gesis Leibniz-Institut für Sozialwissenschaften

SDM Survey Guidelines

Gestaltung von Ratingskalen in Fragebögen

Natalja Menold & Kathrin Bogner

Januar 2015, Version 1.0

Zusammenfassung

Ratingskalen zählen zu den wichtigsten und am häufigsten benutzten Instrumenten der sozialwissenschaftlichen Datenerhebung. Bis heute existiert eine umfangreiche methodische Forschung zur Gestaltung von Ratingskalen und ihren (psycho-) metrischen Eigenschaften. In diesem Beitrag gehen wir auf die einzelnen Aspekte der Fragebogenkonstruktion in Bezug auf die Ratingskalen ein. Dabei werden der Stand der Forschung und Praxiserfahrungen skizziert und Empfehlungen zur Gestaltung von Ratingskalen — soweit solche möglich sind — ausgesprochen.

Zitierung

Menold, Natalja und Bogner, Kathrin (2015). Gestaltung von Ratingskalen in Fragebögen. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (SDM Survey Guidelines). DOI: 10.15465/sdm-sg_015

1. Einleitung

Seit ihrer Einführung durch Thurstone (1929) und Likert (1932) zu Beginn der sozialwissenschaftlichen Forschung Ende der 20er und Anfang der 30er Jahre zählen Ratingskalen zu den wichtigsten und am häufigsten benutzten Instrumenten sozialwissenschaftlicher Datenerhebung. Ratingskalen stellen ein Bewertungskontinuum dar (z. B. Zustimmung, Intensität, Häufigkeit, Zufriedenheit), mit dessen Hilfe unterschiedliche Merkmale und Phänomene in Fragebögen erhoben werden. Die befragten Personen bewerten Inhalte von Fragen und Aussagen (Items), indem sie eine passende Kategorie der Ratingskala ankreuzen. Beispielsweise für die Frage „Wie zufrieden sind Sie — alles in allem — mit Ihrem gegenwärtigen Leben?" aus dem European Social Survey (ESS) wird eine elfstufige Ratingskala von 0 ("äußerst unzufrieden") bis 10 („äußerst zufrieden") vorgegeben. Ratingskalen sind der Gegenstand des vorliegenden Beitrags. Sie stellen nur eine Möglichkeit dar, Antworten in Fragebögen zu dokumentieren. So kommen in Fragebögen andere Formen von Antwortvorgaben vor wie bspw. nominale Kategorienlisten, die jedoch hier nicht im Fokus stehen.

Antworten der befragten Personen werden zunächst als Funktion von zwei grundlegenden Merkmalen von Ratingskalen aufgefasst (Parducci, 1983). Ein Merkmal ist der Variationsbereich, der durch die Skalenpole abgegrenzt ist (sog. Range); das andere Merkmal ist der Differenzierungsgrad, der durch die Anzahl der Antwortkategorien festgelegt ist (sog. Frequency). Damit sind die Anzahl der Kategorien und die Bezeichnungen der Endkategorien für das Verständnis der zu messenden Bewertungsdimension von grundlegender Bedeutung. Als ein Kontext der Fragenbeantwortung können Ratingskalen sowohl zu erwünschten Effekten, wie z. B. Förderung des intendierten Verständnisses von Range und Frequency, als auch zu unerwünschten Effekten führen, z. B zur Akquieszenz (d. h. eine Zustimmung wird unabhängig vom Frageinhalt geäußert). Diese unerwünschten Effekte können durch die Gestaltung von Ratingskalen vermindert werden. Nicht nur die Anzahl der Kategorien und die Kategorienbeschriftungen, sondern auch andere grafische Merkmale von Ratingskalen wie bspw. Skalenorientierung oder Nutzung von Farben, Schriften und Schattierungen, können das Verständnis der Bewertungsdimension beeinflussen („Visual Design", z. B. Tourangeau, Couper & Conrad, 2004; 2007). Es gilt, die befragten Personen zu einer sorgfältigen Beantwortung (sog. Optimizing) zu motivieren. Generell sollte die Aufgabe, Fragen zu beantworten, nicht allzu komplex und schwer sein, aber auch nicht unnötig dazu verleiten, den kognitiven Aufwand zu reduzieren (sog. Satisficing; Krosnick & Alwin, 1987).

Eine umfangreiche methodische Forschung zur Gestaltung von Ratingskalen und ihren (psycho-) metrischen Eigenschaften wurde größtenteils zwischen den 60er und 80er Jahren durchgeführt. Neuere Studien erfolgten überwiegend in Online-Umfragen im Rahmen des „Visual Design"- Ansatzes. Zu nennen sind auch neuere multi-trait-multi-method (MTMM) Untersuchungen, die mit Hilfe von Strukturgleichungsmodellen ausgewertet werden (z. B. Saris & Gallhofer, 2007). Untersucht wurden Gestaltungsaspekte von Ratingskalen wie Anzahl der Kategorien, Verwendung der Skalenmitte, Verwendung verbaler oder numerischer Bezeichnungen, Skalenorientierung und Skalenpolarität (siehe Übersichtsarbeiten von Krosnick & Fabrigar, 1997 oder Menold & Bogner, 2012). Weitere Forschung untersucht die Nutzung spezifischer Bewertungsdimensionen („item specific" oder direkte Skalen), wie z. B. Wichtigkeit, Häufigkeit, Zufriedenheit im Vergleich zu agree-disagree Skalen (Likert-Type), die universell für unterschiedliche Bewertungen genutzt werden können (mehr dazu siehe weiter im Text).

Die Effekte von Ratingskalen wurden im Wesentlichen in Bezug auf die psychometrischen Gütekriterien untersucht. Dazu zählen die Reliabilität (Zuverlässigkeit oder Messgenauigkeiten) und die Validität (Gültigkeit bzw. eine Angabe, inwieweit von Messergebnissen Aussagen über die zu messenden Merkmale möglich sind) (vgl. auch SDM Survey Guidelines Artikel „Reliabilität" (Danner, 2015)). Darüber

hinaus werden Präferenzen der befragten Personen und Schwierigkeiten bei der Beantwortung betrachtet.

Im Folgenden gehen wir auf die einzelnen Aspekte der Fragebogenkonstruktion in Bezug auf die Ratingskalen ein. Dabei werden kurz jeweils der Stand der Forschung skizziert und Empfehlungen zur Gestaltung von Ratingskalen — soweit solche möglich sind — ausgesprochen.

