RatSWD Working Paper Series

www.ratswd.de

RatSWD
Rat für Sozial- und WirtschaftsDaten

230

Qualitätsstandards zur Entwicklung, Anwendung und Bewertung von Messinstrumenten in der sozialwissenschaftlichen Umfrageforschung

Erarbeitet und verfasst von der Arbeitsgruppe Qualitätsstandards

Februar 2014

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Working Paper Series des Rates für Sozial- und Wirtschaftsdaten (RatSWD)

Die *RatSWD Working Papers* Reihe startete Ende 2007. Seit 2009 werden in dieser Publikationsreihe nur noch konzeptionelle und historische Arbeiten, die sich mit der Gestaltung der statistischen Infrastruktur und der Forschungsinfrastruktur in den Sozial-, Verhaltens- und Wirtschaftswissenschaften beschäftigen, publiziert. Dies sind insbesondere Papiere zur Gestaltung der Amtlichen Statistik, der Ressortforschung und der akademisch getragenen Forschungsinfrastruktur sowie Beiträge, die Arbeit des RatSWD selbst betreffend. Auch Papiere, die sich auf die oben genannten Bereiche außerhalb Deutschlands und auf supranationale Aspekte beziehen, sind besonders willkommen.

RatSWD Working Papers sind nicht-exklusiv, d. h. einer Veröffentlichung an anderen Orten steht nichts im Wege. Alle Arbeiten können und sollen auch in fachlich, institutionell und örtlich spezialisierten Reihen erscheinen. Die *RatSWD Working Papers* können nicht über den Buchhandel, sondern nur online über den RatSWD bezogen werden.

Um nicht deutsch sprechenden Nutzer/innen die Arbeit mit der neuen Reihe zu erleichtern, sind auf den englischen Internetseiten der *RatSWD Working Papers* nur die englischsprachigen Papers zu finden, auf den deutschen Seiten werden alle Nummern der Reihe chronologisch geordnet aufgelistet.

Einige ursprünglich in der *RatSWD Working Papers* Reihe erschienenen empirischen Forschungsarbeiten sind ab 2009 in der RatSWD Research Notes Reihe zu finden.

Die Inhalte der *RatSWD Working Papers* stellen ausdrücklich die Meinung der jeweiligen Autor/innen dar und nicht die des RatSWD.

Kontakt: S C I V E R O Verlag | Rat für Sozial- und Wirtschaftsdaten | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

Qualitätsstandards zur Entwicklung, Anwendung und Bewertung von Mess-instrumenten in der sozialwissenschaftlichen Umfrageforschung

Erarbeitet und verfasst von der Arbeitsgruppe Qualitätsstandards

Mitglieder der Arbeitsgruppe:

- Beatrice Rammstedt (Vorsitzende, GESIS - Leibniz-Institut für Sozialwissenschaften)
- Constanze Beierlein (GESIS - Leibniz-Institut für Sozialwissenschaften)
- Elmar Brähler (Universität Leipzig)
- Michael Eid (Freie Universität Berlin)
- Johannes Hartig (Deutsches Institut für Internationale Pädagogische Forschung)
- Martin Kersting (Justus-Liebig-Universität Gießen)
- Stefan Liebig (Universität Göttingen)
- Josef Lukas (Universität Halle)
- Anne-Kathrin Mayer (Leibniz-Zentrum für Psychologische Information und Dokumentation)
- Natalja Menold (GESIS - Leibniz-Institut für Sozialwissenschaften)
- Jürgen Schupp (Deutsches Institut für Wirtschaftsforschung, Berlin)
- Erich Weichselgartner (Leibniz-Zentrum für Psychologische Information und Dokumentation)

Inhaltsverzeichnis

1 Einführung

Der vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Rat für Sozial- und Wirtschaftsdaten (RatSWD) berät seit 2004 die Bundesregierung und die Regierungen der Länder in Fragen der Erweiterung und Verbesserung der Forschungsinfrastruktur für die empirischen Sozial-, Wirtschafts- und Verhaltenswissenschaften (SWV). Ende 2010 hat sich der RatSWD der Fragestellung gewidmet, wie sich die Qualität von Erhebungsinstrumenten in den Sozial- und Wirtschaftswissenschaften, insbesondere in der entsprechenden Umfrageforschung prüfen und sichern lässt. Der RatSWD hat daher auf seiner Sitzung am 9. November 2012 beschlossen, eine Arbeitsgruppe *Qualitätssicherung von Erhebungsinstrumenten* unter Leitung von Prof. Rammstedt einzurichten. Insbesondere war die Berufung der Arbeitsgruppe mit dem Wunsch verbunden, Qualitätsstandards zu definieren, um hierdurch die Qualität zu sichern und zu optimieren. Die Arbeitsgruppe hat sich die Formulierung dieser Standards zum primären Ziel gesetzt. Die vorliegende Publikation stellt diese Standards dar.

1.1 Zur Relevanz der Qualitätssicherung von Messinstrumenten in der sozial- und wirtschaftswissenschaftlichen Umfrageforschung

Forschungsgegenstand der sozial- und wirtschaftswissenschaftlichen Umfrageforschung sind gesellschaftliche Phänomene. Solche Phänomene sind beispielsweise soziale Ungleichheit, Migration, Einstellungen zur Demokratie oder Lebensqualität. Die Betrachtung dieser Phänomene ergibt sich aus der Notwendigkeit, die Mechanismen der Wahrnehmung und Bewertung gesellschaftlicher Phänomene zu identifizieren, um daraus auch Empfehlungen für die Gestaltung des gesellschaftlichen Lebens abzuleiten. Die sozialwissenschaftliche wie auch die wirtschaftswissenschaftliche Umfrageforschung stellen interdisziplinäre Forschungsfelder dar, in denen zur empirischen Beantwortung von Forschungsfragen Umfragedaten verwendet werden. Basis dieser Daten sind zumeist Beantwortungen von Fragebögen z. B. in einem Interview oder in einem webbasierten Survey. Ein Fragebogen enthält Fragen oder Fragekomplexe zur Beschreibung eines Phänomens oder mehrerer unterschiedlicher Phänomene. Fragen, die darauf abzielen ein abgegrenztes Phänomen (Sachverhalt) zu erfassen, werden in dem vorliegenden Dokument als Messinstrument bezeichnet. Unwichtig ist dabei, ob der

Sachverhalt mit einer oder mehreren Fragen erhoben wurde. Vielmehr soll der Begriff „Messinstrument“ verdeutlichen, dass numerische, quantitative Informationen zu einem spezifischen Phänomen als Ergebnis der Datenerhebung mit einem Messinstrument vorliegen[1]. Diese numerischen Informationen (oder quantitative Daten) sind die Basis sozial- und wirtschaftswissenschaftlicher Aussagen in der Umfrageforschung.

Viele Surveys werden als empirische Basis für gesellschaftliche Bestandsaufnahmen, aber auch zur Identifikation politischer Handlungsbedarfe sowie zur Entwicklung von politischen und gesellschaftlichen Interventionen herangezogen. Gerade vor diesem Hintergrund der hohen gesellschaftspolitischen Bedeutung der Umfrageforschung ist die Qualitätssicherung und –optimierung der in Umfragen genutzten Messinstrumente von essentieller Bedeutung. Die Dokumentation der Qualitätsüberprüfung und ihrer Resultate trägt dazu bei, die Forschungsergebnisse nachzuvollziehen und/oder Sekundäranalysen der vorhandenen Daten durchzuführen. Im Laufe der letzten Jahre ist das Bewusstsein für die Qualität von Daten in der Wissenschaftsgemeinschaft sowie in der Gesellschaft generell gestiegen. Beigetragen zu dieser Entwicklung haben unter anderem auch die jüngsten Berichte über wissenschaftliches Fehlverhalten. Nicht zuletzt ist aus forschungsethischen wie auch ökonomischen Gründen dringend geboten, in umfangreichen und kostenintensiven sozialwissenschaftlichen Umfragen Messinstrumente unklarer und teils zweifelhafter Qualität zu vermeiden.

Ausgehend von diesen Überlegungen werden im vorliegenden Dokument Standards für die Qualität von Messinstrumenten vorgeschlagen. Sie sollen in der Gesamtschau eine umfassende Bewertung der Güte von Messinstrumenten ermöglichen und mittelfristig zu einer Steigerung der Qualität von sozial- und wirtschaftswissenschaftlichen Umfragedaten beitragen. Diese Standards sind prinzipiell für alle Beteiligten von Nutzen:

Entwicklerinnen und Entwickler von Messinstrumenten können die Standards als einen Leitfaden der Qualitätssicherung ihres Vorgehens heranziehen, der sich auf alle Schritte des Kon-

[1] Unter „Messung“ wird in der sozialwissenschaftlichen Forschung eine strukturtreue Abbildung der Beziehungen zwischen Objekten mit Hilfe von Zahlen (Messwerten) verstanden (Schnell, Hill & Esser, 2011). Strukturtreu bedeutet hier, dass die Menge von Zahlen (numerisches Relativ) eine bestimmte, zuvor definierte Relation zwischen den Objekten (empirisches Relativ) korrekt abbildet. Werden bspw. die Objekte nach Länge geordnet, so sollten die Zahlen (kleiner – größer) die Beziehung zwischen den Objekten (kürzer – länger) fehlerfrei wiedergeben (vgl. Schnell et al., S. 130).

struktionsprozesses bezieht: Von der Entscheidung über die Notwendigkeit der (teilweisen) Neukonstruktion oder Modifikation eines Messinstruments bis hin zur empirischen Qualitätsüberprüfung und zur adäquaten Dokumentation.

(Potenzielle) Nutzerinnen und Nutzer der Messinstrumente bzw. der mit diesen Instrumenten erhobenen Daten erhalten Hinweise, die ihnen die Beurteilung vorliegender Instrumente erleichtern. Sie können die Standards beispielsweise nutzen, um bei der Planung eigener Datenerhebungen aus verschiedenen Instrumenten dasjenige auszuwählen, das sich im spezifischen Forschungs- und Anwendungskontext am besten bewährt hat. Nur unter der Voraussetzung einer sorgfältigen Dokumentation der Messinstrumente (einschließlich des Vorgehens bei der Instrumentenentwicklung), wie sie in den Standards gefordert wird, sind zudem fundierte Sekundäranalysen der Daten, die mit diesen Messinstrumenten erhoben wurden, möglich.

Neben diesen Beteiligten, die die Hauptadressatinnen und -adressaten des vorliegenden Dokuments sind, profitieren schließlich auch die *Rezipientinnen und Rezipienten der mit den Messinstrumenten gewonnenen Forschungsergebnisse* (z. B. Wissenschaftlerinnen und Wissenschaftler, aber auch gesellschaftliche und politische Entscheidungsträgerinnen und -träger). Ihnen können die Standards als Hilfsmittel zur Beurteilung und Gewichtung der Forschungsergebnisse mit Blick auf deren Aussagekraft und Handlungsrelevanz dienen.

In der Umfragemethodik hat sich das Konzept des *Total Survey Error* (Groves et al., 2004) als sinnvoller Bezugsrahmen zur Bestimmung der Qualitätsmerkmale von Umfragen herauskristallisiert. Grundsätzlich wird dabei zwischen der Qualität der Repräsentation der Grundgesamtheit bzw. Zielpopulation und der Qualität der Messung unterschieden (vgl. Abbildung 1 im Kapitel 2). Ersteres bezieht sich auf die hier nicht im Fokus stehende Frage nach der Güte der Stichprobe und auf die Frage, ob und ggf. in welchem Umfang die in einer Umfrage an einer Stichprobe ermittelten Ergebnisse auf die Grundgesamtheit übertragen werden können. Das Kriterium der Qualität der Messung bezieht sich hingegen auf die eingesetzten Messinstrumente.

