiential integration (OEI): Discovery and development of a new set of trauma therapy techniques. *Journal of Psychotherapy Integration*, 21, 104-171. doi: 10.1037/a0023966

Corrigan, F., & Grand, D. (2013). Brainspotting: recruiting the midbrain for accessing and healing sensorimotor memories of traumatic activation. *Med Hypotheses*, Vol 80, 759-66. Doi: 10.1016/j.mehy.2013.03.005.

Ecker, B., Ticic, R., & Hulley, L. (2016). *Der Schlüssel zum emotionalen Gehirn*. Paderborn: Junfermann Verlag.

Field, A. (2009). *Discovering Statistics using SPSS*. Los Angeles: Sage.

Grand, D. (2013). *Brainspotting*. Kirchzarten: VAK Verlag.

Gysi, J. (2018). Veränderungen im ICD-11 im Bereich Trauma & Dissoziation. *Letzter Zugriff 03.03.2019* - <https://www.jangysi.ch/Therapie,-Medizin,-Journalismus/>

Herman, J. (2010). *Trauma and recovery* (reprint). London UK: Pandora.

Hesse, P. U. (2017). Brainspotting in der Suchttherapie. *Trauma. Zeitschrift für Psychotraumatologie und ihre Anwendungen*, Vol 15 (3), 46-57.

- Hollander, H. E. (2009). ECEM (Eye Closure Eye Movements): Application to Depersonalization Disorder. *American Journal of Clinical Hypnosis*, 52, 95 – 109. doi: 10.1080/00029157.2009.10401701
- Hollander, H. E., & Bender, S. S. (2001). ECEM (Eyes Closure Eye Movements): Integrating Aspects of EMDR with Hypnosis for Treatment of Trauma. *American Journal of Clinical Hypnosis*, 43, 187 – 202. doi: 10.1080/00029157.2001.10404276
- Levine, P. A. (1997). *Waking the tiger, healing trauma: The innate capacity to transform overwhelming experiences*. Berkeley, CA: North Atlantic Books.
- Levine, P. A. (2010). *Sprache ohne Worte*. München: Kösel Verlag.
- Levine, P. A. (2015). *Trauma und Gedächtnis. Die Spuren unserer Erinnerung in Körper und Gehirn*. München: Kösel Verlag.
- Maercker, A. (2013). *Posttraumatische Belastungsstörung*. Berlin: Springer Verlag.
- Marselle, M. R., Irvine, K. N., Lorenzo-Arribas, A., & Warber, S. L. (2016). Does perceived restorativeness mediate the effects of perceived biodiversity and perceived naturalness on emotional well-being following group walks in nature? *Journal of Environmental Psychology*, 46, 217-232. doi: 10.1016/j.jenvp.2016.04.008
- Matthijssen, S. J. M. A, van Beerschoten, L.M. , de Jongh, A., Klugkist, I. G., & van den Hout, M. A. (2018). Effects of "Visual Schema Displacement Therapy" (VSDT), an abbreviated EMDR protocol and a control condition on emotionality and vividness of aversive memories: Two critical analogue studies, *Journal of Behavior Therapy and Experimental Psychiatry*, 63, 48-56. doi: 10.1016/j.jbtep.2018.11.006
- Micic, D., Ehrlichman, H., & Chen, R. (2010). Why do we move our eyes while trying to remember? The relationship between non-visual gaze patterns and memory. *Brain and Cognition*, 74, 210-224. Doi: 10.1016/j.bandc.2010.07.014
- Nijenhuis, E. R. S. (2010). The scoring and interpretation of the SDQ-20 and SDQ-5, *Activitas Nervosa Superior*, Vol. 52, 24–28. Doi: 10.1007/BF03379561
- Ogden, P., Pain, C., & Fisher, J. (2006). A sensorimotor approach to the treatment of trauma and dissociation. *Psychiatric Clinics of North America*, 29, 263–279. doi: 10.1016/j.psc.2005.10.012
- Porges, S. W. (2001). The polyvagal theory: phylogenetic substrates of a social nervous system, *International Journal of Psychophysiology*, 42, 123-146. Doi: 10.1016/S0167-8760(01)00162-3
- Prins, A., Bovin, M. J., Kimerling, R., Kaloupek, D. G., Marx, B. P., Pless Kaiser, A., & Schnurr, P. P. (2015). *The Primary Care PTSD Screen for DSM-5 (PC-PTSD-5)*. *Letzter Zugriff 03.03.2019* - <http://www.ptsd.va.gov/professional/assessment/screens/pc-ptsd.asp>

Sack, M., Zehl, S., Otti, A., Lahmann, C., Henningsen, P., Kruse, J., & Sting, M. (2016). A comparison of dual attention, eye movements, and exposure only during eye movement desensitization and reprocessing for posttraumatic stress disorder: Results from a randomized clinical trial. *Psychother Psychosom*, Vol. 85, 357–365. Doi: 10.1159/000447671

van der Kolk, B. (2014). *The body keeps the score: Brain, mind, and body in the healing of trauma*. New York: Penguin Books.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales; *Journal of Personality and Social Psychology*, Vol. 54, 1063-1070. Doi: 10.1037/0022-3514.54.6.1063

