Leitfaden zur Datenaufbereitung

des

Lehrstuhls Public und Non-Profit Management - Kommunale Verwaltung

Albert-Ludwigs-Universität Freiburg

Stand: 04.2019

Lehrstuhlinhaberin: Prof. Dr. Iris Saliterer

Inhaltsverzeichnis

1. Allgemeines

Dieser Leitfaden soll zur Unterstützung der **Aufbereitung und Analyse insbesondere quantitativer Daten** dienen. Die Datenaufbereitung stellt lediglich einen (Bruch-)Teil eines gesamten Forschungsprozesses dar. Letzterer soll in diesem Kapitel zur Einordnung in Kürze aufgezeigt werden. Zusätzlich sollen ausgewählte Grundlagen der empirischen Sozialforschung vermittelt werden.

1.1 Forschungsprozess

Unabhängig jeglicher Forschungsrichtung untersteht jeder Forschungsprozess zentralen Anforderungen die…

- eine klare Definition des Forschungsziels,
- eine detaillierte Beschreibung der angewandten Verfahren,
- eine sorgfältige und präzise Planung,
- die Angabe von Unvollständigkeiten und Fehler,
- den Gebrauch adäquater Analysemethoden,
- die Beschränkung auf lediglich ergebnisgerechtfertigter Schlussfolgerungen und
- die Erfahrung und Integrität der ForscherIn voraussetzen.

Idealtypisch vollzieht sich der Forschungsprozess nach den folgenden Schritten (siehe Abb. 1).

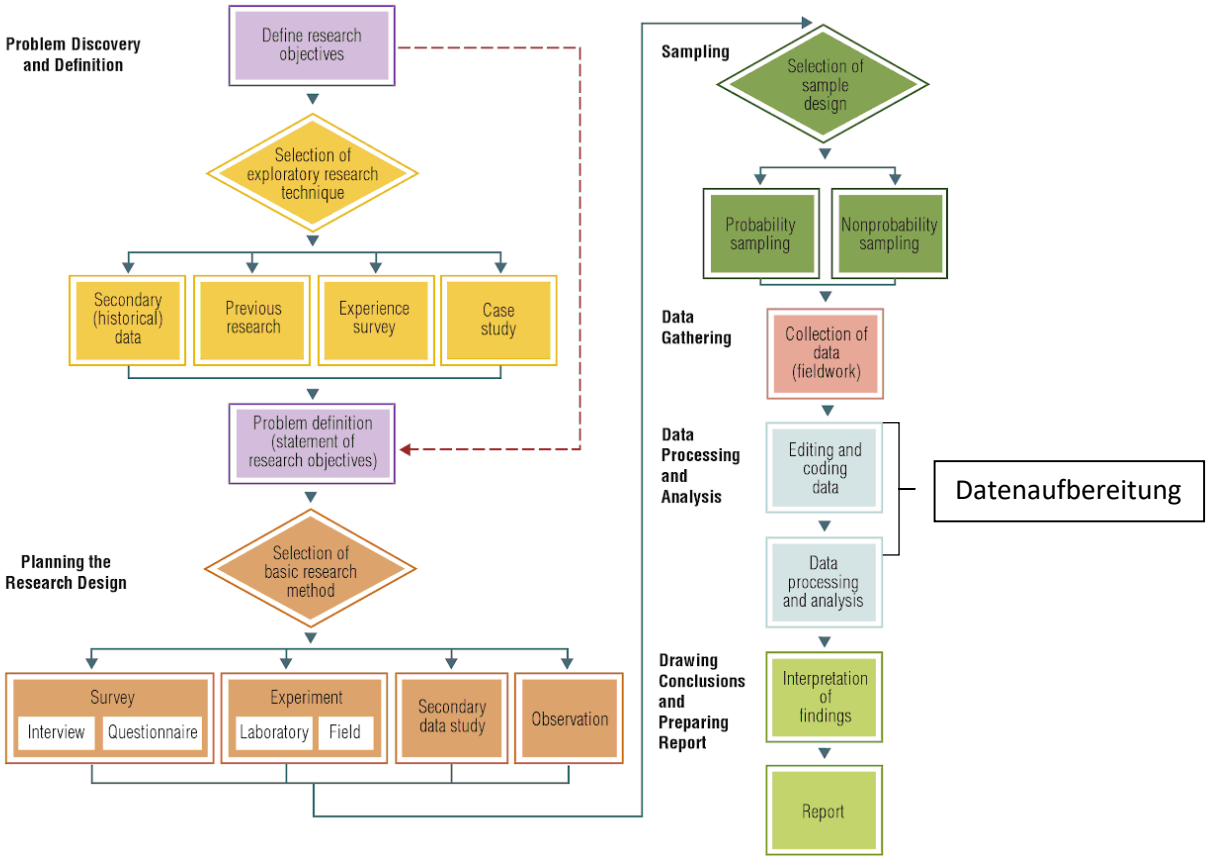


Abbildung 1: Der Forschungsprozess; Quelle: Zikmund et al. 2012, S. 57

1.2 Forschungsdesign und Stichprobendesign

Das Forschungsdesign gibt dem Vorgehen zur Lösung eines bestimmten Forschungsproblems, eine inhaltliche Struktur. Auf diese Weise sind die notwendigen Schritte vorab festgelegt. Grundsätzlich wird unterschieden zwischen: Exploratorischen Forschungsdesigns und konklusiven Forschungsdesigns (letztere können deskriptiv und/oder kausal ausgestaltet sein).

Exploratorische Forschungsdesigns dienen der Sicherstellung erster Einblicke und dem Verständnis eines Forschungsproblems. Es ist demzufolge nicht vorab definiert, welche Informationen benötigt werden. Der Forschungsprozess ist flexibel und weist einen niedrigen Strukturierungsgrad auf. Zum Einsatz kommen i.d.R. qualitative Methoden die keine repräsentativen oder konklusiven Ergebnisse hervorbringen.

Deskriptive Forschungsdesigns dienen der Beschreibung von Charakteristika, Funktionen, Sachverhalten etc. in quantitativer Art und Weise. Informationen über den Untersuchungsgegenstand sind bereits vorhanden und die Forschungsfragen konkretisieren sich in vorab formulierten Hypothesen. Der Forschungsprozess ist fixiert und weist einen hohen Strukturierungsgrad auf. Zum

Einsatz kommen i.d.R. quantitative Methoden, die repräsentative oder konklusive Ergebnisse hervorbringen.

Kausale Forschungsdesigns dienen der Untersuchung von Ursache-Wirkungs-Beziehungen. Dazu müssen experimentelle Methoden zum Einsatz kommen, bei denen eine oder mehrere unabhängige Variablen manipuliert werden, um die kausale Wirkung auf eine oder mehrere abhängige Variablen zu erklären. Während diesem „Treatment" werden idealtypisch alle anderen Variablen kontrolliert und Störvariablen eliminiert.

Zur Strukturierung des Forschungsvorhabens gehört zudem die Frage wer oder was als Untersuchungseinheit definiert ist bzw. auf welche Masse an Untersuchungseinheiten ein Forschungsziel gerichtet ist (Definition der **Grundgesamtheit**). Die Grundgesamtheit gibt gleichzeitig die Antwort, auf wen oder was die Ergebnisse der Untersuchung angewendet werden können. Werden lediglich Teilmengen der Grundgesamtheit untersucht, spricht man von einer **Stichprobe**. Stichproben können sowohl hinsichtlich ihrer Struktur beschrieben und analysiert werden (Deskriptivstatistik) und/oder entsprechend der aufgestellten Hypothesen von den Stichprobenanalysen resultierenden Ergebnissen auf die Grundgesamtheit geschlossen werden (Inferenzstatistik). Um die Stichprobenergebnisse auf die Grundgesamtheit übertragen zu können, muss die Stichprobe repräsentativ sein. Dazu können verschiedene Methoden der Stichprobenbildung (siehe Abb. 2) verwendet werden. Die Auswahl und Beschreibung der verwendeten Methode zur Stichprobenbildung konkretisiert sich im **Stichprobendesign**.

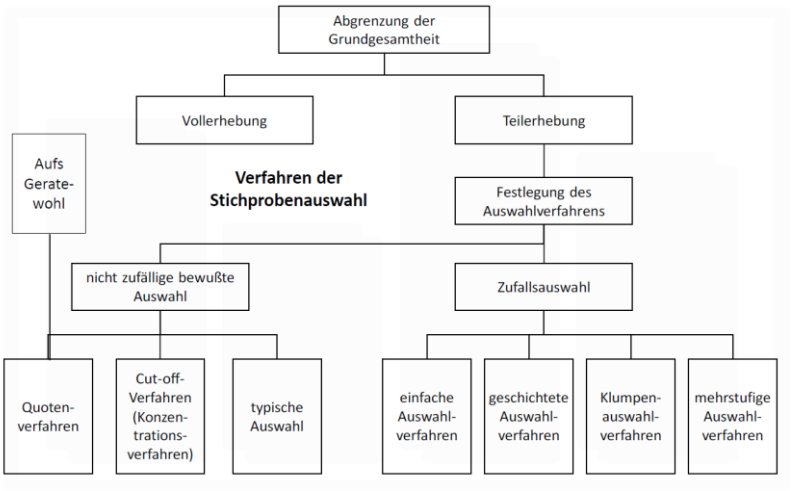


Abbildung 2: Verfahren zur Stichprobenauswahl; Quelle: in Anlehnung an Hammann/Erichson 2006, S.133

1.3 Datenbeschaffung

Das Forschungsdesign setzt deutliche Grenzen für die Art der Datenbeschaffung. Innerhalb dieser Grenzen bestehen allerdings verschieden Herangehensweisen wie die Daten erhoben werden können (siehe Abb. 3).