Die Methoden der Qualitätsbestimmung der Messinstrumente im Rahmen des Total Survey Error entstammen größtenteils der psychologischen Testtheorie (Lord & Novick, 1968; Lienert & Raatz, 1998; vgl. auch Groves et al., 2004; Schnell, Hill & Esser, 2011). In der Psychologie ist es üblich, die Qualität von Tests und Fragebögen zu untersuchen und Informationen über deren Objektivität, Reliabilität und Validität zu veröffentlichen. Demgegenüber erfolgt eine entsprechende Qualitätssicherung in der sozial- und wirtschaftswissenschaftlichen

Umfrageforschung eher selten. So findet man in Umfragen kaum Hinweise oder Dokumentationen dazu, welche Phänomene mit welchen Fragen und mit welcher Qualität gemessen werden. Dieser wenig zufriedenstellende Zustand kann möglicherweise daher rühren, dass in den Sozialwissenschaften – anders als in der Psychologie, wo handlungsorientierte Richtlinien zur Sicherung und Beurteilung der Qualität der psychologischen Tests und Fragebögen vorliegen (z. B. DIN, 2002; Häcker, Leutner & Amelang, 1998; Kersting, 2008) – entsprechende Standards und Richtlinien fehlen.

Die vorliegenden Qualitätsstandards sollen als eine Richtlinie zur Qualitätssicherung von Messinstrumenten in den Sozial- und Wirtschaftswissenschaften dienen. Sie vereinen die Perspektive des Total Survey Error aus der Umfragemethodik und die der psychologischen Messung aus der Testtheorie und Diagnostik. Ziel der Standards ist es, die Qualität aus der prozessualen Sicht der Umfrageplanung und -durchführung vor dem Hintergrund des Total Survey Error zu bestimmen und eine handlungsorientierte Leitlinie zur Qualitätssicherung und -optimierung vorzulegen. Aus dem Konzept des Total Survey Error werden spezifische Annahmen über unterschiedliche Fehlerarten abgeleitet, welche die Qualität von Messinstrumenten in der Umfrageforschung beeinträchtigen können. Basierend auf diesen Annahmen werden Qualitätsstandards bestimmt, welche Wissenschaftlerinnen und Wissenschaftlern sowie Anwenderinnen und Anwendern als Orientierung dienen sollen, um die Qualität der von ihnen entwickelten bzw. eingesetzten Messinstrumente zu überprüfen, zu sichern und zu optimieren.

1.2 Auf welche Messinstrumente beziehen sich die vorliegenden Standards?

Messinstrumente können in Umfragen unterschiedlich realisiert werden. Benutzt werden nicht nur Fragen in typischer Frageform wie beispielsweise die Frage nach der Schichtzugehörigkeit aus dem ALLBUS 2012[2] („Welcher Schicht rechnen Sie sich selbst eher zu – der Unterschicht, der Arbeiterschicht, der Mittelschicht, der oberen Mittelschicht oder der Oberschicht?“. Dabei werden die einzelnen Schichten als Antwortoptionen von der Interviewerin oder dem Interviewer vorgelesen), sondern auch in Form von Aussagen oder Items, wie zum Beispiel die Aussage „Die meisten Politiker interessieren sich in Wirklichkeit gar nicht für die

[2] Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS): http://www.gesis.org/allbus.

Probleme der einfachen Leute“, mit den Antwortalternativen „bin derselben Meinung“ und „bin anderer Meinung“ (ALLBUS 2012). Darüber hinaus werden in Umfragen manchmal auch Testdaten erhoben, wie beispielweise Grundkompetenzen von Schülerinnen und Schülern oder Erwachsenen in den OECD-Studien Programme for the International Student Assessment (PISA; OECD, 2013a; Prenzel, Sälzer, Klieme & Köller, 2013) bzw. Programme for the Assessment of Adult Competencies (PIAAC; OECD, 2013b; Rammstedt, 2013), die mit Hilfe von unterschiedlichen, zu lösenden Aufgaben realisiert werden. Ein Messinstrument besteht daher aus einer oder mehreren Fragen, Items oder Testaufgaben inklusive Antwortvorgaben und dient – wie oben bereits dargestellt – der Erfassung eines einzelnen Phänomens. Ein Fragebogen, wie er üblicherweise in Umfragen benutzt wird, besteht in der Regel aus mehreren Messinstrumenten, die jeweils unterschiedliche Phänomene erfassen.

In sozial- und wirtschaftswissenschaftlichen Umfragen werden zumeist Einstellungen (z. B. Einstellungen zur Demokratie, Europäischen Union oder zur Kindererziehung), Werte (z. B. Solidarität, Toleranz oder Hedonismus) sowie Verhalten (z. B. religiöses Verhalten, Teilnahme an Wahlen oder Gesundheitsverhalten) und Fakten (z. B. Geschlecht, Alter, Bildung und weitere sozio-demografische Informationen) erfasst. Vor allem Einstellungen und Werte sind oft komplex und können nur selten mit einzelnen Items oder Fragen abgebildet werden. Sie werden als nicht direkt beobachtbare Phänomene (oder Konstrukte) zumeist mit Hilfe von mehreren Items oder Fragen erhoben. Auf Grundlage einer Testtheorie (Klassische oder Probabilistische Testtheorie) werden die gewonnenen Antworten in numerische Werte transformiert und zu zusammengesetzten Indikatoren verrechnet. Beispielsweise nehmen Schwartz und Bilsky (1990) im Schwartz Value Survey (SVS) elf universelle Basis-Werte an, die interindividuell unterschiedlich ausgeprägt sein können. Diese Basis-Werte (z. B. Selbstbezogenheit, Stimulierung, Hedonismus usw.) sind so konzipiert, dass sie jeweils zahlreiche spezifische Ziele beinhalten, die wiederum jeweils mit Hilfe von mehreren Items abgefragt werden.

In der Regel wird versucht, Phänomene so effizient wie möglich zu erfassen. So bietet sich bei der Abfrage direkt beobachtbarer Sachverhalte wie einzelner Fakten oder einfacher Verhaltensweisen häufig an, das Phänomen mit nur einer einzigen Frage zu ermitteln, zum Beispiel nach der Anzahl der Kinder zur Bestimmung der Geburtenraten oder nach der Anzahl der Umzüge in einem bestimmten Zeitraum zur Bestimmung der Mobilität. Gerade bei einer

solchen Erhebung von typischerweise interessierenden Verhaltensweisen oder Fakten haben sich in den letzten Jahren Standardinstrumente, wie zum Beispiel die demografischen Standards[3], entwickelt und etabliert, die sowohl für amtliche wie auch wissenschaftsgetragene Erhebungen die Vergleichbarkeit sicherstellen sollen. Um komplexe Sachverhalte zu erfassen, werden zumeist mehrere Fragen erhoben und diese dann anschließend zu einem Index aggregiert. Ein Beispiel hierzu ist die Abbildung des sozio-ökonomischen Status anhand von Bildungs-, Berufs- und Einkommensdaten (z. B. ISEI von Ganzeboom, De Graaf, & Treiman, 1992).

Die Anwendung der Standards bezieht sich auf einzelne Messinstrumente, d.h. Fragen, Testaufgaben oder Itemsets in einem Fragebogen, die zur Erhebung eines abgegrenzten Phänomens verwendet werden. Die Standards dienen demnach nicht zur Beurteilung des gesamten Fragebogens in einer Umfrage, denn, wie oben dargestellt, umfasst ein solcher Fragebogen in der Regel mehrere Messinstrumente. Bei der Zusammenstellung unterschiedlicher Messinstrumente zu einem Fragebogen sollte auf weitere Aspekte wie die Sukzession der Darbietung, Gliederung durch Überschriften, Länge des Fragebogens, eine passende Platzierung von Instruktionen usw. geachtet werden. Diese – sehr wichtigen – Aspekte eines Fragebogens sind jedoch nicht der Gegenstand dieser Qualitätsstandards. Zur Erarbeitung und Beurteilung von Fragebögen soll an dieser Stelle auf die entsprechende Basisliteratur verwiesen werden (Dillman, Smyth & Christian, 2009; Schnell, 2012).

Messinstrumente in der Umfrageforschung können mit unterschiedlicher Zielsetzung eingesetzt werden (Hussy, Schreier & Echterhoff, 2010). So kann eine Fragestellung primär auf die *Beschreibung* und *Quantifizierung* eines Phänomens abzielen, zum Beispiel zur Bestimmung der Wahlbeteiligung, der Beschäftigungsstatistiken oder der Zufriedenheit mit der Arbeit der Bundesregierung. Alternativ können Zusammenhänge zwischen unterschiedlichen Phänomenen, wie zum Beispiel zwischen Arbeitslosigkeit und Delinquenz von Jugendlichen, oder Gruppenunterschiede bezüglich bestimmter Phänomene wie beispielsweise Geschlechterdifferenzen im Einkommen untersucht werden. Des Weiteren kann ein Messinstrument eingesetzt werden, um Phänomene mit Hilfe von Theorien zu *erklären.* Hierfür werden zunächst An-

[3] http://www.gesis.org/unser-angebot/studien-planen/demographische-und-regionale-standards (Zugriff am 8.1.2014)

nahmen über die Ursachen bestimmter Entwicklungen formuliert und diese Annahmen auf ihre empirische Gültigkeit mit Hilfe des Messinstruments überprüft. Das Ergebnis der Messung kann wiederum dazu führen, dass die ursprünglichen theoretischen Annahmen und / oder die Konzeption der Phänomene modifiziert und verändert werden müssen. Analog können, ebenfalls auf Basis von Erklärungen, *Vorhersagen* oder Prognosen über die Entwicklung bestimmter Phänomene abgeleitet werden. Nicht zuletzt kann auch beabsichtigt sein, Folgen von *Veränderungen* zu untersuchen, zum Beispiel Evaluationsstudien, die Interventions-, Rehabilitations- oder Präventionsmaßnahmen begleiten, um ihre Wirksamkeit zu ermitteln und deren Folgen abzuschätzen.

Die hier formulierten Standards finden Anwendung auf die in Frage stehenden Messinstrumente unabhängig von dem mit ihnen verfolgten Ziel, von ihrem Forschungsdesign und unabhängig von der primären wissenschaftlichen Disziplin.

1.3 Hinweise zur Nutzung der Qualitätsstandards

Die Qualitätsstandards spiegeln Anforderungen an die eingesetzten Messinstrumente und deren Dokumentation wider, die mindestens erfüllt sein müssen, um eine Beurteilung der Instrumentenqualität zu ermöglichen. Zu den jeweils übergeordneten Qualitätsstandards wurden spezifische Leitfragen formuliert. Diese Leitfragen können Wissenschaftlerinnen und Wissenschaftler sowie Anwenderinnen und Anwender nutzen um systematisch zu prüfen, welche Aspekte der Qualitätssicherung bei der Instrumentenentwicklung und -anwendung Beachtung fanden.

Zur Begründung der vorliegenden Qualitätsstandards werden Hintergrundinformationen aus der sozialwissenschaftlichen Methodik und der psychologischen Testtheorie herangezogen und es wird – ebenfalls unter Rückgriff auf diese Informationen – auf empirische Methoden und Vorgehensweisen zur Qualitätssicherung verwiesen. Dabei ist jedoch keine detaillierte Darstellung der einzelnen Methoden möglich, hierzu sollte bei Bedarf die an den einzelnen Stellen genannte einschlägige Literatur konsultiert werden.

Bei der Verwendung der Standards sollte eine mechanische und isolierte Betrachtung vermieden werden. Vielmehr sollten die Standards unter Berücksichtigung der Zielsetzung und der Besonderheiten der konkreten Studie herangezogen werden in der das Messinstrument eingesetzt wird.