Weathers, F. W., Blake, D. D., Schnurr, P. P., Kaloupek, D.D., Marx, B. P. & Keane, T. M. (2013). *Clinician-administered PTSD scale for DSM-5 (PTSD CAPS)*. Letzter Zugriff 03.03.2019 - <https://www.ptsd.va.gov/professional/assessment/adult-int/caps.asp>

Wolpe, J. (1969). *The Practice of Behavior Therapy*. New York: Pergamon Press.





**Anschriften der Autoren**

ZeE	Zentrum für empirische Evaluationsmethoden c/o Humboldt-Universität zu Berlin Institut für Psychologie Rudower Chaussee 18, 12489 Berlin email: vorstand@z-e-e.de homepage: <a href="http://www.z-e-e.de">http://www.z-e-e.de</a>
Dr. Rebekka Gerlach	Humboldt-Universität zu Berlin Institut für Psychologie Rudower Chaussee 18, 12489 Berlin <a href="mailto:rebekka.gerlach@psychologie.hu-berlin.de">rebekka.gerlach@psychologie.hu-berlin.de</a>
Prof. Dr. Elke van der Meer	Humboldt-Universität zu Berlin Institut für Psychologie Rudower Chaussee 18, 12489 Berlin <a href="mailto:vdMeer@rz.hu-berlin.de">vdMeer@rz.hu-berlin.de</a>
Birgit Nimke-Sliwinski	Berliner Stadtreinigungsbetriebe Anstalt des öffentlichen Rechts Ringbahnstr. 96, 12103 Berlin
Rainer Foerges	AWB Abfallwirtschaftsbetriebe Köln GmbH Maarweg 271, 50825 Köln
Prof. Dr. Reinhard Beyer	Humboldt-Universität zu Berlin Institut für Psychologie Rudower Chaussee 18, 12489 Berlin <a href="mailto:reinhard.beyer@psychologie.hu-berlin.de">reinhard.beyer@psychologie.hu-berlin.de</a>
Marc Hochstrate	Zentrum für empirische Evaluationsmethoden Rudower Chaussee 18, 12489 Berlin <a href="mailto:marc.hochstrate@campus.tu-berlin.de">marc.hochstrate@campus.tu-berlin.de</a>
Marius Fahrner	Touro College Berlin Am Rupenhorn 5, 14055 Berlin <a href="mailto:dozent.fahrner@t-online.de">dozent.fahrner@t-online.de</a>
Prof. Dr. Birgit Wolf	Touro College Berlin Am Rupenhorn 5, 14055 Berlin <a href="mailto:birgit.wolf@touroberlin.de">birgit.wolf@touroberlin.de</a>
Prof. Dr. M. Luzi Beyer	Methoden der quantitativen Forschung im Sozial- und Gesundheitswesen Alice Salomon Hochschule Berlin Alice-Salomon-Platz 5, 12627 Berlin <a href="mailto:luzi.beyer@ash-berlin.de">luzi.beyer@ash-berlin.de</a>

Dr. Karin Schüttpelz-Brauns	Medizinische Fakultät Mannheim der Universität Heidelberg Theodor-Kutzer-Ufer 1-3, 68167 Mannheim katrin.schuettpelz-brauns@medma.uni-heidelberg.de
Prof. Dr. Inga Großmann	HMKW Hochschule für Medien, Kommunikation und Wirtschaft University of Applied Sciences Ackerstraße 76, 13355 Berlin inga_grossmann@gmx.de
Prof. Dr. Bodo Krause	Humboldt-Universität zu Berlin Institut für Psychologie Rudower Chaussee 18, 12489 Berlin bkrause@cms.hu-berlin.de
Dr. Majken Bieniok	LARA e.V. Projekt für Frauen mit Komplextumatisierung Beratung – Vernetzung - Bildung Fuggerstr. 19, 10777 Berlin MAJKEN:BIENIOK@LARA-BERLIN.DE
Dipl. Psych. Alexander Reich	Akademie für integrative Traumatherapie Käthe-Niederkirchner-Str. 23, 10407 Berlin alexander.reich@akademie-traumatherapie.de
Dipl. Soz. Wiss. Peter Uwe Hesse	Harzkrankenhaus Dorothea Christiane von Erleben Station P 7 Trauma Thiestr. 7 – 10, 38889 Blankenburg ZENTRUM.PB@WEB.DE