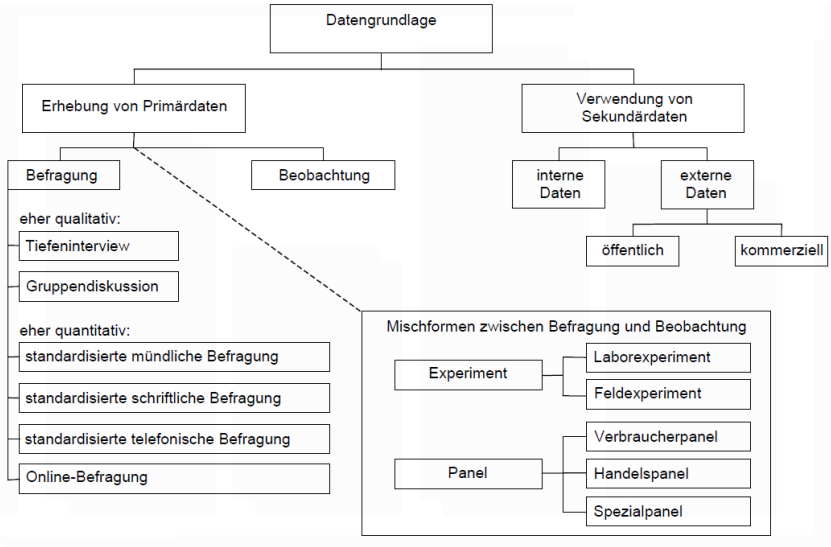


Abbildung 3: Arten der Datenbeschaffung; Quelle Homburg 2017, S.264

Im Fall der Primärerhebung kommt es im Rahmen der Datenbeschaffung zur Erhebung von Merkmalsausprägungen von interessierenden Merkmalen der Merkmalsträger in der Stichprobe. Dazu müssen geeignete Erhebungs- bzw. Messinstrumente erstellt bzw. verwendet werden.

1.4 Messung: Variablen, Skalierung, Skalenniveau, Skalierungsverfahren

Das **Erhebungsinstrument** ist das Instrument mit dem die gesuchten Daten erfasst werden sollen (z.B. ein standardisierter Fragebogen). In diesem Zusammenhang ist der Begriff der **Skalierung** von zentraler Bedeutung. Unter Skalierung versteht man die Entwicklung eines Maßstabs (/Skala) zur Messung der Merkmalsausprägungen eines Merkmals (/Variable) bei den betrachteten Merkmalsträgern (/Personen/Objekte).

Messen wird in der empirischen Sozialforschung als das systematische Aufzeichnen von empirischen Sachverhalten bezeichnet. Das Ergebnis bzw. Output einer Messung sind Daten. Das Messen von physikalischen Größen, wie bspw. Größe oder Gewicht erfolgt über festgelegte Maßeinheiten, wie Meter oder Kilogramm und stellt keine besondere Herausforderung dar. Zur Messung

(Quantifizierung) empirischer Merkmale (z.B. „Selbstwirksamkeit“) stehen allerdings keine Maßeinheiten zur Verfügung. Aus diesem Grund kommt die Skalierung zum Einsatz. Die Skalierung transformiert qualitative Merkmale in quantitative Größen (siehe Abb. 4).

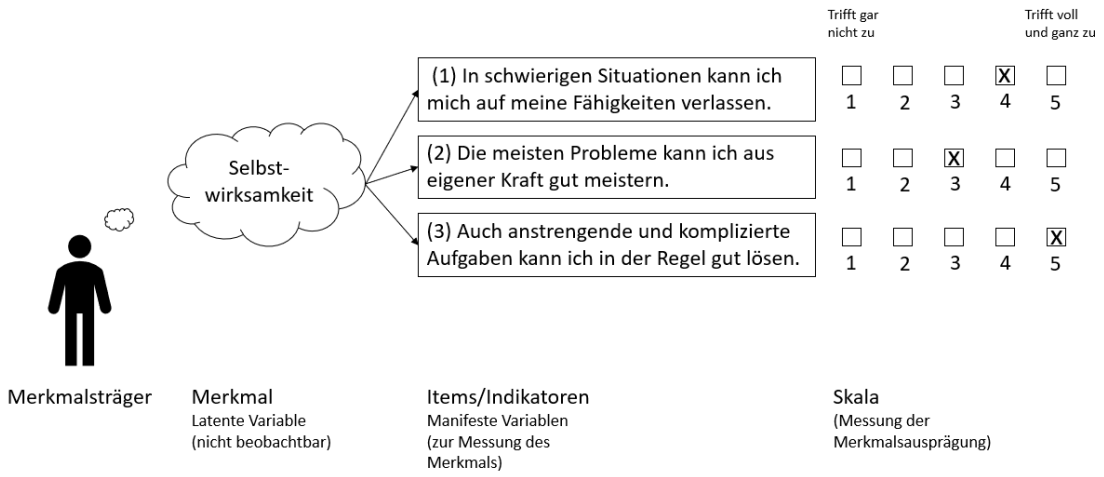


Abbildung 4: Beispiel: Messung des Merkmals „Selbstwirksamkeit“ (ASKU); Quelle Messinstrument: GESIS 2012, Eigene Darstellung

Im Beispiel in Abbildung 4 wurden zur Messung der Variable „Selbstwirksamkeit“ die Antworten des Probanden A auf den 5-Punkt Likert Skalen angekreuzt (Antworten: 4, 3, 5). Nach der vollzogenen Aufbereitung der gewonnen Daten, können die Werte zu einem einzigen Skalenwert aggregiert werden. Dieser bildet dann die gemessene Ausprägung der Selbstwirksamkeit des Probanden A ab. Für die Aggregation stehen verschiedene Methoden zur Verfügung. Beispielsweise können die einzelnen Werte summiert (= 12), ein Durchschnittswert gebildet (= 4) oder ein Faktorwert mittels Faktoranalyse berechnet werden.

Je nachdem was gemessen werden soll bzw. wie das Erhebungsinstrument gestaltet ist, unterscheidet sich das **Skalenniveau** der Daten. Das Skalenniveau bestimmt die mathematischen Eigenschaften einer Skala (sowie der gemessenen Variable) und damit den Informationsgehalt der zu erhebenden Daten. Es werden vier verschiedene Skalenniveaus unterschieden: Nominalniveau, Ordinalniveau, Intervallniveau und Rationiveau (siehe Abb. 5).

Zunahme des Informationsgehaltes		Messniveau	Mathematische Eigenschaften der Messwerte	Beschreibung der Messwerteigenschaften	Beispiele
	nicht-metrische Daten	**Nominalniveau**	$A = A \neq B$	Klassifikation: Die Messwerte zweier UEn sind identisch oder nicht identisch	Zweiklassig: Geschlecht (männlich/weiblich) Mehrklassig: Betriebstyp (Discounter/ Verbrauchermarkt/ Supermarkt)
		Ordinalniveau	$A > B > C$	Rangordnung: Messwerte lassen sich auf einer MD als kleiner/größer/ gleich einordnen	Präferenz- und Urteilsdaten: z. B. Marke X gefällt mir besser, gleich gut, weniger als Marke Y
	metrische Daten	**Intervallniveau**	$A > B > C$ und $A - B = B - C$	Rangordnung und Abstandsbestimmung: Die Abstände zwischen Messwerten sind angebbar	Intelligenzquotient Kalenderzeit
		Rationiveau (Verhältnis-Skala)	$A = x \cdot B$	Absoluter Nullpunkt: Neben Abstandsbestimmung können auch Messwertverhältnisse berechnet werden	Alter Jahresumsatz

Abbildung 5: Skalenniveaus und ihre Eigenschaften; Quelle: Berekoven et al. 2009

Nach der Messung und Datenaufbereitung ist das Skalenniveau dafür entscheidend, welche Analysen durchgeführt werden können. Im Bereich der deskriptiven Analysen können mit nominal skalierten Daten bspw. keine Mediane oder Mittelwerte berechnet werden. Ordinal skalierte Daten lassen zwar die Berechnung eines Medians zu, aber auch hier kann kein arithmetisches Mittel berechnet werden. Zentral ist hierbei, dass es sich bei nominal und ordinal skalierten Daten um nicht-metrische Daten handelt. Bei intervall- und verhältnisskalierten Daten handelt es sich um metrische Daten.

Die Unterscheidung der Art und Weise, wie durch Skalen verschiedene Daten gemessen werden variiert mit dem angewandten **Skalierungsverfahren**. Skalierungsverfahren sind formalisierte Methoden zur Konstruktion von Skalen, die als Messinstrumente dienen (siehe Abb. 6).

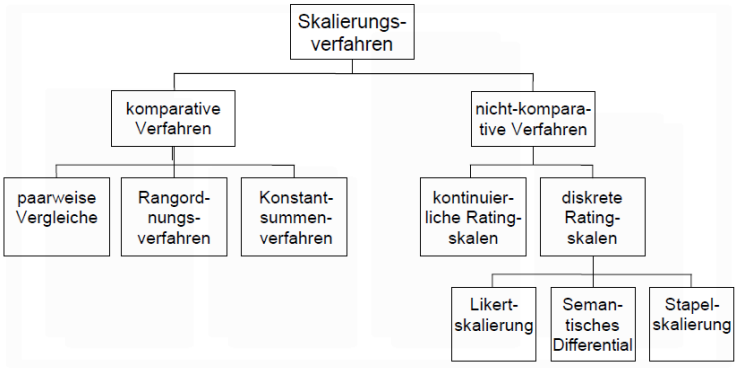


Abbildung 6: Klassifikation der Skalierungsverfahren (in Anlehnung an Malhotra 2015); Quelle: Homburg (2017) S. 313

1.5 Gütekriterien

Die Güte der durch die Messung erzeugten Daten hängt in entscheidender Weise von der Qualität des Messvorganges, insbesondere des Messinstrumentes ab. Zur Gütebeurteilung von Messvorgängen werden in der Regel drei Kriterien verwendet: **Objektivität**, **Reliabilität** und **Validität** (siehe Abb. 7).