Abschließend soll angemerkt werden, dass die Qualität der eingesetzten Messinstrumente eine notwendige, jedoch nicht hinreichende Voraussetzung für die Aussagekraft einer Untersuchung in Bezug auf die o.g. Ziele der sozial- und wirtschaftswissenschaftlichen Umfrageforschung ist. Hierzu muss zusätzlich auf die Qualität des Forschungsdesigns und der zugrunde gelegten Theorie sowie die hinreichende Repräsentation der interessierenden Grundgesamtheit geachtet werden.

1.4 Aufbau des Dokumentes

Das Dokument gliedert sich in drei Kapitel. Folgend auf dieses einführende Kapitel, werden in Kapitel 2 Qualitätskonzepte sozial- und wirtschaftswissenschaftlicher Messinstrumente im Kontext des Total Survey Error definiert. Dabei werden zunächst zentrale Aspekte des Total Survey Error-Ansatzes vorgestellt. Daraus werden die Schritte der Qualitätssicherung abgeleitet, die als Gliederung für die im Kapitel 3 vorgestellten Qualitätsstandards dienen. Ein besonderes Augenmerk der Qualitätsstandards liegt auf Hinweisen zur Dokumentation der Messinstrumente einschließlich der theoretischen und empirischen Argumente, die ihre Konstruktion und Anwendung rechtfertigen sollen, da durch die systematische und vollständige Dokumentation eine fundierte Qualitätsbeurteilung erst ermöglicht wird.

2 Qualitätsmerkmale sozialwissenschaftlicher Messinstrumente aus der Perspektive des Total Survey Error

In diesem Kapitel wird der Total Survey Error-Ansatz vorgestellt. Die Darstellung erfolgt im Wesentlichen in Anlehnung an Groves und Kollegen (2004) beziehungsweise an Groves und Lyberg (2010). Aus der Darstellung zum Total Survey Error-Ansatzes werden Schritte zur Qualitätssicherung der sozialwissenschaftlichen Messinstrumente abgeleitet.

Der Total Survey Error-Ansatz geht zunächst von typischen Schritten der Planung und Durchführung von Umfragen aus. Es wird angenommen, dass unterschiedliche Fehler im Prozess der Umfrageplanung und -durchführung die Qualität der erhobenen Daten beeinträchtigen können. Fehlerquellen können dabei sowohl die fehlerbehaftete Messung eines interessierenden Phänomens als auch die nicht perfekte und somit fehlerbehaftete Repräsentativität der Stichprobe sein. Die Fehler in Hinblick auf die Messung bzw. die Repräsentation der Grundgesamtheit können dabei in verschiedenen Phasen der Umfrageplanung und -durchführung auftreten. Abbildung 1 veranschaulicht die verschiedenen Phasen der Umfrageplanung und -

durchführung (dargestellt als Rechtecke). Die in diesem Ablauf möglichen Fehler bzw. die zu sichernden Qualitätskonzepte sind als Ovale in Abbildung 1 dargestellt.

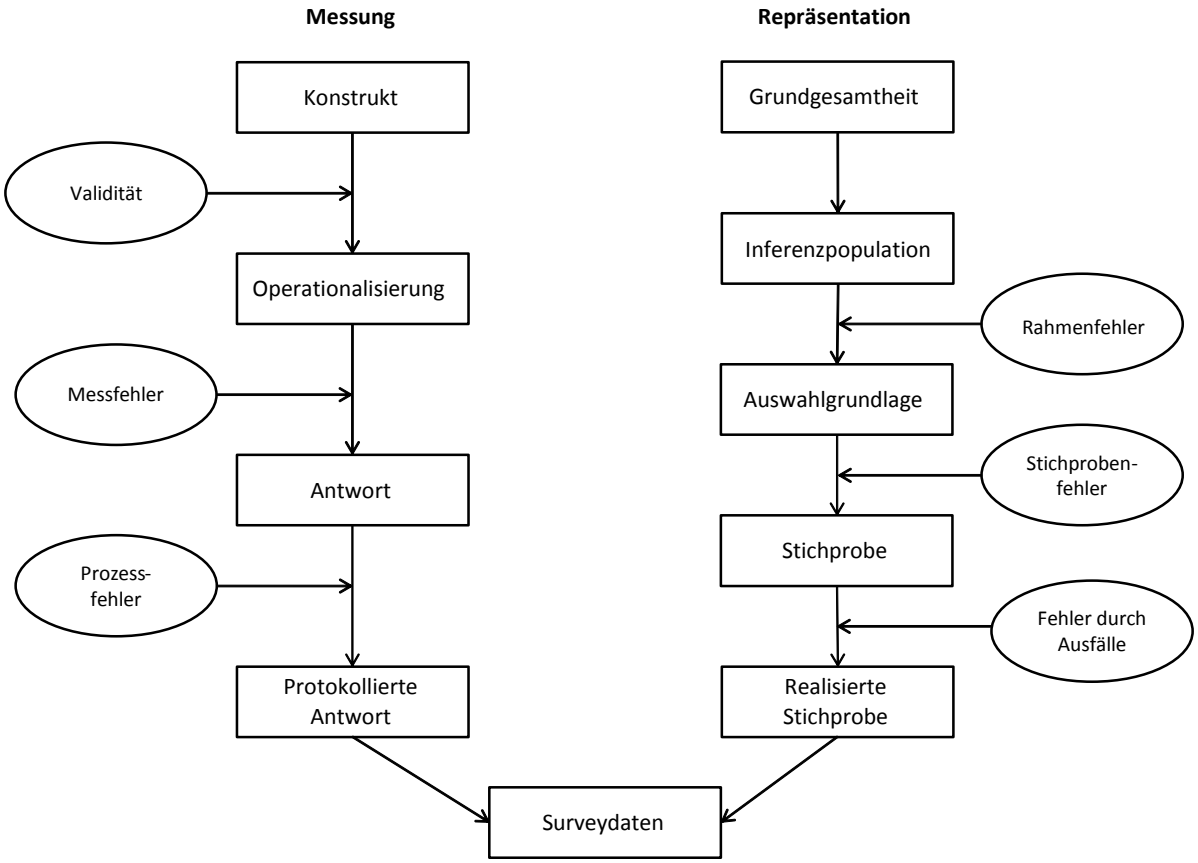

Abbildung 1. „Total Survey Error" nach Groves und Lyberg (2010).

Da sich das vorliegende Dokument ausschließlich auf die Qualitätssicherung der Messinstrumente konzentriert, wird im Folgenden nur auf den Bereich der *Messung* eingegangen (Abbildung 1, linker Strang). Für den Bereich der Repräsentation existieren in der internationalen wie nationalen Umfrageforschung bereits detaillierte und etablierte Standards, die im Anhang dieses Dokuments referenziert werden.

Im ersten Schritt der *Messung* (vgl. Abb. 1) wird die Forschungsfrage formuliert und es wird festgelegt, welche Phänomene untersucht werden sollen. Beispielsweise könnte die Forschungsfrage fokussieren, wie konservativ die CDU-Wähler sind, wie stark die Deutschen die Idee der Demokratie unterstützen oder wie umweltbewusst die Deutschen sind. Die zu untersuchenden Phänomene sind in der Regel nicht direkt beobachtbar; ihre Benennung hat zunächst die Eigenschaft eines (allgemeinen) Begriffs. Im Hinblick auf ihre Messung nennt man

solche Begriffe Konstrukte. Konstrukte (z. B. „Konservatismus“, „prodemokratische Einstellung“ oder „Umweltbewusstsein“) sind demnach zunächst Ideen, die die zu untersuchenden Phänomene benennen. Mit der Festlegung des zu untersuchenden *Konstruktes* (s. Abb.1) wird die erste Phase der Messung abgeschlossen.

In der zweiten Phase werden die festgelegten Konstrukte genauer definiert, abgegrenzt und es wird bestimmt, welche Indikatoren (Fragen oder Items, beobachtbare Reaktionen der Befragten) zur Erhebung des Konstrukts herangezogen werden. Diese Phase wird als *Operationalisierung*[4] in Abbildung 1 bezeichnet. Im Rahmen der Operationalisierung sollte definiert werden, wie nicht beobachtbare Phänomene im Rahmen der Befragung beobachtet beziehungsweise gemessen werden können. Als Operationalisierung wird daher die Menge hinreichend genauer Anweisungen bezeichnet, nach denen die untersuchten Personen als Träger von Merkmalen, die das Konstrukt unterscheidet, mit Hilfe von Surveydaten (Abb. 1) beschrieben werden können (vgl. Diekmann, 2007, Schnell et al., 2011). Ziel der Operationalisierung ist es, das interessierende Konstrukt möglichst genau, d.h. inhaltlich treffend, abzubilden.

Die *Validität* im Rahmen des Total Survey Error-Ansatzes bezeichnet die Korrespondenz zwischen dem interessierenden Konstrukt und der dazugehörigen Operationalisierung. Die Validität ist nach dieser Perspektive gegeben, wenn von der vorgenommenen Operationalisierung ein Rückschluss auf das zu erhebende Konstrukt möglich ist. So wäre beispielsweise empirisch zu belegen, dass das Item „Ich bin stolz, ein Deutscher zu sein“ ein Indikator des Konstrukts „Nationalstolz“ darstellt und nicht ein anderes Merkmal wie beispielsweise Nationalismus misst. Umgekehrt ist die Validität gemindert, wenn bereits aufgrund der vorgenommenen Operationalisierung die Generalisierung auf das zu erhebende theoretische Konstrukt eingeschränkt oder nicht gegeben ist, weil beispielsweise die Items oder Fragen das Konstrukt nur partiell oder gar nicht erfassen. Diese Korrespondenz zwischen der Operationalisierung und dem Konstrukt gilt es zu prüfen und nachzuweisen.

---

[4] Groves et al. (2004) und Groves und Lyberg (2010) verwenden den Begriff „Measurement“ sowohl für diese Phase der Umfragedurchführung als auch für den Bereich der Messung. Wir haben die konkrete Phase der Durchführung einer Umfrage mit dem Begriff der Operationalisierung, der sich in der deutschen Literatur zu Umfrageforschung etablierte, übersetzt. Der Begriff der „Operationalisierung“ stellt unserer Ansicht nach eine treffende Bezeichnung der entsprechenden Phase („Measurement“) im Total Survey Error-Ansatz dar.

Eine allgemeine Definition der Validität in der psychologischen Testtheorie gibt ein umfassenderes Verständnis der Validität wieder, die jedoch hohe Übereinstimmungen mit dem hier dargestellten Validitätsbegriff aufweist. Demnach ist Validität dann gegeben, wenn intendierten Interpretationen der Messwerte (im vorliegenden Falle basierend auf der Operationalisierung) durch theoretische und/oder empirische Belege gerechtfertigt werden können (Kane, 2013). Die Operationalisierung stellt die Basis für die weiteren Phasen im Prozess der Messung dar. Durch eine möglichst fehlerfreie Durchführung der Messung in jeder Phase und durch empirische Untersuchungen zu den jeweiligen Fehlereinflüssen wird sichergestellt, dass bei der Nutzung von Umfragedaten weitgehend zutreffende Schlussfolgerungen hinsichtlich des untersuchten Konstrukts möglich sind.