**Aus der Reihe „Empirische Evaluationsmethoden“ sind bereits erschienen:**

**Band 1**

- |                       |   |
|-----------------------|---|
| B. Krause             | Probleme der Veränderungsmessung im Rahmen der Evaluationsmethodik                              |
| T. Bachmann           | Strukturgleichungsmodelle in der Veränderungsmessung  |
| U. Gauger             | Versuchspläne zur Interventionsevaluation   |
| P. Metzler, B. Krause | Methodischer Standard bei Studien zur Therapieevaluation  |
| R. Beyer              | Methoden zur Evaluation von sprachlichen Äußerungen und Texten bezüglich ihrer Verständlichkeit |

**Band 2**

- |                       |  |
|-----------------------|--|
| B. Krause             | Grundprobleme der Nutzung nicht-metrischer Daten in der Evaluationsforschung   |
| H.-J. Lander, M. Huth | Die statistische Beurteilung von Interventionseffekten mittels einer kategorialen Prä- Posttest- Analyse                       |
| S. Klein              | Modelle zum Vergleich unterschiedlicher Ordinal-Skalen am Beispiel des semantischen Differential                               |
| T. Raykov             | On Structural Equation Model Identification  |
| R. Beyer              | Nutzung experimenteller Methoden zur Prüfung diagnostischer Fragestellungen unter Verwendung von Sprachverstehensanforderungen |

**Band 3**

- |                          |  |
|--------------------------|--|
| B. Krause, S. Klein      | Validierung von Modellen der qualitativen Prä- und Posttest-Analyse  |
| S. Klein                 | Vergleich von Item-Antwort-Modellen zur Qualitativen Veränderungsmessung   |
| U. Gauger                | Alternative Analysetechniken für Prä-Posttest-Designs (Vergleich mehrerer Gruppen)   |
| K.H. Schmidt, P. Metzler | Werden psychische Erkrankungen bei der Beurteilung von Pflegestufen nach dem Pflegeversicherungsgesetz hinreichend berücksichtigt? – Evaluation mit Hilfe einer neuen Skala zur Beurteilung der Pflegebedürftigkeit von Senioren (BPS) |
| R. Beyer                 | Identifizierung von Alterseffekten bei der Nutzung von Vor-  |

wissen im Kontext von Sprachverstehensanforderungen

A. Bergner

Einfluß des Erlebens und der Verarbeitung von Fehlgeburten auf den Verlauf einer neuen Schwangerschaft, die Geburt und die frühe Mutter-Kind-Beziehung

#### **Band 4**

R. Beyer, T. Guthke,  
B. Kothe

Identifikation von spezifischen kognitiven Defiziten bei links- und rechtshemisphärischen Verletzungen unter Verwendung von Sprachverstehensanforderungen

G. Kaul

Die Kompatibilität zwischen der Messung kognitiver Leistungsparameter und der subjektiv erinnerbaren Beschwerdehäufigkeit bei erwerblich bedingter, neurotoxischer Belastung durch Lösungsmittel

S. Künstner, O. Sezer

Die Bedeutung der zeitlichen Perspektive für die Übereinstimmung von Lebensqualitätsurteilen bei Krebspatienten aus Arzt- und Patientenperspektive

H.-J. Lander, M. Huth

Beziehungen zwischen einem regressionsanalytischen und einem varianzanalytischen Zugang zur Veränderungsmessung

S. Klein

IRT-Modelle in der Veränderungsmessung für dichotome Daten

O. B. Walter

Die Hauptkomponentenanalyse als Ansatz zur Auswertung von Grid-Matrizen

#### **Band 5**

H.-J. Lander, M. Huth,

Ein methodischer Ansatz zur Erfassung und zum Vergleich von Wissensstrukturen nach ihren strukturdimensionalen Eigenschaften

R. Beyer, T. Guthke

Integration von Vorwissen beim Sprachverstehen und differentielle Effekte in Abhängigkeit vom Lebensalter (Senioren)

B. Bergmann, P. Zehrt

Transferorientierte Evaluation von Trainingsprogrammen zur Entwicklung von Störungsdiagnosekompetenz in flexiblen Fertigungsprozessen

S. Künstner

Substitutionsregeln - Grenzen linearer Strukturgleichungsmodelle bei der Prüfung konkurrierender theoretischer Annahmen