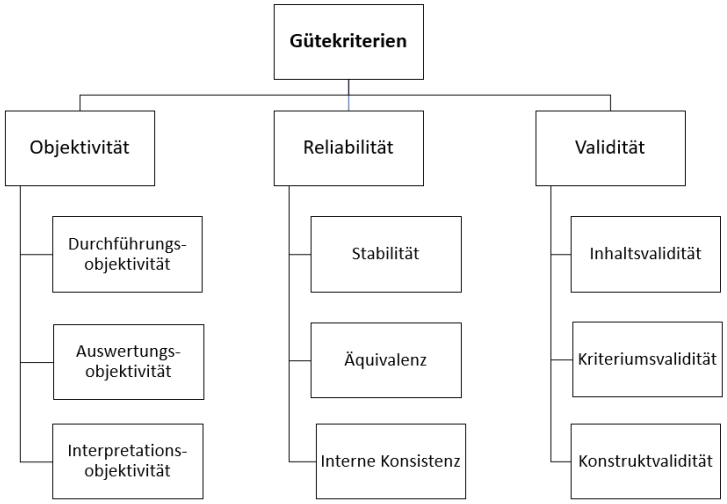


Abbildung 7: Gütekriterien (in Anlehnung an Malhotra 2009); Quelle: Eigene Darstellung

Objektivität:

Objektivität liegt vor, wenn die aus dem Messvorgang resultierenden Messergebnisse unabhängig vom Durchführenden sind. Das bedeutet, dass mehrere Personen, die unabhängig voneinander die Messergebnisse registrieren, zum gleichen Ergebnis gelangen. Es lassen sich drei Arten der Objektivität unterscheiden:

Durchführungsobjektivität ist dann gegeben, wenn das Verhalten bzw. Antwortverhalten der Versuchsperson im Rahmen der Datenerhebung nicht durch den/die ForscherIn beeinflusst wird.

Auswertungsobjektivität ist gegeben, wenn der/die ForscherIn keine oder nur sehr wenige Freiheitsgrade bei der Auswertung der Messergebnisse besitzt.

Interpretationsobjektivität liegt vor, sofern der/die ForscherIn keinen Spielraum bei der Interpretation der Messergebnisse besitzt.

Reliabilität:

Reliabilität beschreibt die Zuverlässigkeit, d.h. die formale Genauigkeit einer Messung. Sie beschreibt den Grad, zu dem das Messverfahren frei von Zufallsfehlern ist. Ein Messinstrument ist unter der Voraussetzung konstanter Messbedingungen dann reliabel, wenn die Messwerte präzise und stabil, d.

h. bei wiederholter Messung reproduzierbar sind. Der Grad der Reliabilität einer Messung lässt sich durch den Standardfehler ausdrücken, der ein Streumaß ist und damit angibt, um wie viel die Messwerte bei wiederholter Messung um einen Mittelwert liegen. Es geht also darum ob der Messvorgang stabil, äquivalent und intern konsistent sind.

Stabilität liegt vor, wenn das gleiche Messinstrument zu unterschiedlichen Einsatzzeitpunkten die gleichen Ergebnisse liefert. Stabilität kann mit der **Test-Retest-Methode** überprüft werden. Dabei wird das Messinstrument bei identischen Probanden zu unterschiedlichen Zeitpunkten verwendet und die Messwerte auf signifikante Mittelwertunterschiede (und Varianzunterschiede) untersucht. Bestehen solche Unterschiede, liegt keine Stabilität vor.

Äquivalenz liegt vor, wenn identische Probanden auf zwei verschiedenen Messinstrumenten bzw. Skalenformen zur Messung des gleichen Merkmals, hoch korrelierte (inhaltlich gleiche) Antworten geben. Äquivalenz kann mit der **Paralleltest-Methode** überprüft werden. Dabei werden zwei äquivalente Messinstrumente mit unterschiedlichen Items den gleichen Probanden zu unterschiedlichen Zeitpunkten vorgelegt und die Ergebnisse auf hohe Korrelationen untersucht. Sind keine hohen Korrelationen auszumachen, liegt keine Äquivalenz vor.

Interne Konsistenz liegt vor, wenn alle Items eines Messinstruments zusammen gut dazu geeignet sind, den gleichen Sachverhalt zu messen. Es wird betrachtet inwieweit unterschiedliche Bestandteile einer Skala konsistent in ihrer Aussage über die gemessenen Charakteristika sind. Interne Konsistenz kann mit der **Split-Half-Methode** überprüft werden. Dabei wird das Messinstrument in zwei Hälften geteilt und die Skalenwerte beider Teilskalen werden anschließend miteinander korreliert. Es wird sozusagen eine Hälfte der erhaltenen Ergebnisse mit der anderen Hälfte verifiziert. Sind keine hohen Korrelationen auszumachen, liegt keine interne Konsistenz vor. Ein geläufiges Maß für die Beurteilung der internen Konsistenz ist „**Cronbachs Alpha (α)**“. Der Wert von Cronbachs Alpha kann zwischen 0 und 1 variieren. Als Faustregel kann interne Konsistenz für Werte ≥0,7 (akzeptabel) bzw. ≥0,8 (gut) angenommen werden (Vgl. Cho und Kim 2015). Allerdings ist darauf zu achten, dass die Werte für Cronbachs Alpha mit der Anzahl der Items des Messinstruments steigen.

Validität:

Validität oder Gültigkeit eines Messverfahrens ist gegeben, sofern es gelingt, den eigentlich interessierenden Sachverhalt tatsächlich zu erfassen. Es wird also genau das gemessen, was man messen möchte (materielle Genauigkeit). Konkret kann eine Messung als valide bezeichnet werden, wenn sie zusätzlich zur Freiheit von Zufallsfehlern (Reliabilität) frei von systematischen Messfehlern ist. Es besteht also ein Zusammenhang zwischen Objektivität, Reliabilität und Validität. Objektivität ist

die Voraussetzung für Reliabilität und diese muss wiederum für die Validität eines Messinstrumentes sichergestellt sein. Zu unterscheiden sind drei verschiedene Validitätsformen:

Inhaltsvalidität liegt vor, wenn das Messinstrument bestmöglich dazu geeignet ist, die zu messende (latente) Variable zu erfassen. Inhaltsvalidität kann nicht über quantitative Maßzahlen beschrieben werden. In der Regel sind eine subjektive Plausibilitätsprüfung bzw. ein Vergleich der Definition des zu messenden Sachverhalts mit den Items des Messinstruments zu vollziehen (z.B. Sind Noten tatsächlich dazu geeignet, um die Leistung von Studierenden zu messen?).

Kriteriumsvalidität liegt vor, wenn die Ergebnisse eines Messinstruments für eine bestimmte (latente) Variable (bspw. Berufseignung) mit einem korrespondierenden Außenkriterium (beruflicher Erfolg) validiert werden können, d.h. hoch korrelieren. Man spricht in diesem Zusammenhang von einem sogenannten Korrelationsschluss.

Konstruktvalidität liegt vor, wenn eine quantitative Überprüfung zeigt, dass das Messinstrument die latente Variable (und keine andere Variable) sowie deren gesamten Charakteristika (und keine Charakteristika anderer Variablen) misst. Konstruktvalidität wird anhand der Konvergenzvalidität und der Diskriminanzvalidität überprüft. **Konvergenzvalidität** beschreibt den Grad, indem die gebündelten Items eines Messinstruments die gleiche (latente) Variable messen. Konvergenzvalidität wird in der Regel mit der Maßzahl **„Average Variance Extracted" (AVE)** gemessen, die eine Relation der durch die Variable erklärten Varianz zur Gesamtvarianz der Messung darstellt. Der Wert der AVE kann zwischen 0 und 1 variieren und Werte ≥0,5 gelten als konvergenzvalide (Vgl. Carlson und Herdman 2012). **Diskriminanzvalidität** beschreibt den Grad, indem die gebündelten Items eines Messinstruments unterschiedliche (latente) Variablen messen. Diskriminanzvalidität wird in der Regel mit dem **„Fornell-Larcker-Kriterium"** überprüft, welches durch einen Vergleich überprüfen soll, ob die AVE einer latenten Variablen höher, als jede quadrierte Korrelation mit einer anderen Variablen ist (Vgl. Fornell und Larcker 1981). Ist dies der Fall, kann von Diskriminanzvalidität ausgegangen werden.

2. Forschungsdokumentation

Ziel der Forschungsdokumentation ist es, den vollständigen Forschungsprozess, insb. den Prozess Datenerhebung, der Datenaufbereitung und -bereinigung sowie im Falle einer Publikation empirischer Ergebnisse ggf. auch die durchgeführten Analysen **für sich und andere** nachvollziehbar und rekonstruierbar zu machen. Zur Forschungsdokumentation gehören:

- Die Erstellung des Codeplans (siehe Kapitel 3.2)
- Die Erstellung eines Methodenberichts:
  - ✓ Beschreibung des vollständigen Prozesses der Datenerhebung
  - ✓ Beschreibung der Erhebungsinstrumente (inkl. Quellen verwendeter Items)
  - ✓ Beschreibung der Grundgesamtheit und Stichprobenauswahl
  - ✓ Beschreibung der Anzahl der Ausfälle aus jeweils bestimmten Ausfallgründen
  - ✓ Beschreibung der Erhebungstechnik und -dauer sowie Anfang und Ende der Feldphase
  - ✓ Beschreibung der Interviewer- und Datenkontrolle
- Dokumentation aller Datenbearbeitungsschritte und Datenmanipulationen
  - ✓ Dokumentation der Datenerfassung
  - ✓ Beschreibung der Fehlerdiagnose und Fehlerkorrektur
  - ✓ Beschreibung durchgeführter Prozeduren wie Umcodierungen oder Berechnungen
  - ✓ Beschreibung der Bildung neuer Variablen

Sämtliche Datenbearbeitungsschritte und Datenmanipulationen können in Form von kommentierter Syntax bzw. Do-Files der entsprechenden Statistiksoftware erfolgen.