In der dritten Phase der Umfrageplanung und -durchführung werden die in Phase zwei entwickelten Fragen durch die Befragungsteilnehmerinnen und -teilnehmer beantwortet (*Antwort* in Abb. 1). Fehler im Rahmen dieses Beantwortungsprozesses können insbesondere dadurch entstehen, dass Fragen schwer verständlich formuliert, zu abstrakt, nicht eindeutig oder aus einem anderen Grund schwer zu beantworten sein können. Darüber hinaus können sich auch die Befragungsteilnehmerinnen und -teilnehmer in ihren kognitiven Fähigkeiten, ihrer Motivation oder auch anderen für die eigentliche Fragebeantwortung irrelevanten Eigenschaften unterscheiden. Diese Eigenschaften können zu störenden Einflüssen bei der Fragebeantwortung führen und sie so verzerren. Eine weitere mögliche Fehlerquelle stellt der Erhebungsmodus dar: Messinstrumente können mittels unterschiedlicher Medien (z. B. Papier, Telefon, face-to-face durch die Interviewerin bzw. den Interviewer, elektronisch auf Bildschirm) administriert werden. Der Erhebungsmodus per se oder auch der Einsatz unterschiedlicher Modi innerhalb einer Untersuchung können das Antwortverhalten beeinflussen. So können beispielsweise die Teilnehmerinnen und Teilnehmer mit den verschiedenen Medien unterschiedlich erfahren sein oder der Grad der Anonymität wird unterschiedlich bei den verschiedenen Modi wahrgenommen.

Der Messfehler (Abb. 1) ist definiert, als Differenz zwischen der aufgrund der Operationalisierung des Konstrukts (ausformulierte Fragen bzw. Items im Fragebogen) erwarteten Antwort und der noch nicht protokollierten Antwort der befragten Person, die mündlich geäußert wurde oder als mentale Repräsentation entstanden ist. Die Fehler, bzw. Differenzen, die während der Fragenbeantwortung entstehen, können systematisch und nicht systematisch sein. Systematischer Fehler – auch als Bias bezeichnet - bedeutet, dass die Antworten aller befragten Personen oder einer definierten Teilmenge dieser Personen in gleicher Weise feh-

lerbehaftet sind. Sozial erwünschtes Antworten kann eine Form des Bias sein. Zum Beispiel beantworten alle befragten Personen eine Frage zum Drogenkonsum nicht wahrheitsgemäß, indem sie sozial erwünscht weniger Drogenkonsum berichten als real stattfand (Underreporting). Messfehler werden als unsystematisch bezeichnet, wenn sie bei einzelnen Personen zum Beispiel aufgrund von Effekten der Erhebungssituation unterschiedlich ausfallen. So können Antworten fehlerhaft sein, wenn die Befragung in einer unruhigen, ablenkenden Umgebung stattgefunden hat. Der unsystematische Messfehler entspricht der Definition des Messfehlers in der Klassischen Testtheorie und wird im Zusammenhang mit der Bestimmung der Reliabilität (siehe Kapitel 3.4) untersucht.

In einer letzten Phase der Umfragedurchführung wird die Antwort der befragten Person protokolliert, entweder durch sie selbst oder auch durch die Interviewerin bzw. den Interviewer. Als Ergebnis liegt die *protokollierte Antwort* vor (Abb. 1). Diese Antwort kann durch Editierung beim Protokollieren von der zu äußernden oder geäußerten Antwort abweichen. So kann eine befragte Person Abitur als höchsten Bildungsabschluss angeben; die Interviewerin bzw. der Interviewer verwechselt aber versehentlich die zu kodierenden Nummern und notiert fälschlicherweise den Code für ein Fachabitur. Auch können in selbstadministrierten Fragebögen die Antwortkategorien nicht schlüssig sein, was ebenfalls eine fehlerhafte Protokollierung der Antwort zur Folge haben kann. Weitere Abweichungen können unter anderem durch Datenbereinigung, Korrekturen aufgrund von Plausibilitätsprüfungen (z. B. wenn ein 12-Jähriger einen Hochschulabschluss berichtet oder ein Mann eine Schwangerschaft), Zusammenfassung der Werte und Imputationen der fehlenden Werte entstehen. Der *Prozessfehler* als Qualitätskonzept (Abb. 1) ist definiert als systematische Abweichung zwischen (a) der gebildeten oder geäußerten Antwort einer Person und (b) dem protokollierten sowie danach, während der Datenbearbeitung, modifizierten numerischen Wert (*Surveydaten* in Abb. 1).

Zusammengefasst ist die Qualität der Messung entsprechend der Total Survey Error-Perspektive dann gewährleistet, wenn eine Nutzung der resultierenden (Survey-)Daten eine möglichst fehlerfreie Schlussfolgerung auf die wahren Ausprägungen des interessierenden Konstrukts bei den untersuchten Personen ermöglicht (siehe Abb. 1).

Aus der Perspektive des Total Survey Error lassen sich die folgenden Aspekte der Qualitätssicherung für die Messung ableiten:

1) Konstruktdefinition und –operationalisierung im Zuge der Instrumentenentwicklung
2) Bestimmung der Validität

3) Bestimmung des Messfehlers; unterteilt in:
    a. Bestimmung und Minimierung systematischer Effekte, die aufgrund der Spezifika der Erhebungssituation (z. B. Modus der Datenerhebung) entstehen (methodenspezifische Effekte)
    b. Bestimmung und Minimierung des Umfanges des unsystematischen Messfehlers durch die Reliabilitätsbestimmung.
4) Bestimmung und Minimierung des Prozessfehlers bei der Protokollierung der Daten und Datenaufbereitung.

Im folgenden Kapitel werden, in Anlehnung an die hier definierten Aspekte der Qualitätssicherung, Qualitätsstandards formuliert.

3 Qualitätsstandards

Die in diesem Kapitel vorgestellten sechs Qualitätsstandards (Übersicht s. Kasten 1) wurden in Anlehnung an etablierte Methoden und Best Practices in der Umfrageforschung entwickelt und werden ergänzt um Vorgehensweisen zur Qualitätssicherung und –optimierung.

Kasten 1: Übersicht der sechs Qualitätsstandards

Standard 1: Instrumentenentwicklung

Ziel und Zweck der Instrumentenentwicklung sind genannt und das methodische Vorgehen bei der Entwicklung des Messinstruments ist dokumentiert.

Standard 2: Validität

Die für die jeweilige Untersuchung zentrale Interpretation der mit einem Instrument erfassten Messwerte ist explizit formuliert und es sind Belege angeführt, mit denen diese Interpretation gestützt wird.

Standard 3: Minimierung methodenspezifischer Effekte

Mögliche methodische Einflüsse, die das Antwortverhalten der Befragungsteilnehmerinnen und -teilnehmer systematisch beeinflussen, sind im Rahmen der Instrumentenentwicklung thematisiert und geprüft und die Ergebnisse sind dokumentiert.

Standard 4: Reliabilität

Eine Prüfung der Reliabilität ist erfolgt, die gewählte Art der Reliabilitätsprüfung ist begründet und die Reliabilitätskennwerte sind dokumentiert und bewertet.

Standard 5: Minimierung des Prozessfehlers

Standardisierte Instruktionen und Verfahrenshinweise zur Durchführung der Datenerhebung und -auswertung sind vorhanden und begründet.

Standard 6: Weitere Qualitätsmerkmale

Angaben zur Ökonomie, zur Zumutbarkeit des Instruments und zur Aktualität der psychometrischen Kennwerte sind vorhanden.

3.1 Instrumentenentwicklung

Hintergrund

Der erste Schritt zur Verbesserung der Qualität von Instrumenten der Umfrageforschung besteht in einer umfassenden und exakten Dokumentation des methodischen Vorgehens bei der Instrumentenentwicklung: Die einzelnen Entwicklungsschritte und die dabei getroffenen Entscheidungen müssen sorgfältig dokumentiert und begründet sein; es muss dargelegt werden, zu welchem Zweck und für welche Zielpopulation das Instrument entwickelt und im Hinblick auf welche Fragestellungen es getestet wird; darüber hinaus soll beschrieben werden, für welche Erhebungsmodi das Verfahren geeignet, bzw. überprüft ist. Diese Informationen sind nicht nur für eine Beurteilung der mit einem Messinstrument erzielten Ergebnisse, sondern auch im Rahmen von Wiederverwendungen des Messinstruments wie auch von Sekundäranalysen der ursprünglichen Daten von hoher Bedeutung.

Grundsätzlich gilt, dass in der Umfrageforschung, aus Gründen der besseren Vergleichbarkeit von Befunden sowie der höheren Ökonomie, möglichst auf bestehende Instrumente zurückgegriffen wird. Solche erprobten und etablierten Instrumente finden sich beispielsweise in wissenschaftlichen Fachdatenbanken[5] sowie Online-Archiven sozialwissenschaftlicher und

[5] Z.B. PSYNDEX Tests (http://www.zpid.de/index.php?wahl=PSYNDEX&uwahl=Tests; Zugriff am 8.1.2014)

psychologischer Erhebungsverfahren[6]. Bei der Wiederverwendung eines bestehenden Messinstrumentes kann dessen Qualität mit Hilfe der vorliegenden Standards bewertet werden. Wenn vorhandene Instrumente genutzt werden, sollten deren Herkunft und die gegebenenfalls vorgenommenen Modifikationen sorgfältig dokumentiert sein.

Falls die existierenden Instrumente nur in einer anderen Sprache vorliegen und sie für den eigenen Einsatz übersetzt werden, sollten in diesem Rahmen die verfügbaren Übersetzungsstandards[7] beachtet werden. Generell gilt, dass im Falle substanzieller Modifikationen oder einer Übersetzung die Qualität der modifizierten Instrumente erneut empirisch (ggf. auch theoretisch) belegt werden muss.

Liegen für den zu untersuchenden Forschungsgegenstand keine Instrumente vor oder können diese aus methodischen oder inhaltlichen Gründen nicht verwendet werden, gilt es Instrumente neu zu entwickeln. Bei einer Neuentwicklung eines Messinstruments sollte zunächst die Notwendigkeit eben dieser begründet werden. Eine Neuentwicklung ist indiziert, wenn keine Messinstrumente vorhanden sind oder vorhandene Messinstrumente, im konkreten Untersuchungskontext, nicht geeignet sind, weil sie beispielsweise andere Zielpopulation ansprechen, eine zu große Anzahl an Fragen aufweisen, bestimmten Qualitätsanforderungen nicht genügen oder weil die Fragen keine aktuellen Sachverhalte erfassen.

Im Zuge der Neuentwicklung eines Messinstruments muss zunächst das interessierende Phänomen beschrieben und definiert werden (Konstruktdefinition). Die Konstruktdefinition soll möglichst aus einer inhaltlichen sozial- oder wirtschaftswissenschaftlichen Theorie abgeleitet werden (Kerntheorie). So basiert das Konstrukt der menschlichen Basis-Werte (Schwartz Value Survey, Schwartz & Bilsky, 1990) auf Theorien über formale Merkmale der Werte als abstrakte, inhaltlich distinkte Motive. Die Inhalte der Werte werden aus Motivationstheorien abgeleitet.

---

[6] Z. B. die Zusammenstellung sozialwissenschaftlicher Items und Skalen ZIS von GESIS (http://www.gesis.org/unser-angebot/daten-erheben/zis-ehes/; Zugriff am 8.1.2014) und das Elektronische Testarchiv des ZPID (http://www.zpid.de/index.php?wahl=products&uwahl=frei&uuwahl=userlog: Zugriff am 8.1.2014)

[7] Siehe die Richtlinien für die Übersetzung von Messinstrumenten in kulturvergleichenden Umfragen, auf die an dieser Stelle nicht näher eingegangen werden soll (Translation in Guidelines for Best Practices in Cross-Cultural Surveys: http://ccsg.isr.umich.edu/index.cfm; Zugriff am 8.1.2014)

Zusätzlich zu der Kerntheorie wird ein weiteres theoretisches Hilfsgebilde benötigt, das angibt, wie die abstrakten Begriffe in beobachtbare Sachverhalte überführt werden. Dieses Hilfsgebilde, auch Hilfstheorie genannt (Schnell et al., 2011), enthält Aussagen über die Operationalisierung des Konstrukts, welche wiederum empirisch überprüft werden können. Fehlt der Forscherin bzw. dem Forscher das detaillierte Wissen über das interessierende Phänomen, so dass keine begründeten Aussagen über die Operationalisierung des Konstrukts abgeleitet werden können, kann dieses Wissen beispielsweise auch im Rahmen qualitativer Vorstudien (z. B. Beobachtung, qualitative Interviews) gewonnen werden. Unter der Einbeziehung der Hilfstheorie wird, zu Beginn der Instrumentenentwicklung, das passende Messmodell (im Rahmen der Klassischen oder Probabilistischen Testtheorie) bestimmt sowie die Methode der Itementwicklung oder eine Skalierungsmethode gewählt.

Für die Entwicklung der Items oder Fragen des Messinstruments sollten Konstruktionsprinzipien festgelegt und begründet werden. Ein verbreitetes Verfahren besteht darin, für jeden einzelnen Inhaltsbereich (Dimension) des Messinstruments möglichst mehrere Indikatoren zu finden. Zum Beispiel zählen als Indikatoren für Fremdenfeindlichkeit ethnische Stereotype und diskriminierende Verhaltensweisen gegenüber Mitgliedern bestimmter ethnischer Gruppen. Auf Basis dieser Indikatoren wird dann der sogenannte Itempool, ein Satz von Items oder Fragen, gebildet der die entsprechenden Indikatoren repräsentieren soll. Um aus diesem Itempool nun die am besten geeigneten Fragen oder Items auszuwählen, stehen verschiedene Methoden zur Auswahl: Zum einen können die Items einer Stichprobe der Zielpersonen zur Bearbeitung vorgelegt werden. Anschließend werden Itemanalysen durchgeführt und Items anhand der Itemkennwerte (Schwierigkeit, Trennschärfe) ausgewählt. Zweitens kann die Wahl der Items argumentativ durch ausgewählte Expertinnen und Experten wie folgt beurteilt werden: Bilden die Items die Menge der möglichen Items hinreichend ab? Stehen die Inhalte in einem angemessenen Verhältnis zueinander? Sind irrelevante Inhalte enthalten? Die Expertenurteile können dabei in Einzelinterviews oder in Fokusgruppen erhoben werden. Alle genannten Methoden können auch in Kombination verwendet werden.

Zusätzlich zur Entwicklung und Auswahl der Fragen oder Items sollte die Wahl der Antwortvorgaben begründet werden (Warum wurde beispielsweise eine fünfstufige Rankingskala benutzt? Weshalb wurden bestimmte Antwortalternativen für das Messinstrument ausgewählt?). Dabei sollte darauf geachtet werden, dass die Antwortvorgaben auf die Zielgruppe passen sowie eindeutig und vollständig sind (Dillman et al., 2009; Krosnick & Fabrigar, 1997).

Die Entwicklerinnen und Entwickler der Messinstrumente sollten empirische Nachweise über die Qualität des Messinstruments erbringen. Bei der Entwicklung neuer Messinstrumente sollten zunächst empirische Studien (wie z. B. Pretests) zur Entwicklung und Validierung sorgfältig geplant werden. Die Untersuchungen zur Validierung von Messwertinterpretationen erfolgen idealerweise nicht mit derselben Stichprobe, die zur Entwicklung genutzt wurde (Entwicklungsstichprobe), sondern werden in einer oder mehreren separaten Stichproben durchgeführt (Validierungsstichprobe/n). Können aufgrund der mangelnden Ressourcen keine Validierungsuntersuchungen an neuen Stichproben durchgeführt werden, ist dies in der Dokumentation des Messinstruments anzugeben.

Standard 1:

Ziel und Zweck der Instrumentenentwicklung sind genannt und das methodische Vorgehen bei der Entwicklung des Messinstruments ist dokumentiert.

Leitfragen zur Instrumentenentwicklung

1. Ist das Ziel der Entwicklung des Messinstruments angegeben? (Grundlagenforschung/Angewandte Forschung; Beschreiben, Erklären, Vorhersagen, Veränderungsmessung)
2. Ist angegeben, für welche Zielpopulation(en) das Messinstrument entwickelt wurde?
3. Ist das methodische Vorgehen bei der Entwicklung des Messinstruments dargelegt? Gibt es Angaben zur Personenstichprobe, bei der das Messinstrument eingesetzt wurde und zum Modus der Datenerhebung?
4. Liegt eine Definition des zu messenden Phänomens bzw. Konstrukts vor? Wurde eine Abgrenzung von benachbarten/ähnlichen Phänomenen vorgenommen?
5. Ist angegeben, anhand welcher sozialwissenschaftlichen Theorie das Phänomen definiert und seine Spezifikation vorgenommen wurde? Sind ggfs. einschlägige Studien zur Überprüfung der Kerntheorie genannt? Sind ggfs. die Ergebnisse dieser Studien, die für die Konstruktdefinition von Bedeutung sind, überblicksweise dargelegt?
6. Wird die gewählte Testtheorie genannt und das verwendete Messmodell beschrieben? Wird die Wahl in Bezug auf die Hilfstheorie begründet? Sind ggfs. bisherige Studien zur Überprüfung des Messmodells genannt sowie ihre zentralen Erkenntnisse dargelegt?
7. Wird die Notwendigkeit einer Neuentwicklung begründet?
8. Im Fall der Modifikation eines vorhandenen Messinstruments: Ist die Originalquelle ange-

geben? Sind die vorgenommenen Modifikationen vollständig aufgeführt? Werden die Modifikationen begründet?

9. Im Fall der Übersetzung eines vorhandenen Messinstruments: Ist das Vorgehen bei der Übersetzung beschrieben? Welche Standards wurden bei der Übersetzung angewandt?
10. Erfolgte die Erzeugung der Items nach festgelegten Regeln? Wird die gewählte Skalierungsmethode (z. B. Likert-, Thurstone-, Guttman-, Raschskalen) begründet bzw. ist sie anhand der Angaben zur Hilfstheorie nachvollziehbar? Sind die Regeln für die systematische Auswahl der Items angegeben und begründet?
11. Wurden bei der Entwicklung des Messinstruments Expertenurteile im Rahmen der Auswahl der Items herangezogen? Sind der fachbezogene Ausbildungsstand und die Erfahrung der Experten in der Dokumentation zum Messinstrument angegeben? Sind die Einschätzungen der Experten beschrieben und der Grad der Übereinstimmung zwischen den Experten angegeben?
12. Sind die Antwortvorgaben beschrieben? Ist ihre Wahl begründet? Sind ggfs. die einschlägigen Studien mit Ergebnis genannt?
13. Wurden bei der Entwicklung des Messinstruments Validierungsstudien durchgeführt? Sind die Ergebnisse dieser Studien dargelegt?
14. Wurden Instrumentenentwicklung und -validierung an unterschiedlichen Stichproben vorgenommen?
15. Ist angegeben, für welche wissenschaftlichen Zwecke das Messinstrument verwendet werden kann? Sind ggf. Einschränkungen der Ergebnisinterpretation genannt, die aus den Einschränkungen im Rahmen der Operationalisierung, Entwicklung des Messinstruments und der Validierungsergebnisse resultieren?

3.2 Validität

Hintergrund

Das Gütekriterium der Validität bezieht sich auf die *Gültigkeit der Interpretationen und Verwendungen* von Messwerten. Zusammenfassend kann Validität betrachtet werden als „ein integriertes bewertendes Urteil über das Ausmaß, in dem Angemessenheit und Güte von Interpretationen und Maßnahmen auf Basis von Testwerten oder anderen diagnostischen Verfahren durch empirische Belege und theoretische Argumente gestützt sind“ (Messick,

1989, S. 13, Übersetzung Hartig et al., 2012, S. 114). Ältere Ansätze zum Gütekriterium der Validität betrachten diese als eine Eigenschaft eines Messverfahrens. Demgegenüber gehen moderne Ansätze davon aus, dass Validität kein Attribut eines Messverfahrens per se ist. Stattdessen können lediglich spezifische Interpretationen und Verwendungen der resultierenden Messwerte gültig (valide) sein. Im Laufe der Entwicklung des Validitätsbegriffs wurden verschiedene Aspekte der Validität differenziert, insbesondere *Kriteriumsvalidität, Inhaltsvalidität* und *Konstruktvalidität* (zur Geschichte des Begriffs siehe Hartig et al., 2012; Kane, 2001, 2013). Während diese Aspekte zeitweise als separate Alternativen zur Untersuchung des Gütekriteriums Validität betrachtet wurden, entwickelte sich die *Konstruktvalidität* (Cronbach & Meehl, 1955) zum integrierenden, übergreifenden Konzept. Neuere Ansätze nehmen eine Differenzierung nach „Validitätsarten“ nicht mehr vor, es steht die Güte der *Argumentation* im Mittelpunkt, mit der spezifische Messwert-Interpretationen gerechtfertigt werden können (*Validity as evaluation argument*, Cronbach, 1988; *An argument-based approach to validation*, Kane, 1992).

Nach Kane (2001) können sich Interpretationen von Messwerten unter anderem beziehen auf

1. das *Bewerten* von Messergebnissen,
2. das *Verallgemeinern* eines Ergebnisses über die Inhalte der konkret eingesetzten Fragen oder Aufgaben hinaus,
3. das *Extrapolieren* im Sinne von Schlussfolgerungen auf interessierende Phänomene außerhalb des Testkontextes,
4. das (kausale) *Erklären* eines Testwertes mit Bezug auf ein zugrundeliegendes theoretisches Konstrukt sowie
5. das Treffen weiterführender *Entscheidungen* als Konsequenz aus den Messwerten.

Kane (2013) verwendet im Kontext der Validierung von Messwert-Interpretationen den Begriff des Interpretations-/Nutzungs-Arguments (*interpretation/use argument*, IUA). Das Interpretations-/Nutzungs-Argument enthält eine explizite Aussage darüber, wie Messwerte interpretiert und benutzt werden sollen, und durch welche Argumente diese spezifische Interpretation gerechtfertigt werden kann. Eine spezifische Interpretation eines Messwertes kann nach Kane in dem Umfang als valide betrachtet werden, in dem das jeweilige Interpretations-/Nutzungs-Argument in sich schlüssig ist, die intendierte Interpretation vollständig repräsentiert wird und in dem die in der Argumentation enthaltenen Annahmen angemessen belegt sind.

Im Rahmen der Messung in der Umfrageforschung sollten dementsprechend zunächst explizit diejenigen Interpretationen und Verwendungen der mit einem Instrument erfassten Messwerte

benannt werden, die für das jeweilige Untersuchungsziel am bedeutsamsten sind. Darauf aufbauend sollten empirische und / oder theoretische *Belege* angeführt werden, mit denen diese Interpretationen gestützt werden können. Charakteristisch für das Gütekriterium der Validität ist, dass seine Überprüfung kein Routineverfahren darstellt, bei dem immer auf die gleichen Methoden zurückgegriffen werden kann. Die *Validierung* von Messwertinterpretationen stellt vielmehr theoriegeleitete Forschung dar, mit der spezifische Interpretationen gestützt, aber auch falsifiziert werden können (Hartig et al., 2012; Kane, 2001).

