Antwortskalen in standardisierten Befragungen

51

Axel Franzen

51.1 Einleitung

Jeder, der eine standardisierte Befragung plant, muss sich früher oder später überlegen, welche Fragen gestellt werden sollen, wie die Fragen formuliert sein sollen und wie die Antwortkategorien zu gestalten sind. Die Antworten auf diese drei Fragen sind nicht unabhängig, sondern hängen eng miteinander zusammen. In der Regel ist es sinnvoll in der genannten Reihenfolge vorzugehen, also sich erst mit den Inhalten und der Formulierung von Fragen zu beschäftigen und danach die Konstruktion der Antwortskalen vorzunehmen. Ein Forscher ist aber gut beraten, schon bei der Formulierung der Fragen an die spätere Datenauswertung zu denken, und spätestens bei der Festlegung der Antwortkategorien stehen Überlegungen zum Skalenniveau und den statistischen Analysemöglichkeiten im Vordergrund. Deshalb ist es durchaus sinnvoll, sich unabhängig von den Inhalten und Formulierungen der Fragen mit den Konstruktionsmöglichkeiten von Antwortkategorien zu beschäftigen. Die Wahl der Antwortskalen beeinflusst nämlich die Formulierung der Fragen. Aus diesem Grund ist es auch sinnvoll, allgemeine Überlegungen zur Formulierung von Fragen und zur Konstruktion der Antwortskalen in zwei Kapitel zu trennen, wie das in diesem Handbuch erfolgt (Porst, Kapitel 50 in diesem Band).

Die Fragen, die man in ein Erhebungsinstrument aufnimmt, hängen natürlich vom Untersuchungsgegenstand, dem Befragungsmodus und der Zielpopulation ab, an die sich die Befragung wendet. Fragebögen sind deshalb vor allem an die spezifischen Untersuchungszwecke anzupassen und das gilt natürlich auch für die Antwortskalen. Es ist deshalb nicht ganz einfach, allgemeine Regeln zur Konstruktion von Antwortskalen zu formulieren. Vielleicht wird das Thema deshalb in den einschlägigen Lehrbüchern zur empirischen Sozialforschung auch sehr kurz gehalten (Diekmann 2011; Schnell et al. 2011), lediglich die auf die Umfrageforschung spezialisierten Lehrbücher von Faulbaum et al. (2009) und Schnell (2012) beschäftigen sich ausführlicher mit Antwortskalen. In einer älteren Auflage des Lehrbuchs von Schnell et al. findet sich sogar die Aussage „Die Literatur zu

Response-Errors umfasst die langweiligsten Texte der empirischen Sozialforschung“ (2008: 420 [8. Auflage]). Vielleicht haben die Autoren sogar recht. Aber nichtsdestotrotz stehen Forscher häufig vor der Frage, wie eine Antwortskala auszusehen hat, um möglichst geringe Antwortverzerrungen auszulösen und dieses Kapitel versucht, hier einige Hilfestellungen zugeben.[1]

Bei der Durchführung einer Befragungsstudie kommen häufig weitere Aspekte hinzu, die man in Betracht ziehen sollte. Viele Untersuchungen werden erst dann so richtig interessant, wenn sie einen Trend beschreiben oder die Beobachtung der Untersuchungseinheiten über einen gewissen Zeitraum erlauben. Ein Trend- oder Paneldesign erfordert natürlich, dass die exakt gleichen Frageformulierungen und die exakt gleichen Antwortskalen wiederholt werden (Mochmann, Schupp, Kapitel 14 und 73 in diesem Band). Ansonsten ist es unmöglich, die realen Veränderungen von den methodisch induzierten Veränderungen zu trennen. Untersuchungen gewinnen in der Regel auch durch die Möglichkeit internationaler Vergleiche. Es kann sehr aufschlussreich sein, Trends einer Organisation oder einer Gesellschaft mit den Veränderungen in anderen Gesellschaften oder Organisationen zu vergleichen. Überlegungen zur Vergleichbarkeit können und sollten daher die Wahl von Antwortskalen mitbestimmen und können dazu führen, dass suboptimale Antwortskalen übernommen werden bzw. den Konstrukteur eines neuen Fragebogens mit schwierigen Abwägungsproblemen konfrontieren.