Hinweis: Vermeiden Sie die Bearbeitung lediglich einer einzigen Datei, in der Sie Bearbeitungsschritte vollziehen und lediglich überspeichern. Idealerweise werden für jeden Bearbeitungsschritt neue Dateien (über „Speichern unter“) mit entsprechendem Datum erstellt und ältere Dateien auf diese Weise archiviert.

3. Vom Fragebogen zum Datensatz

Wenn Forschungsproblem, Forschungsziel, Forschungsdesign, Stichprobendesign (kurzum: alle Vorarbeiten) umfassend konkretisiert und dokumentiert wurden, kann die Datenerhebung vollzogen werden. Grundsätzlich können dazu die unterschiedlichsten Verfahren angewandt werden. Dieser Leitfaden beschränkt sich auf die Datenerhebung mittels Befragung in Form standardisierter Fragebogen. Von besonderer Wichtigkeit ist die vorausschauende Planung folgender Aktivitäten, denn bereits vor der Erstellung des Fragbogens ist zu bedenken, dass sich jeder Arbeitsschritt auf den folgenden Arbeitsschritt auswirkt. Insofern wirkt sich die Gestaltung des Fragbogens auf die Art und die Qualität der damit erhobenen Daten aus. Diese wirken sich wiederum auf die Möglichkeiten der Datenanalyse und deren Verfahren aus. Letztendlich wird die Qualität Ergebnisse der Untersuchung schon durch die Vorarbeiten determiniert.

3.1 Erstellung des Fragebogens

Bei der Konzipierung des Fragbogens ist darauf zu achten, dass der Fragenbogen bereits auf die Datenerfassung, die Datenaufbereitung und die Datenauswertung verweist. Es lohnt sich, schon bei der Fragebogengestaltung festzulegen, mit welchen Verfahren und mit welchem Programm die Daten später ausgewertet werden sollen. Die Durchführung mancher Auswertungsverfahren ist nicht möglich, wenn der Fragebogen nicht ein bestimmtes Format aufweist. Unter anderem ist zu beachten, dass bestimmte Verfahren der Datenanalyse ein bestimmtes Skalenniveau voraussetzen.

Beispiel:

- Hören Sie Radio? Antwortkategorien: Ja / Nein -> Nominalskala
- Wie oft hören Sie Radio? Antwortkategorien: Nie / Sehr selten / Selten / Oft / Sehr oft -> Ordinalskala
- Wie viele Stunden pro Tag hören Sie Radio? Antwortkategorien: 0 bis 24 Stunden -> Ratioskala

Grundsätzlich sollte hinsichtlich der Analyseverfahren zusätzlich darauf geachtet werden, bei gleichen Skalenniveaus auch gleiche Antwortkategorien (Anzahl und Wording) zu verwenden bzw. eine starke Variation zu vermeiden.

Üblicherweise sollte der erstellte Fragebogen einem Pre-Test unterzogen werden. Dabei wird insbesondere beurteilt, wie verständlich der Fragebogen ist, inwieweit die Befragten über ausreichende Informationen verfügen, um die Fragen zu beantworten, inwieweit die Antwortkategorien bei geschlossenen Fragen alle relevanten Aspekte umfassen und wie viel Zeit die vollständige Beantwortung des Fragebogens tatsächlich in Anspruch nimmt.

3.2 Erstellung des Codeplans

Die Erstellung des Codeplans sollte bei standardisierten Befragungsmethoden und späterer quantitativer Analyseverfahren bereits vor der Datenerhebung erfolgen. Der Codeplan schreibt in expliziter Weise vor, wie und welche Zahlenwerte den Antwortkategorien im Fragenbogen zugeordnet werden. Der Codeplan bietet dadurch ein unerlässliches Instrument der Dokumentation und Interpretation. Er richtet sich an zwei Zielgruppen: Personen, die die Daten in den Datensatz eingeben, entnehmen dem Codeplan, wie sie bestimmte Angaben in Zahlen umsetzen sollen. Personen, die die Daten auswerten, entnehmen dem Codeplan, wie bestimmte Zahlen im Datensatz zu interpretieren sind.

Im Codeplan enthalten sind folgende Aspekte:

1) Zuordnung und Bezeichnung der Variablen zu den Fragen (Variable Labels)
2) Zuordnung von Zahlen zu den Antwortkategorien (Value Labels)
3) Umgang und Bezeichnung von fehlenden Antworten (Missing Values)

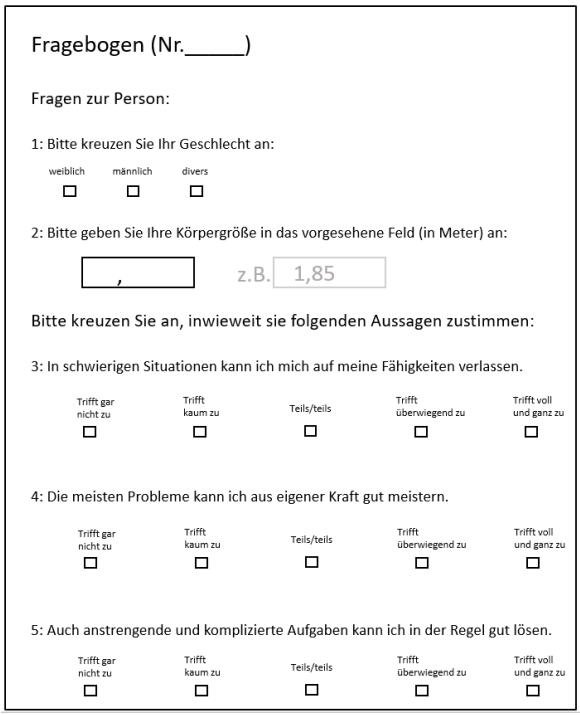
Hinweise:

Zu 1): Benennen Sie die Variablen kurz, aussagekräftig und mit Bezug zum Fragebogen (z.B. für Frage 1 im Fragebogen, Merkmal „Geschlecht“ →„Var_01_Sex“);

Zu 2): Verwenden Sie möglichst kleine Zahlenwerte für kategoriale Größen (z.B. Merkmal Geschlecht „weiblich / männlich / divers“ → „0 / 1 / 2“; Vergeben Sie hohe Werte (z.B. „5“ auf 5-Punkt Likert Skala) für hohe Antwortoptionen (z.B. starke Zustimmung) und achten Sie auf die Interpretierbarkeit; Bei stetigen Größen muss auf Einheitlichkeit bezüglich der Dezimalstellen (Trennung durch Komma oder Punkt) geachtet werden;

Zu 3): Fehlende Werte sollten Einheitlich codiert werden (z.B. „9999“ oder „.“ etc.)

Im Folgenden wird ein Fragebogen mit dem entsprechendem Codeplan dargestellt (siehe Abb. 8).



Fragebogen (Nr._____)

Fragen zur Person:

1: Bitte kreuzen Sie Ihr Geschlecht an:

weiblich ☐ männlich ☐ divers ☐

2: Bitte geben Sie Ihre Körpergröße in das vorgesehene Feld (in Meter) an:

_____, z.B. 1,85

Bitte kreuzen Sie an, inwieweit sie folgenden Aussagen zustimmen:

3: In schwierigen Situationen kann ich mich auf meine Fähigkeiten verlassen.

Trifft gar nicht zu ☐ | Trifft kaum zu ☐ | Teils/teils ☐ | Trifft überwiegend zu ☐ | Trifft voll und ganz zu ☐

4: Die meisten Probleme kann ich aus eigener Kraft gut meistern.

Trifft gar nicht zu ☐ | Trifft kaum zu ☐ | Teils/teils ☐ | Trifft überwiegend zu ☐ | Trifft voll und ganz zu ☐

5: Auch anstrengende und komplizierte Aufgaben kann ich in der Regel gut lösen.

Trifft gar nicht zu ☐ | Trifft kaum zu ☐ | Teils/teils ☐ | Trifft überwiegend zu ☐ | Trifft voll und ganz zu ☐

Frage im Fragebogen	Variable (Merkmal)	Variable Label	Value Labels
0	Person (Nr.)	ID	Laufende Nr. (ab 1)
1	Geschlecht	Var_01_Sex	weiblich = 0 männlich = 1 divers = 2 missing value = 999
2	Körpergröße	Var_02_Kgr	stetiger Wert in Meter (mit Komma) missing value = 999
3	Selbstwirksamkeit 1	Var_03_Eff_1	Trifft gar nicht zu = 1 Trifft kaum zu = 2 Teils/teils = 3 Trifft überwiegend zu = 4 Trifft voll und ganz zu = 5 missing value = 999
4	Selbstwirksamkeit 2	Var_04_Eff_2	Trifft gar nicht zu = 1 Trifft kaum zu = 2 Teils/teils = 3 Trifft überwiegend zu = 4 Trifft voll und ganz zu = 5 missing value = 999
5	Selbstwirksamkeit 3	Var_05_Eff_3	Trifft gar nicht zu = 1 Trifft kaum zu = 2 Teils/teils = 3 Trifft überwiegend zu = 4 Trifft voll und ganz zu = 5 missing value = 999

Abbildung 8: Beispielfragebogen und -codeplan; Quelle: Eigene Darstellung

<u>Hinweis:</u> Es kann für die folgenden Schritte der Datenerfassung, Datenbereinigung und Datenanalyse hilfreich sein, einen Blanko-Fragebogen (rein zu Dokumentations- und Übersichtszwecken) so zu beschriften, dass zu jeder Frage die entsprechenden Variablen Labels und Value Labels geschrieben werden. Auf diese Weise entsteht eine Art Zusammenführung von Fragebogen und Codeplan, die für schnelles Nachschlagen und Abgleichen (insb. bei der Datenerfassung) sehr gut geeignet sein kann.

<u>Kleiner Exkurs (qualitative Daten):</u>

Werden keine standardisierten Fragebogen mit direkt quantifizierbaren Antwortkategorien verwendet wie es bspw. bei Tiefeninterviews oder Experteninterviews der Fall ist, müssen die „freien Antworten“ zunächst transkribiert werden. Das Transkribieren beschreibt den Prozess der ganzheitlichen Niederschrift des Gesprochenen, d.h. die Verschriftlichung nach einem vorab bestimmten Regelwerk. Infolgedessen liegen qualitative Daten vor, wie sie ebenfalls bei Beobachtungen oder offenen Fragen im Allgemeinen entstehen können. Für eine umfassende Analyse sollte die Codierung auch bei qualitativen Daten vollzogen werden. Hierbei erfolgt die Zuordnung nicht länger quasi-automatisch, sondern erfordert die Interpretation durch die codierende Person. Der Prozess der Codierung vollzieht sich hier als ein mehrstufiger Prozess:

1. Grobe Überprüfung der Daten, auf sich wiederholende Regelmäßigkeiten. Diese Regelmäßigkeiten können daraufhin die Basis für ein erstes Kategoriensystem bilden;

2. Zuordnung der einzelnen Aussagen zu den entdeckten Kategorien;
3. Beurteilung der Zuordnung anhand der Zielkriterien „interne Homogenität“ (die Daten innerhalb der Kategorien sollen sich möglichst ähneln) und „externe Heterogenität“ (die Kategorien sollen sich möglichst gut unterscheiden);
4. Sind diese Kriterien nicht erfüllt oder lassen sich viele Daten nicht oder nicht eindeutig zuordnen, sollte die Kategorisierung überarbeitet werden.

So gefundene Kategorisierungen sollten die folgenden Kriterien erfüllen:

- Die Kategorien sollten in sich konsistent sein und alle wesentlichen Facetten des untersuchten Phänomens abdecken;
- Die vorhandenen Daten sollten alle in das Kategoriensystem einzuordnen sein;
- Das Kategoriensystem sollte durch andere unabhängige Personen reproduziert werden können.

Hinweis: Die hier beschriebenen Ausführungen zur Codierung qualitativer Daten stellen lediglich einen kurzen Überblick und keine abschließende Verfahrensbeschreibung dar. Werden qualitative Daten aufbereitet sollte zusätzliche Literatur zu Rate gezogen werden.

3.3 Datenerfassung

Die Datenerfassung beschreibt den tatsächlichen Schritt der Übertragung der ausgefüllten Fragenbögen in ein digitales Verarbeitungsprogramm wie Excel, i.d.R. jedoch in ein Statistikprogramm wie SPSS, Stata oder R. Grundsätzlich sind hierbei zwei Verfahrensweisen aufgrund der vorangegangen Erhebungsmethodik (analog/computergestützt) zu unterscheiden. Wurden analoge Fragebogen verwendet (paper & pencil survey), sind die Daten manuell zu digitalisieren (in SPSS einzugeben). Wurden computergestützte Verfahren wie bspw. Online-Fragebogen, CATI, CAPI verwendet, wird der Arbeitsschritt der Datenerfassung zwar nicht manuell vollzogen, entfällt jedoch keineswegs, da auch hierbei die Daten in einem passenden Format von dem Statistikprogramm erfasst werden müssen. Die folgenden Ausführungen orientieren sich an der Verwendung des Statistikprogramms SPSS. Das grundsätzliche Vorgehen ist jedoch auf andere Programme übertragbar.

Fall A): Analoger Fragebogen

Hinweis: Spätestens zu diesem Zeitpunkt sollten Sie alle ausgefüllten Fragenbogen mit einer fortlaufenden Zahl (sog. ID-Nr. (Identifikationsnummer)) nummerieren. Damit gewährleisten Sie, dass Sie bei Auffälligkeiten im Datensatz oder fehlenden Werten jederzeit im Originalfragebogen nachsehen können.

Bei der Datenerfassung muss grundsätzlich zwischen der Variablenansicht (Variablenbeschreibung) und der Datenansicht (Datensatz) unterschieden werden. Es ist stets auf eine strikte inhaltliche und gedankliche Trennung dieser beiden Ansichten zu achten. Die Variablenansicht erfasst und beschreibt alle Variablen, die in Form von Fragen im Fragebogen erfasst werden. Die Datenansicht erfasst die dazugehörigen Angaben der Probanden zu den Variablen des Fragebogens. Personenidentifizierende Daten (insb. Namen, Initialen, Adressen) dürfen auf keinen Fall in der Datei erscheinen.

In einem **ersten Schritt** sind alle im Fragebogen auftauchenden Variablen in der Variablenansicht zu erfassen. Hierbei kommt der zuvor erstellte Codeplan zum Einsatz. Im Prinzip ist der Codeplan lediglich in das Eingabefenster der Variablenansicht manuell zu übertragen. Als erste Variable sollte stets die ID-Nr. definiert werden. In SPSS sind folgende Voreinstellungen zur Beschreibung jeder Variable zu bearbeiten:

- **Name:** Kürzel der Variable (Variable Label) (entsprechend Codeplan).
- **Typ:** Datentyp der Variable (≠ Skalenniveau). Folgende Auswahl kann getroffenen werden:

  Numerisch: eine Variable, deren Inhalt Zahlen sind (z. B. 33). Es werden Zahlen in das Datenfile eingegeben, deshalb ist auch die Voreinstellung numerisch.

  Komma: eine (numerische) Variable, deren Werte mit Komma, als Tausender-Trennzeichen und Punkt als Dezimaltrennzeichen angezeigt werden (z. B. „2,111.48“)

  Punkt: eine (numerische) Variable, deren Werte mit Punkt als Tausender-Trennzeichen und Komma als Dezimaltrennzeichen angezeigt werden (z. B. „27.000,11“)

  Wissenschaftliche Notation: eine (numerische) Variable, z. B. deren Werte mit einem E und einer Zehnerpotenz mit Vorzeichen angezeigt werden (z. B. „1,03E + 003“)

  Datum: eine (numerische) Variable, deren Werte in einem Datums- oder Uhrzeitformat angezeigt werden (z. B. „29-10-1911“, Format lässt sich aus einer Liste auswählen)

  Dollar: eine (numerische) Variable mit einem führenden Dollarzeichen ($), deren Werte mit einem Komma als Tausender-Trennzeichen und einem Punkt als Dezimaltrennzeichen angezeigt werden

  Spezielle Währung: eine (numerische) Variable, deren Werte in einem wählbaren Währungsformat angezeigt werden

  String: eine alphanumerische Variable, deren Inhalt Buchstaben und Zahlen sind. Die Länge des Strings wird bei der Variablendefinition festgelegt (lässt sich aber auch später noch ändern). Groß- und Kleinbuchstaben werden als unterschiedliche Zeichen gewertet.

  Eingeschränkt numerisch: ganze Zahl mit führenden Nullen.
- **Spaltenformat mit den Dezimalstellen:** In der Regel kann hier die Voreinstellung mit acht Zeichen für das Spaltenformat und zwei für die Dezimalstellen übernommen werden.
- **Variablenlabels:** Nähere inhaltliche Beschreibung der Variable (entsprechend Codeplan).
- **Wertelabels:** Zuordnung von Codes zu den Ausprägungen der Variable (Value Labels) (entsprechend Codeplan).
- **Fehlende Werte:** Zuordnung der Codes zu fehlenden Werten der Variable (entsprechend Codeplan).

- **Spalten:** Breite der entsprechenden Spalte in der Datenansicht der Datei. In der Regel kann hier die Voreinstellung übernommen werden.
- **Ausrichtung:** Ausrichtung der Werte (rechtsbündig / linksbündig). In der Regel kann hier die Voreinstellung übernommen werden.
- **Messniveau:** Skalenniveau der Variable. SPSS unterscheidet allerding nur zwischen nominalen, ordinalen und metrischen Skalenniveaus (Messniveaus). Intervall- bzw. ratioskalierte Variablen werden als metrisch bezeichnet und nicht differenziert.
- **Rolle:** Hier können bestimmte Variablenrollen festgelegt werden. In der Regel kann hier die Voreinstellung übernommen werden.

Sind alle Variablen entsprechend des Codeplans in der Variablenansicht definiert, erfolgt in einem **zweiten Schritt** der Wechsel in die Datenansicht zur Eingabe der Daten. Eine Zeile entspricht einer Person. Eine Spalte entspricht einer Variablen. Es können also für jede einzelne Person die Daten in einer Zeile abgelesen werden oder pro Spalte die Ausprägungen einer Variablen.

Fall B): Digitaler Fragebogen

Wurde ein computergestütztes Verfahren zur Datenerhebung verwendet, werden die Daten durch die Übernahme des Ausgabedatensatzes der Erhebung in den Zieldatensatz (SPSS) überführt. Dabei vollzieht sich der **Import einer „.sav"-Datei** (SPSS Format) relativ problemlos. Dennoch sind jegliche o.g. Schritte zur Beschreibung der einzelnen Variablen durchzuführen bzw. zu überprüfen.

Der **Import einer Excel-Datei** setzt einige Vorarbeit voraus, denn die vorhandene Excel-Datei muss zuerst für die Datenimport aufbereitet werden. Dafür sind folgende Punkte zu kontrollieren:

- Die Exceldatei enthält keine Formeln, Beschreibungen, leere Zeilen / Spalten, Grafiken etc.
- Sämtliche personenidentifizierenden Daten müssen entfernt werden
- Die erste Zeile des Datenblatts sollte die Variablenbezeichnungen (Spaltenbezeichnungen) enthalten (keine Sonderzeichen)
- Ab der Zweiten Zeile beginnt der Datensatz mit dem ersten Untersuchungsobjekt (Proband)
- Fehlende Werte entsprechen leeren Zellen und leere Zellen entsprechen fehlenden Werten oder werden bereits codiert (z.B. 999)

Grundsätzlich bietet SPSS einen Import-Assistenten (insb. für weitere Dateiformate), weshalb auf die technischen Aspekte des Datenimports hier nicht näher eingegangen wird. Dennoch ist zu bedenken, dass eine hohe Qualität des Datensatzes vor dem Datenimport anfallende Nacharbeiten nach dem Import erheblich reduzieren kann.

3.4 Nachkontrolle der Daten

Wurden alle Daten in das Zielformat (hier SPSS) übertragen bzw. eingegeben, lohnt es sich das nun entstandene Datenfile einer kurzen Überprüfung zu unterziehen. Dabei sollten Sie sich folgende Fragen stellen und mittels Betrachtung des Datenfiles beantworten können:

1) Sind alle im Fragebogen vorkommenden Variablen (Fragen) vollständig erfasst?
2) Sind alle Fragebogen (Personen) erfasst?
3) Sind alle Werte der einzelnen Fragebogen erfasst?
4) Sind alle fehlenden Werte entsprechend der verwendeten Codierung vollständig erfasst?

Hinweis: Bei Unstimmigkeiten ist ein Blick in den ausgefüllten Originalfragebogen unausweichlich.

Sofern die Überprüfung abschlossen ist, ist der Rohdatensatz fertiggestellt. Es empfiehlt sich den Rohdatensatz in Form einer Sicherheitskopie und als solchen gekennzeichnet abzuspeichern. Dadurch kann beim nächsten Schritt (der Datenaufbereitung) jederzeit auf den Rohdatensatz zurückgegriffen werden.

4. Datenaufbereitung

Liegt das vollständig erfasste und überprüfte Datenfile vor, können damit noch keine statistischen Analysen durchgeführt werden. Dafür müssen die Daten zunächst einer Aufbereitung unterzogen werden. Dieser Arbeitsschritt ist von der vorangegangenen Datenkontrolle, die vornehmlich die Vollständigkeit überprüft, zu unterscheiden. Selbst wenn alle Daten vollständig vorliegen, kann der Rohdatensatz dennoch Fehler beinhalten (z.B. falsch ausgefüllte Fragebogen) oder die Variablen liegen nicht in der zur Analyse benötigten Form vor. Bei der Datenaufbereitung werden statistische Methoden herangezogen, mit denen die Plausibilität der erhobenen Daten überprüft und potentielle Fehler sichtbar gemacht werden können. Konnten Fehler aufgedeckt werden ist eine Fehlerdiagnose zu erstellen und eine mögliche Fehlerkorrektur zu überprüfen.

Hinweis: Die Datenaufbereitung sollte keinesfalls übergangen werden, denn sie ist für die Qualität der Ergebnisse von großer Bedeutung. Es mag sein, dass in vier von fünf Erhebungen keine nennenswerten Fehler im Datensatz zu entdecken sind. Doch angesichts der Tatsachen, dass immer häufiger "fremde" Daten re-analysiert werden und dass bei "eigenen" Studien die Erhebung oft an Dritte delegiert wird, sollte es selbstverständlich sein, dass man sich der Qualität der Daten versichert.

4.1 Plausibilitätstest

Plausibilitätsüberlegungen mögen methodisch zwar wenig anspruchsvoll erscheinen, nichtsdestotrotz setzten sie umfassende Kenntnisse bezüglich des gesamten Forschungsprozesses voraus. Es müssen logische Schlussfolgerungen getroffen werden, indem gleichzeitig das beschriebene Skalenniveau und die möglichen Antworten stets mitgedacht werden.

Hinweis: Eine Ausgabe und Analyse der deskriptiven Statistiken entspricht diesem Vorgehen größtenteils, dennoch ist eine systematische Überprüfung, entsprechend der dargebotenen Reihenfolge vorzuziehen.

1. **Wertelabels überprüfen**

   SPSS bietet die Funktion, in der Datenansicht zwischen den Variablenwerten (z.B. 0, 1, 2) und den Wertelabels (z.B. weiblich, männlich, divers) zu wechseln. Dadurch können mögliche Codierungs- und Eingabefehler relativ schnell aufgedeckt werden. Ist beispielsweise zwischen vielen „weiblich“, „männlich“ oder „divers“ eine Zahl zu erkennen, liegt ein Fehler vor, denn ein Wert ohne Wertelabel ist häufig ein nicht zulässiger Wert.

2. **Berechnen und Sichten von Streuungsmaßen**

   Eine einfache Methode zur Überprüfung der Daten ist die Analyse von Streuungsmaßen wie der Varianz, der Standardabweichung, dem Quartilsabstand der Spannweite und der Perzentilwerte.

Ein Vergleich der Verteilung der Variablenwerte mit den möglichen und logischen Antwortkategorien über Streuungsmaße ist insb. zur Erkennung von Ausreißern geeignet. Beispielsweise kann die Variable „Alter“ über die Spannweite relativ einfach auf Unstimmigkeiten (z.B. von 2 bis 911) überprüft werden.

3. Berechnen und Sichten von Häufigkeitsverteilungen

In einem zweiten Schritt können über Häufigkeitsverteilungen über alle Variablen weitere Unstimmigkeiten aufdeckt werden. Zur Überprüfung der Daten müssen die Bedeutungen der Variablen (bzw. dessen mögliche Ausprägungen) und die entsprechenden Häufigkeiten verglichen werden. Beispielsweise sollte bei der ID-Nr. in der Spalte für absolute Häufigkeiten eine ,,2" oder ,,3" nicht vorkommen. Ferner dürfen bei Variablen mit kategorialen Ausprägungen (bspw. 1, 2, 3) keine Häufigkeiten für andere Zahlen bestehen. Bei Variablen mit vielen Ausprägungen (bspw. Alter) macht es Sinn sich auf Häufigkeiten hoher und niedriger Werte (Extremwerte) zu konzentrieren. Auf diese Weise können unplausible Werte (z.B. Alter „5“ oder „234“) entdeckt werden. Oft sind Informationen auch nicht nachweislich falsch, sondern nur unwahrscheinlich (z. B. eine zweistellige Anzahl Kinder im Haushalt, ein extrem hohes Einkommen oder ein extrem hohes Alter). Ansonsten sind generell hohe Anteile fehlender Werte von Interesse.

4. Berechnen und Vergleichen von Häufigkeitsverteilungen

Das Abgleichen zweier Häufigkeitstabellen von Variablen, die miteinander in Beziehung stehen, kann einen ersten Hinweis geben, ob Daten plausibel sind. Wenn z. B. die Anzahl derjenigen, die angeben, ein eigenes Gehalt zu beziehen, wesentlich von der Anzahl derer abweicht, die angeben, erwerbstätig zu sein, dann sind vermutlich falsche Angaben gemacht worden. Der Vergleich von Häufigkeitsverteilungen sollte allerdings mit interpretatorischer Vorsicht durchgeführt werden.

5. Berechnen und Sichten von Kreuztabellen

Über Kreuztabellen kann der Datensatz auf Fehler überprüft werden, welche erst bei der Betrachtung mehrerer Variablen gleichzeitig aufgedeckt werden können. Dabei empfiehlt es sich Variablen die logisch miteinander in Beziehung stehen, gegeneinander aufzutragen. Als Beispiel könnten die Variablen „vorhandene Kinder“ und „Alter der Kinder“ genannt werden. Bei Befragten, die angegeben haben kinderlos zu sein, sollte das entsprechende Feld in der Kreuztabelle „Alter der Kinder“ keine Fälle enthalten. Nicht nur fehlerhafte Filterführungen, sondern auch inhaltliche Auffälligkeiten können so identifiziert werden, etwa wenn die angegebene Größe der Wohnung nicht mit deren Miethöhe korrespondiert.

4.2 Fehlerdiagnose und Fehlerkorrektur

Die zuvor beschriebenen Plausibilitätstests dienen dazu, mögliche Fehler zu identifizieren. Konnten Fehler identifiziert werden, stellen sich zwei wichtige Fragen:

1. Worin besteht der Fehler und wie kommt er zustande? → Fehlerdiagnose
2. Wie ist mit dem Fehler umzugehen? → Fehlerkorrektur

Die **Fehlerdiagnose** lässt sich kaum durch einen formalen Prozess beschreiben. Sind Fehler identifiziert sollte als erster Schritt immer ein Abgleich mit dem Originalfragebogen erfolgen. Letztendlich ist für die Fehlerdiagnose und die darauffolgende **Fehlerkorrektur** eine Einzelfallbetrachtung durchzuführen. Grundsätzlich stehen bei der Einzelfallbetrachtung drei mögliche Optionen zur Auswahl:

a) Wert nicht löschen:
Wenn nicht zweifelsfrei nachgewiesen kann, dass ein Wert fehlerhaft ist darf er nicht gelöscht werden. Trotzdem sollte auf den vermuteten Fehler z. B. in Form einer Flag-Variable hingewiesen werden. Flag-Variablen erhalten bei Vorliegen einer Auffälligkeit den Wert 1 und anderenfalls den Wert 0.

b) Wert löschen:
Wenn zweifelsfrei nachgewiesen werden kann, dass ein Wert fehlerhaft ist aber der wahre Wert ungeklärt bleibt.

c) Wert ändern:
Wenn zweifelsfrei nachgewiesen werden kann, dass ein Wert fehlerhaft ist und der wahre Wert ohne Zweifel geklärt werden kann.

<u>Hinweis:</u> Die Fehlerdiagnose und insb. jede Fehlerkorrektur muss ausnahmslos dokumentiert werden!

Bei der Einzelfallbetrachtung können Überlegungen hinsichtlich Fehlerarten und potenzieller Fehlerquellen hilfreich sein. **Fehlerarten** können als an den Daten sichtbare **Symptome** beschrieben werden. Diese Symptome sollten zuerst überprüft werden. **Fehlerquellen** beschreiben die **Ursachen** dieser Symptome.

Fehlerart	Beispiel	Prüfung	Korrekturmöglichkeit
Wert außerhalb des gültigen Bereichs	Geschlecht = 3	Häufigkeitsverteilung sichten	Vergleich mit Originalfragebogen oder Nacherhebung, sonst: Missing Value
Wert außerhalb des realistischen Bereichs	Anzahl Zimmer = 14	Häufigkeitsverteilung sichten	Vergleich mit Originalfragebogen oder Nacherhebung, sonst: Kennzeichnung des Falles mit einer Flag-Variablen
Ungültig fehlender Wert	Geburtsjahr = (leeres Feld)	Häufigkeitsverteilung sichten	Vergleich mit Originalfragebogen oder Nacherhebung, sonst: Missing Value
Inkonsistenz innerhalb des Fragebogens	pers. Einkommen = „2500 €“ HH. Einkommen = „1800 €“	Kreuztabelle sichten	Vergleich mit Originalfragebogen oder Nacherhebung, sonst: Kennzeichnung des Falles mit einer Flag-Variablen
Ungültiger Wert trotz Filterführung (Spezialfall einer Inkonsistenz)	Filterfrage = „kinderlos“ Alter des 1. Kindes = „9“	Kreuztabelle sichten	Variable über alle Fälle recodieren, sonst: Missing
Widerspruch zu bestimmten anderen Fällen (z.B. bei Befragung von Paaren)	Mann: „Frau nimmt Pille.“ Frau: „Wir verhüten nicht.“	Fälle matchen und Kreuztabellen sichten	Kennzeichnung des Falles mit einer Flag-Variablen
Widerspruch gegenüber der Gesamtheit oder Informationen außerhalb der Datenerhebung	Pers. X: Kindergeld = „60 €“ andere Pers./ Recherche: „184 €“	Häufigkeitsverteilung sichten	falls möglich: Korrektur anhand anderer Informationen, sonst: Missing

Abbildung 9: Fehlerdiagnose und -korrektur nach Fehlerarten; Quelle: Lück und Landrock S.407

Fehlerquelle	Nachweis/Prüfung	Korrekturmöglichkeit
Fehler im Erhebungsinstrument	Vergleich mit Originalfragebogen	Falls möglich: Variable über alle Fälle recodieren, sonst: Vergleich mit Originalfragebogen, sonst: Missing
Reaktivität des Befragten	Geringe Abweichung in typische Richtung, evtl. im Zusammenhang zu verstehen (Kreuztabelle sichten)	Keine Korrekturmöglichkeit, da im Einzelfall nicht zu identifizieren
Irrtum durch Befragten	Evtl. im Zusammenhang zu verstehen (Kreuztabelle sichten)	Falls möglich: Korrektur anhand der anderer Informationen, sonst: Missing
Interpretation des Fragebogens	Vergleich mit Originalfragebogen, Häufung von Fehlern bei einem Interviewer (Kreuztabelle mit Interviewer-Nr.), Rücksprache mit Interviewer	Falls möglich: Variable über alle Fälle recodieren, sonst: Vergleich mit Originalfragebogen, sonst: Missing
Kommunikation zw. Interviewer und Befragtem	Rücksprache mit Interviewer	Falls möglich: Korrektur anhand der mündlichen Aussage, sonst: Missing
Bewusst falsche Auskunft durch den Befragten	Häufung von eindeutigen Fehlern bei einem einzelnen Fall, Rücksprache mit Interviewer	Missing
Fälschung von Interviews durch den Interviewer	Häufung von Fehlern bei einem Interviewer (Kreuztabelle mit Interviewer-Nr.), Rücksprache mit Interviewer	Missing
Technischer Fehler bei der Datenerfassung	Vergleich mit Originalfragebogen	Erneute Datenerfassung (Einlesen)
Tippfehler bei der Datenerfassung	Einzelner, typischer Fehler (z.B. Zahlendreher), Vergleich mit Originalfragebogen	Korrektur anhand des Originalfragebogens

Abbildung 10: Fehlerdiagnose und -korrektur nach Fehlerquellen; Quelle: Lück und Baur S. 75

6. Transformationen

Bei der Transformation geht darum, Informationen, die im Datenerhebungsprozess gewonnen wurden, entweder zu verändern, zu erweitern oder zu verdichten. Häufig sind Variablen nicht in der gewünschten Form, sodass erst durch Berechnungen (bspw. durch Summation mehrerer Variablen) die benötigten Testwerte entstehen. Zusätzlich besteht in bestimmten Fällen die Notwendigkeit die Ausprägungen einer oder mehrerer Variablen neu zu definieren (bspw. Altersausprägungen in Altersklassen umcodieren). Grundsätzlich kommen dabei zwei Hauptanwendungen in Betracht:

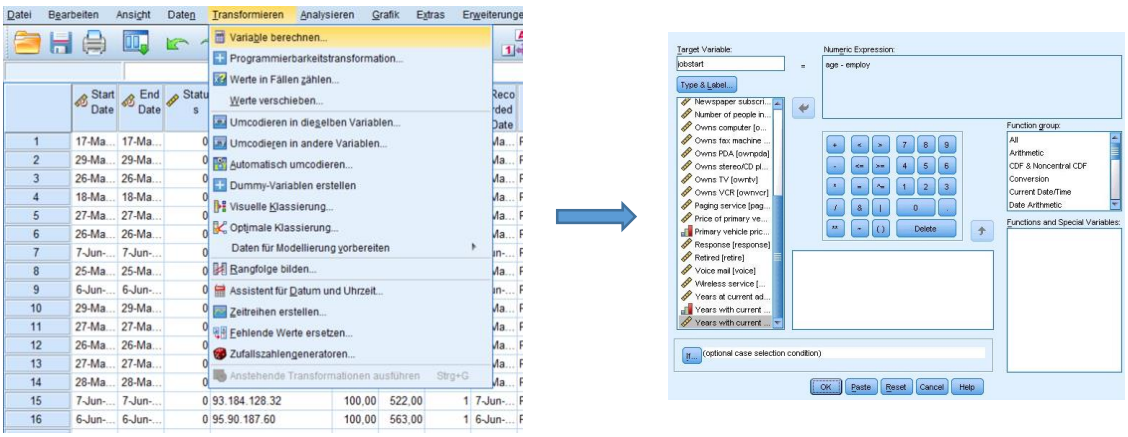
a) Berechnen von Variablen
b) Umcodieren (recodieren) von Variablen

<u>Hinweis:</u> Bei der Transformation sollte stets darauf geachtet werden, dass die transformierten Variablen in ihrer Variablenbezeichnung stets von der ursprünglichen Bezeichnung abweichen. Insbesondere bei der Umcodierung ist IMMER eine ANDERE VARIABLE zu erstellen. Wird die Umcodierung in derselben Variable vorgenommen, sind die ursprünglichen Werte irreversibel überschrieben. Auch an dieser Stelle sollte jeder Transformationsvorgang dokumentiert werden.

6.1 Berechnen von Variablen

Um mathematische Operationen mit einer oder mehreren Variablen durchzuführen stellt SPSS den Befehl „*Variable berechnen*“ bereit („*Transformieren*“ – „*Variable berechnen*“). Ein Hauptanwendungsfall stellt das Aggregieren mehrerer Variablen (/Items) zu einer Variable (/latentes Konstrukt) dar.

Wird die Anwendung „*Variable berechnen*“ in SPSS ausgewählt, öffnet sich eine Dialogbox in der die Zielvariable bezeichnet werden kann sowie die vorhanden Variablen für mathematische Operationen ausgewählt werden können.

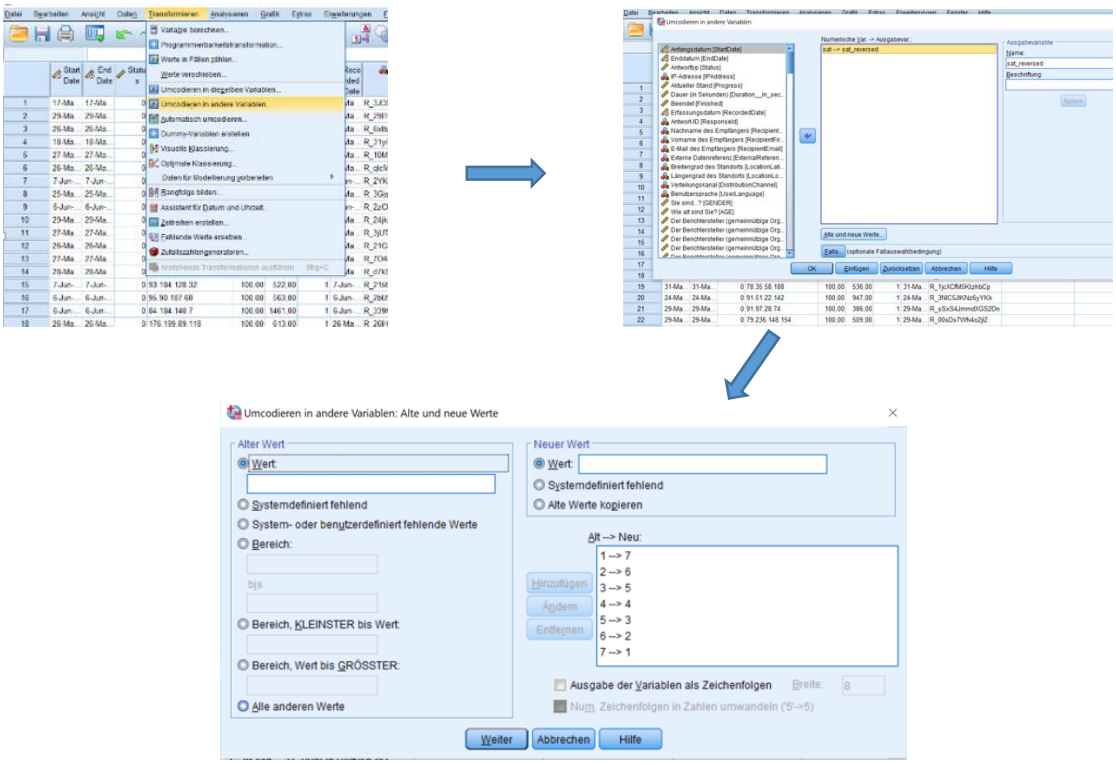


6.2 Umcodieren einer Variablen

Um Variablenwerte bzw. -kategorien zu modifizieren stellt SPSS den Befehl *„Umcodieren in andere Variablen“* bereit (*„Transformieren“ – „Umcodieren in andere Variablen* bereit*“*). Hauptanwendungsfälle sind:

- **Dichotomisieren von Variablen** -> Bsp. Variable „Höhe der Nebeneinkünfte“ (xxx €) wird in „Nebeneinkünfte nein oder ja“ (nein=0, ja=1) recodiert.
- **Zusammenfassen von Wertebereichen** -> Bsp. Variable „Anzahl Berufsjahre“ (xxx) wird in „Berufserfahrungsklassen“ (geringe Erfahrung: 0 bis 3 Jahre=0, mittlere Erfahrung: 4 bis 7 Jahre=1, hohe Erfahrung: ab 8 Jahren=2) recodiert.
- **Umpolung von Variablen** -> Bsp. Die Variable „Zufriedenheit mit dem Produkt“ wird mit einer 7-Punkt Skala erhoben. Die Antwort „vollkommen zufrieden“ wurde mit 1, „vollkommen unzufrieden mit 7, etc., codiert. In der Regel kann die Interpretierbarkeit bzw. die Möglichkeiten zur Datenanalyse erhöht werden, wenn eine höhere Zustimmung zu einer Aussage einem höheren Wert zugewiesen wird.

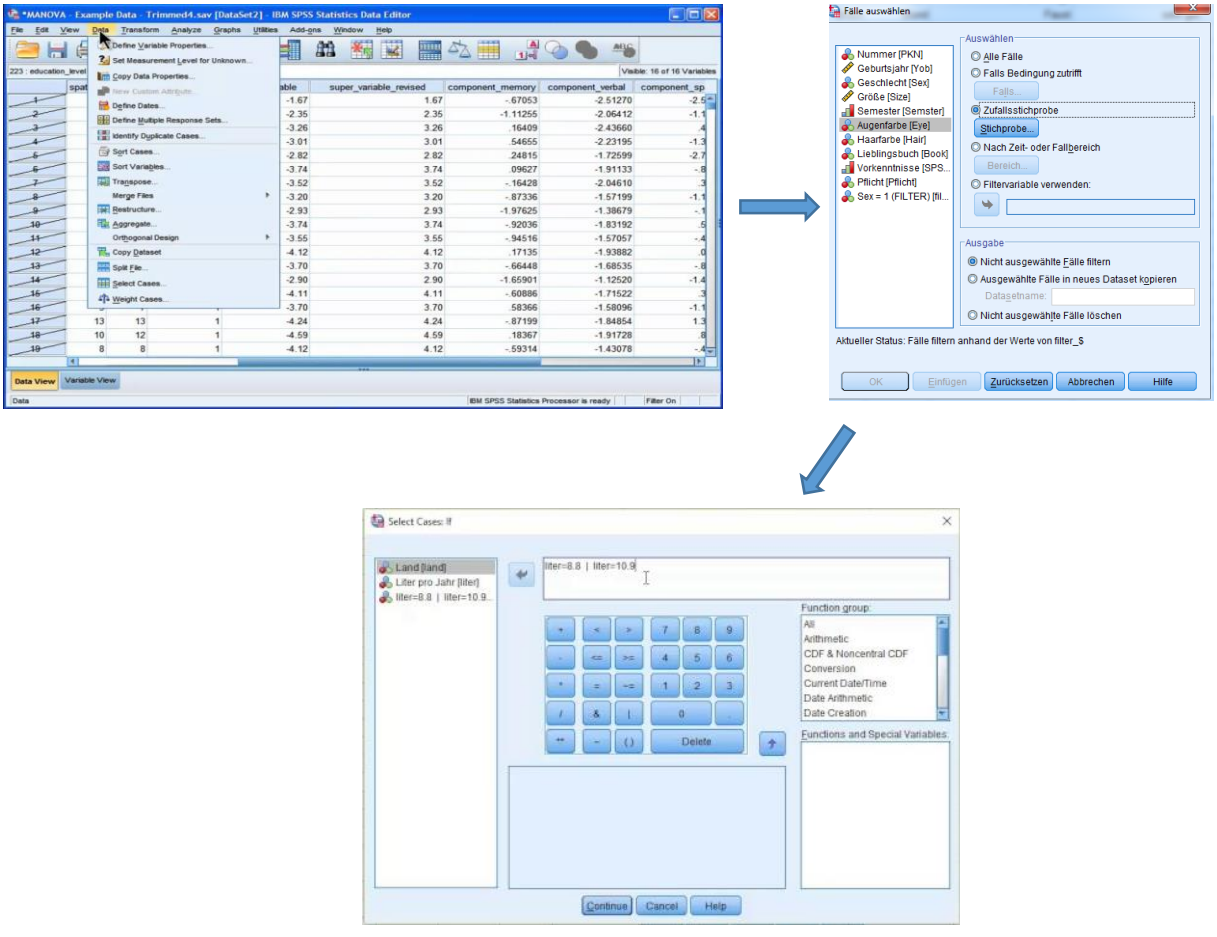
Wird die Anwendung *„Umcodieren in andere Variablen“* in SPSS ausgewählt, öffnet sich eine Dialogbox in der die ursprünglichen Variablen ausgewählt und die neuen Ausgabevariablen bezeichnet werden können. Die eigentliche Wertetransformation kann unter *„Alte und neue Werte“* festgelegt werden. Hierzu öffnet sich eine weitere Dialogbox, in der alte durch neue Codierungen ersetzt werden können. Soll die Umcodierung nur bei bestimmten Fällen bzw. Bedingungen erfolgen, können unter *„Falls…“* die notwendigen Bedingungen formuliert werden.



7. Fälle auswählen

Häufig ist für einzelne Analysen nicht der gesamte Datensatz relevant. Sollen nur Teilgruppen bzw. nur bestimmte Merkmalsausprägungen untersucht werden, z.B. nur die männlichen Probanden, bietet SPSS die Funktion *„Fälle auswählen“ („Daten“ –„Fälle auswählen“ – „Falls Bedingung zutrifft“).*

Wird die Anwendung *„Fälle auswählen“* in SPSS ausgewählt, öffnet sich eine Dialogbox in der die Variablen ausgewählt werden können. Durch Auswählen der Option *„Falls Bedingung zutrifft“* und Klicken auf *„Falls…“* öffnet sich eine weitere Dialogbox, in der die relevanten Ausprägungen bei den interessierenden Variablen festgelegt werden können (Bspw. Variable „Geschlecht“ =0, für lediglich weibliche Probandinnen und Variable „Berufserfahrungsklassen“ = 2, für hohe Erfahrung).



Literaturverzeichnis

Akremi, Leila/Baur, Nina/Fromm, Sabine (2011): Datenanalyse mit SPSS für Fortgeschrittene 1. Datenaufbereitung und uni- und bivariate Statistik. 3. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften/Springer Fachmedien Wiesbaden GmbH Wiesbaden.

Albers, Sönke/Klapper, Daniel/Konradt, Udo/Walter, Achim; Wolf, Joachim (2009): Methodik der empirischen Forschung. 3. Auflage. Wiesbaden: Gabler Verlag.

Beierlein, Constanze/Kemper, Christoph/Kovaleva, Anastassiya/Rammstedt, Beatrice. (2013). Short Scale for Measuring General Self-efficacy Beliefs (ASKU).

Berekoven, Ludwig/Eckert, Werner/Ellenrieder, Peter (2009): Marktforschung. Methodische Grundlagen und praktische Anwendung. 12. Auflage. Wiesbaden: Gabler Verlag/GWV Fachverlage GmbH Wiesbaden.

Carlson, Kevin D./Herdman, Andrew O. (2012). Understanding the Impact of Convergent Validity on Research Results. In: Organizational Research Methods, 15(1), 17–32.

Cho, Eunseong/Kim, Seonghoon (2015). Cronbach's Coefficient Alpha: Well Known but Poorly Understood. In: Organizational Research Methods, 18(2), 207.

Fornell, Claes/Larcker, David F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. In: Journal of Marketing Research, 18(1), 39-50.

Häder, Michael (2006): Empirische Sozialforschung. Eine Einführung. Wiesbaden: VS Verl. für Sozialwiss.

Hammann, Peter/Erichson, Bernd (2006): Marktforschung. 4. Auflage. Stuttgart: Lucius & Lucius.

Homburg, Christian (2017): Marketingmanagement. Strategie - Instrumente - Umsetzung - Unternehmensführung. 6. Auflage. Wiesbaden: Springer Gabler.

Lück, Detlev/Landrock, Uta (2014): Datenaufbereitung und Datenbereinigung in der quantitativen Sozialforschung. In: Nina Baur und Jörg Blasius (Hg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS, S. 397–409.

Malhotra, Naresh K. (2010): Marketing research. An applied orientation. Sixth edition, global edition. Boston, Amsterdam, Dubai: Pearson.

Raab-Steiner, Elisabeth/Benesch, Michael (2015): Der Fragebogen. Von der Forschungsidee zur SPSS-Auswertung. 4. Auflage. Wien: Facultas-Verlag.

Zikmund, William G./Babin, Barry J. (2012): Essentials of marketing research. Boston: Nelson Education.