Aus der Perspektive des Total Survey Error ist in der sozialwissenschaftlichen Umfrageforschung vor allem die oben unter 4. genannte erklärende, theoriebasierte Interpretation von Messwerten als Indikatoren für zugrundeliegende Konstrukte zentral. Diese Interpretation ist eng mit dem von Cronbach und Meehl (1955) geprägten Begriff der *Konstruktvalidität* verbunden. Die Prüfung einer theoriebasierten Messwertinterpretation kann erfolgen, indem aus der Theorie abgeleitete Vorhersagen darüber geprüft werden, wie die Messwerte mit anderen Variablen in Zusammenhang stehen. Dabei können vielfältige, sowohl experimentelle als auch korrelative Untersuchungsdesigns zum Einsatz kommen. Idealtypisch hierfür ist die von Cronbach und Meehl (1955) vorgeschlagene theoretische Ableitung eines *nomologischen Netzwerkes*, in dem die Beziehungen eines Konstrukts zu anderen Konstrukten beschrieben werden. Solche Beschreibungen sind nach Verständnis der sozialwissenschaftlichen Literatur als Aussagen in Hilfstheorien enthalten (Schnell et al., 2011; Krebs & Menold, in Druck). Es kann dann geprüft werden, ob die Messwerte, die ein Konstrukt repräsentieren sollen, empirisch die aus der Theorie zu erwartenden Zusammenhänge mit Messwerten für die anderen Konstrukte aufweisen.

Auch wenn die konstruktbezogene Interpretation im Fokus der sozial- und wirtschaftswissenschaftlichen Umfrageforschung steht, können darüber hinaus andere Messwertinterpretationen bedeutsam sein und dementsprechend eine Überprüfung erfordern. *Verallgemeinernde* Interpretationen von Messwerten setzen voraus, dass die verwendeten Fragen oder Aufgaben ein breiteres Inhaltsgebiet (und damit alle denkbaren Fragen oder Aufgaben) angemessen repräsentieren. Dieser Anspruch, der in älteren Arbeiten häufig mit dem Begriff der *Inhaltsvalidität* verbunden ist, wird häufig untersucht, indem die Testinhalte Expertinnen und Experten zur Beurteilung vorgegeben werden. Dies wird z. B. für Schulleistungstests praktiziert, die Lehrplaninhalte oder spezifische Bildungsstandards angemessen abbilden sollen (z. B. Harsch, Pant & Köller, 2010; Pant, Stanat, Pöhlmann & Böhme, 2013). Ein Beispiel aus der sozialwissenschaftlichen Forschung stellen Demografische Standards dar, die durch einen Exper-

tenkreis aus der sozialwissenschaftlichen Methodenforschung und amtlichen Statistik entwickelt werden. Wenngleich dies oft die einzige Möglichkeit ist, empirische Belege für eine verallgemeinernde Interpretation zu gewinnen, werden diese auf Beurteilungen basierenden Untersuchungen oft auch kritisch betrachtet (Guion, 1977; Kane, 2001; vgl. auch Krebs & Menold, in Druck).

Ebenfalls bedeutsam für die sozialwissenschaftliche Umfrageforschung sind *extrapolierende Interpretationen* von Messwerten, nämlich dann, wenn Prognosen angestrebt werden. Der Schluss von Messwerten auf interessierende Phänomene außerhalb des Kontextes eines bestimmten Messinstruments ist mit einem der ältesten Validitätsbegriffe, der *Kriteriumsvalidität*, verbunden (vgl. auch Rammstedt, 2010). Empirische Belege für die Validität derartiger schlussfolgernder Messwertinterpretationen werden häufig über die Zusammenhänge der Messwerte mit den relevanten Außenkriterien untersucht. So kann z. B. der Zusammenhang zwischen den Ergebnissen in einem Leistungstest und Schulerfolg untersucht werden, oder Außenkriterien aus amtlichen Statistiken können verwendet werden, um die Güte von Prognosen auf Basis von Fragebogendaten zu untersuchen. Ebenso kann spezifisches Verhalten (z. B. die Wahl konservativer Parteien) als Kriterium für mit einem Messinstrument erfasste politische Einstellungen verwendet werden. In jedem Fall müssen die gewählten Kriterien in einem sinnvollen Zusammenhang zur angestrebten Interpretation der Messwerte stehen und ihrerseits im spezifischen Kontext reliabel und valide erfasst werden können. Kasten 2 gibt eine Übersicht über eine Auswahl von empirischen Methoden, mit denen Belege für spezifische Messwert-Interpretationen generiert werden können.

Kasten 2: Methoden zur Generierung von Belegen für spezifische Messwert-Interpretationen

Erklärende, konstruktbezogene Interpretationen

- **Untersuchung korrelativer Zusammenhänge** mit Messinstrumenten, die **gleiche oder theoretisch verwandte Konstrukte** erfassen und demzufolge hohe Zusammenhänge mit den Messwerten aufweisen sollten („konvergente Validität“). So korrelieren die Messergebnisse mit der Skala zu politischer Entfremdung von Schwartz (1973) mit Skalen zur Erfassung politischer Unwirksamkeit (*Inefficacy*) positiv und substantiell (Übersicht in Robinson, Shaver & Wrightsman, 1999).
- **Untersuchung korrelativer Zusammenhänge** mit Messinstrumenten, **die Konstrukte erfassen, von denen das Zielkonstrukt abzugrenzen ist** („diskriminante Validität“). Beispielsweise sollten die Ergebnisse zur Erfassung von Konservatismus mit einem Messinstrument gering mit den Ergebnissen von Messinstrumenten

korrelieren, die Liberalismus erfassen (Robinson, Shaver & Wrightsman, 1999).

- Untersuchung der Konsistenz verschiedener Messmethoden (z. B. Papier- und Internetbasierte Befragung, Selbst- und Fremdbeurteilung; unterschiedliche Messinstrumente zur Messung ähnlicher Konstrukte) für dasselbe Konstrukt mit dem **Multitrait-Multimethod-Ansatz** (MTMM) (Campbell & Fiske, 1959). Dabei werden Korrelationen untersucht und miteinander verglichen, die dieselben und unterschiedliche Merkmale mit ähnlichen und unterschiedlichen Methoden (Messinstrumenten) erfassen.
- Untersuchung der aus einer Theorie abgeleiteten **dimensionalen Struktur eines Messinstruments** mit konfirmatorischen Faktorenanalysen oder Methoden der Item-Response-Theorie (z. B. die im Vorfeld definierten zwei Facetten von Nationalstolz – Nationalismus und Patriotismus – werden in einer zweifaktoriellen Struktur repliziert).

Verallgemeinernde Interpretationen

- **Expertenurteile** zur Repräsentation des interessierenden Konstrukts oder Gegenstandsbereiches durch die verwendeten Fragen oder Aufgaben im Messinstrument.
- **Dokumentenanalysen** (z. B. von Lehrplänen) zur Prüfung der Abdeckung eines Gegenstandsbereichs durch die verwendeten Fragen oder Aufgaben.

Extrapolierende Interpretationen

- **Querschnittliche Untersuchung der Zusammenhänge** von Messwerten mit Außenkriterien (z. B. Angaben zum gesunden Essverhalten entsprechen den zeitgleich erhobenen Gesundheitsindikatoren, oder eine Rechts-Links-Skala unterscheidet zwischen Wählern linker und rechter Parteien; Groves et al., 2004; „Übereinstimmungsvalidität“, „konkurrente Validität“).
- **Vorhersage von zeitlich nachgeordneten Ereignissen** mit den Messwerten (z. B. die zu einem Zeitpunkt erhobene Parteipräferenz korrespondiert mit dem späteren Wahlverhalten; „Vorhersagevalidität“, „prädiktive Validität“).
- **Retrospektive Untersuchung der Zusammenhänge** von Messwerten mit tatsächlichen, früher eingetretenen Ereignissen (z. B. Wahlverhalten bei der letzten Bundestagswahl; „retrograde Validität“).
- **Untersuchung der Stabilität von Zusammenhängen** zwischen Messwerten und Außenkriterien in verschiedenen Teilpopulationen (z. B. hängt Schulerfolg bei Jungen und Mädchen gleichermaßen mit Leistungsmotivation zusammen?; „Differenzielle Validität“).

Zusammenfassend sollte die Wahl der Validierungsmethode(n) im Einzelnen durch die Interpretation der Messwerte begründet sein, die für eine spezifische Untersuchung am bedeutsamsten ist. Wenn z. B. die Beschreibung einer Population bezogen auf ein bestimmtes Konstrukt im Mittelpunkt steht, muss belegt werden, dass die Messwerte tatsächlich als Indikatoren für dieses Konstrukt interpretiert werden können. Wenn Prognosen angestrebt werden, sollte belegt werden, dass die hierzu herangezogenen Messwerte tatsächlich mit den vorherzusagenden Phänomenen in Zusammenhang stehen.

Standard 2:

Die für die jeweilige Untersuchung zentrale Interpretation der mit einem Instrument erfassten Messwerte ist explizit formuliert und es sind Belege angeführt, mit denen diese Interpretation gestützt wird.

Leitfragen zur Validität

1. Wird in der Dokumentation zum Messinstrument explizit ausgeführt, wie die mit einem Instrument erfassten Messwerte interpretiert und verwendet werden sollen?
2. Wird in der Dokumentation zum Messinstrument die Argumentation, auf der diese Interpretation aufbaut, nachvollziehbar dargelegt?
3. Werden empirische / und oder theoretische Belege für die Annahmen aufgeführt, auf denen die Argumentation aufbaut?
4. Sind empirische Befunde, die als Belege herangezogen werden, an derselben Zielpopulation gewonnen worden wie diejenige, auf die sich die aktuelle angestrebte Interpretation bezieht?
5. Wird dazu Stellung genommen, welche Annahmen in der Validitäts-Argumentation am wenigsten gut belegt sind?

3.3 Minimierung methodenspezifischer Effekte

Hintergrund

In diesem Abschnitt geht es um die Abschätzung und Minimierung systematischer Messfehler vor dem Hintergrund des Total Survey Error. Systematische Messfehler entstehen, wenn die Antworten aller Befragungsteilnehmerinnen und -teilnehmer oder einer Teilmenge der Be-

fragten systematisch von der Antwort abweichen, die man als eine „ideale Antwort“ erwarten würde.

Grundsätzlich können unterschiedliche Elemente der Realisierung des Messinstruments, z. B. die Itemformulierung oder die Wahl der Antwortformate, zu unerwünschten systematischen Effekten führen (z. B. Dillman et al., 2009; Faulbaum, Prüfer & Rexroth, 2009; Krosnick & Presser, 2010, Tourangeau, Rips & Rasinski, 2000). Auch sogenannte Kontexteffekte zählen zu systematischen Effekten, die durch die Reihenfolge der Fragen im Fragebogen oder durch andere Elemente des Fragebogens (wie z. B. Bilder), entstehen (Sudman, Bradburn, & Schwarz, 1996). Ferner können sich bei der Item- und Fragenformulierung Teilgruppen der Zielpopulation in unterschiedlichem Maße angesprochen fühlen, beispielsweise aufgrund der spezifischen Wortwahl, die das Gebot der vorurteilsfreien Forschung verletzt.

Eine weitere Quelle systematischer Abweichungen stellen unterschiedliche Modi der Datenerhebung dar (face-to-face, Telefon, schriftlich-postalisch, webbasiert). Instruktionen und Hilfestellungen für die Interviewerinnen und Interviewer und für die Befragungsteilnehmerinnen und –teilnehmer dienen der Minimierung der unerwünschten Effekte. Zusätzlich soll auf das Layout der Darbietung des Messinstruments geachtet werden. Die Instruktionen und das Layout sollten in Bezug auf den gewählten Modus der Datenerhebung optimal sein (für Anforderungen und Realisierungsbeispiele siehe Schnell, 2012). Bei paralleler Verwendung unterschiedlicher Erhebungsmodi, sogenannter Mixed-Mode Erhebungen, sollte auf eine einheitliche Darbietung in den verschiedenen Modi geachtet werden. Dieser Anspruch kann jedoch nicht immer erfüllt werden. Beispielsweise müssen je nach Modus spezifische Modifikationen vorgenommen werden, um einen optimalen Frage-Antwort Prozess innerhalb eines Erhebungsmodus zu gewährleisten (Dillman et al., 2009).