Sind die Probleme und Ziele hinsichtlich zeitlicher und internationaler Vergleichbarkeit geklärt, dann stellen sich bei der Konstruktion von Antwortskalen mindestens sechs Fragen:

1. Wie viele Antwortkategorien soll die Skala enthalten?
2. Soll eine gerade oder ungerade Anzahl verwendet werden?
3. Ist die Beschriftung aller Kategorien besser als nur die Bezeichnung der Endpunkte?
4. Sollte die Skalenbeschriftung bipolar oder unipolar erfolgen?
5. Sollte eine Skala positiv (zustimmend) oder negativ (ablehnend) beginnen?
6. Sollte die Verbalisierung itemspezifisch oder standardisiert vorgenommen werden?

Die Methodenforschung hat sich intensiv mit diesen Fragen beschäftigt. Dabei geht es darum, das Abschneiden unterschiedlicher Varianten von Skalen in Hinblick auf die zentralen Gütekriterien Objektivität, Reliabilität und Validität empirisch zu testen (Krebs/Menold, Kapitel 30 in diesem Band). Objektivität ist in standardisierten Befragungsstudien in der Regel kein Problem. Es gibt zwar in mündlichen Befragungen Interviewereffekte (Glantz/Michael, Kapitel 21 in diesem Band), aber es gibt keine plausiblen Hinweise, dass solche Probleme durch die Gestaltung von Antwortskalen wesentlich beeinflusst werden. Antwortskalen können sich dagegen in Bezug auf ihre Reliabilität unterscheiden. Gibt man den Befragten die Möglichkeit, ihre Meinungen anhand vielstufiger Skalen differenziert zu äußern, dann sollte eigentlich auch die Gefahr größer sein, dass bei Messwiederholungen

[1] Offensichtlich haben die Autoren ihre Meinung geändert. Wenigstens fehlt die zitierte Aussage in der überarbeiteten 9. Auflage.

Abweichungen auftreten und die Test-Retest-Reliabilität im Vergleich zu wenig Antwortmöglichkeiten geringer ausfällt (Eifler, Kapitel 11 in diesem Band). Erstaunlicherweise lässt sich aber empirisch das Gegenteil beobachten: Mehr Antwortkategorien erhöhen in der Regel die Reliabilität.

Weil Antwortskalen den Befragten Hinweise zur Interpretation der Fragen liefern können, dies allerdings auch in Bezug auf die vermeintlich erwünschte Antwort, beeinflussen sie natürlich auch die Validität (Krebs/Menold, Kapitel 30 in diesem Band). Allerdings sind Untersuchungen zur Reliabilität sehr viel einfacher zu bewerkstelligen (z.B. durch einfache Messwiederholungen) als Studien zur Validität, und letztere sind daher eher selten. Darüber hinaus lässt sich aber auch vermuten, dass die Validität eines Messinstruments in der Regel stärker von der Frage selbst und ihrer Formulierung abhängt als von der verwendeten Antwortskala. So gesehen erscheinen Tests zur Reliabilität auch aus inhaltlichen Erwägungen in der Methodenforschung zu Antwortskalen den wichtigsten Stellenwert einzunehmen.

Der zweite Abschnitt des Kapitels versucht, die wichtigsten Forschungsergebnisse zu den oben beschriebenen sechs Fragen zusammenzufassen. Einige Ergebnisse scheinen hinreichend konsistent und zuverlässig zu sein, so dass sich daraus Regeln für die Konstruktion von Antwortskalen formulieren lassen. Diese Regeln werden in der Zusammenfassung beschrieben.