S. Klein

Pros und Cons zur Untersuchung von Differential Item Functioning

- M. L. Beyer Die Robustheit von Mittelwertvergleichen
- U. Gauger Vergleich verschiedener Techniken zur Item-Selektion
- Band 6**
- R. Beyer, Kompensation von Defiziten im Arbeitsgedächtnis bei Senioren durch Rückgriff auf die Wissensrepräsentation im Langzeitgedächtnis
- G. Kaul Der autokinetische Lichttest und sein Einsatz unter experimentell kontrollierten Bedingungen, zur Beobachtung von Effekten eines elektromagnetischen Feldes
- S. Künstner, O. Sezer Die Vergleichbarkeit von Lebensqualitäts-Urteilen: Eine Multi-Trait Multi-Method Analyse über den EORTC QLQ-C30, den SF-36 und den FLIC
- B. Krause Bewertung von Schätzurteilen – Ratingskalen und Präferenzskalen zur Erfassung der Lebensqualität
- S. Künstner Parceling von Itemmengen bei linearen Strukturgleichungsmodellen
- J. Wolf Ein experimenteller Ansatz zur Erfassung von Variablen der Leistungsmotivation
- U. Scheidereiter, J. Wolf Wie gut sind Clusteranalysen?
- A. Frey Anwendungsperspektiven adaptiver Teststrategien bei Evaluationsstudien
- Band 7**
- B. Krause Methodische Grundlagen zur Evaluation verkehrspsychologischer Rehabilitation
- T. Loßnitzer, A. Frey Modellierung intraindividuelle Veränderungen mit Strukturgleichungsmodellen
- S. Ivanov Homogenität und Veränderung: Anwendung der latenten Klassenanalyse im Zusammenhang von Veränderungsmessung
- S. Klein Effektstärken für multivariate Zielgrößen von Evaluationsstandards in der Psychotherapieforschung
- S. Wolf Kriteriumorientierte Therapieevaluation. Eine Reanalyse
- A. Dubrowsky, U. Gauger Identifikation von Bedienprofilen bei Klimaautomaten im PKW anhand von Zeitreihenclusteranalyse: ein möglicher Weg?

- L. Beyer Der Vergleich von Skalierungsverfahren bei der Beurteilung der Einstellung zu politischen Parteien
- A. Frey, T. Loßnitzer Circumplexmodellierung emotionaler Selbstbeschreibung
- Band 8**
- P. Metzler Skalierung durch Validierung - Ein Beitrag zur Testkonstruktion für die klinische Psychologie und Neuropsychologie
- A. Frey, S. Klein Vergleich von drei Verfahren zur Modellierung von Modatoreffekten bei qualitativen Moderatorvariablen
- J. Uher, L. Beyer Denken in Implikationen: Der Einfluss des Abstraktionsgrades und der Negation der Aussagen auf das Erkennen der Korrektheit konditionaler Schlussschemata
- R. Beyer, T. Guthke, E. van der Meer, E. Ferstl Einfluss von spezifischen Hirnläsionen auf die Aktivierung und kontrollierte Bereitstellung von Vorwissen
- N. Graf Evaluation eines Expertensystems zur Gestaltung von Assistenzsystemen
- Band 9**
- G. Kaul Der Einsatz eines balancierten Periodenversuchsplans zur Suche nach der Stecknadel im Hauhaufen: Gibt es eine begründete Hypersensibilität auf elektromagnetische Felder?
- K. Brauns Anwendung der Prä-Posttest-Analyse von Lander zur Evaluation von ANNASS, einem Konzept zur Sprachförderung von 3- bis 4-jährigen Kindern
- N. Graf Evaluation der Wirksamkeit und Gebrauchstauglichkeit des Expertensystems GUIDEAS
- A.-M. Hetze, B. Bergmann, E. Dittmann Inhaltsabhängigkeit von Lehrveranstaltungen auf die studentische Bewertung: Ein Vergleich grundlagen- vs. anwendungsorientierter Kurse des Dresdner integrativen problemorientierten Lernens (DIPOL)
- R. Beyer, E.v.d.Meer, J. Türke, R. Reichmayr Merkmale einer Metropole – psychologische Untersuchungen zur Klärung eines unscharfen Begriffs
- P. Schepler Gewinne sind nicht gleich Gewinne: Evaluation des postoperativen neuropsychologischen Outcomes bei Epilepsiepatienten

- B. Krause Äquidistante Ratingskalen? Zugänge und Formen ihrer Prüfung
- Band 10**
- P. Schepler Zusammenhang von subjektiven und objektiven Gedächtnismaßen
- R. Beyer,  
E. van der Meer,  
R. Gerlach Differenzierung minimalistischer und maximalistischer Inferenzansätze beim Sprachverstehen
- B. Krause Äquidistante Ratingskalen ? – Teil 2  
Differenzierung der Ursachen bei Anpassungsproblemen
- D. Heintze, E. Mäthner Probleme und Lösungsansätze bei der Beurteilung der Effektivität von Trainingsmaßnahmen
- K. Brauns,  
Z. M. Nouns Der Progress Test Medizin
- U. Feige-Klische Evaluation der Qualitätsentwicklung an einer Berliner Grundschule
- K. Nepp Modell zur Wirksamkeit von Schulberatung zum Schulprogramm im Land Berlin
- T. Loßnitzer,  
S. Born,  
Boris Schmidt Lehrveranstaltungsevaluation- an der Friedrich-Schiller-Universität Jena – Eine Analyse der Dimensionalität der eingesetzten Fragebögen
- Band 11**
- R. Gerlach  
R. Beyer,  
E. van der Meer, Identifikation und Nutzung von semantischen Relationen beim Sprachverstehen
- S. Prüfrock  
R. Beyer Empirische Prüfung eines Leseverständnistests für Erwachsene
- A. Herbon Messen von Emotionen zur Anwendung in der Mensch-Maschine-Interaktion
- M.L. Beyer  
D. Heintze Ein Urwald an Informationen – Lernen und Entscheiden unter Unsicherheit
- M. Barth  
St. Freutel Analyse und Bewertung von Ratingskalen über Itemmengen eines Gegenstandsbereiches
- B. Krause Zur Konstruktion von messtheoretisch begründeten Ratingskalen