Systematische Messfehler können mit Hilfe von kognitiven Pretests oder Split-ballot Experimenten untersucht und aufgedeckt werden. Dabei können Effekte der Frageformulierung, der Wahl der Antwortformate oder auch mögliche Kontexteffekte untersucht werden. Beim kognitiven Pretesting wird mit Hilfe qualitativer Methoden untersucht, wie die Befragungsteilnehmerinnen und -teilnehmer die Items/Fragen verstehen und beantworten. Diese Methode ermöglicht es zu ermitteln, ob Verzerrungen durch Verständnisprobleme oder während der Itembeantwortung entstehen. Bei Split-ballot Experimenten handelt es sich um randomisierte Experimente, bei denen unterschiedliche Varianten eines Messinstruments und ihrer Realisierung miteinander verglichen werden, so dass gegebenenfalls Verzerrungen als Unterschiede in

den empirischen Werten zwischen den untersuchten experimentellen Gruppen sichtbar werden.

In der empirischen Sozialforschung hat sich das kognitive Pretesting als Methode des Testens von Fragebögen durchgesetzt (Faulbaum et al., 2009; Schuman & Presser, 1981). Die Durchführung eines kognitiven Pretests bei der Entwicklung und Optimierung eines Messinstruments stellt daher einen Mindeststandard dar. Ein zusätzlicher Test zum Messinstrument mit Hilfe von Split-ballot Experimenten bietet sich im Rahmen von quantitativen Standard-Pretests an und kann bei vorhandenen Ressourcen zusätzlich zu kognitiven Pretests durchgeführt werden.

Standard 3:

Mögliche methodische Einflüsse, die das Antwortverhalten der Befragungsteilnehmerinnen und -teilnehmer systematisch beeinflussen, sind im Rahmen der Instrumentenentwicklung thematisiert und geprüft und die Ergebnisse sind dokumentiert.

Leitfragen zur Minimierung methodenspezifischer Effekte

1. Wird in der Dokumentation zum Messinstrument beschrieben, wie eine optimale Item-/ Frageformulierung sichergestellt wurde?
2. Sind modusspezifische Instruktionen zum Messinstrument vorhanden? Gibt es Hinweise zum Layout des Messinstruments in einem bestimmten Erhebungsmodus?
3. Sind Untersuchungsergebnisse beschrieben, die die Notwendigkeit modusspezifischer Unterschiede in der Instruktion oder dem Layout eines Messinstruments begründen?
4. Wurde im Rahmen der Instrumentenentwicklung die Verständlichkeit der Items geprüft, bspw. mittels kognitiver Pretests? Werden Methoden und Ergebnisse dieser Prüfung sowie die daraus resultierenden Modifikationen des Messinstruments dokumentiert und begründet?
5. Wurden im Rahmen der Instrumentenentwicklung Split-ballot Experimente durchgeführt? Werden Methoden und Ergebnisse dieser Experimente sowie die daraus resultierenden Modifikationen des Messinstruments dokumentiert und begründet?

3.4 Reliabilität

Hintergrund

Die unsystematische Varianz über Messungen eines Merkmals hinweg reflektiert die Ungenauigkeit der Messung und somit die mangelnde Reliabilität des Messinstruments. Die Reliabilität beschreibt daher die Genauigkeit, mit der ein Messinstrument ein Merkmal misst und ist in der Klassischen Testtheorie als Anteil der Varianz der wahren Werte an der Varianz der beobachtbaren Werte definiert. Bei hoher Reliabilität unterscheiden sich die Ergebnisse bei wiederholten Messungen kaum. Die Korrelation zwischen wiederholten Messungen desselben Merkmals wäre bei perfekter Reliabilität somit 1. Ist das Instrument hingegen gar nicht reliabel, korrelieren die Messungen zu null.

Die Reliabilität ist ein Maß, das angibt, inwieweit sich beobachtbare Unterschiede zwischen Personen auf wahre (messfehlerfreie) Unterschiede zurückführen lassen (Eid & Schmidt, 2014). Wenn bei einer Studie die Untersuchung von Unterschieden zwischen Personen im Fokus steht, werden daher entsprechend hohe Anforderungen an die Reliabilität der Messinstrumente gestellt. Die Reliabilität eines Messinstruments ist auch für das Kriterium der Validität (s. Abschnitt 3.2) von großer Bedeutung, da die Interpretationen von Messwerten von deren Genauigkeit abhängen (Kane, 2013).

Die Bestimmung der Reliabilität setzt voraus, dass ein Merkmal mehrmals gemessen wird. Hierzu kann man sich unterschiedlicher Verfahren bedienen (Eid & Schmidt, 2014; Groves et al., 2004; Rammstedt, 2010; Schnell et al., 2011):

- Test-Retest-Methode (auch Reinterview): Das Messinstrument wird derselben Stichprobe zweimal in einem angemessenen Zeitabstand unter vergleichbaren Bedingungen vorgelegt. Dies setzt voraus, dass das zu messende Merkmal stabil ist, sonst ist die unsystematische „Unreliabilität“ der Messung mit der systematischen Instabilität des Merkmals konfundiert.
- Paralleltest-Methode: Erfassung des Merkmals mit zwei verschiedenen, jedoch äquivalenten (oder parallelen) Messinstrumenten.
- Split-half-Methode (nur bei Messinstrumenten, die aus mehreren Items bestehen): Aufteilung der Items eines Messinstruments in zwei Teile.
- Interne Konsistenz (nur bei Messinstrumenten, die aus mehreren Items bestehen). Die Konsistenzanalyse stellt eine Erweiterung der Split-Half-Methode dar, in dem nicht nur zwei Hälften eines Messinstruments gebildet werden, sondern sämtliche Items betrachtet werden.

Die Vielfalt der Verfahren trägt der Unterschiedlichkeit der zu messenden Merkmale und den Unterschieden bei der Realisierung der Messinstrumente Rechnung. Auf Grundlage dieser Verfahren kann die Reliabilität mit unterschiedlichen Koeffizienten bestimmt werden. Diese Koeffizienten setzen voraus, dass bestimmte Anforderungen an die wiederholten Messungen erfüllt sind. Diese werden in Messmodellen formuliert und lassen sich mit empirischen Tests überprüfen (Eid & Schmidt, 2014). Die Auswahl eines Messmodells hängt von dem Skalenniveau der Items und derjenigen des zu messenden Konstrukts (latente Variable) ab. Man kann hierzu auf Modelle der Item-Response-Theorie (z. B. Latent-Trait-Modell), Latent-Class-Modelle und auf Modelle der konfirmatorischen Faktorenanalysen zurückgreifen (Eid & Schmidt, 2014; Rost, 2004). Sind z. B. die wiederholten Messungen mindestens essentiell τ-parallel (d. h. sie messen eine einzelne Facette bzw. Dimension eines Merkmals mit gleicher Diskrimination und Fehlervarianz), kann die Retest-, Paralleltest- und Split-half-Reliabilität einfach anhand der Korrelation beider Messungen bestimmt werden. Für die in Umfragen verbreiteten Messungen eines Merkmals mit lediglich einer Frage oder einem Item, kommen die Retest- und Paralleltest-Methode in Frage. Die Einschränkungen dieser Methoden bestehen darin, dass die Retestmethode nur bei zeitlich stabilen Merkmalen verwendet werden kann; bei der Paralleltest-Methode ergibt sich die Schwierigkeit der Konstruktion einer weiteren Frage, die eine äquivalente Messung ermöglicht. Als Mindeststandard sollte die Reliabilität mit mindestens einer Methode überprüft und berichtet werden, wobei die Voraussetzungen für die Wahl der Methode der Reliabilitätsbestimmung erfüllt sein sollten. Unterschiedliche Methoden der Reliabilitätsschätzung sind u.a. bei Eid und Schmitt (2014) ausführlicher dargestellt.

Standard 4:

Eine Prüfung der Reliabilität ist erfolgt, die gewählte Art der Reliabilitätsprüfung ist begründet und die Reliabilitätskennwerte sind dokumentiert und bewertet.

<u>Leitfragen zur Reliabilität</u>

1. Werden Reliabilitätswerte zum Messinstrument berichtet?
2. Wird die Wahl des Verfahrens inklusive des entsprechenden Maßes für die Reliabilitätsbestimmung begründet?

3. Inwieweit ist gesichert, dass die Voraussetzungen einer Methode zur Bestimmung der Reliabilität erfüllt sind?

4. Werden die Stichprobe und die Untersuchung, die zur Ermittlung der Reliabilität verwendet wurde, beschrieben?

5. Wird das Ergebnis der Reliabilitätsüberprüfung bewertet und werden Hinweise in Bezug auf die Verwendung des Messinstruments und die Ergebnisinterpretation gegeben? Werden bei geringer Reliabilität Einschränkungen im Hinblick auf die Verwendung des Messinstruments genannt?

3.5 Prozessfehler

Hintergrund

Die Prozessfehler, die die Qualität eines Messinstruments beeinträchtigen können, können zum einen während der Datenerhebung, zum anderen durch Übertragungsfehler der Daten und im Zuge der Datenbearbeitung entstehen. Im Einzelnen wird zumeist unterschieden zwischen Interviewereffekten, Effekten der Datenerhebungssituation aufgrund des gewählten Datenerhebungsmodus, Kodierfehlern und Fehlern bei der Datenbereinigung und Gewichtung (Groves et al., 2004). Anstelle des Begriffs Prozessfehler wird in der Psychologie in diesem Kontext der Begriff der Objektivität verwendet.

Eine Verringerung von Prozessfehlern bzw. eine Erhöhung der Objektivität wird durch eine möglichst weitgehende Standardisierung des Messinstruments (einschließlich seiner Instruktionen, Auswertungs- und Interpretationshinweise) und der Erhebungsbedingungen erreicht. Bei der Durchführung von Befragungen sollte daher darauf geachtet werden, dass die Erhebungssituation möglichst gering zwischen den befragten Personen variiert. Bei der Verwendung von Messinstrumenten mit multiplen Indikatoren sollten Auswertungshinweise gegeben werden, die detailliert beschreiben, wie die verschiedenen Indikatoren zu einem Index aggregiert werden. Die entsprechenden Voraussetzungen zur Indexbildung sollten im Rahmen der Validierung und der Reliabilitätsprüfung von den Instrumentenentwicklerinnen und -entwicklern sichergestellt werden. So sollte bei einer Zusammenfassung der einzelnen Indikatoren zu einem Skalenwert (Mittelwert oder Summe) die Voraussetzung der Eindimensionalität erfüllt werden (Bühner, 2011; Schnell et al., 2011).

Standard 5:

Standardisierte Instruktionen und Verfahrenshinweise zur Durchführung der Datenerhebung und -auswertung sind vorhanden und begründet.

Leitfragen zur Prüfung des Prozessfehlers

1. Sind *modusspezifische* Instruktionen zur Durchführung der Datenerhebung mit dem Messinstrument vorhanden?
2. Sind die Instruktionen und Verfahrenshinweise zur Durchführung der Datenerhebung mit dem Messinstrument so gestaltet, dass verschiedene Personen in der Lage sind, das Messinstrument allein aufgrund dieser Verfahrenshinweise in vergleichbarer Weise anzuwenden?
3. Sind Regeln angegeben, wie mit vorhersehbaren Nachfragen der Befragten/Interviewer umgegangen wird?
4. Sind ggfs. spezifische technische Vorgaben/Anforderungen (bei Computerunterstützung: Computer Assisted Personal Interviewing [CAPI], Computer Assisted Telephone Interviewing [CATI], Computer Assisted Self-Interviewing [CASI], Computer Assisted Web Interviewing [CAWI)]) aufgeführt?
5. Sind ggfs. die möglichen apparativen Störungen bei der Durchführung (je nach Modus) und Möglichkeiten ihrer Behebung angegeben?
6. Sind Instruktionen zur Indexbildung bei Messinstrumenten mit multiplen Indikatoren enthalten? Wird berichtet, inwieweit die Voraussetzungen für die Indikatorzusammenfassung oder Indexbildung erfüllt sind?

3.6 Weitere Qualitätsmerkmale

Hintergrund

Über die dargestellten Qualitätskriterien hinaus sind weitere in der Psychologie etablierte Qualitätsmerkmale, die sogenannten Nebengütekriterien, auch für Instrumente von Bedeutung, die in sozial- und wirtschaftswissenschaftlichen Umfragen Anwendung finden. Im Vordergrund dieser weiteren Qualitätsmerkmale steht nicht die Qualität der Messung, sondern die mit der Anwendung eines Messinstruments verbundenen Voraussetzungen, nämlich die Ökonomie, die Zumutbarkeit und die Aktualität des Instruments[8].

Die *Ökonomie* eines Messinstruments wird darüber bestimmt, wie lang bzw. kurz dessen Durchführungszeit und wie einfach dessen Handhabung ist. So wäre bei vergleichbaren Messinstrumenten mit ähnlichem Inhaltsbereich und gleicher Messqualität eine Skala zu bevorzugen, die weniger Items enthält und hierdurch schneller zu beantworten ist. Im Kontext der sozial- und wirtschaftswissenschaftlichen Umfrageforschung ist der Aspekt der Ökonomie von zentraler Bedeutung, da die Durchführung von Bevölkerungsumfragen mit hohen Kosten verbunden ist und auf Seiten der Befragten einen beträchtlichen Zeitaufwand für Forschungszwecke bedeutet. In dem stark begrenzten Zeitrahmen einer Umfrage müssen daher nach Möglichkeit mehrere Inhaltsbereiche abgefragt werden. Hierzu sind Messinstrumente notwendig, die trotz reduzierter Anzahl der Items und Fragen eine hohe Messqualität gewährleisten.[9]

Die *Zumutbarkeit* beschreibt, dass für die Befragungsteilnehmerinnen und -teilnehmer durch das Messinstrument keine unnötige Belastung körperlicher oder psychischer Art entsteht.

---

[8] Ein weiteres Nebengütekriterium ist die Nützlichkeit. Sie beschreibt, inwieweit ein praktisches Bedürfnis für ein Erhebungsinstrument besteht. Die Nützlichkeit ist dann gering, wenn ein neues Instrument Inhalte misst, die mit bereits existierenden Erhebungsinstrumenten erhoben werden können. Ein entsprechender Qualitätsstandard wurde in dem vorliegenden Dokument bereits im Zusammenhang mit der Instrumentenentwicklung eingeführt. (s. Standard 1, Leitfrage 7).

[9] In Kooperation mit dem Deutschen Institut für Wirtschaftsforschung (DIW) wurden im Rahmen eines Projekts bei GESIS - Leibniz-Institut für Sozialwissenschaften standardisierte Kurzskalen zur Erfassung psychologischer Merkmale entwickelt. Die Kurzskalen inklusive ihrer Dokumentation sind für interessierte Wissenschaftler/-innen frei verfügbar und können unter folgender Adresse abgerufen werden: http://www.gesis.org/kurzskalen-psychologischer-merkmale/kurzskalen/ (Zugriff am 9.1.2013).

Entwicklerinnen und Entwickler von Messinstrumenten sollten sich zur Sicherung der Zumutbarkeit an den in den Sozialwissenschaften etablierten ethischen Standards orientieren.[10]

Die *Aktualität* bezieht sich auf die empirischen Ergebnisse zur Sicherung der Messqualität. Die Angaben zur Validität und zu den weiteren Aspekten der Messqualität eines Instruments sollen aktuell sein beziehungsweise in regelmäßigen Rhythmen aktualisiert werden. Diese Rhythmen sollten in der Regel acht Jahre nicht überschreiten (vgl. DIN 33430 in DIN 2002; Kersting, 2008).

Standard 6:

Angaben zur Ökonomie, zur Zumutbarkeit des Instruments und zur Aktualität der psychometrischen Kennwerte sind vorhanden.

Leitfragen zur Prüfung der Ökonomie, Zumutbarkeit und Aktualität

1. Sind die Durchführungszeit und der Auswertungsaufwand genannt? Inwieweit stehen diese in einer sinnvollen Relation zur Messintention?
2. Ist die Datenerhebung mit dem Messinstrument für die Befragungsteilnehmerinnen und -teilnehmer schadensfrei? Können aufgrund der Datenerhebung mit dem Messinstrument körperliche und/oder psychische Belastungen der befragten Personen entstehen?
3. Sind die Zeitangaben zu den letzten Untersuchungen zur Messqualität vorhanden?

[10] Ethik-Kodex der Deutschen Gesellschaft für Soziologie (DGS) und des Berufsverbandes Deutscher Soziologen (BDS): http://www.soziologie.de/index.php?id=19; ADM Codex: http://www.adm-ev.de/fileadmin/user_upload/PDFS/Erklaerung_2008.pdf (Zugriff am 8.1.2014).

4 Literatur

Bühner, M. (2011). Einführung in die Test- und Fragebogenkonstruktion. München: Pearson Studium.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the Multimethod-Multitrait Matrix. *Psychological Bulletin, 56*, 833-853.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Diekmann, A. (2007). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen* (18. Auflage). Reinbek bei Hamburg: Rowohlt.

Dillman, D. A., Smyth, J. D. & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys. The tailored design method*. Wiley: New Jersey.

DIN. (2002). DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen. Berlin: Beuth.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion.* Göttingen: Hogrefe.

Faulbaum, F., Prüfer, P. & Rexroth, M. (2009). Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität. Wiesbaden: VS Verlag.

Ganzeboom, H. B. G., De Graaf, P. M. & Treiman, D. J. (1992): A Standard International Socio-Economic Index of Occupational Status. *Social Science Research, 21* (1), 1-56.

Groves, R. M, Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2004). *Survey methodology*. New Jersey: Wiley.

Groves, R. M. & Lyberg, L. (2010). Total Survey Error: Past, present and future. *Public Opinion Quarterly, 74*, 849-879.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10.

Häcker, H., Leutner, D. & Amelang, M. (Hrsg.) (1998). *Standards für pädagogisches und psychologisches Testen.* Supplementum 1/1998 der Diagnostica und der Zeitschrift für Differentielle und Diagnostische Psychologie. Bern: Hogrefe.

Harsch, C., Pant, H. A. & Köller, O. (Eds.) (2010). Calibrating standards-based assessment tasks for English as a first foreign language: standard-setting procedures in Germany. Münster: Waxmann

Hartig, A., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion. 2. Auflage* (S. 143-171). Heidelberg: Springer Verlag.

Hussy, W., Schreier, M. & Echterhoff, G. (2010). *Forschungsmethoden in Psychologie und Sozialwissenschaften*. Heidelberg: Springer-Verlag.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Kane, M. T. (2013). Validating the interpretations and the uses of test scores. *Journal of Educational Measurement, 50,* 1-73.

Kersting, M. (2008). Qualität in der Diagnostik und Personalauswahl - Der DIN Ansatz. Göttingen: Hogrefe.

Krebs, D. & Menold, N. (in Druck). Gütekriterien quantitativer Sozialforschung. In J. Blasius & N. Baur (Hrsg.). *Methoden der Sozialforschung*.

Krosnick, J. A. & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), Survey measurement and process quality (pp. 141-164). New York: Wiley.

Krosnick, J. A. & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of Survey Research* (2nd edition, pp. 263-313). West Yorkshire, England: Emerald Group.

Lienert, G. A. & Raatz, U. (1998) *Testaufbau und Testanalyse.* Weinheim: Beltz PVU.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition, pp.13-103). New York: American Council on Education/Macmillan.

OECD (2013a). PISA 2013 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy. Paris: OECD.

OECD (2013b). OECD Skills Outlook 2013: First results from the Survey of Adult Skills. Paris: OECD.

Pant, H. A., Stanat, P., Pöhlmann, C. & Böhme, K. (2013). Die Bildungsstandards im allgemeinbildenden Schulsystem. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 13–22). Münster: Waxmann.

Prenzel, M., Sälzer, C., Klieme, E. & Köller, O. (2013). *PISA 2012: Fortschritte und Herausforderungen in Deutschland.* Münster: Waxmann.

Rammstedt, B. (2010). Messen und Skalieren in den Sozialwissenschaften, Gütekriterien (Reliabilität, Validität, Objektivität). In C. Wolf & H. Best (Hrsg.) *Handbuch der sozialwissenschaftlichen Datenanalyse (S.239-258).* Wiesbaden: VS Verlag.

Rammstedt, B. (Hrsg.) (2013). Grundlegende Kompetenzen Erwachsener im internationalen Vergleich - Ergebnisse von PIAAC 2012. Münster: Waxmann.

Robinson, J. P., Shaver, P. R., Wrightsman, L. S. (1999). *Measures of political attitudes*. San Diego: Academic Press.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2. Aufl.). Bern: Huber.

Schnell, R. (2012). Survey-Interviews. Methoden standardisierter Befragungen. Wiesbaden: VS-Verlag.

Schnell, R., Hill, P. B. & Esser, E. (2011). *Methoden der empirischen Sozialforschung.* München: R. Oldenbourg Verlag.

Schuman, H. & Presser, S. (1981). Questions and answers in attitude surveys: Experiments in question form, wording, and context. New York: Academic Press.

Schwartz, D. C. (1973). Political alienation and political behavior. Chicago: Aldine.

Schwartz, S. & Bilsky, W. (1990). Toward a theory of the universal content and structure of values: Extensions and cross cultural replications. *Journal of Personality and Social Psychology, 58*, 878-891.

Sudman, S., Bradburn, N. M. & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The psychology of survey response.* Cambridge: Cambridge University Press.

Anhang A

Quellen zur Qualitätssicherung im Bereich der Repräsentation der Umfragedaten

Rahmenfehler und Stichprobenfehler:

ADM Arbeitskreis Dt. Markt- u. Sozialforschungsinst. e.V.; AG.MA Arbeitsgemeinschaft Media-Analyse e.V. (Hrsg.). (1999). *Stichproben-Verfahren in der Umfrageforschung: eine Darstellung für die Praxis.* Opladen: Leske und Budrich.

Arbeitsgruppe Regionale Standards (Hrsg.) (2013). *Regionale Standards*, Band 12, GESIS Schriftenreihe: http://www.gesis.org/unser-angebot/studien-planen/demographische-und-regionale-standards/

Gabler, S. (2010): Stichprobenziehung. In H. Holling & B. Schmitz (Hrsg.). *Handbuch Statistik, Methoden und Evaluation. Handbuch der Psychologie*, Bd. 13 (S. 27-36). Göttingen: Hogrefe.

Gabler, S. & Ganninger, M. (2010). Gewichtung. In Ch. Wolf & H. Best (Hrsg.). *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 143-164). Wiesbaden: VS Verl. für Sozialwissenschaften.

Fehler durch Ausfälle:

AAPOR (Ed.) (2011). *Standard Definitions. Final Dispositions of Case Codes and Outcome Rates for Surveys.* Verfügbar unter: http://aapor.org/Content/NavigationMenu/AboutAAPOR/StandardsampEthics/StandardDefinitions/StandardDefinitions2011.pdf [Zugriff am 30.01.2014].

Groves, R. M. & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.