51.2 Die Gestaltung von Antwortkategorien

51.2.1 Offene versus geschlossene Fragen

Antwortskalen liefern den Befragten Hinweise zur Interpretation und der Bedeutung von Fragen. Darüber hinaus können sie auch Hinweise über die vermeintlich sozial erwünschte Antwort liefern und damit die Verzerrung des Antwortverhaltens beeinflussen. Ein bekanntes Beispiel stammt aus einer Studie von Schwarz et al. (1985). In einem randomisierten Labor-Experiment wurden den Befragten zwei Versionen von Antwortskalen bei einer Frage nach ihrem täglichen Fernsehkonsum vorgelegt. In einer Version begann die vorgegebene Antwortskala mit einem kleinen Intervall (0 bis ½ Stunde, mehr als ½, aber weniger als 1 Stunde etc.), in der anderen Version mit einem großen Intervall. In der letzteren Version ergab sich ein bedeutend höherer berichteter Fernsehkonsum als bei der ersten Variante. Weil die Schwarz-Studie schon etwas in die Jahre gekommen ist, kleine Fallzahlen aufweist und in einem Labor stattfand, haben wir die Studie in einer Online-Befragung unter Studierenden der Universität Bern im Jahr 2012 wiederholt. Die Frage lautete „Wie viele Stunden pro Woche sind Sie durchschnittlich sportlich aktiv? (Zählen Sie bitte auch Fortbewegungen mit dem Velo und zu Fuß dazu)“. Die Frage war in einen Online-Fragebogen zur Studienmotivation und zum Studierverhalten integriert, wobei je eine unterschiedliche Version von Antwortmöglichkeiten randomisiert den Befragten vorgelegt wurde. In der ersten Variante begannen die Antwortkategorien mit „bis zu einer

Stunde", „ca. zwei Stunden" und so weiter bis zu „mehr als 5 Std.". In der zweiten Variante startete die Skala mit „bis zu 5 Stunden", „ca. 6 Stunden" und so weiter bis zu „mehr als 9 Stunden". Die Ergebnisse sind in Tab. 51.1 dargestellt.

Während in der ersten Variante 78% der Befragten angaben, bis zu 5 Stunden sportlich aktiv zu sein, sind es in der zweiten Variante 44,5%, ein Unterschied von mehr als 30 Prozentpunkten. Die Studie repliziert damit die Ergebnisse der Schwarz-Studie und zeigt darüber hinaus, dass Antwortskalen auch außerhalb von Labors und selbst bei nichtheiklen Fragen die Befragungsergebnisse beeinflussen können. Offensichtlich orientieren sich Befragte nicht nur bei sensitiven Fragen an der mittleren Kategorie einer vorgegebenen Antwortskala, weil sie dieses Verhalten als erwünscht oder erwartet interpretieren. Falls möglich, sollte aus diesem Grund auf die Vorgabe von Antwortkategorien verzichtet und stattdessen offenen numerischen Antwortmöglichkeiten der Vorzug gegeben werden. Für die offene numerische Abfrage spricht auch, dass die natürlichen Einheiten angegeben werden und damit die Variable auf dem höchstmöglichen Skalenniveau gemessen wird. Verhaltensfragen wie nach der Anzahl der Theaterbesuche im letzten Halbjahr oder der Anzahl der sozialen Kontakte sollten daher offen gestellt werden. Offene numerische Antwortmöglichkeiten liefern in diesen Fällen genaue Daten auf Absolutskalenniveau.

Die Vermeidung von Effekten der sozialen Erwünschtheit und die Realisierung des höchstmöglichen Skalenniveaus würden auch bei der Frage nach dem Einkommen eine offene Abfrage empfehlenswert erscheinen lassen. Allerdings treten bei der Einkommensfrage in der Regel die höchsten Antwortverweigerungen in einer Befragung auf. Offene numerische Abfragen können den Eindruck erwecken, dass es der Interviewer ganz genau wissen will und sie können damit die Verweigerungsrate bei der Einkommensfrage noch erhöhen (Engel/Schmidt, Kapitel 23 in diesem Band). Die Vermeidung von Item-Nonresponse kann natürlich ein sehr guter Grund sein, bei der Einkommensfrage doch geschlossene Antworten vorzugeben.

Tab. 51.1 Effekte von Antwortskalen: Wie viele Stunden pro Woche sind Sie durchschnittlich sportlich aktiv?