K. Brauns Zugänge zur Evaluation eines nicht-parametrischen  
Personen-Fit-Index

### **Band 12**

C. Platho Einfluss interindividueller Unterschiede auf Bewertung  
und Präferenz von Automobildesign

R. Gerlach, R. Beyer,  
E. van der Meer Evaluation von Begriffshierarchien bezüglich ihrer An-  
schaulichkeit und ihrer Merkmalscharakteristik

M. Bieniok, R. Beyer,  
E. van der Meer Interkultureller Vergleich des Metropolenbegriffs - eine  
Internetstudie zu Anpassungsproblemen

L. Beyer, K. Brauns Eintopf oder klare Brühe – Aussagekraft und Robust-  
heit von Skalensystemen

M. Barth Erfassung extrinsischer Regulationstypen

S. Kunert Wirkungsanalyse eines kognitiven Lernwerkzeugs –  
Quantifizierung von Unterschieden in Effektgrößen  
mittels zweifaktorieller ANOVA

S. Schwarz, L. Beyer Validierung der deutschen MADRS für das Kindes-  
und Jugendalter

D. Heintze, M. Steinbach Möglichkeiten und Grenzen der Objektivierung projek-  
tiver diagnostischer Instrumente am Beispiel des  
Satzergänzungsverfahrens

### **Band 13**

S. Kain, J. Nachtwei Die Rolle von Kontrollvariablen in der Human Factors  
Forschung – Ein bewährtes Konzept in einem  
modernen Anwendungsfeld

C. Schermuly Die systematische Verhaltensbeobachtungsmethode in  
der empirischen interaktions- und Gruppenforschung –  
Chancen, Probleme und neue Entwicklungen

B. Krause Zur praktischen Bedeutsamkeit der Modellierung mit  
latenten Variablen

M. Bieniok, R. Beyer, E. v.  
d. Meer Die gewünschte Metropole – Ein Vergleich von Berlin,  
Paris und London

R. Gerlach, R. Beyer, E. v.  
d. Meer Vorwissenseinfluss beim Satz- und Textverstehen

S. Kunert Value Contribution of Human Resource Management



**Band 14**

- J. Nachtwei, S. Kain                      Über die ökonomische Validität in Untersuchungen zur Mensch-Technik-Interaktion
- G. Kaul, J. Keitel                            Zur intraindividuellen Stabilität einer visuellen Wahrnehmungsillusion des autokinetischen Effektes
- M. Bieniok, R. Beyer, E.v.d.Meer                            Die ideale Metropole
- M. Hochstrate                                Überlegungen zu einer zeitbasierten Projektwertermittlung
- L. Beyer, J. Hamann K. Schüttpelz-Brauns                            Hält das Rasch-Modell was es verspricht?
- J. Hamann, L. Beyer Schüttpelz-Brauns                            Welchen Beitrag zum Personenfit leistet die K. probalistische Testtheorie? Eine Weiterführung des Beitrags „Eintopf oder klare Brühe“ für ordinale Daten
- B. Krause                                        Die Rolle der Poweranalyse im empirischen Forschungsprozess
- J. Nachtwei                                      Bericht über die thematische Sitzung „Ökologische Validität“
- B. Krause                                        Ökologische Validität – Einführung

**Band 15**

- Schölmerich, F., Nachtwei, J., Schermuly, C.C.                            Evaluating the quality of assessment centers used in employee selection – Development of a Benchmark for Assessment Center Diagnostics (BACDi)
- Stade, M., Meyer, C., Niestroj, N. Nachtwei, J.                            (Not) everybody´s darling: value and prospects of multiple linear Regression Analysis and Assumption Checking
- Krause, B.                                        Probleme der Interpretation experimenteller Befunde
- Bieniok, M., Beyer, R., van der Meer, E.                            Die ideale Metropole
- Beyer, L., Schüttpelz-Brauns, K.                            Rasch-Modell vs. Lambdabasierte Reliabilitätskoeffizienten

Stein, M., Beyer, L.  
Schüttpelz-Brauns, K., Populationsunabhängigkeit des Rasch-Modells

Großmann, I., Beyer, L. Kontrollierte Interventionsstudie zur Kurzzeit-Intensiv-Schematherapie. Evaluation des Hoffman-Quadrinity-Prozesses

### **Band 16**

R. Beyer, S. Pröfrock,  
M. Bieniok, R. Gerlach,  
C. Hintze, E. van der Meer Differenzierung der emotionalen und Kognitiven Beanspruchung mit Hilfe physiologischer Parameter