2. Anzahl der Antwortkategorien

Die Anzahl der Antwortkategorien ist eine für das Verständnis der Bewertungsdimension sehr wichtige Eigenschaft von Ratingskalen, weil sie entsprechend dem Range-Frequency-Modell den Differenzierungsgrad einer Ratingskala festlegt (siehe Einleitung).

In Übersichtsarbeiten kommen Krosnick und Kollegen (Krosnick & Fabrigar, 1997; Krosnick & Presser, 2010) zum Schluss, dass eine optimale Messung — in Bezug auf die Reliabilität, Validität und den Differenzierungsgrad — mit fünf bis sieben Kategorien erreicht werden kann. Die Präferenzen der befragten Personen liegen ebenso in diesem Bereich (Krosnick & Fabrigar, 1997). Dieser Befund wird so erklärt, dass bei zu vielen Kategorien die Bedeutung der einzelnen Kategorien weniger klar ist, was die Schwierigkeit bei der Beantwortung erhöht. Bei einer geringeren Anzahl an Kategorien differenzieren die Ratingskalen nicht ausreichend.

Es gibt jedoch neuere Studien, die einen linearen Zusammenhang zwischen der Anzahl der Kategorien und den psychometrischen Gütekriterien fanden, d. h. mit steigender Anzahl an Kategorien steigt die Messqualität (z.B. Saris & Gallhofer, 2007; Pajares, Hartley, & Vahante, 2001; Preston & Colman, 2000). Die maximale Anzahl der getesteten Kategorien variiert in diesen Studien zwischen 10, 11 und 100 Kategorien. In der Praxis scheint sich dennoch die Faustregel „fünf bis sieben" in Bezug auf die Skalenbreite durchgesetzt zu haben, zumal solche Skalen einfacher verbalisiert werden können (siehe den nächsten Abschnitt).

Fazit und Empfehlung: Im Einklang mit der Mehrheit der Forschungsarbeiten, die für die Regel „fünf bis sieben" sprechen, empfehlen wir diese Anzahl an Kategorien.

3. Kategorienlabels

Bei der Nutzung der Kategorienlabels stellt sich die Frage, ob nur die Endkategorien verbalisiert, die den Range der Ratingskala markieren, oder auch verbale Beschriftungen für jede Kategorie verwendet werden sollen. Zusätzlich zu verbalen Beschriftungen werden in Ratingskalen numerische Marker sehr häufig genutzt.

Übersichtsarbeiten kommen zum Schluss, dass vollverbalisierte Ratingskalen die Reliabilität bei wiederholten Messungen (Test-Retest-Reliabilität) und die Validität erhöhen (Krosnick & Fabrigar, 1997; Maitland, 2009). Saris und Gallhofer (2007) und Menold, Kaczmirek, Lenzner und Neusar (2014) fanden, dass vollverbalisierte Antwortskalen auch dann die Reliabilität erhöhen, wenn diese auf einem Messzeitpunkt basieren (cross-sectional; split-half Methode). Zusätzlich präferieren befragte Personen vollverbalisierte Ratingskalen (z. B. Wallsten, Budescu, & Zwick, 1993; Zaller, 1988). Darüber hinaus reduziert die Vollverbalisierung unerwünschte Effekte anderer visueller Elemente im Fragebogen (z. B. Toepoel & Dillman, 2011). Der positive Effekt der Vollverbalisierung wird dadurch erklärt, dass die Bedeutung der Kategorien so klarer ist — im Vergleich zu den Skalen, die keine verbalen Beschriftungen

verwenden. Insbesondere Personen mit geringer Bildung profitieren von der Vollverbalisierung (Krosnick & Fabrigar, 1997).

Numerische Labels, die in Umfragen sehr oft verwendet werden, sind – theoretisch – wenig günstig, Zum einen können sie für unterschiedliche befragte Personen unterschiedliche Bedeutung haben, die wiederum mit der Bedeutung der Kategorien der Ratingskala inkongruent sein kann (z. B. Glücks- und Unglückszahlen oder auch Bedeutung von Noten). Zum anderen ist es wenig natürlich und selbstverständlich, Selbst- oder Fremdbeschreibungen anhand von Zahlen vorzunehmen (Krosnick & Fabrigar, 1997). Es gibt nur wenige Studien, die verbale und numerische Labels in Ratingskalen vergleichen, die Ergebnisse dieser Studien unterstützen jedoch die Annahmen von Krosnick und Fabrigar (1997) (Windschitl & Wells, 1996; Christian, Parsons, & Dillman, 2009).

Die verbalisierten Ratingskalen sollten den folgenden Anforderungen entsprechen: Erstens sollten sie präzise sein. Zweitens sollten die Ratingskalen ausbalanciert sein. Ausbalancierte Ratingskalen sind symmetrisch bzw. haben die gleiche Anzahl positiver und negativer Kategorien. Drittens sollten sie allgemein verständlich oder universell sein, und viertens sollten die Abstände zwischen den Kategorien in einer Ratingskala gleich (äquidistant) sein. Eine solche Gestaltung von Ratingskalen stellt keineswegs eine triviale Aufgabe dar. So zeigten Pollack, Friedman und Presby (1990) in Bezug auf die universelle Verständlichkeit, dass sich die emotionale Färbung und Extremität der Verbalisierungen in Ratingskalen auf die Beantwortung von Fragen auswirken. Außerdem führen bestimmte Verbalisierungen stärker zu schiefen Verteilungen als andere verbale Bezeichnungen (French-Lazovik & Gibson, 1984). Worcester und Burns (1975) zeigten, dass verbal gegensätzliche Bezeichnungen (Antonyme) nicht unbedingt als Antonyme in ihrer Anordnung in Ratingskalen wahrgenommen werden. Weiterhin variieren die prozentualen Angaben zu Quantifizierungen (z.B. „rare", „unlikely", „possible") sehr stark zwischen unterschiedlichen Studien (Theil, 2002), was auf ihr unterschiedliches Verständnis hindeutet.

In den 80er und 90er Jahren wurde eine Reihe an Arbeiten zur Entwicklung universell verwendbarer verbaler Markierungen für unterschiedliche Bewertungsdimensionen – wie Häufigkeit, Evaluation oder Wahrscheinlichkeit – für verschiedene Wortarten wie Adjektive oder Adverbien durchgeführt, jedoch überwiegend im englischsprachigen Raum (eine Übersicht findet sich bei Clark, 1990). Für die Formulierung in der deutschen Sprache liegt eine Arbeit von Rohrmann (1978) vor, der für unterschiedliche Bewertungsdimensionen Verbalisierungen vorschlug.

Fazit und Empfehlung:

Die Ergebnisse der bisherigen Forschung sprechen eher für die Verwendung vollverbalisierter Ratingskalen und weniger für die Verwendung von endverbalisierten Ratingskalen (z. B. mit numerischen Markern). Die vollverbalisierten Ratingskalen können insbesondere die befragten Personen mit geringer formaler Bildung unterstützen. Die oben genannten Arbeiten zur Entwicklung von vollverbalisierten Ratingskalen können bei der Auswahl von Verbalisierungen genutzt werden. Eine Vollverbalisierung ist in Kombination mit einer moderaten Anzahl an Kategorien (siehe oben) praktikabel.

4. Skalenpolarität

Skalenpolarität unterscheidet zwischen uni- und bipolaren Ratingskalen. Bei bipolaren Ratingskalen werden zwei gegensätzliche Dimensionen abgebildet, wie z. B. positiv - negativ; lehne ab – stimme zu; bei unipolaren Ratingskalen wird ein Kontinuum von einer geringen bis hohen Ausprägung dargestellt, wie z. B. gar nicht zufrieden - sehr zufrieden.

Saris und Gallhofer (2007) untersuchten die Passung zwischen der Polarität von Bewertungsdimensionen und der Polarität von Ratingskalen. Sie definierten eine Bewertungsdimension

als unipolar, wenn hier eine gegensätzliche Dimension nicht denkbar ist, z.B. für den Grad der Zufriedenheit oder des Glücks existieren sprachliche Gegensätze (unzufrieden - zufrieden, unglücklich - glücklich), während für die Dimension „Häufigkeit" kein solcher Gegensatz existiert, d. h. man kann nur zwischen einem "nie" und „immer" unterscheiden. Demensprechend wären für solche Bewertungsdimensionen wie „Häufigkeit" unipolare und für die „Zufriedenheit" und „Glück" bipolare Ratingskalen zu nutzen. Die Autoren fanden jedoch keinen Effekt einer solchen Passung auf die Gütekriterien.

Es gibt jedoch keine einheitliche Definition der Polarität in der Literatur. So fassen Krosnick und Fabrigar (1997) die Wichtigkeit als eine unipolare Bewertungsdimension auf, während sie nach der Definition von Saris und Gallhofer (2007) als eine bipolare Bewertungsdimension einzuordnen wäre. Andere Autoren definieren die Polarität durch die Nutzung von numerischen Markern: negative und positive Zahlen sind in bipolaren Ratingskalen zu finden, Zahlen von Null bis eine höhere Ausprägung in unipolaren Ratingskalen (Moors, Kieruj & Vermunt, in press). In Bezug auf die Nutzung von negativen Zahlen wurde gezeigt, dass die befragten Personen den negativen Skalenbereich meiden und positivere Antworten produzieren (Schwarz et al., 1991; Schaeffer & Barker, 1995).

Da viele Bewertungsdimensionen sowohl uni- als auch bipolar Realisierung zulassen, stellt sich die Frage nach grundsätzlichen Vorteilen der bipolaren Ratingskalen. Hierzu zeigten Krebs (2012) und Menold (2013), dass die Wahl der uni- oder bipolaren Bezeichnungen in Ratingskalen die psychometrischen Eigenschaften von Multi-Item Messungen und die Messwerte der latenten Variablen beeinflussen kann. Für eine abschließende Empfehlung uni- oder bipolare Ratingskalen zu nutzen ist jedoch weitere Forschung notwendig.

Fazit und Empfehlung: Generell sind die Effekte der Skalenpolarität wenig untersucht, so dass bisher keine eindeutigen Empfehlungen hierzu möglich sind, außer dass negative numerischen Marker systematische Effekte – positivere Antworten – erzeugen können und deshalb zu vermeiden sind.

5. Skalenorientierung

Bei der Skalenorientierung geht es um die Entscheidung, ob man mit der geringsten bzw. negativen Ausprägung beginnt und mit der höchsten bzw. positiven Ausprägung endet (aufsteigende Reihenfolge) oder eine umgekehrte – absteigende – Reihenfolge wählt. So wäre zwischen der Reihenfolge „trifft überhaupt nicht zu", „trifft wenig zu", „trifft teilweise zu", „trifft ziemlich zu", „trifft voll und ganz zu" oder einer umgekehrten Reihenfolge zu entscheiden.

Während bei der vertikalen Darstellung von Antwortkategorien stärkere Reihenfolge-Effekte auftreten können (Primacy- und Recency- Effekte), bei welchen die erst- oder letztpräsentierten Kategorien bevorzugt gewählt werden (Krosnick & Alwin, 1987), sind solche Effekte bei einer horizontalen Präsentation in Ratingskalen nur geringfügig ausgeprägt (Tourangeau, Rips & Rasinski, 2000). Dabei wird die Kategorie am linken Ende der Ratingskala häufiger gewählt, im Vergleich zur Kategorie am rechten Ende, unabhängig von der gewählten Skalenorientierung. Dieser Effekt wird als „general primacy effect" bezeichnet. Allerdings wurden in neueren Studien Zusammenhänge zwischen der Skalenorientierung und dem Primacy-Effekt festgestellt. So fand Toepoel (2008) den Primacy-Effekt nur für die aufsteigende Skalenorientierung, d. h. hier wurde die negative Kategorie am linken Ende der Ratingskala häufiger gewählt, im Vergleich zur positiven Kategorie. Hingegen fanden andere Autoren (Hofmans et al., 2007; Krebs & Hoffmeyer-Zlotnick, 2010) den Primacy-Effekt ausschließlich für die absteigende Reihenfolge.

Es liegen nur wenige Studien vor, die die Auswirkung der Skalenorientierung auf die Messqualität untersuchen, wobei keine Effekte gezeigt werden konnten (Saris & Gallhofer, 2007; Krebs & Hoffmeyer-Zlotnik, 2010).

Fazit und Empfehlung: Aus den gemischten Ergebnissen zum „Primacy Effekt" lassen sich zurzeit keine Empfehlungen zur Wahl der Skalenorientierung ableiten. Da bisher keine Effekte der Skalenorientierung auf die Gütekriterien gefunden werden konnten, können die Forschenden sowohl die auf- als auch absteigende Reihenfolge bei Ratingskalen verwenden bzw. es wäre im Zweifelsfall mit Hilfe von kognitiven Pretests (s. SDM Survey Guidelines Artikel „Kognitives Pretesting" (Lenzner, Neuert & Otto, 2015)) eine passende Skalenorientierung zu wählen.

6. Skalenmitte

Bei der Gestaltung von Antwortskalen muss entschieden werden, ob eine Mittelkategorie angeboten werden soll oder nicht. Dabei ist zuerst auf die oben beschriebene Polarität der Ratingskala zu achten, denn in Abhängigkeit von dieser bringt eine Mittelkategorie unterschiedliche Positionen einer befragten Person zum Ausdruck: Bei bipolaren Ratingskalen kann die Mittelkategorie sowohl Indifferenz („weder noch") als auch Ambivalenz („teils - teils") ausdrücken (Kaplan, 1972; Dubois & Burns, 1975). Diese Uneindeutigkeit erschwert die Interpretation der Mittelkategorie in bipolaren Ratingskalen sowohl für den Befragten als auch für den Forscher. Bei einer unipolaren Ratingskala steht die Mittelkategorie für eine mittlere Position, was mittels der Verwendung von Labels wie „mittlere Zustimmung" oder „trifft mäßig zu" Ausdruck findet.

Neben der Polarität müssen bei der Skalengestaltung drei weitere mögliche Fehlerquellen, entstehend aus der Vorgabe bzw. der Nichtvorgabe einer Mittelkategorie, gegeneinander abgewogen werden:

Erstens, eine Mittelkategorie kann eine Einladung für jene befragte Personen darstellen, die ein Satisficing-Verhalten zeigen. Das sind meist unmotivierte oder ermüdete Personen, die die Mittelkategorie wählen, um den kognitive Aufwand der Fragebeantwortung zu reduzieren, und nicht weil diese ihrer tatsächlichen Einstellung entspricht. Vielmehr tendieren die meisten befragten Personen in eine Richtung der Skala und würden, wenn keine Mittelkategorie angeboten wäre, diese Einstellung auch berichten (Krosnick, 1991). Verschiedene experimentelle Studien kommen zu dem Ergebnis, dass die Vorgabe einer mittleren bzw. neutralen Kategorie die Tendenz zur Mitte verstärken und zu einer weniger gründlichen Beantwortung führen kann (Kalton, Robert, & Holt, 1980; Krosnick & Fabrigar, 1997; Schumann & Presser, 1981; Saris & Gallhofer, 2007). Bishop, Oldendick, Tuchfarberund und Bennett. (1980) und O'Muircheartaigh, Krosnick und Helic (1999) fanden, dass Befragte, die das Thema als nicht sehr wichtig oder als uninteressant einstuften oder dazu eine niedrige Einstellungsstärke besaßen, häufiger die Mittelkategorie wählten. Ein Zusammenhang zwischen der Häufigkeit der Wahl der Mittelkategorie und dem Umfang an Wissen zu dem Thema wurde allerdings nicht gefunden (O'Muircheartaigh et al., 1999). Auch ein Zusammenhang zwischen Bildung und der Wahl der Mittelkategorie wurde in verschiedenen Studien nicht bestätigt (Kalton, Roberts & Holt, 1980; Schuman & Presser, 1981; O'Muircheartaigh et al., 1999; Krosnick, Narayan, & Smith, 1996).

Zweitens, stehen den Befragten, die aus Gründen des Satisficing die Mittelkategorie wählen, solche entgegen, die tatsächlich eine neutrale bzw. mittlere Position gegenüber dem Thema besitzen. Wird ihnen eine Skala ohne Mittelkategorie vorgegeben, können sie ihre neutrale bzw. mittlere Einstellung nicht korrekt zum Ausdruck bringen. Es stellt sich daher die Frage, ob diese Befragten zufällig oder systematisch eine andere Antwortkategorie nutzen und ob somit systematische Fehler entstehen. O'Muircheartaigh et al. (1999) zeigten, dass sich die Reliabilität und Validität von Skalen unter Hinzunahme einer Mittelkategorie erhöht. Zudem kommen verschiedene Studien zu dem Schluss, dass

befragte Personen bei fehlender Mittelkategorie nicht zufällig eine andere Kategorie, sondern systematisch eine Kategorie in der Nähe der eigentlichen Skalenmitte wählen (Krosnick, 2002; Schumann & Presser, 1981). Daher empfehlen Krosnick und Presser (2010) eine Mittelkategorie anzubieten.

Drittens besteht die Möglichkeit, dass Befragte ohne eine Einstellung zu dem Thema aus Gründen der sozialen Erwünschtheit die Mittelkategorie wählen, anstatt eine fehlende Einstellung zu berichten. Dadurch wird aus den gewonnenen Daten zum einen der Anteil an Personen mit einer Einstellung zu dem Thema überschätzt und zum anderen die Ordinalitätsannahme der Skala verletzt, da die Mittelkategorie dann nicht mehr neutrale bzw. mittlere, sondern auch fehlende Einstellungen repräsentiert (u.a. Sturgis, Roberts & Smith, 2014). Kulas, Stachowski und Haynes (2008) kommen mit einer Test-Retest-Studie zu der Erkenntnis, dass Befragte die Mittelkategorie häufig als Ersatz für eine fehlende „Don't-Know"-Katgeorie nutzen. Allerdings nahm dieses Verhalten keinen Einfluss auf die Validität und Reliabilität der untersuchten Persönlichkeitsskalen. Kulas et al. (2008) empfehlen daher eine Mittelkategorie in Ratingskalen anzubieten. Auch Sturgis et al. (2014) konnten mittels follow-up Probes zeigen, dass ein großer Anteil jener Befragten, welche die Mittelkategorie gewählt hatten, eigentlich keine Einstellung zu dem Thema besitzt. Sie bezeichnen diese Mittelkategorie-Antworten als „face saving don't knows". Anders als Kulas et al. (2008) finden sie aber, dass eine Behandlung der „face-saving don't knows" als „Don't Know"-Antworten die Verteilungen der untersuchten Items signifikant verändert. Zudem zeigen ihre Ergebnisse, dass vor allem unter solchen Befragten eine Tendenz zum Berichten von „face-saving don't knows" besteht, die der Meinung sind, dass sie zu wichtigen Themen eine Einstellung besitzen und berichten sollten, wodurch ein systematischer Fehler in den Daten eingeführt wird. Die follow-up Probes zeigen aber auch, dass der andere Teil der befragten Personen die Mittelkategorie wählt, weil er tatsächlich eine neutrale Einstellung zu dem Thema besitzt. Daher empfehlen Sturgis et al. (2014), die Mittelkategorie in Ratingskalen anzubieten, um zu verhindern, dass Personen mit neutraler Einstellung zu einer inhaltlich falschen Antwort gezwungen werden.

Fazit und Empfehlung: Die Ergebnisse zu Effekten der Mittelkategorie zeigen zwar, dass befragte Personen sie nicht nur wie erwünscht im Falle einer mittleren bzw. neutralen Einstellung wählen, sondern auch aus Gründen des Satisficing oder der sozialen Erwünschtheit. Trotzdem empfehlen die meisten Forscher, diese Kategorie anzubieten, um zu vermeiden, dass Befragte mit einer mittleren bzw. neutralen Einstellung auf andere Kategorien ausweichen und die Daten somit systematisch verzerren.

7. Nicht-inhaltliche (DK-) Kategorie

In der wissenschaftlichen Diskussion um die Verwendung von nicht-inhaltlichen Antwortkategorien in Ratingskalen, sog. ‚Weiß-nicht-', keine Meinung oder ‚DK-' Kategorien, werden zwei gegenläufige Standpunkte vertreten: Nach der klassischen Position wird die generelle Vorgabe von DK-Kategorien[1] empfohlen, da die Annahme lautet, dass befragte Personen ohne relevante Einstellung sich ansonsten gedrängt sehen, eine inhaltliche Antwort zu geben, also zufällig eine inhaltliche Antwortkategorie wählen anstatt das Fehlen einer relevanten Einstellung zu kommunizieren (z.B. Katz, 1942; Payne, 1950; Vaillancourt, 1973; Schuman & Presser, 1981; Converse & Presser, 1986). Vertreter der neueren Position hingegen argumentieren, dass die Vorgabe von DK-Kategorien problematisch ist, da nicht nur befragte Personen ohne eine relevante Einstellung, sondern auch solche mit einer relevanten Einstellung diese Kategorie aufgrund von Satisficing auswählen (z.B. Gilljam & Granberg, 1993; Krosnick & Fabrigar,

[1] In diesem Kapitelabschnitt wird der Begriff DK-Kategorie als Synonym für die verschiedenen nicht-inhaltlichen Antwortkategorien wie *keine Meinung*, *Weiß-nicht*, *Kann ich nicht sagen* benutzt.

1997). Ferner können befragte Personen mittels einer DK-Antwort die Äußerung von nicht sozialerwünschten Einstellungen vermeiden oder auch dann die DK-Kategorie wählen, wenn sie die Frage nicht verstehen oder mit den Antwortvorgaben nicht klar kommen (Krosnick & Fabrigar, 1997). Auch können Befragte die Vorgabe einer DK-Kategorie dahingehend interpretieren, dass elaboriertes Wissen zur Beantwortung der Frage erforderlich ist, was zu Unsicherheit und somit zur Wahl der DK-Kategorie führen kann (Hippler & Schwarz, 1989).

Als Alternative zu DK-Kategorien wird die Verwendung von vorgeschalteten Filterfragen betrachtet. Bei diesem Vorgehen wird erhoben, ob eine Person zu einem bestimmten Thema eine Einstellung besitzt. Ist dies der Fall, wird diese Einstellung detailliert abgefragt, anderenfalls wird die nächste Frage gestellt. Ziel dieser Filterung ist, befragte Personen ohne Einstellung nicht zu einer falschen inhaltlichen Antwort zu drängen und somit die Datenqualität zu verbessern. Ein Vergleich von gefilterten und ungefilterten Fragen zeigt allerdings, dass die Rate an DK-Antworten bei der gefilterten Version um 20 - 25 % höher liegt als bei der ungefilterten (Schuman & Presser, 1981). Die Formulierung des Filters nimmt dabei erheblichen Einfluss auf die Verneinung einer existierenden Einstellung: Ist die Filterfrage eher allgemein formuliert (z. B. „Besitzen Sie eine Einstellung zu dem Thema?"), berichten Befragte eher das Vorhandensein einer Einstellung als wenn die Filterfrage die Notwendigkeit einer intensiven Auseinandersetzung mit dem Thema impliziert, um eine Antwort geben zu können (z. B. „Haben Sie genügend über das Thema nachgedacht/darüber gelesen, um eine Meinung dazu zu haben?") (z. B. Bishop, Oldendick& Tuchfarber, 1983; Hippler & Schwarz, 1989, Krosnick & Abelson, 1991; Fowler & Cannell, 1996). Dieser Effekt der Filterfrageformulierung wird umso stärker, je abstrakter oder weniger bekannt das inhaltliche Thema der Frage ist (Bishop et al., 1983).

Bezüglich der Datenqualität zeigte Andrews (1984), dass Skalen mit einer DK-Kategorie eine höhere Validität sowie geringere Methodeneffekte und Fehlervarianzen erreichen als Skalen ohne DK-Kategorie. Andere experimentelle Studien kommen hingegen zum Ergebnis, dass der Ausschluss von DK-Kategorien die Datenqualität nicht beeinflusst (z. B. Poe et al.,1988; Alwin & Krosnick, 1991; McClendon & Alwin, 1993; Krosnick et al., 2002). In einer Wahlstudie wurden schließlich exaktere Wahlprognosen erzielt, wenn Befragte, die eine DK-Antwort gewählt hatten, anschließend unter Nachdruck um eine inhaltliche Antwort gebeten wurden (Visser, Krosnick, Marquette & Curtin, 2000). Möglicherweise spielen der Frageinhalt und der Grad der Ausdifferenzierung der Einstellung bei der Nutzung der DK-Kategorie eine Rolle.

Die Form der Vorgabe einer DK-Kategorie unterscheidet sich entsprechend dem Erhebungsmodus. Für schriftliche oder andere papierbasierte Umfragen muss eine Entscheidung getroffen werden, ob eine DK-Kategorie explizit angeboten wird oder nicht. In persönlichen oder telefonischen Interviews hingegen besteht die Wahl zwischen der expliziten Vorgabe einer DK-Kategorie oder der Akzeptanz durch den Interviewer einer vom Befragten selbständig geäußerten fehlenden Einstellung als DK-Antwort. Oftmals wird dem Interviewer vorgegeben, bei solch einer Äußerung einmalig nachzufragen, dann aber die selbständig geäußerte DK-Antwort als solche aufzunehmen. In interaktiven Computer-assistierten Umfragen gibt es verschiedene technische Umsetzungsmöglichkeiten der DK-Vorgabe: Entweder wird eine DK-Kategorie explizit angeboten oder der Befragte wird bei Nichtbeantwortung der Frage mittels einer direkt folgenden Nachfrage gebeten, doch eine Antwort zu geben oder die DK-Antwort zu bestätigen (implizite Vorgabe). Es gibt auch noch eine Kombination der beiden Vorgehensweisen, nämlich die explizite Vorgabe einer DK-Kategorie und eine direkt folgende Nachfrage, sofern der Befragte weder eine der inhaltlichen noch die DK-Kategorie gewählt hat. DeRouvray & Couper (2002) finden die geringste Rate an Item-Nonresponse für das Design in dem keine DK-Kategorie explizit vorgegeben wurde, die befragten Personen aber bei der Nichtbeantwortung durch eine nachgeschaltete Nachfrage die Möglichkeit hatten, eine DK-Antwort zu bestätigen.

<u>Fazit und Empfehlung:</u>

Bei der Entscheidung, eine DK-Kategorie vorzugeben oder auf diese zu verzichten, sollten Frageinhalt, der Modus der Datenerhebung und die Zielgruppe berücksichtigt werden. So müssen die Forscher entscheiden, inwieweit es bei einer Zielgruppe problematisch sein kann, keine DK-Kategorie anzubieten. Ist man sicher, dass die befragten Personen eine Antwort kennen, kann auf die DK-Option verzichtet werden.

8. Likert-Type und direkte Ratingskalen

Likert-Type Skalen sind Ratingskalen mit Zustimmung als Bewertungsdimension, z.B. „stimme zu – lehne ab" oder „stimme überhaupt nicht zu – stimme voll und ganz zu". Solche Likert-Type Skalen sind in Umfragen sehr beliebt, möglicherweise aufgrund einer kompakteren Darstellung der Fragen als Item-Matrix. Insbesondere kommen bipolare „stimme zu – lehne ab" Ratingskalen sehr häufig vor, sie werden bspw. als Äquivalent des englischsprachigen „agree – disagree" bei Übersetzungen und in interkulturellen Umfragen verwendet. So werden im ESS in Deutschland die Items „Gesetze sollten unter allen Umständen immer befolgt werden." oder „Politische Parteien, welche die Demokratie stürzen wollen, sollten verboten werden." mit „stimme zu – lehne ab" als Übersetzung des „agree-disagree" im Basisfragebogen bewertet.

Likert-Type Skalen können universell bei unterschiedlichen Aussagen in sog. Itembatterien genutzt werden. Beispielsweise wird im International Social Programme (ISSP) 2012 eine Ratingskala „stimme voll und ganz zu", „stimme zu", „weder noch", „stimme nicht zu" und „stimme überhaupt nicht zu" für die Bewertung der folgenden Aussagen genutzt (Terwey & Baltzer, 2013):

> Ein alleinstehender Elternteil kann sein Kind genauso gut großziehen wie beide Eltern zusammen.
>
> Ein Paar, bei dem beide Frauen sind, kann ein Kind genauso gut großziehen wie ein Mann und eine Frau.
>
> Ein Paar, bei dem beide Männer sind, kann ein Kind genauso gut großziehen wie ein Mann und eine Frau.

Eine Alternative wäre hier, die Erziehungsgüte zu skalieren, z. B. „Wie gut kann ein alleinstehender Elternteil sein Kind erziehen?". Die Bewertungsdimension wäre hier eine evaluative, z.B. „sehr schlecht – sehr gut". Diese Art der Nutzung von Ratingskalen wird als „Item-Specific Response Options" (Saris, Krosnick & Shaeffer, 2010) oder als „direkte" Abfrage bezeichnet.

Einige Studien zeigen, dass agree-disagree Ratingskalen Akquieszenz fördern (Billiet & McClendon, 2000). Krosnick und Presser (2010) führen Ergebnisse zusammen, die zeigen, dass die Akquieszenz auch bei „richtig-falsch" und „ja-nein" Vorgaben sehr wahrscheinlich ist. Aus diesem Grund wird vorgeschlagen, direkte Ratingskalen zu nutzen, insbesondere bei den Aussagen, wie „Ich bin häufig traurig" oder „Kurze Wartezeiten beim Arzt sind mir wichtig". Bei solchen Aussagen ist es einfacher, direkt die Häufigkeit oder die Wichtigkeit bewerten zu lassen, wodurch auch die Messgüte erhöht wird. Dies konnte bereits für unterschiedliche Länder (inklusive Deutschland) im ESS gezeigt werden (Saris et al., 2010).

<u>Fazit und Empfehlung:</u> Empirische Befunde legen nahe, direkte Ratingskalen zu nutzen. Agree-disagree (oder „stimme zu – lehne ab") Ratingskalen wären zu vermeiden, da sie höhere Zustimmungsraten erhalten als direkte Ratingskalen.

9. Graphische Darstellung von Skalen

Experimentelle Studien zeigen, dass graphische Elemente von Ratingskalen befragte Personen in ihrem Antwortverhalten systematisch beeinflussen können, da nicht nur verbale, sondern auch nonverbale, visuelle Fragebogenelemente zur Interpretation und Beantwortung der Fragen herangezogen werden (z. B. Smith, 1995; Christian & Dillman, 2004; Tourangeau, Couper & Conrad, 2004; Tourangeau, Couper & Conrad, 2007; Christian, Parsons & Dillman, 2009; Toepoel & Couper, 2011).

Für vertikale vs. horizontale Darstellungen von Ratingskalen wurden in experimentellen Studien signifikant unterschiedliche Antwortverteilungen beobachtet (z. B. Friedman & Friedman, 1994; Toepoel et al., 2009), allerdings konnte keine eindeutige Effektrichtung ausgemacht werden. Verschiedene Studien belegen aber, dass in vertikal dargestellten Ratingskalen Primacy-Effekte auftreten und daher eher einer horizontale Darstellungsform genutzt werden sollte (Tourangeau et al., 2000).

Ein wesentliches Element der visuellen Gestaltung von Ratingskalen ist der Skalenmittelpunkt, weil sich die befragten Personen bei der Interpretation der Skala daran orientieren. Es gibt Designs bei denen eine Diskrepanz zwischen inhaltlichem und visuellem Skalenmittelpunkt besteht, zum Beispiel wenn eine Weiß-nicht-Kategorie als weitere Kategorie angehängt und nicht mittels eines Trennstriches oder Leerraums von den inhaltlichen Kategorien abgetrennt wird, oder weil ungleichmäßige Abstände zwischen den Skalenkategorien vorliegen. Ein experimenteller Vergleich von Ratingskalen mit und ohne Übereinstimmung des inhaltlichen und visuellen Mittelpunktes findet signifikant unterschiedliche Antwortverteilungen (z. B. Tourangeau et al., 2004; Christian et al., 2009).

Tourangeau et al. (2004) fanden, dass bei unterschiedlicher farblicher Schattierung der extremen Antwortkategorien (Blautöne auf der ablehnenden Seite, Rottöne auf der zustimmenden Seite der Skala), befragte Personen den ablehnenden Bereich eher meiden, als wenn gleiche farbliche Schattierungen (z .B. nur Blautöne) der Extrempole genutzt werden. Diese Effekte traten aber nicht mehr auf, wenn die Skala mit verbalen Etiketten versehen wurde.

Fazit und Empfehlung: Generell empfehlen wir, zurückhaltend mit der Verwendung nicht aufgabenbezogener grafischer Elemente wie Farben, Schattierungen oder Symbole bei Ratingskalen zu sein, da sie zu unerwünschten Effekten führen können. Wichtig ist, dass anhand der graphischen Darstellung die Skalensymmetrie und die Äquidistanz der Kategorien widergegeben werden, z. B. sollten die nicht-inhaltlichen Kategorien von dem Rest der Ratingskala visuell abgehoben werden. Schließlich sollten Ratingskalen horizontal ausgerichtet werden, um den Primacy-Effekt zu reduzieren.

Literaturverzeichnis

Alwin, D. F. & Krosnick J. A. (1991). The reliability of survey attitude measurement: The influence of question and response attributes. *Sociological Methods and Research, 20*, 139-181.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural equation approach. *Public Opinion Quarterly, 48*, 409-448.

Billiet, J. & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling, 7 (4):* 608-628.

Bishop, G. F., Oldendick R. W. & Tuchfarber, A. J. (1983). Effects of filter questions in public opinion surveys. *Public Opinion Quarterly*, *47*, 528-546.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J. & Bennett, S. E. (1980). Pseudo-opinions on public affairs. *Public Opinion Quarterly, 44*, 198-209.

Christian, L. M. & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly, 68* (1), 57-80.

Christian, L. M., Parsons, N. L. & Dillmann, D. A. (2009). Designing scalar questions for web surveys. *Sociological Methods and Research, 37*, 393-425.

Clark, D. A. (1990). Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology: Research and Reviews, 9*(3), 203-235.

Converse, J. M. & Presser, S. (1986). *Survey questions*. Beverly Hills: Sage Publications, Inc.

Danner, D. (2015). *Reliabilität – die Genauigkeit einer Messung*. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (SDM Survey Guidelines). DOI: 10.15465/sdm-sg_011

DeRouvray, C. & Couper, M. P. (2002). Designing a strategy for reducing "no opinion" responses in Web-based surveys. *Social Science Computer Review, 20*, 3-9.

Dubois, B. & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement, 35*, 869-884.

Fowler, F. J. & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 15–36). San Francisco: Jossey-Bass.

French-Lazovik, G. & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8*(1), 49-57.

Friedman, L. W. & Friedman, H. H. (1994). A comparison of vertical and horizontal rating scales. *The Mid-Atlantic Journal of Business, 30,* 107-202.

Gilljam, M. & Granberg, D. (1993). Should we take the "Don't Know" for an answer? *Public Opinion Quarterly, 57*(3), 348-357.

Hippler, H. J. & Schwarz, N. (1989). "No opinion" filters: A cognitive perspective. *International Journal of Public Opinion Research, 1*, 77-87.

Hofmans, J., , Theuns, P., Baekelandt, S., Mairesse, O. Schillewaert, N., & Cools, W. (2007). Bias and changes in perceived intensity of verbal qualifiers affected by scale orientation. *Survey Research Methods 1*, 97–108.

Kalton, G., Robert, J. & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician, 29,* 65-78.

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement. A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77, 361-372.

Katz, D. (1942). Do interviewers bias pool results? *Public Opinion Quarterly, 6,* 248-268.

Krebs, D. (2012). The impact of response format on attitude measurement. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, Theories, and Empirical Applications in the Social Sciences* (pp. 105-113). Wiesbaden: VS Verlag für Sozialwissenschaften.

Krebs, D. & Hoffmeyer-Zlotnik, J. H. P. (2010). Positive first or negative first? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 118-127.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. A. (2002). The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be. In R. M. Groves, ,Don A. Dillman, John N. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 88-100). New York: Wiley-Interscience.

Krosnick, J. A. & Abelson, R. P. (1991). The case for measuring attitude strength in surveys. In J. M. Tanur (Ed.), *Questions about survey questions (pp. 177-203)*. New York: Russell Sage.

Krosnick, J. A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*, 201-219.

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: John Wiley & Sons, Inc.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Rudd, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of "no opinion" response options on data quality: non-attitude reduction or an invitation to satisfice? *The Public Opinion Quarterly, 66*, 371-403.

Krosnick, J. A., & Presser S. (2010). Question and Questionnaire Design. Peter V. Marsden und James D.
Wright (eds.), *Handbook of Survey Research*, (pp. 264-313). Bingley, UK: Emerald.

Krosnick, J. A., Narayan, S. S. & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. In M. T. Braverman & J. K. Slater (Eds.), *Advances in survey research* (pp. 29-44). San Francisco: Jossey-Bass.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (second edition) (pp. 263-313). Bingley, UK: Emerald Group.

Kulas, J. T., Stachowski, A. A. & Haynes, B. A. (2008). Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology, 22*(3), 251-259.

Lenzner, T.; Neuert, C. & Otto, W. (2015). *Kognitives Pretesting*. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (SDM Survey Guidelines). DOI: 10.15465/sdm-sg_010

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 1-55.

Maitland, A. (2009). How many scale points should I include for attitudinal questions? *Survey Practice 06*. AAPOR e-journal.

McClendon M. J. & Alwin, D. F. (1993). No-opinion filters and attitude measurement reliability. *Sociological Methods Research, 21*, 438-464.

Menold, N. (2013). Does the polarity of rating scales matter? How unipolar, bipolar and mixed rating scales affect measurements with latent variables. Paper presented at the *5th Conference of the European Survey Research Association (ESRA)*, 15-19 July, 2013.

Menold, N. & Bogner, K. (2012). Antwortskalen in sozialwissenschaftlichen Umfragen: Theoretische Modelle, Stand der Forschung und Forschungsperspektiven. In H.-G. Soeffner, (Hrsg.): *Transnationale Vergesellschaftungen. Verhandlungen des 35. Kongresses der Deutschen Gesellschaft für Soziologie in Frankfurt am Main 2010* (CD-ROM).Wiesbaden: VS Verlag.

Menold, N., Kaczmirek, L., Lenzner, T. & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods, 26*(1), 21-39.

Moors, G., Kieruj, N., & Vermunt, J. K. (in press). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology.*

O'Muircheartaigh, C., Krosnick, J. A. & Helic, A. (1999). Middle alternatives, acquiescence, and the quality of questionnaire data. Paper presented at the *annual meeting of the American Association for Public Opinion Research*, St. Petersburg, Florida.

Pajares, F., Hartley, J. & Vahante, G. (2001): Response format in writing self-efficacy assessment: Greater discrimination increases prediction. *Measurement and Evaluation in Counseling and Development, 33,* 214-221.

Parducci, A. (1983). Category ratings and the relational character of judgment. In H. G. Geissler, H. F. J. M. Bulfart, E. L. H. Leeuwenberg & V. Sarris, *Modern Issues in Perception* (pp. 262-282). Berlin: VEB Deutscher Verlag der Wissenschaften.

Payne, S. L. (1950). Thoughts about meaningless questions. *Public Opinion Quarterly, 14,* 687-696.

Poe, G. S., Seeman, I., McLaughlin, J., Mehl, E. & Dietz, M. (1988). Don't know boxes in factual questions in a mail questionnaire. *Public Opinion Quarterly, 52*, 212-222.

Pollack, S., Friedman H. H., & Presby L. (1990). Two salient factors in the construction of rating scales: Strength and direction of anchoring adjectives. *International Conference of Measurement Errors in Surveys*. Tucson, Arizona, November 11-14, p. 57.

Preston, C. C. & Colman, A. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15.

Rohrmann, B. (1978): Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, *1978*, 222-245.

Saris, W. E. & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. S*urvey Research Methods, 4,* 61-79.

Schaeffer, N. C., & Barker, K. (1995). Issues in using bipolar response categories: Numeric labels and the middle category. Paper presented at the *annual meeting of the American Association for Public Opinion Research,* Ft. Lauderdale, FL, May 23, 1995.

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

Schwarz, N., Bless, H., Bohner, G., Harlacher, U., & Kellenbenz, M. (1991). Response scales as frames of reference: The impact of frequency range on diagnostic judgment. *Applied Cognitive Psychology*, 5, 37-50.

Smith, T. W. (1995). Little things matter: A sampler of how differences in questionnaire format can affect survey responses. *Proceedings of the American Statistical Association, Survey Research Methods Section*: 1046-1051.

Sturgis, P., Roberts, C. & Smith, P. (2014). Middle alternatives revisited: How the neither/nor response acts as a way of saying "I don't know"? *Sociological Methods & Research, 43*(1), 15-38.

Terwey, M. & Baltzer, S. (2013): Variable Report ALLBUS / Allgemeine Bevölkerungsumfrage der Sozialwissenschaften 2012. ZA-Nr. 4614. Köln: GESIS, GESIS - Variable Reports; No. 2013/16.

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis. *Journal of Risk Research, 5*(2), 177-186.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review*, *36*(3), 222-241.

Toepoel, V. (2008). *A Closer Look at Web Questionnaire Design*. Tilburg: Tilburg University Press.

Toepoel, V. & Couper, M. P. (2011). Can verbal instructions counteract visual context effects in web surveys? *Public Opinion Quarterly*, *75*(1), 1-18.

Toepoel, V., Das, M. & van Soest, A. (2009). Design of web questionnaires: The effect of layout in rating scales. *Journal of Official Statistics, 25*, 509-528.

Toepoel, V. & Dillman, D.A. (2011). Words, numbers, and visual heuristics in web surveys: Is there a hierarchy of importance? *Social Science Computer Review, 29*(2), 193-207.

Tourangeau, R., Couper, M. P. & Conrad, F. G. (2004). Spacing, position and order. Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly, 68*, 368-393.

Tourangeau, R., Couper, M. P. & Conrad, F. G. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly, 71*, 91-112.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Vaillancourt, P. M. (1973). Stability ofchildren's survey responses. *Public Opinion Quarterly*, 37(3), 373-387.

Visser, P. S., Krosnick, J. A., Marquette, J. F. & Curtin, M. F. (2000): Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In P. Lavrakas& M. W. TRaugott (eds.), *Election polls, the news media, and democracy*. New York: Chatham House.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*, 176-190.

Windschitl, P. D. & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343-364.

Worcester, R. M. &. Burns, T. R. (1975). A Statistical Examination of the Relative Precision of Verbal Scales. *Journal of the Market Research Society 17* (3), 181–197.

Zaller, J. R. (1988). Vague questions vs. vague minds: Experimental attempts to reduce measurement error. Paper presented at the *annual meeting of the American Political Science Association*, Washington, D.C.