Version 1	N	in %	Version 2	N	in %
bis zu 1 Std.	167	11,1	bis zu 5 Std.	690	44,5
ca. 2 Std.	287	19,1	ca. 6 Std.	280	18,1
ca. 3 Std.	305	20,3	ca. 7 Std.	213	13,7
ca. 4 Std.	210	14,0	ca. 8 Std.	118	7,6
ca. 5 Std.	201	13,4	ca. 9 Std.	78	5,0
mehr als 5 Std.	330	22,0	mehr als 9 Std.	171	11,0
Summe	1500	100%		1550	100%

51.2.2 Zur Anzahl der Antwortkategorien

Bei vielen Fragen sind allerdings offene numerische Antwortkategorien nicht einsetzbar. Dies trifft auf Fragen zu, bei denen es um die Erhebung von Einstellungen, Werten, Meinungen oder Einschätzungen geht. So wäre zum Beispiel die Frage nach der Sorge um den Umweltschutz oder nach der wirtschaftlichen Lage eines Landes mit einer offenen numerischen Antwort nur schwer umsetzbar. Außerdem ist hier im Unterschied zur Messung des Einkommens oder des Fernsehkonsums eine hohe Genauigkeit oft nicht möglich und auch nicht immer notwendig. Aus diesem Grund sind solche Fragen mit geschlossenen Antwortkategorien zu versehen. Hierbei wird zwischen ungeordneten und geordneten Antwortvorgaben unterschieden. Ein Beispiel für Fragen mit ungeordneten Antwortkategorien ist die Frage „Was machen Sie in Ihrer Freizeit?", die dann von den Antwortvorgaben „Sport", „lesen", „Musik" etc. gefolgt wird. Aus ungeordneten Antwortvorgaben resultieren in der Regel, wie auch in diesem Fall, Nominalskalen. Die Gestaltung von ungeordneten Antwortkategorien lässt sich kaum vom Inhalt der Frage trennen. Bei ungeordneten Antwortkategorien gelten nur drei Regeln, nämlich dass die Antwortkategorien erschöpfend, disjunkt und nicht zu umfangreich sein sollten.

Anders sieht es dagegen bei geordneten Antwortvorgaben aus, bei denen in der Regel das Ausmaß an Zustimmung oder Ablehnung zu einer Aussage, oder die Intensität der Beliebtheit oder des Vertrauens von Personen oder Organisationen gemessen wird. Hierbei stellt sich die Frage, wie viele Abstufungen geschlossene Antwortskalen (auch Ratingskalen genannt) enthalten sollen. In der Praxis (z.B. im ALLBUS 2012) sind fast alle Möglichkeiten anzutreffen, beginnend mit zwei Antwortkategorien über dreistufige, vierstufige, fünfstufige oder bis zu 10er und 100er Skalen. Im World Values Survey (WVS) werden Einstellungen und Werte überwiegend mittels vierstufiger Skalen erhoben. Im International Social Survey Programme (ISSP) finden sich dagegen hauptsächlich fünfstufige Antwortskalen, was die Vergleichbarkeit der Ergebnisse natürlich erschwert (Franzen/Vogl 2013).

In der methodischen Forschung wird vor allem untersucht, welche Skalen in Test-Retest-Untersuchungen die höchste Reliabilität aufweisen. Zunächst erscheint es plausibel, dass eine hohe Anzahl an Antwortkategorien den Befragten mehr Optionen eröffnet und eine feinere Abstufung und damit eine genauere Messung zulässt. Viele Abstufungen erhöhen aber auch die kognitiven Anforderungen an die Befragten. Nach Miller (1956) kann die optimale Anzahl an Antwortkategorien durch die Regel „sieben plus/minus zwei" beschrieben werden. Dieses Ergebnis wird auch durch neuere Studien bestätigt (Preston/Colman 2000, Kieruj/Moors 2010, Svensson 2000). Abgeraten wird demnach von Skalen mit weniger als fünf Antwortkategorien, die paradoxerweise über eine geringere Reliabilität verfügen. Skalen mit mehr als 7 oder 9 Abstufungen bringen dagegen keine weiteren Vorteile. Zusätzlich wird davon ausgegangen, dass Befragte mit höherem Bildungshintergrund besser in der Lage sind, mit ausdifferenzierten Skalen umzugehen. Für allgemeine Bevölkerungsbefragungen sollten dagegen eher weniger Kategorien verwendet werden (Weijters et al. 2010). Einen Einfluss auf die Abstufungen hat dabei auch der Befragungsmodus. Schriftliche Befragungen erlauben mehr Antwortkategorien, während