D.S. Blume, L.M. Beyer Evaluation einer Beratungsstelle für Stalkingopfer

N. Schagen, L.M. Beyer Der Einfluss von Anforderungen und Ressourcen auf das Arbeitsengagement unter Studierenden

B. Köppen, P. Schepler Psychologische Beratung in der Wohnungslosenhilfe - Ergebnisse einer Effektstudie

M.L. Beyer, I. Großmann Phänomenologische und Funktionale Evaluation Spielen zusammen

N. Johr, S. Udelhoven,  
C. Meyer, J. Nachtwei Auf dem Prüfstand: Assessment und Development Center bei Unternehmensgründern

### **Band 17**

M. Bieniok, R. Beyer,  
E. van der Meer Potenzial von Metropolenmerkmalen zur Befriedigung von Bedürfnissen der Bewohner und dessen Berücksichtigung in der Berliner Stadtplanung

R. Gerlach, R. Foerges,  
E. van der Meer, R. Beyer,  
B. Nimke-Sliwinski, Ursachen von Littering und Evaluation Von Interventionsmaßnahmen

S. Kunert, S. Spittier Entwicklung eines Klassifikationssystems der Methoden der Arbeits- und Organisationspsychologie

I. Großmann Entwicklung einer Korrekturformel für Verzerrungseffekte in nicht anonymen Untersuchungsbedingungen

A. Voigt, K. Elbe,  
S. Uedelhoven,  
C. von Bernstorff,  
J. Nachtwei Entwicklung eines Kategoriensystems zur Überführung Von textbasierten Ergebnissen in numerische Werte Anhand eines Assessment Centers zur Entrepreneur-Diagnostik

- K. Schüttpelz-Brauns,  
M.L. Beyer Die weiß-nicht-Option beim Rasch-Modell  
-Segen oder Fluch?
- B. Krause Bewertungsaspekte psychologischer Messmodelle
- Band 18**
- R. Gerlach, R. Beyer,  
R. Foerges,  
E. van der Meer,  
B. Nimke-Sliwinski Evaluation von Maßnahmen gegen Littering-Verhalten  
im öffentlichen Raum
- S. Tirre, S. Kunert Trainingsevaluation in der Praxis: Konzeption und  
Evaluation eines Wissenschaftlich gestützten Ansatzes
- K. Schüttpelz-Brauns,  
L. Beyer Kann die Güte des Summenwertes als Fähigkeits-  
schätzung im Rahmen des Rasch-Modells mit Hilfe von  
Bootstrap- Goodness of-Fit-Statistiken nachgewiesen  
werden?
- A. Voigt, E. Schröder,  
C. von Bernstorff,  
K. Elbe, J. Nachtwei Unternehmerische Eignung von Existenzgründern aus  
der Erwerbslosigkeit am Beispiel eines Assessment  
Centers des Lotsendienstes Brandenburg -  
Eine Analyse auf Aufgaben und Kompetenzebene
- S. Bedenk, W. Scholl Leitfadengestützte Experteninterviews mit Top-Managern
- T. Ludwig, J. Bar,  
B. Priemer Ein Instrument zur Erfassung von Argumentation beim  
Experimentieren in der Physik
- B. Krause Zur Replikation wissenschaftlicher Berufe  
Forschungsmethodische Aspekte zur aktuellen  
öffentlichen Diskussion

**Band 19**

- A. Freitag, S. Kunert  
A. Tiede, K. Waak Kontext-, Personen- und Prozessvariablen im  
Innovationsmanagement: Möglichkeiten  
und Grenzen der inferenzstatistischen  
Übereinstimmungsanalyse
- L. Beyer  
K. Schüttpelz-Brauns Alpha- Maximierung – das Verfahren, das  
Artefakte schafft
- M. Hochstrate Vertrauen und Wertschätzung – Messung  
verteilter Transaktionen
- M. Meißner, J. Nachtwei,  
C. von Bernstorff Big Data = Big Chance: Wie sich Datenmengen  
im HR-Bereich (effizient und sicher) nutzen lassen
- R. Gerlach, R. Beyer,  
D. Bussmann, Maßnahmen zur Sensibilisierung der Bevölkerung  
für die Stadt-Sauberkeit am Beispiel des Einsatzes

E. van der Meer von Koordinatoren („Kümmerern“) in Hamburg

B. Krause Vor- und Nachteile der Modellierung  
psychologischer Messungen

### **Band 20**

R. Gerlach,  
E. van der Meer,  
B. Nimke-Sliwinski Berlin wird Sauber! Pilostudie zur Wahrnehmung  
und Bewertung von Sauberkeit in Berlin

M. Bieniok Mentale Repräsentation von urbanen Lieblingsplätzen

P. Majunke,  
L. Beyer „Cultural Diversity“ als Instrument der  
Personalentwicklung

L. Beyer,  
K. Schüttpelz-Brauns Alpa-Maximierung – Ergebnisse aus Re- Analysen

J. Walf,  
J. Nachtwei Qualitative Inhaltsanalyse: Ein Methodenbaukasten an  
einem Forschungsbeispiel aus der Vertriebspsychologie