in mündlichen (face-to-face) oder telefonischen Befragungen die Teilnehmer besser mit weniger Kategorien zurechtkommen. Alles in allem sprechen die einschlägigen Ergebnisse dafür, in allgemeinen Bevölkerungsbefragungen fünfstufige oder siebenstufige Skalen zu verwenden (vgl. auch Faulbaum et al. (2009) für eine ähnliche Empfehlung).

51.2.3 Gerade oder ungerade Antwortskalen?

Ausführlich wird in der Literatur auch die Frage diskutiert, ob eine gerade oder ungerade Anzahl von Antwortkategorien die Reliabilität erhöht. Eine gerade Anzahl von Antwortkategorien zwingt die Befragten zu einer eher zustimmenden oder ablehnenden Entscheidung. Ungerade Antwortmöglichkeiten enthalten dagegen eine Mitte. Diese ist allerdings nicht immer eindeutig zu interpretieren. Sie könnte auch von Befragten gewählt werden, die zum Thema keine Meinung haben, was dann zu Fehlmessungen führen würde. Eine gerade Anzahl birgt dagegen die Gefahr, der Akquieszenz Vorschub zu leisten, also der Tendenz der Befragten, eine Frage im Zweifelsfall zu bejahen. Auch dies würde einen Messfehler verursachen. Empirisch sprechen die Ergebnisse eher für ungerade Skalen. Bei einem Vergleich zwischen 5-stufigen und 6-stufigen Antwortkategorien konnte Moors (2008) allerdings keine wesentlichen Unterschiede erkennen. O'Muircheartaigh et al. (2000) berichten dagegen eine bessere Reliabiliät für Skalen mit einer mittleren Kategorie.

Um zu vermeiden, dass Befragte die mittlere Kategorie wählen, weil sie damit ihre Meinungslosigkeit zum Ausdruck bringen wollen, wird gelegentlich die explizite Verwendung einer „weiß nicht" Kategorie empfohlen (Converse/Presser 1986). Dies wird in der einschlägigen Literatur aber sehr kritisch gesehen (vgl. den ausführlichen Literaturüberblick in Krosnick 1999). Eine explizite „weiß nicht" Kategorie erhöht den Anteil an Befragten, die angeben, keine Meinung zu einem Thema zu haben. Dieser höhere Anteil ist aber nur teilweise auf wirkliche Meinungslosigkeit zurückzuführen. Darüber hinaus wird die „weiß nicht" Option aber auch gerne genutzt, um den kognitiven Aufwand zu reduzieren, eine Frage zu verstehen und zu beantworten. So wird sie beispielweise von einigen Befragten selbst bei völlig unterschiedlichen Frageinhalten konsistent genutzt. Das explizite Hinweisen auf eine „weiß nicht" Option könnte Interviewten auch suggerieren, dass sie Experten sein sollten, um Fragen zu beantworten. Insbesondere Befragte mit geringem Bildungshintergrund fühlen sich deshalb entmutigt oder verunsichert. Insgesamt zeigen die Studien, dass explizite „weiß nicht" Kategorien die Reliabilität nicht erhöhen.

51.2.4 Die Beschriftung der Antwortkategorien

Eine Beschriftung oder Verbalisierung aller Kategorien einer Antwortskala ist nur bei Skalen bis zu 7 oder 9 Kategorien sinnvoll. Bei längeren Skalen ist es schwierig, eine angemessene verbale Abstufung zu konstruieren. In diesem Fall werden in der Regel nur die Endpunkte mit z.B. „stimme sehr stark zu" oder „stimme überhaupt nicht zu" bezeichnet.

Auf den ersten Blick erscheinen nummerierte Antwortkategorien präzisere Abstufungen zu ermöglichen als verbale Differenzierungen. Empirische Studien zeigen aber, dass die semantische Kennzeichnung der einzelnen Antwortkategorien zu einer besseren Reliabilität führt (z.B. Krosnick 1999, Tourangeau et al. 2007). Falls eine Beschriftung aufgrund der Abstufungen möglich ist, dann sollte sie auch gewählt werden.

Bei der Verbalisierung der Kategorien von Ratingskalen ist darauf zu achten, dass die Bezeichnungen möglichst gleiche Intervallabstände nahelegen. Das ist bei fünf oder siebenstufigen Skalen leicht durchführbar. Die Antwortkategorien bei Zustimmungsfragen lauten dann z.B. „stimme sehr zu", „stimme eher zu", stimme teilweise zu/teilweise nicht", „stimme eher nicht zu" und „stimme überhaupt nicht zu". Bei siebenstufigen Skalen würde man die Skala um extremere Endpunkte z.B. „stimme absolut zu" und „stimme absolut nicht zu" erweitern. Falls in schriftlichen Befragungen gleichzeitig numerische Werte präsentiert werden, dann sollten sie gleichsinnig vergeben werden, also in unserem Fall die Ziffern 5 der Ausprägung „stimme sehr zu" und die Ziffer 1 der Ausprägung „stimme überhaupt nicht zu" zugeordnet werden. Streng genommen handelt es sich bei Ratingskalen um ordinale Messniveaus. Ob die Skalenabstände darüber hinaus gleich sind und die Skala damit Intervallskalenniveau aufweist, ist eine empirische Frage, die durch geeignete statistische Verfahren (z.B. Item-Response-Theorie) getestet werden kann (Geiser/Eid 2010, Rost 2004). Es empfiehlt sich auch, bei der statistischen Auswertung die Ergebnisse der Analyseverfahren für intervallskalierte Daten (z.B. OLS-Regressionen) mit Ergebnissen für ordinale Daten (z.B. logistische Regressionen) zu vergleichen (Blasius/Baur, Kapitel 79 in diesem Band). Auf der sicheren Seite ist man dann, wenn unterschiedliche Analyseverfahren zu vergleichbaren Ergebnissen kommen.

51.2.5 Bipolare oder unipolare Beschriftung?

Eine bipolare Beschriftung von Antwortskalen verwendet das semantische Differential (Gegensatzpaare). Diese sogenannten Polaritätsprofile werden gerne in der Psychologie und der Marktforschung eingesetzt, in der es häufig um die Einschätzung bzw. Zuschreibung von Eigenschaften zu Personen oder Marken geht. Beispiele für Gegensatzpaare sind „stark – schwach", „richtig – falsch" oder „gesund – ungesund". In der soziologisch orientierten Sozialforschung sind Polaritätsprofile eher selten anzutreffen. Bei einer unipolaren Bezeichnung wird dagegen die sprachliche Abstufung eindimensional erreicht. Werden die Kategorien etwa mit „stimme sehr stark zu" bis „lehne sehr stark ab" bezeichnet, so handelt es sich um eine bipolare Beschriftung. Unipolar wird die Abstufung mit den Labels „stimme sehr stark zu" bis „stimme überhaupt nicht zu" erreicht. Befragte kommen bei einer Befragung besser mit möglichst einfachen Antwortskalen zurecht, die schnell erlernt werden können und die nicht bei jeder Frage neu beachtet werden müssen. Diese Regel, Befragungen möglichst einfach zu halten, spricht daher für die Verwendung unipolarer Skalen. Dies scheinen auch empirische Studien zu bestätigen (Schaeffer/Presser 2003).

Die Beschriftung durch semantische Differenziale kann ebenso wie eine unipolare Bezeichnung mit positiven numerischen Werten kombiniert werden (beispielweise von 1 bis 7) oder aber auch negative Ziffern umfassen (z.B. von -3 bis +3). Empirische Studien haben aber gezeigt, dass Befragte zusätzlich zu der Verbalisierung den verwendeten Ziffern eine Bedeutung zuschreiben. So werden negative Ziffern eher gemieden als positive Zahlen (Schwarz et al. 1991). Gleichmäßige Abstufungen werden daher besser erreicht, wenn unipolare Formulierungen auch mit einem positiven Wertebereich der Ziffern kombiniert werden.