B. Krause Neuere Ansätze zur Beurteilung von Homogenität  
in Item- Antwort- Modellen

### **Band 21**

R. Gerlach,  
E. van der Meer,  
B. Nimke-Sliwinski Wahrnehmung von Sauberkeit im öffentlichen Raum –  
eine Replikationsstudie

M. Bieniok Lernstrategien und Motivation von Studierenden in der  
Lernumgebung Universität

K. Roters,  
J. Nachtwei,  
C. von Bernstorff Aller Anfang ist schwer – oder ? Fünf Möglichkeiten,  
Anforderungs- Analysen im Vertrieb zu verbessern

L. Beyer,  
K. Schüttpelz-Brauns Alpa-Maximierung – Ergebnisse aus Re-Analysen  
mit kleinen Stichprobenumfängen

M. Stricker,  
J. Nachtwei,  
D. Liebenow Qualität in der Kompetenzmodellierung – Entwicklung  
eines Kriterienkatalogs zur Erstellung des Benchmarks für  
Kompetenzmodellierung und – modell (BeKom)

B. Krause Zum Dauerthema Summenscore bei dichotomen Items

**Band 22**

- |  |  |
|--|--|
| R. Gerlach, R. Beyer,<br>E. van der Meer,<br>B. Nimke-Sliwinski    | Längsschnittuntersuchung zu Ursachen<br>und zur Vermeidung von Littering-<br>Verhalten im öffentlichen Raum  |
| R. Stein, F. Dittelbach,<br>R. Beyer, A. Nowarra,<br>D. Winkelmann | Pilotstudie zur Evaluation des Ansatzes<br>„Drei Statt Dry“ zur Behandlung<br>alkoholabhängiger Menschen   |
| L. Birkenbach,<br>J. Nachtwei                                      | Persönlichkeit und Profession – Der<br>Zusammenhang von Persönlichkeit und<br>Kriterien beruflicher Bewährung bei HR<br>Professionals und Business Coaches |
| M. Fahrner,<br>B. Wolf   | Analysis of Skills at Entry and Overall<br>Academic Performance in Undergraduate<br>Management Studies from 2005/2006 to<br>2015/2016                      |
| I. Großmann,<br>A. K. Grimberghe,<br>A. Hottung                    | Predicting relationship quality based on<br>personality traits:<br>A machine learning approach to improve<br>reproducibility                               |
| K. Schüttpelz-Brauns,<br>L. Beyer                                  | Hast Du nicht alle Daten in der Bank?<br>Ersetz' sie einfach! oder<br>wie zuverlässig sind Imputationsverfahren?   |
| B. Krause  | Questionable Research Practices –<br>Anmerkungen zur aktuellen Diskussion  |

**Aus der Reihe „Wissenschaftliche Veranstaltungen“ sind  
bereits erschienen:****Band 1**

- |                      |   |
|----------------------|---|
| H.J. Lander, H. Huth | Wissenschaftliche Würdigung von Hermann<br>Ebbinghaus   |
| B. Krause            | Methoden der Beobachtung und Modellierung<br>elementarer Lern- und Gedächtnisprozesse                                   |
| L. Beyer             | Einfluss struktureller Komponenten auf das Erlernen<br>logischer Funktionen   |
| U. Gauger            | Modellierung des Erwerbs logischer Funktionen: Ein<br>Vergleich von menschlichem Lernverhalten mit<br>neuronalen Netzen |
| S. Klein             | Modellierung von Veränderungen in der Zeit  |

## **Band 2**

- B. Krause 10 Jahre Zentrum für empirische Evaluationsmethoden
- H.-J. Lander, M. Huth Ebbinghaus war nicht der Erste. Ein Vorläufer Herrmann Ebbinghaus namens Francis E. Nipher
- B. Krause Herrmann Ebbinghaus auch ein Begründer der empirischen Evaluationsforschung?
- S. Klein Die Verwendung von Personen Response Curves (PCR) zur Messung des Personenfits in Latent-Trait-Modellen
- C.Kaul, R. Beyer, B. Stummvoll Anwendung eines Messwiederholungsplanes zur Prüfung spezifischer Bedingungswirkungen auf elementare kognitive Prozesse beim Sprachverstehen
- N. Graf Können Softwareentwickler von Psychologen profitieren? – Konzept der Evaluation eines Expertensystems zur Gestaltung von Assistenzsystemen
- L. Beyer Entwicklung und Evaluation eines online – Tests zur Personenselektion

## **Band 3**

- E. Sommerfeld Grußwort des Präsidiums der Leibniz-Sozietät der Wissenschaften zu Berlin
- D. Dörner Über die Notwendigkeit einer Theoretischen Psychologie
- W. Ebeling Über Helmholtz und die Interdisziplinarität
- T. Raykov The Power of latent Variable Modeling in Measurement and Behavioral Research
- G. Kaul Von Rangierproblemen zu elektromagnetischen Feldern
- S. Ebisch Eine Gleichung mit zwei Unbekannten? – Zu den Gründern des Psychologischen Instituts Carl Stumpf und Kurt Gottschaldt
- J. Blauwitz Von Affen, Tönen und Problemen – Verflechtungen des Institutes für Psychologie mit der Akademie der Wissenschaften
- B. Krause Reflexionen aus meinem Wirken für eine naturwissenschaftlich orientierte Psychologie

**Band 4**

- N. Arsenyeva                      Adaptiver Test für Computer Literacy
- R. Reckin, M. Brucks            Methodenintegration in der Human-Centred-Design-Forschung am Beispiel der Entwicklung einer Software-Oberfläche für ältere Menschen im Rahmen Des Forschungsprojektes SMILEY
- K. Mura                             Successful Industrial Training with Serious Games: Challenges of Evaluation
- E. van der Meer, R. Beyer  
B. Nimke-Sliwinski,  
H. Hagendorf, R. Gerlach,  
A. Pannekamp                    Einsatz von physiologischen Messtechniken (Blickbewegung) zur Identifikation von Strategien bei der Bewertung der Sauberkeit in Großstädten
- G. Kaul                             Hilfe, der Effekt ist weg
- B. Krause                         Methodische Anforderungen an psychologische Mess- und Testverfahren in der Mensch- Technik-Forschung

**Band 5**

- H.-J. Rothe, A.-M. Metz        Ist die Brechnung der testdiagnostischen Gütekriterien auf die Gütebestimmung von bedingungsbezogenen arbeitsanalytischen Verfahren übertragbar?
- E. Witruk                         Historische Aspekte der Pädagogischen Psychologie an der Universität Leipzig
- B. Krause                         Quo vadis ZeE? 20 Jahre Zentrum für empirische Evaluationsmethoden
- G. Ricken                         Lernverlaufsdagnostik in inklusiven Schulen
- K. Schüttpelz-Brauns          Medizinische Ausbildungsforschung-Methoden und Herausforderungen
- O. Bajewski, J. Nachtwei,  
H. Lohse-Bossenz                Was sie nicht wissen, ein Ozean? Status-quo in der Management-Diagnostik
- S. Kunert                         Notwendigkeit und Umsetzungsformen von Evaluation psychologischer Dienstleistungen

**Band 6**

- B. Krause Methodische Grundlagen der Therapieerfolgskontrolle (Zum Anliegen der Jahrestagung 2016)
- K. Schüttpelz-Brauns,  
I. Beyer Angewandte Interventionseffekanalyse im Panel - Ein Anwendungsbeispiel in der medizinischen Ausbildungsforschung
- G. Ricken Alle Lehrkräfte auf Inklusiv Schulen vorbereiten - Ein Anwendungsbeispiel in der medizinischen Ausbildungsforschung
- B. Krause Methodische Probleme und Entwicklungen in der Interventionseffektforschung
- R. Brenke Methodenwandel in der physiotherapeutischen und Saunaforschung im letzten Jahrhundert
- K.-L. Resch Physiotherapeutische Forschung: Mehr als Metaanalysen und Randomisierung
- W. Seidel Funktionelle Schmerzmedizin – Behandlung von Schmerzerkrankungen des Bewegungssystems
- W. Siems Biochemische Untersuchungen zur Interpretation thermotherapeutischer Effekte
- H. Meffert, H. Piazena Neuere Ergebnisse zu Wirkmechanismen Von Sauna und Infrarot – Hyperthermi

**Band 7**

- D. Dörner Darwin gegen Watson – Die aristotelische Seele und Psychologie als Verhaltensforschung
- W. Sommer, G. Ouyang,  
C. Zhou Methodische Entwicklungen in der EEG- Forschung
- S. Kunert Didaktische Entwicklungstrends in der Psychologie: Erweitern, Kombinieren und Bewahren
- B. Krause Entwicklungen und Probleme bei der Nutzung von Modellierungsmethoden
- E. Witruk Psychologische Methoden der Diagnostik und Der Intervention für Legastheniker



R. Beyer, R. Gerlach,  
B. Nimke-Sliwinski,  
E. van der Meer

Einsatz psychologischer und psychophysiologischer  
Methoden in der Angewandten Forschung

G. Kaul

Metaanalytische Erkenntnisse empirischer Forschungen  
zu psychischen Belastungsfaktoren der Arbeit

A.-M. Metz,  
H.-J. Rothe

Zur Methodik der Beurteilung von Gesundheits-  
gefährdungen infolge psychischer Arbeitsbelastung

K. Schüttpels-Brauns

Umgang mit nicht-seriösen Teilnehmern  
bei nichtbestehensrelevanten Tests.