51.2.6 Von positiv zu negativ oder umgekehrt?

Bei beiden Formaten (unipolar oder bipolar) stellt sich die Frage, ob eine Antwortskala zuerst die positiven oder zustimmenden Kategorien verwenden oder aber mit den negativen oder ablehnenden Alternativen beginnen sollte. Hierzu gibt es eine Hypothese aus der kognitiven Psychologie. Optimalerweise sollten Befragte zunächst alle Antwortmöglichkeiten abwarten (lesen oder hören) und dann diejenige wählen, die ihrer Meinung am besten entspricht. Diese Maximierung oder Optimierung des Antwortverhaltens wird aber häufig zugunsten von „satisficing" aufgegeben. Beim satisficing wählen Befragte unmittelbar diejenige Alternative, die als erstes akzeptabel erscheint. Ein solches Verhalten würde implizieren, dass die Reihenfolge einer Antwortskala (z.B. von positiv zu negativ) von Bedeutung sein kann. Krosnick (1999) führt eine Vielzahl von Studien auf, in denen Reihenfolgeeffekte besonders bei ungeordneten mehrkategorialen Antwortskalen auftraten. Werden unterschiedliche Kategorien als Antwort vorgelegt, dann werden besonders in selbstadministrierten Befragungen die zuerst genannten Antwortmöglichkeiten häufiger gewählt („primacy effect"). In telefonischen und mündlichen Interviews werden dagegen die zuletzt vorgelesenen Antworten besser im Gedächtnis behalten und häufiger gewählt, was als „Recency-Effekt" bezeichnet wird. Bei Ratingskalen (geordnete Antwortvorgaben) scheinen diese Effekte dagegen weniger stark ausgeprägt zu sein und die Reihenfolge für das Antwortverhalten keine entscheidende Rolle zu spielen (Krebs/Hoffmeyer-Zlotnik 2010).

51.2.7 Standardisierte Antwortskalen oder itemspezifische Skalen?

Im Prinzip spricht die Regel, Befragungen möglichst einfach zu halten, für die Verwendung der gleichen Antwortskalen bei allen Fragen. Die Fragen sollten also so formuliert sein, dass beispielsweise immer eine bestimmte (z.B. fünfstufige) Zustimmungsskala verwendet werden kann. Dennoch zeigen einige Studien, dass itemspezifische Antwortskalen die Reliabilität der Messung erhöhen können (Krebs 2011). Itemspezifische Antwortskalen sind möglichst genau an die Formulierung der Frage angepasst und erleichtern dadurch das Verständnis der Antwortskalen. Die Verwendung immer gleicher Skalen könnte auch zur Ermüdung der Befragten führen. Allerdings sind die Evidenzen für itemspezifische Antwortskalen noch unsicher.

51.3 Zusammenfassung

Zusammenfassend lassen sich aus dem Literaturüberblick sieben Regeln zur Gestaltung von Antwortkategorien ableiten:

1. Falls möglich, sollten keine geschlossenen Antwortkategorien vorgegeben werden, sondern die Antworten offen numerisch erhoben werden.
2. Antwortkategorien sollten disjunkt und erschöpfend sein.
3. Antwortkategorien sollten das höchstmögliche Skalenniveau realisieren.
4. Die optimale Anzahl an Antwortkategorien ist sieben plus/minus zwei.
5. Antwortkategorien sollten ungerade sein.
6. Falls möglich, sollte jede Antwortkategorie beschriftet sein.
7. Zur Vermeidung von Antwortmustern sollte die Reihenfolge von positiv zu negativ formulierten Antwortkategorien wechseln.

Aber: keine Regel ohne Ausnahme. Mit Ausnahme von Regel zwei kann es gute Gründe geben, von den formulierten Empfehlungen abzuweichen. Die Einkommensfrage wurde als mögliche Ausnahme schon erwähnt. Auch bei Fragen zur Zufriedenheit mit verschiedenen Lebensbereichen oder der allgemeinen Lebenszufriedenheit haben sich beispielweise 11er Skalen (von 0 bis 10) etabliert und bewährt. Die zweite Regel ist dagegen eine Mindestanforderung an Antwortskalen, bei der keine Ausnahmen gemacht werden sollten. Schließlich ist darauf zu achten, dass die Gestaltung von Antwortskalen vom Befragungsmodus und natürlich von der Zielpopulation abhängen. Es ist deshalb empfehlenswert, neue Erhebungsinstrumente vor der Feldphase einem gründlichen Pretest zu unterziehen.

Literatur

Converse, Jean M./Presser, Stanley (1986): Survey Questions. Beverly Hills: Sage

Diekmann, Andreas (2011): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. 5. Aufl. Reinbek bei Hamburg: Rowohlt

Faulbaum, Frank/Prüfer, Peter/Rexroth, Margit (2009): Was ist eine gute Frage? Die systematische Evaluation der Fragequalität. Wiesbaden: VS Verlag

Franzen, Axel/Vogl, Dominikus (2013): Acquiescence and the Willingness to Pay for Environmental Protection: A Comparison of the ISSP, WVS, and EVS. In: Social Science Quarterly 94: 637-659

Geiser, Christian/Eid, Michael (2010): Item-Response-Theorie. In: Wolf/Best (Hg.): 311-332

Kieruj, Natalia D./Moors, Guy (2010): Variations in Response Style Behavior by Response Scale Format in Attitude Research. In: International Journal of Public Opinion Research 22: 320-342

Krebs, Dagmar (2011): Vortrag am Jahrestreffen der European Social Research Association (ESRA) in Lausanne, Schweiz

Krebs, Dagmar/Hoffmeyer-Zlotnik, Juergen H.P. (2010): Positive First or Negative First? In: Methodology 6: 118-127

Krosnick, Jon A. (1999): Survey Research. In: Annual Review of Psychology 50: 537-567

Miller, George A. (1956): The Magical Number Seven Plus or Minus Two: Some Limits on our Capacity for Processing Information. In: Psychological Review 63: 81-97

Moors, Guy (2008): Exploring the Effect of Middle Response Category on Response Style in Attitude Measurement. In: Quality and Quantity 42: 779-794

O'Muircheartaigh, Colm/Krosnick, Jon A./Helic, Armin (2000): Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data. Working paper. University of Chicago

Preston, Carolyn C./Colman, Andrew M. (2000): Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. In: Acta Psychologica 104: 1-15

Rost, Jürgen (2004): Lehrbuch Testtheorie – Testkonstruktion. Bern: Huber

Schaeffer, Nora C./Presser, Stanley (2003): The Science of Asking Questions. In: Annual Review of Sociology 29: 65-88

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

Schnell, Rainer/Hill, Paul B./Esser, Elke (2011): Methoden der empirischen Sozialforschung. 9. Auflage. München: Oldenbourg

Schwarz, Norbert/Hippler, Hans-J./Deutsch, Brigitte/Strack, Fritz (1985): Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgements. In: Public Opinion Quarterly 49: 388-395

Schwarz, Norbert/Knäuper, Bärbel/Hippler, Hans J./Noelle–Neumann, Elisabeth/Clark, Leslie (1991): Rating Scales: Numeric Values May Change the Meaning of Scale Labels. In: Public Opinion Quarterly 55: 570-582

Svensson, Elisabeth (2000): Concordance between Ratings Using Different Scales for the Same Variable. In: Statistics in Medicine 19: 3483-3496

Tourangeau, Roger/Couper, Mick P./Conrad, Frederick (2007): Color, Labels and Interpretive Heuristics for Response Scales. In: Public Opinion Quarterly 71: 91-112

Weijters, Bert/Cabooter, Elke/Schillewaert, Niels (2010): The Effect of Rating Scale Format on Response Style: The Number of Response Categories and Response Category Labels. In: International Journal of Research in Marketing 27: 236-247

Wolf, Christof/Best, Henning (Hg.) (2010): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag