Einflüsse unterschiedlicher Formen der Verbalisierung von Antwortskalen auf das Antwortverhalten von Befragungspersonen

Antje Rosebrock
University of Mannheim (Germany)

Stephan Schlosser
University of Göttingen (Germany)

Jan Karem Höhne
University of Mannheim (Germany)
Universitat Pompeu Fabra (Spain)

Steffen M. Kühnel
University of Göttingen (Germany)

Zusammenfassung
Antwortskalen sind ein zentraler Bestandteil standardisierter Fragebögen. In der Surveyforschung wird häufig zwischen endpunktbenannten und vollverbalisierten Antwortskalen unterschieden. Bei ersteren werden nur die beiden Pole, bei letzteren alle Kategorien einer Antwortskala verbalisiert. Bei vollverbalisierten Skalen ist eine häufige Empfehlung, dass die Antwortkategorien äquidistant sind, was bedeutet, dass die Antwortkategorien den Wertebereich der Skala gleichmäßig abdecken. In diesem Beitrag vergleichen wir vollverbalisierte Skalen mit nicht-äquidistant erscheinenden Antwortkategorien, vollverbalisierte Skalen mit äquidistant erscheinenden Antwortkategorien und endpunktbenannte Skalen. Wir unterscheiden dabei zwischen der Beobachtungsebene, d.h. den von den Befragten ausgewählten Antwortkategorien, und der latenten Ebene, den nicht direkt beobachteten Positionen der Befragten auf der latenten kontinuierlich-metrischen Dimension. Die Ergebnisse zeigen, dass die unterschiedlichen Antwortskalen zu Unterschieden im empirischen Antwortverhalten führen. Diese Unterschiede sind nicht nur Folge einer unterschiedlichen Beziehung zwischen latenter Dimension und empirischer Antwortskala, sondern auch Folge unterschiedlicher Verteilungen auf der latenten Dimension.

Keywords: Antwortverhalten, Äquidistanz, Balancierung, Onlinebefragung und Ratingskalen

1 Einleitung und Hintergrund

Antwortskalen dienen in standardisierten Befragungen dazu, Einstellungen, Meinungen und das Verhalten von Befragungspersonen zu erfassen. Dabei müssen nach Krosnick und Presser (2010) Antwortskalen vier grundlegende Voraussetzungen gewährleisten, damit

This document is a preprint and thus it may differ from the final version: Rosebrock, A., Schlosser, S., Höhne, J.K., & Kühnel, S.M. (2019). Einflüsse unterschiedlicher Formen der Verbalisierung von Antwortskalen auf das Antwortverhalten von Befragungspersonen. In N. Menold & T. Wolbring (Eds.), Qualitätssicherung sozialwissenschaftlicher Erhebungsinstrumente (pp. 65–102). Wiesbaden: VS Verlag. DOI: 10.1007/978-3-658-24517-7_3

Befragungspersonen in der Lage sind, eine passende Antwortkategorie zu wählen: 1) Die Antwortkategorien sollten das gesamte inhaltliche Kontinuum abdecken; 2) angrenzende Antwortkategorien sollten sich inhaltlich nicht überschneiden; 3) die Antwortkategorien sollten über verschiedene Messungen hinweg präzise und verständlich sein; 4) alle Befragungspersonen sollten die Antwortkategorien in gleicher Weise interpretieren. Sind diese Voraussetzungen nicht gegeben, beeinflusst dies möglicherweise die Antworten von Befragungspersonen in unerwünschter Form, sodass Antwortqualität und Vergleichbarkeit abnehmen.

Die Gestaltung von Antwortskalen kann die mentale Urteilsbildung und somit auch die finale Antwort von Befragungspersonen beeinflussen. Bei der Konstruktion von Antwortskalen sind deswegen verschiedene Aspekte zu berücksichtigen, die einerseits zu positiven Effekten wie der Verbesserung des Frageverständnisses führen, andererseits aber auch negative Effekte zur Folge haben können, wie inhaltsunabhängige Antworttendenzen (DeCastellarnau 2017; Höhne und Krebs 2018; Menold und Bogner 2015). Aus diesem Grund sollten insbesondere die folgenden Aspekte bei der Antwortskalenkonstruktion Berücksichtigung finden (siehe Höhne und Krebs 2018): 1) die Entscheidung über die Verwendung einer geraden oder ungeraden Anzahl an Antwortkategorien; 2) die Entscheidung über die Anzahl möglicher Antwortkategorien; 3) die Entscheidung, ob eine ansteigende oder absteigende Reihenfolge der Antwortkategorien gewählt wird; 4) die Entscheidung, ob numerische und/oder verbale Labels verwendet werden.

Insbesondere die Verbalisierung von Antwortkategorien ist ein bedeutsamer Punkt in der Skalenkonstruktion, da Antwortkategorien eine verbale Orientierungs- sowie Verständnishilfe bieten und somit eine wichtige Determinante bei der mentalen Urteilsbildung sind (Ostrom und Gannon 1996; Parducci 1983). Im Grunde kann zwischen vollverbalisierten und endpunktbenannten Antwortskalen unterschieden werden. Während vollverbalisierte Skalen für jede Antwortkategorie eine entsprechende Verbalisierung beinhalten, verfügen endpunktbenannte Skalen lediglich an den Endpunkten über Verbalisierungen.

Obwohl endpunktbenannte Skalen in der Praxis sehr häufig verwendet werden, entspricht dies nicht den gängigen Empfehlungen aus der Surveyforschung. Stattdessen raten verschiedene Autoren zur Verwendung vollverbalisierter Skalen, da diese den Befragungspersonen mehr Informationen über die Abstufung zwischen den beiden Endpunkten liefern und somit kognitiv leichter zu verarbeiten sind (Johnson et al. 2005; Weng 2004). Daher wird oft argumentiert, dass vollverbalisierte Antwortskalen insbesondere für Befragungspersonen mit geringeren kognitiven Fähigkeiten eine Unterstützung bei der Beantwortung von Fragen darstellen (Krosnick und Fabrigar 1997; Rohrmann 1978). Tatsächlich zeigen empirische Befunde, dass vollverbalisierte Skalen im Gegensatz zu endpunktbenannten Skalen eine höhere Reliabilität und Validität erzielen (Coromina und Coenders 2006; DeCastellarnau 2017; Krosnick und Berent 1993; Krosnick und Presser 2010; Menold 2017). Zudem scheinen Befragungspersonen vollverbalisierte Skalen auch zu bevorzugen (Dickinson und Zellinger 1980; Zaller 1988).

Rohrmann (1978) zufolge erfordert die Verwendung vollverbalisierter Antwortskalen jedoch detaillierte Kenntnisse über die Angemessenheit der verwendeten Graduierungsbegriffe (Antwortkategorien). Er kann dies über empirische Analysen zum Verständnis von Graduierungsbegriffen zu Häufigkeiten, Intensitäten, Wahrscheinlichkeit und Bewertungen zeigen. Angemessen erscheinen Graduierungsbegriffe, wenn sie in der Lage

sind, die Bedeutung jeder einzelnen Kategorie im Kontext der Gesamtskala zu verdeutlichen (French-Lazovik und Gibson 1984; Parducci 1983). Die Problematik, adäquate Verbalisierungen zu vergeben, steigt aber insbesondere mit der Länge der Antwortskala (Krosnick und Presser 2010).

Sowohl die Position einer Antwortkategorie im Kontext der Gesamtskala als auch ihre spezifische Verbalisierung auf dem zugrundeliegenden Inhaltskontinuum kann das Antwortverhalten von Befragungspersonen beeinflussen (Friedman et al. 1981; Klockars und Yamagishi 1988). In diesem Zusammenhang unterscheidet Parducci (1983) zwischen zwei Dimensionen von Antwortskalen: dem Variationsbereich (Range) und ihrem Differenzierungsgrad (Frequency). Der Variationsbereich wird durch die Endkategorien festgelegt, und der Differenzierungsgrad ergibt sich aus der Anzahl der Antwortkategorien. Parducci (1983) geht zudem davon aus, dass Befragungspersonen eine subjektive Vorstellung der Spannweite des inhaltlichen Kontinuums haben. Diese subjektive Vorstellung wird von Befragungspersonen auf die Endkategorien der Antwortskala projiziert. Die mittleren Antwortkategorien geben dann einen Hinweis darauf, wie das inhaltliche Kontinuum zu differenzieren ist. Wichtig ist demnach, dass Antwortskalen über eindeutig verankerte Endpunkte (siehe Saris und Gallhofer 2007) sowie möglichst gleichabständige (äquidistante) Antwortkategorien verfügen. Diese Anforderungen sind auch für die Datenanalyse notwendig, bei der häufig metrisches Messniveau vorausgesetzt wird. Die Annahme einer äquidistanten Wahrnehmung durch die Befragungspersonen sollte daher nicht verletzt sein, auch wenn die entsprechenden Voraussetzungen nur selten geprüft werden (Lantz 2013; Rohrmann 1978).

Eine optimale Gestaltung von Antwortkategorien im Sinne von Parduccis Range-Frequency Ansatzes unterstützt die Vereinfachung kognitiver Antwortprozesse, da keine Divergenzen im Verständnis der Antwortkategorien zu erwarten sind. Einerseits kann dies verbal über die Wahl inhaltlich äquidistanter Kategorien und andererseits visuell über die optimale Anordnung von Antwortkategorien erfolgen. Hierzu gehört z.B. auch die Separierung von substantiellen und nicht-substantiellen Antwortkategorien (siehe Tourangeau et al. 2004). Beide Aspekte (Verbalisierung und Visualisierung) können die Wahrnehmung der Äquidistanz von Antwortkategorien wechselseitig beeinflussen (Friedman et al. 1981; Lantz 2013; Menold und Bogner 2015).

Neben der Studie von Rohrmann (1978) existieren nur wenige (jüngere) Untersuchungen, die sich mit Problemen der Äquidistanz empirisch auseinandersetzen. Auch der verwandte Aspekt der Balancierung – die Positionierung einer neutralen Kategorie in der konzeptionellen Mitte der Skala sowie das Vorhandensein gleich vieler „negativer“ und „positiver“ Kategorien – wird nur selten behandelt (Friedman et al. 1981; Liao 2014). Allerdings sind die Begriffe Äquidistanz und Balancierung nur bedingt abgrenzbar.[1]

Vorliegende Arbeiten zu Unterschieden zwischen verschieden verbalisierten Skalen beschränken sich mehrheitlich auf den Vergleich vollverbalisierter und endpunktbenannter Antwortskalen (Lantz 2013; Menold et al. 2014; Menold und Kemper 2015; Menold und Tausch 2016; Moors et al. 2014). In Bezug auf Äquidistanz kommt Lantz (2013) dabei zum Ergebnis, dass auch endpunktbenannte Skalen von den Befragungspersonen nicht grundsätzlich als äquidistant wahrgenommen werden.

Im vorliegenden Beitrag untersuchen wir mittels Split-Ballot-Experimenten in einer

[1] Dieser Punkt wird in der Diskussion aufgegriffen.

Onlinebefragung, ob und inwiefern sich verschiedene Antwortskalenformen auf die beobachteten und latenten Antwortverteilungen auswirken. Ausgangspunkt sind Antwortskalen von Fragen aus verschiedenen sozialwissenschaftlichen Bevölkerungsbefragungen, wie dem International Social Survey Programme (ISSP) und der Allgemeinen Bevölkerungsbefragung Sozialwissenschaften (ALLBUS), wobei die ursprünglichen Antwortskalen nach unserer Einschätzung das Kriterium der Äquidistanz vermutlich nicht erfüllten. Wir haben die Antwortskalen bei unveränderter Frageformulierung daher neu formuliert und dabei versucht, Äquidistanz zu erreichen. Neben dieser modifizierten vollverbalisierten Version wurde eine weitere Version der Antwortskala generiert, bei der nur die Endpunkte verbalisiert wurden. Die Verbalisierungen orientieren sich dabei an der modifizierten vollverbalisierten Version, übernehmen also deren Pole.

Bisherige Arbeiten beschränken sich auf die Analyse der Häufigkeitsverteilungen der empirischen Antworten sowie auf Vergleiche von Mittelwertdifferenzen als Folge unterschiedlicher Antwortskalen (bspw. Friedman et al. 1981). Bei der Berechnung von Mittelwerten wird jedoch ungeprüft metrisches Skalenniveau unterstellt. Demgegenüber berücksichtigen wir in diesem Beitrag die Ordinalität der Antwortskalen und betrachten die vorgegebenen Antwortskalen als grobe Messungen einer metrisch-kontinuierlichen (latenten) Antwortdimension (siehe Kühnel 1993). Wir unterscheiden somit explizit zwischen der eigentlich interessierenden, nicht beobachtbaren Antwort auf einer latenten Antwortdimension und der ausgewählten Antwortkategorie auf der vorgegebenen Antwortskala.

Im Folgenden werden wir zunächst Hypothesen formulieren und das experimentelle Forschungsdesign vorstellen. Daran anschließend beschreiben wir die realisierte Stichprobe sowie die Analysestrategie. Es folgt die Darstellung der Ergebnisse unserer Analysen und eine abschließende Diskussion.

2 Forschungshypothesen

Wir haben Fragen aus verschiedenen nationalen und internationalen Befragungen ausgewählt, die das Kriterium der Äquidistanz vermutlich nicht erfüllen. In dieser Studie wird untersucht, ob und inwiefern Modifikationen der Antwortskalen Einfluss auf die beobachteten Antwortverteilungen, auf die Beziehungen zwischen beobachteter Antwortskala und der latenten Version und auf die latenten Antwortverteilungen haben. Unsere Erwartungen bezüglich dieser Auswirkungen lassen sich in drei Hypothesen zusammenfassen:

H1a: Wir erwarten, dass sich die beobachteten Verteilungen der Antworten bei unterschiedlichen Verbalisierungen der Antwortskalen deutlich unterscheiden.

Gäbe es keine Unterschiede, dann hätten unterschiedliche Antwortskalen keine Konsequenz für das Antwortverhalten. Das widerspricht ganz offensichtlich den oben berichteten Befunden und würde auch Empfehlungen für Antwortskalen obsolet machen.

H1b: Wir erwarten zudem, dass sich die beobachteten Antwortverteilungen stärker zwischen zwei vollverbalisierten Antwortskalen mit unterschiedlich verbalisierten Antwortkategorien unterscheiden als zwischen einer vollverbalisierten und einer endpunktbenannten Skala, bei denen die Endpunkte identisch verbalisiert sind.

Diese Vermutung ist eine Konsequenz der berichteten Befunde, dass unterschiedliche Antwortkategorien bei vollverbalisierten Antwortskalen unterschiedliches Antwortverhalten und unterschiedliche Abstände zwischen den Antwortkategorien implizieren. Bei voll- und endpunktbasierten Skalen mit identischen Polen ist zumindest der Range der beobachteten Antwortskala gleich, so dass hier mit geringeren Unterschieden zu rechnen ist, als wenn auch die Endpunkte unterschiedlich verbalisiert sind.

H2: Wir erwarten, dass unterschiedliche Verbalisierungen der Antwortskalen primär die Beziehung zwischen latenter Antwortskala und beobachteter Antwortskala beeinflussen.

H3: Wir erwarten zudem, dass unterschiedliche Verbalisierungen der Antwortskalen keine Auswirkungen auf die Antwortverteilungen auf der latenten Dimension haben.

Die zweite und dritte Hypothese beziehen sich auf die Art und Weise, wie unterschiedliche Antwortskalen wirken. Dass sie die Beziehung zwischen latenten Antwortskalen und beobachtetem Antwortverhalten beeinflussen, ergibt sich etwa aus den Befunden von Rohrmann (1978), nach denen unterschiedliche Verbalisierungen zu unterschiedlichen Abständen zwischen den Antwortkategorien führen. Eine solche Aussage ist überhaupt nur sinnvoll, wenn von einem unbeobachteten und stabilen metrischen Kontinuum ausgegangen wird, dem sich eine gemessene kategoriale Antwortskala annähern soll. Die in Hypothese 3 postulierte Stabilität ist somit eine Voraussetzung dafür, dass über ein Item auch bei unterschiedlichen Antwortvorgaben die gleiche Dimension gemessen wird. Falls jedoch die Verteilung (entgegen H3) auf der latenten Antwortdimension bei ansonsten unveränderten Bedingungen variiert, dann misst ein Item mit verschiedenen Antwortskalen jeweils etwas Unterschiedliches. Antworten auf der Basis unterschiedlicher Antwortvorgaben sind dann aufgrund fehlender Messäquivalenz nicht mehr vergleichbar.

3 Methoden

3.1 Forschungsdesign

Für das Experiment wurden in einer Onlinebefragung die Befragungspersonen zufällig einer von drei Experimentalgruppen zugeordnet.[2] Die erste Gruppe (n = 868) beantwortete vier Fragen mit den originalen Antwortskalen. Die zweite Gruppe (n = 897) beantwortete identische Fragen mit modifizierten, vollverbalisierten Antwortskalen. Die dritte Gruppe (n = 871) beantwortete ebenfalls identische Fragen mit modifizierten und ausschließlich endpunktbenannten Antwortskalen, wobei die Endpunkte mit den Formulierungen in der modifizierten vollverbalisierten Antwortskala übereinstimmen. Da die Zuordnung zu den Gruppen randomisiert ist, können sich die „wahren“ Positionen zu den in den Fragen thematisierten Dimensionen höchstens zufällig unterscheiden.

3.2 Beschreibung der Fragen

Im Folgenden werden die ursprünglichen Fragen und die fragespezifischen Modifikationen vorgestellt, und es wird diskutiert, welche Folgen wir aus den unterschiedlichen Varianten der

[2] Die leicht unterschiedlichen Fallzahlen in den Gruppen sind Folge des in Abschnitt 4 beschriebenen Ausschlusses von Fällen nach der Datenerhebung.

Antwortskalen erwarten.

Tabelle 1. Allgemeine Zufriedenheit

	Ganz allgemein: Würden Sie sagen, Sie sind zur Zeit …			
Original	sehr glücklich	ziemlich glücklich	nicht sehr glücklich	überhaupt nicht glücklich
Modifiziert (vollverbalisiert)	sehr glücklich	eher glücklich	eher nicht glücklich	überhaupt nicht glücklich
Modifiziert (endpunktbenannt)	sehr glücklich			überhaupt nicht glücklich

Diese Frage wurde der deutschsprachigen Fragebogenversion der fünften Welle des World Value Survey (WVS 2006) entnommen. Wir vermuten bei der originalen Antwortskala eine größere Distanz zwischen den Antwortkategorien „nicht sehr glücklich“ und „überhaupt nicht glücklich“ als zwischen „sehr glücklich“ und „ziemlich glücklich“. Durch die Reformulierungen erwarten wir eine Angleichung der Distanzen.

Tabelle 2. Lebensziele

	Denken Sie jetzt einmal an Ihre persönliche Situation. Haben sich – einmal alles zusammengenommen – Ihre Vorstellung über das, was Sie im Leben erreichen wollen, bisher …			
Original	mehr als erfüllt	erfüllt	nicht ganz erfüllt	überhaupt nicht erfüllt
Modifiziert (vollverbalisiert)	ganz und gar erfüllt	eher erfüllt	eher nicht erfüllt	überhaupt nicht erfüllt
Modifiziert (endpunktbenannt)	ganz und gar erfüllt			überhaupt nicht erfüllt

Die Frage zu den Lebenszielen wurde aus der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS 2010) entnommen. Problematisch ist hier insbesondere die Distanz zwischen den Kategorien „nicht ganz erfüllt“ und „überhaupt nicht erfüllt“ im Vergleich zu „mehr als erfüllt“ und „erfüllt“. Während die Verbalisierung der ersten drei Kategorien nur eine sehr geringe Differenzierung impliziert, scheint uns der interpretative Abstand zwischen den beiden letzten Kategorien deutlich größer zu sein. Darüber hinaus erscheint „mehr als erfüllt“ nur bedingt als Endpunkt geeignet zu sein, da unklar ist, um wieviel „mehr“ es sich handelt.

Tabelle 3. Politisches Interesse

	Wie stark interessieren Sie sich im Allgemeinen für Politik?				
Original	sehr stark	stark	mittelmäßig	weniger stark	überhaupt nicht
Modifiziert (vollverbalisiert)	sehr stark	stark	mittelmäßig	schwach	sehr schwach
Modifiziert (endpunktbenannt)	sehr stark				sehr schwach

Die Frage zum politischen Interesse wurde der German Longitudinal Election Study

(GLES 2009) entnommen. Es ist zu vermuten, dass insbesondere der Abstand zwischen den Antwortkategorien „mittelmäßig“, „weniger stark“ und „überhaupt nicht“ anders wahrgenommen wird als zwischen „mittelmäßig“, „stark“, und „sehr stark“ (Originalversion). Die Kategorie „weniger stark“ liegt vermutlich inhaltlich näher an der Antwortkategorie „mittelmäßig“ als an der Antwortkategorie „stark“. Zu erwarten ist auch, dass die Änderung von einer unipolaren zu einer bipolaren Skala die Antwortverteilung beeinflusst.

Tabelle 4. Gesundheitszustand

	Alles in allem betrachtet, würden Sie sagen, Ihre Gesundheit ist …				
Original	ausgezeichnet	sehr gut	gut	mittelmäßig	schlecht
Modifiziert (vollverbalisiert)	sehr gut	gut	mittelmäßig	schlecht	sehr schlecht
Modifiziert (endpunktbenannt)	sehr gut				sehr schlecht

Diese Frage wurde aus dem Fragebogen des International Social Survey Programme (ISSP 2011) entnommen. Insgesamt betrachtet ist anzunehmen, dass die Antwortkategorien als nicht-äquidistant gesehen werden können, weil uns die Abstände auf dem positiven Kontinuum deutlich kleiner erscheinen als der Abstand zwischen „mittelmäßig“ und „schlecht“. Problematisch erscheint zudem, dass nicht das gesamte negative inhaltliche Kontinuum der bipolaren Antwortskala abgedeckt ist. Implizit bedarf es einer Erweiterung, wie „sehr schlecht“ und „außerordentlich schlecht“.

4 Beschreibung der Stichprobe

Die Grundlage dieser Studie bilden Daten, die im April und Mai 2017 an zwei deutschen Universitäten erhoben wurden. Die Befragungspersonen wurden mittels einer E-Mail um ihre Teilnahme an einer Onlinebefragung gebeten. Das Anschreiben in der E-Mail beinhaltete Informationen zum Studieninhalt, dem Ablauf und Umfang der Befragung sowie den Link zur Onlinebefragung.

Insgesamt wurden 34.856 Studierende eingeladen, an der Befragung teilzunehmen. 3.309 der eingeladenen Studierenden besuchten die erste Fragebogenseite (9,49%). Für die Analysen wurden 187 Personen ausgeschlossen, die lediglich die erste Seite mit dem Begrüßungstext besucht haben. Weitere Personen wurden ausgeschlossen, weil die Befragung vor der ersten studienrelevanten Frage abgebrochen wurde (n = 356) oder weil keine substantiellen Antworten auf die studienrelevanten Fragen gegeben wurden (n = 8). Bei 122 Personen war die Voraussetzung Deutsch als Muttersprache nicht erfüllt. Die nachfolgenden Analysen beziehen sich somit auf Angaben von 2.636 Befragungspersonen.

Von den Befragungspersonen haben 65,4% (n = 1.724) über einen PC oder Laptop an der Befragung teilgenommen, und 34,6% (n = 912) über ein mobiles Endgerät, wie ein Smartphone oder ein Tablet.[3] Die Befragungspersonen waren zum Zeitpunkt der Befragung

[3] Bei allen folgenden Analysen wird auf eine getrennte Darstellung nach Endgeräten verzichtet, da keine randomisierte Zuordnung erfolgt ist. Bei den Analysen zeigen sich geringe Unterschiede für die Befragungspersonen mit mobilen Endgeräten in Bezug auf die Standardfehler/Signifikanzen der Unterschiede in den latenten Verteilungen, die jedoch nichts an der substantiellen Interpretation der Ergebnisse ändern und sehr wahrscheinlich auf die unterschiedlichen Fallzahlen zurückzuführen sind.

zwischen 18 und 60 Jahre alt.[4] Das durchschnittliche Alter beträgt 24,9 Jahre mit einer Standardabweichung von 4,5 Jahren. 61,9% der Befragungspersonen sind weiblich und 92,1% haben zuvor bereits mindestens einmal an einer Onlinebefragung teilgenommen. Chi-Quadrat-Tests zeigen keine statistisch signifikanten Unterschiede zwischen den drei Experimentalgruppen in Bezug auf Geschlecht, Alter und Befragungserfahrung.

5 Analytisches Vorgehen

Aufgrund der offensichtlichen Ordinalität der Antwortvorgaben unterscheiden wir zwischen dem empirischen Antwortverhalten auf einer vorgegebenen ordinalen Antwortskala und der eigentlich interessierenden latenten metrisch-kontinuierlichen Antwortdimension, die mit einer Frage und deren Antwortvorgaben erfasst werden soll. Die Unterscheidung zwischen beobachtetem Wert y_i einer Untersuchungseinheit i auf eine Frage Y und dem eigentlich interessierenden „wahren" Wert (True Score) θ_i findet sich bereits in der klassischen Testtheorie (KTT), bei der jeder beobachtete Wert die Summe aus dem True Score und einem zufälligen Messfehler ε_i ist (Moosbrugger 2012). Streng genommen impliziert die Additivität der KTT, dass True Scores, Messfehler und beobachtete Werte auf einer gemeinsamen kontinuierlich-metrischen Dimension erfasst werden.

Wenn wir von ordinalen Antworten ausgehen, entspricht dies eher der Logik der Item Response Theory (IRT), bei der die Wahrscheinlichkeit einer realisierten Antwortkategorie eines binären oder ordinalen Items eine Funktion von itemspezifischen Parametern und der Position der Befragungsperson auf der interessierenden latenten Dimension, also dessen True Score, ist (Geiser und Eid 2010). Sowohl bei der KTT als auch bei der IRT werden für eine Stichprobe von Befragungspersonen Messungen unterschiedlicher Items für eine Dimension benötigt, um Aussagen über die Messqualität treffen zu können.[5]

Wir analysieren hier jedoch jedes Item getrennt für sich. Zur Modellierung der Beziehung zwischen der ordinalen vorgegebenen Antwortskala und der zugrunde liegenden latenten Antwortdimension θ formulieren wir ein ordinales Probitmodell, nach dem die Wahrscheinlichkeit, dass eine Befragungsperson i bei dem Item Y die Kategorie k von insgesamt K vorgegebenen ordinalen Kategorien (Antwortvorgaben) wählt, von der Position θ_i der Person auf der latenten Dimension und itemspezifischen Schwellenwerten τ_k (k = 0,1, … , K-1 mit $\tau_0 = -\infty$, $t_{K-1} = +\infty$) abhängt (Jöreskog 1994; Millsap und Yun-Tein 2004; Muthén 1984):

$$Pr(y_i = k) = Pr(\tau_{k-1} \leq \theta_i < \tau_k) = \Phi(\theta_i < \tau_k) - \Phi(\theta_i < \tau_{k-1}).$$

Die Differenzen zwischen zwei Schwellenwerten τ_k und τ_{k-1} eines Items definieren dabei den

[4] 65 (2,47%) der Befragungspersonen haben angegeben, dass sie zum Zeitpunkt der Befragung älter als 35 Jahre waren. Möglicherweise handelt es sich um Studierende des dritten Lebensalters, Personen in einer Weiterbildung oder Berufstätige in einem Teilzeitstudium.

[5] Alternativ können auch Messwiederholungen betrachtet werden, die aber das Problem aufweisen, dass Messwiederholungen nicht unabhängig voneinander sind und dass Messfehler von Veränderungen der Position auf der latenten Dimension über die Zeit unterschieden werden müssen.

Bereich, den die Antwortkategorie k auf der latenten Dimension θ umfasst.[6] Bei gleichabständigen Antwortkategorien verteilen sich die Schwellenwerte τ_1 bis t_{K-1} gleichabständig auf der latenten Dimension.

Da sich die Hypothesen H1a und H1b nur auf die beobachteten Antwortskalen beziehen, kann deren Prüfung ohne Schätzung des ordinalen Probitmodells erfolgen. Stattdessen testen wir die Antworten der Befragungspersonen auf ein Item mit Pearsons Chi-Quadrat-Test auf Gleichheit der Verteilungen in den drei Gruppen. Nach Hypothese H1a erwarten wir dabei signifikante Unterschiede zwischen den Gruppen. Wird diese Hypothese verworfen, bedeutet dies, dass es bei dem analysierten Item vermutlich keine Rolle spielt, welche der drei Antwortskalen verwendet wird. Dann erübrigen sich weitere Analysen, weil gleiche Verteilungen auf der Ebene der beobachteten Antworten, bezogen auf das ordinale Probitmodell, implizieren, dass sich weder die Schwellenwerte noch die latenten Verteilungen unterscheiden. Eine Ablehnung von H1a impliziert somit auch die Ablehnung von H1b und H2 und eine Annahme von H3.

In H1b postulieren wir größere Unterschiede zwischen den beiden vollverbalisierten Antwortversionen als zwischen der vollverbalisierten Version und der endpunktbenannten Version mit identisch benannten Endpunkten. Für die Prüfung vergleichen wir das Ergebnis des Chi-Quadrat-Tests auf Gleichheit der beiden vollverbalisierten Antwortskalen mit dem Chi-Quadrat-Test auf Gleichheit der vollverbalisierten Originalformulierung und der endpunktbenannten Antwortskala. Die Hypothese betrachten wir als widerlegt, wenn das empirische Signifikanzniveau (p-Wert) des ersten Tests nicht kleiner ist als das des zweiten Tests.[7]

Für die Prüfung von H2 und H3 müssen wir für ein Item das oben vorgestellte ordinale Logitmodell schätzen. Aus Identifikationsgründen ist es jedoch nicht möglich, sowohl die Schwellenwerte τ_k eines Items als auch die individuellen Positionen θ_i auf der latenten Dimension zu schätzen. Solange nur eine Stichprobe vorliegt, können zudem nur die K-1 Schwellenwerte geschätzt werden, wenn angenommen wird, dass die individuellen Positionen standardnormalverteilt sind.[8]

Da wir jedoch im experimentellen Design drei Gruppen (Stichproben) haben, deren latente Verteilung der True Scores auf der latenten Dimension θ als Folge der

---

[6] Ein Schwellenwert entspricht der Itemschwierigkeit in einem IRT mit polytomen geordneten Kategorien. Bei der IRT geben die Itemschwierigkeiten allerdings nicht die Grenze zwischen zwei aufeinanderfolgenden Kategorien an, sondern die Wahrscheinlichkeit, dass eine Kategorie mit einer Wahrscheinlichkeit von 50% überschritten wird. Anstelle von K-1 Schwellenwerten gibt es dann K-1 Schwierigkeitsparameter.

[7] Streng genommen müsste inferenzstatistisch geprüft werden, ob sich die Antwortverteilungen zwischen vollverbalisierten und endpunktbenannten Antwortskalen ähnlicher sind als die Antwortverteilungen zwischen den beiden vollverbalisierten Skalen. Wir verzichten auf die recht komplexe inferenzstatistische Absicherung (etwa über den simultanen Test auf Gleichheit mehrerer Parameter eines spezifischen loglinearen Modells) und betrachten nur die Unterschiede bei Pearsons Chi-Quadrat-Statistik, da diese eine monotone Funktion des Zusammenhangsmaß Cramérs V zwischen Gruppenzugehörigkeit und Antwortverhalten ist. Um die leicht unterschiedlichen Fallzahlen zu berücksichtigen, betrachten wir nicht nur die Teststatistik selbst, sondern den daraus berechneten p-Wert.

[8] Wenn anstelle der Normalverteilung eine andere Verteilung angenommen wird, ergibt sich ein anderes Modell, z.B. bei logistisch verteilten True Scores ein ordinales Logitmodell. Aufgrund fehlender Informationen ist es nicht möglich, die tatsächliche Verteilung zu bestimmen. Bei nur einer Stichprobe und einem Messzeitpunkt bzw. Item sind zudem der Mittelwert und die Varianz der latenten Positionen der Befragten nicht identifiziert, weswegen von standardnormalverteilten True Scores ausgegangen wird.

Randomisierung – wie in H3 formuliert - identisch sein sollte, können wir im simultanen Gruppenvergleich Unterschiede zwischen den Gruppen testen. Unter der Annahme gleicher latenter Verteilungen der True Scores in den Gruppen (H3) lassen sich die Schwellenwerte vergleichen und auf Unterschiede testen. Werden mindestens zwei Schwellenwerte über Gruppen gleichgesetzt, lassen sich zudem die Gleichheit der Mittelwerte und Varianzen der True Scores und damit H3 testen.

Für die Berechnung der Modellparameter verwenden wir Weighted Least Squares (WLS)-Schätzer mit dem Programm Mplus Version 6.12 (Muthén und Muthén 2012). Wie bei der ML-Schätzung eines Probitmodells sind die Schätzer bei korrekter Modellspezifikation konsistent und asymptotisch erwartungstreu (Mullahy 1990). Ein Vorteil gegenüber der ML-Schätzung besteht darin, dass restriktive Modelle leichter zu schätzen sind, da keine multivariaten Normalverteilungen berechnet werden müssen.

Zunächst wird für jede Frage ein gerade identifiziertes (saturiertes) Modell mit frei variierenden Schwellenwerten und gleichen Mittelwerten und Varianzen der True Scores geschätzt (M1). Die dann möglichen Unterschiede zwischen den Schwellenwerten geben Hinweise für die Wahrnehmung der unterschiedlichen Antwortkategorien durch die Befragungspersonen. Unterstellt wird dabei die Gültigkeit von H3 (Gleichheit der Verteilungen auf der latenten Ebene). Aufgrund des experimentellen Designs ist das Zutreffen von H3 zu erwarten. Um gleichwohl H3 zu testen, schätzen wir ein zweites Modell (M2), in dem alle Schwellenwerte gleichgesetzt sind. In diesem Modell können dann Differenzen zwischen den Mittelwerten und Varianzen von θ in den drei Gruppen geschätzt werden. Da alle Items vier oder fünf Antwortkategorien haben, und damit drei bzw. vier Schwellenwerte aufweisen, ist das Modell M2 im Unterschied zu M1 überidentifiziert, sodass die Übereinstimmung von Modell und Daten mit einem Goodness-of-Fit-Test geprüft werden kann.

Ist der Modellfit befriedigend und unterscheiden sich die Mittelwerte und Varianzen nicht, dann entspricht dieser Test Pearsons Chi-Quadrat-Test zur Prüfung von H1a. Der Unterschied besteht darin, dass hier die Ordinalität der Antwortskalen berücksichtigt wird, was bei Pearsons Chi-Quadrat-Test nicht der Fall ist. Die Ablehnung von H1a – also keine Unterschiede – impliziert gleichzeitig die Ablehnung von H1b und H2 sowie die Annahme von H3.

Unterscheiden sich dagegen bei befriedigendem Modellfit Mittelwerte und/oder Varianzen, bedeutet dies, dass sich die unterschiedlichen Antwortskalen ausschließlich auf die zu messende Beurteilungsdimension auswirken, während die Schwellenwerte sich nicht unterscheiden. Die Beziehung zwischen latenter Dimension und ordinaler Messung ist also gleich, was im Widerspruch zu H2 steht. Formal entspricht die Gleichsetzung der Schwellenwerte über die drei Gruppen skalarer Invarianz (van de Vijver 2003).[9] Wenn sich bei hinreichendem Modellfit die latenten Mittelwerte und/oder Varianzen unterscheiden, bedeutet dies, dass sich als Folge der unterschiedlichen Antwortskalen die Mittelwerte verschieben bzw. die Unterschiedlichkeit der Bewertungen zu- bzw. abnehmen. Da sich aufgrund des experimentellen Designs die Verteilung der True Scores zwischen den Gruppen

[9] Bei ordinalen Variablen sind Mittelwert und Varianz nicht definiert. Sind Schwellenwerte gleich, bedeutet dies, dass die latenten metrischen Variablen, deren ordinale Messungen vorliegen, formal die gleiche Einheit und den gleichen Nullpunkt haben, was in der Diskussion um Messinvarianz als skalare Invarianz bezeichnet wird.

aber eigentlich nicht unterscheiden sollte, folgt, dass trotz formaler skalarer Invarianz des Messinstruments durch unterschiedliche Antwortvorgaben unterschiedliche True Scores generiert werden.[10] Dies ist aber nur möglich, wenn die Kombination aus identischer Itemformulierung und unterschiedlichen Antwortvorgaben für die Befragungspersonen Unterschiedliches messen. H3 ist dann widerlegt.

Fittet das Modell hingegen nicht, spricht dies gegen die Annahme gleicher Schwellenwerte. Dann ist davon auszugehen, dass die Antwortkategorien das Schwellenwertmodell beeinflussen, was für H2 sprechen würde. Ist dies der Fall, setzen wir zur genaueren Analyse der Unterschiede verschiedene Schwellenwerte über die Gruppen frei. Solange mindestens zwei Schwellenwerte gleichgesetzt bleiben, können die übrigen Schwellenwerte, bei gleichzeitiger Schätzung von Mittelwerten und latenten Verteilungen, geschätzt werden. Wenn sich die latenten Mittelwerte und Varianzen trotz Freigabe von Schwellenwerten weiterhin unterscheiden, spricht dies nicht nur für Unterschiede in der Interpretation der Antwortkategorien (Unterstützung von H2), sondern auch dafür, dass zusätzlich eine unterschiedliche Reaktion auf die durch die jeweilige Fragenformulierung vorgegebene Beurteilungsdimension ausgelöst wird (Widerlegung H3).[11]

6 Ergebnisse

6.1 Allgemeine Zufriedenheit

Die Frage nach der allgemeinen Zufriedenheit zeigt deutliche Unterschiede in den beobachteten Verteilungen zwischen den drei Versionen der Antwortvorgaben (Tabelle 5). Die Antwortkategorie „sehr glücklich“ wird in beiden modifizierten Versionen häufiger ausgewählt als in der Originalversion. Dagegen werden die beiden mittleren Antwortvorgaben der vierstufigen Skala im Falle der Originalversion insgesamt seltener ausgewählt. Die Unterschiede bei den Antwortverteilungen zwischen beiden modifizierten Versionen und der ursprünglichen Version sind signifikant, die endpunktbenannte und die vollverbalisierte Version unterscheiden sich jedoch nicht signifikant. Bei dieser Frage werden daher sowohl H1a wie H1b nicht widerlegt.

[10] Dies weist darauf hin, dass das für metrische Variablen entwickelte Konzept der metrischen oder skalaren Invarianz bei ordinalen Variablen nicht hinreichend ist.

[11] Aufgrund der sehr ausführlichen Beschreibung der analytischen Vorgehensweise sowie Platzgründen verzichten wir auf die zusätzliche Darstellung des Mplus Codes.

Tabelle 5. Verteilungen und kumulierte Verteilungen zur Frage zur allgemeinen Zufriedenheit

	Ganz allgemein: Würden Sie sagen, Sie sind zur Zeit …			
Original	sehr glücklich	ziemlich glücklich	nicht sehr glücklich	überhaupt nicht glücklich
% (kum. %)	16.1	66.6 (82.6)	16.1 (98.7)	1.3 (100)
Modifiziert (vollverbalisiert)	sehr glücklich	eher glücklich	eher nicht glücklich	überhaupt nicht glücklich
% (kum. %)	21.2	60.9 (82.1)	15.5 (97.5)	2.5 (100)
Modifiziert (endpunktbenannt)	sehr glücklich			überhaupt nicht glücklich
% (kum. %)	25.3	59.8 (85.1)	13.4 (98.5)	1.5 (100)

Anmerkungen: $\chi^2(6) = 27.5$, $p < .01$; Original n = 864, Modifiziert (vollverbalisiert) n = 893, Modifiziert (endpunktbenannt) n = 871; Original vs. Modifiziert (vollverbalisiert) $\chi^2(3) = 11.7$, $p < .01$; Modifiziert (vollverbalisiert) vs. Modifiziert (endpunktbenannt) $\chi^2(3) = 6.6$, $p = .085$; Original vs. Modifiziert (endpunktbenannt) $\chi^2(3) = 23.0$, $p < .001$.

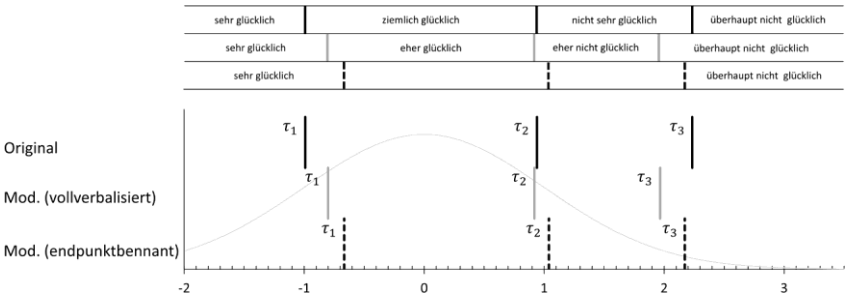

Abbildung 1. Schwellenwerte und latente Antwortverteilungen bei der allgemeinen Zufriedenheit (M1)

Abbildung 1 zeigt die Position der Schwellenwerte und die geschätzte Normalverteilung auf der latenten Antwortdimension für das Probitmodell M1. Bei der hier unterstellten Annahme gleicher Antwortverteilungen zwischen den drei Gruppen zeigen sich eher geringe Unterschiede zwischen den Schwellenwerten (Abbildung 1). Die Vermutung liegt also nahe, dass diese gleichgesetzt werden können. Geschieht dies, (Abbildung 2), und werden dabei die in M1 gleichgesetzten Mittelwerte und Varianzen der latenten Antwortverteilungen zwischen den Gruppen freigegeben, ergibt sich eine sehr gute Übereinstimmung des Modells mit den Daten (Modellfit M2: $\chi^2(2) = .605$, $p = .739$; RMSEA = .001). Daraus kann geschlossen werden, dass sich die Schwellenwerte zwischen den Gruppen nicht unterscheiden, was wie oben beschrieben, Hypothese H2 widerlegt.

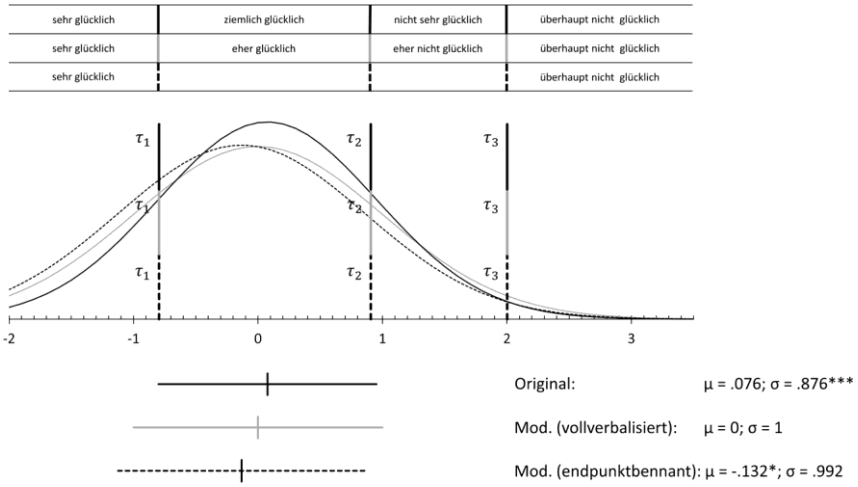

Abbildung 2. Schwellenwerte und latente Antwortverteilungen bei gleichgesetzten Schwellenwerten zur Frage zur allgemeinen Zufriedenheit (M2)

Anmerkungen: *: $p < .05$, **: $p < 0.01$, ***: $p < .001$

Gleichzeitig ergeben sich nun aber signifikante Unterschiede in den latenten Verteilungen. Gegenüber der Gruppe mit modifizierten vollverbalisieren Antwortkategorien ist die Varianz in der Gruppe mit den originalen Antwortkategorien signifikant geringer ($p < .001$). Zudem ist auch der Mittelwert bei den Befragungspersonen, denen die endpunktbenannte Antwortskala vorgelegt wurde, signifikant geringer ($p < .05$). Aufgrund der experimentellen Zuordnung zu den Gruppen können diese latenten Unterschiede nicht auf unterschiedliche Eigenschaften der Gruppen zurückgeführt werden, sondern müssen Folge der Unterschiede in den Antwortskalen sein. Damit ist H3 widerlegt. Inhaltlich bedeutet dies, dass unterschiedliche Antwortkategorien bei der Frage nach der Zufriedenheit nicht nur die beobachteten Antworten, sondern auch die latenten Antworten beeinflussen. Bei unterschiedlichen Antwortskalen kann das Ausmaß der Zufriedenheit zwischen zwei Stichproben nicht verglichen werden.

Dieses Ergebnis zeigt sich auch in Chi-Quadrat-Differenzentests, wenn jeweils zwei der drei Gruppen gegeneinander getestet werden, wobei in den liberaleren Modellen (M2a-M2c) ausschließlich die Schwellenwerte und in den restriktiveren Modellen (M3a-M3c) zusätzlich die latenten Verteilungen gleichgesetzt werden. Bei allen drei Vergleichen verschlechtern sich die Modellanpassungen bei gleichgesetzten latenten Verteilungen signifikant (Tabelle 6). Es muss davon ausgegangen werden, dass sich ausschließlich die latenten Verteilungen, aber nicht das Messmodell zwischen den Gruppen unterscheiden.

Wird beim Vergleich der Gruppen mit den endpunktbenannten und den vollverbalisierten Antwortskalen das Modell mit gleichen Schwellenwerten und gleichen latenten Verteilungen (M3a: $\chi^2(3) = 6.631$, $p = .085$) mit einem liberaleren Modell (M4a: $\chi^2(2) = 5.853$, $p = .054$) verglichen, in welchem bei gleichen latenten Verteilungen nur die äußersten Schwellenwerte gleichgesetzt sind, der mittlere Schwellenwert dagegen zwischen den beiden Gruppen variieren kann, zeigen sich keine signifikanten Unterschiede (Differenzentest M4a-M3a: $\chi^2(1) = 0.778$, $p = .378$). Die Verbalisierung der beiden mittleren Kategorien scheint also zu keinen deutlichen Unterschieden in den Schwellenwerten zu führen. Damit bestätigt auch dieser Test das Ergebnis, dass sich bei der Frage nach der Lebenszufriedenheit nicht das Messmodell, sondern ausschließlich die Verteilung der latenten

Antworten zwischen den Gruppen unterscheiden. Die unterschiedlichen Antwortkategorien induzieren bei dieser Frage ganz offensichtlich nicht nur auf der Beobachtungsebene, sondern auch auf der latenten Ebene andere Bewertungen der Lebenszufriedenheit.

Tabelle 6. Chi-Quadrat-Differenzentests zur Überprüfung der Invarianz der latenten Verteilungen zwischen den Gruppen in den Modellen M2(a-c) (all- gemeine Zufriedenheit)

	M2(a-c) gleichgesetzte Schwellenwerte		M3(a-c) M2 + gleichgesetzte lat. Verteilung		M3-M2 (a-c)		
	χ^2	df	χ^2	df	χ^2	df	p
a) Mod. (vollverbalisiert) vs. Mod. (endpunktbenannt)	0.369	1	6.631	3	6.262	2	.044
b) Original vs. Mod. (vollverbalisiert)	0.491	1	11.615	3	11.124	2	.003
c) Original vs. Mod. (endpunktbenannt)	0.005	1	22.944	3	22.939	2	< .001

6.2 Lebensziele

Bei der zweiten Frage zu den Lebenszielen zeigt sich ebenfalls ein deutlicher Unterschied in den beobachteten Antwortverteilungen zwischen der Originalversion und den beiden modifizierten Versionen (Tabelle 7). Die zweite Antwortkategorie wird sowohl bei der modifizierten endpunktbenannten als auch bei der modifizierten vollverbalisierten Skala häufiger ausgewählt als in der Originalversion. Dafür werden in dieser Gruppe nicht einmal halb so oft die dritte und die vierte Antwortkategorie gewählt. Die Unterschiede zwischen den beobachteten Antwortverteilungen der ursprünglichen Version und beiden modifizierten Versionen sind signifikant ($p < .001$). H1a wird auch für diese Frage durch unsere Ergebnisse unterstützt. Da sich zudem keine signifikanten Unterschiede zwischen der voll- und der endpunktbenannten Version zeigen ($p = .232$), unterstützt dies ebenfalls unsere Erwartungen gemäß H1b.

Tabelle 7. Verteilungen und kumulierte Verteilungen zu der Frage zu Lebenszielen

	Denken Sie jetzt einmal an Ihre persönliche Situation. Haben sich – einmal alles zusammengenommen – Ihre Vorstellung über das, was Sie im Leben erreichen wollen, bisher …			
Original	mehr als erfüllt	erfüllt	nicht ganz erfüllt	überhaupt nicht erfüllt
% (kum. %)	6.9	37.9 (44.7)	48.6 (93.3)	6.7 (100)
Modifiziert (vollverbalisiert)	ganz und gar erfüllt	eher erfüllt	eher nicht erfüllt	überhaupt nicht erfüllt
% (kum. %)	12.4	62.8 (75.2)	22.1 (97.3)	2.7 (100)
Modifiziert (endpunktbenannt)	ganz und gar erfüllt			überhaupt nicht erfüllt
% (kum. %)	14.8	58.4 (73.2)	23.3 (96.5)	3.5 (100)

Anmerkungen: $\chi^2(6) = 226.0$, $p < .01$; Original n = 861, Modifiziert (vollverbalisiert) n = 892, Modifiziert (endpunktbenannt) n = 866; Original vs. Modifiziert (vollverbalisiert) $\chi^2(3) = 170.7$, $p < .001$; Modifiziert (vollverbalisiert) vs. Modifiziert (endpunktbenannt) $\chi^2(3) = 4.3$, $p = .232$; Original vs. Modifiziert (endpunktbenannt) $\chi^2(3) = 148.6$, $p < .001$

In Abbildung 3 sind die Unterschiede zwischen den Schwellenwerten bei Annahme gleicher Verteilungen auf der latenten Dimension (M1) dargestellt. Hier zeigt sich, dass sich die Schwellenwerte bei der originalen Antwortskala gegenüber den modifizierten Versionen kaum unterscheiden. Entgegen unserer Erwartungen sind jedoch die (über die Schwellenwerte erfassten) Abstände zwischen den beiden mittleren Antwortkategorien bei den modifizierten Skalen deutlich größer als bei der von uns als nicht-äquidistant angenommenen Originalversion. Die Gleichsetzung aller Schwellenwerte über die drei Gruppen führt zu einem unzureichenden Modellfit (M2: $\chi^2(2) = 38.583$, $p < .001$; RMSEA = .145), weswegen wir hier auf die graphische Darstellung verzichten.

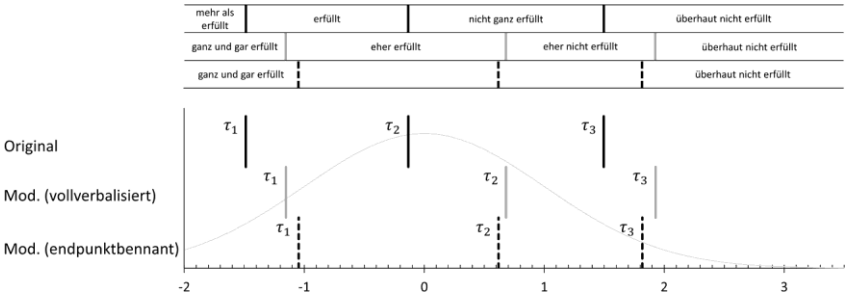

Abbildung 3. Schwellenwerte und latente Antwortverteilung bei der Frage zu Lebenszielen (M1)

Der größte Unterschied zwischen den Antwortskalen sollte sich in Bezug auf den mittleren Schwellenwert ergeben. So kann vermutet werden, dass, wie in Abbildung 3, die ersten beiden Schwellenwerte bei der Originalskala einen geringeren Abstand haben als in den modifizierten Versionen, da die Verbalisierung der modifizierten vollverbalisierten Version deutlich stärker zwischen den Antwortkategorien differenziert als die Originalversion. Tatsächlich führt die Freisetzung des zweiten Schwellenwerts τ_2 zwischen der zweiten und dritten Antwortkategorie in der Originalversion und die Gleichsetzung aller übrigen Schwellenwerte zu einem akzeptablen Modellfit (M4: $\chi^2(1) = .248$, $p = .619$; RMSEA = .001) (Abbildung 4). Dies unterstützt Hypothese H2.

Gleichzeitig zeigt sich aber auch ein signifikanter Mittelwertunterschied bei der latenten Skala und den modifizierten Antwortskalen bei der latenten Skala, was zwischen der originalen Antwortskala und den modifizierten Antwortskalen, was im Widerspruch zu Hypothese H3 steht. Zwischen den beiden modifizierten Antwortskalen unterscheiden sich die Antwortverteilungen auf der latenten Dimension hingegen nicht. Da zwischen diesen beiden Gruppen sowohl das Messmodell als auch die Verteilungen gleichgesetzt werden können, widerlegt dies Hypothese H2 beim Vergleich der vollverbalisierten und der endpunktbenannten Antwortskalen. H3 trifft für alle drei Antwortskalen zu. Die entsprechenden Differenztests zum Vergleich der Modelle mit gleichgesetzten/freien latenten Verteilungen sind in Tabelle 8 dargestellt und bestätigen diese Ergebnisse.

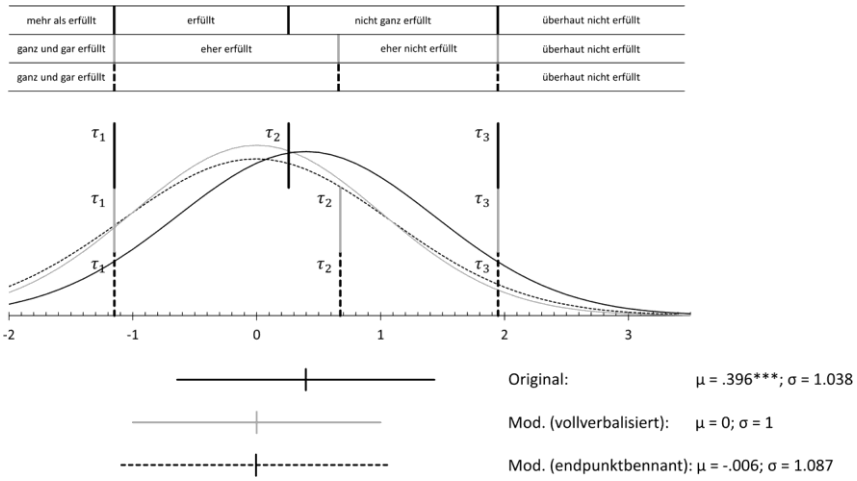

Abbildung 4. Schwellenwerte und latente Antwortverteilungen bei freigesetztem mittleren Schwellenwert zur Frage zu den Lebenszielen (M4)
Anmerkungen: *: p < .05, **: p < 0.01, ***: p < .001

Tabelle 8. Chi-Quadrat-Differenzentests zur Überprüfung der Invarianz der latenten Verteilungen in den Modellen M2(a-c) und M3(a-c) (Lebensziele)

	M2(a-c) gleichgesetzte Schwellenwerte		M3(a-c) M2 + gleichgesetzte lat. Verteilung		M3-M2 (a-c)		
	χ^2	df	χ^2	df	χ^2	df	p
a) Mod. (vollverbalisiert) vs. Mod. (endpunktbenannt)	0.248	1	4.289	3	4.041	2	.133
b) Original vs. Mod. (vollverbalisiert)	31.427	1	169.584	3	138.157	2	< .001
c) Original vs. Mod. (endpunktbenannt)	25.394	1	148.409	3	123.015	2	< .001

6.3 Politisches Interesse

Bei der Frage nach politischem Interesse zeigen sich bei den beobachteten Verteilungen zwischen allen drei Gruppen signifikante Unterschiede (Unterstützung H1a). Die relativ größte Übereinstimmung gibt es hier zwischen der Originalversion und der vollverbalisierten Modifikation, bei denen die jeweils ersten drei Antwortkategorien identisch verbalisiert sind. Im Unterschied zu den beiden zuvor betrachteten Fragen unterscheiden sich (gemessen über die Chi-Quadrat-Statistik) hier auch die vollverbalisierte modifizierte und die endpunktbenannte Version stärker, was Hypothese H1b widerspricht. Dies könnte daran liegen, dass der modifizierte untere Extrempunkt (Politisches Interesse: „sehr schwach“) vom absoluten Nullpunkt „gar kein Interesse“ abweicht (Tabelle 9).

Tabelle 9. Verteilungen und kumulierte Verteilungen zu den Fragen hinsichtlich politischem Interesse

	Wie stark interessieren Sie sich im Allgemeinen für Politik?				
Original	sehr stark	stark	mittelmäßig	weniger stark	überhaupt nicht
% (kum. %)	8.3	26.2 (34.5)	35.7 (70.1)	24.5 (94.7)	5.3 (100)
Modifiziert (vollverbalisiert)	sehr stark	stark	mittelmäßig	schwach	sehr schwach
% (kum. %)	11.1	22.5 (33.6)	39.9 (73.5)	18.2 (91.7)	8.3 (100)
Modifiziert (endpunktbenannt)	sehr stark				sehr schwach
% (kum. %)	13.5	26.1 (39.5)	27.1 (66.7)	20.2 (86.9)	13.1 (100)

Anmerkungen: $\chi^2(8) = 74.0$, $p < .01$; Original n = 864, Modifiziert (vollverbalisiert) n = 894, Modifiziert (endpunktbenannt) n = 870; Original vs. Modifiziert (vollverbalisiert) $\chi^2(4) = 21.8$, $p < .001$; Modifiziert (vollverbalisiert) vs. Modifiziert (endpunktbenannt) $\chi^2(4) = 36.5$, $p < .001$; Original vs. Modifiziert (endpunktbenannt) $\chi^2(4) = 52.5$, $p < .001$

In Modell M1 mit verschiedenen Schwellenwerten zwischen den Gruppen (Abbildung 5) unterscheiden sich, wie aufgrund der unterschiedlichen Verbalisierung vermutet werden kann, vor allem die Schwellenwerte zwischen den beiden geringsten Ausprägungen des politischen Interesses. Bei Gleichsetzung aller Schwellenwerte zwischen den Gruppen (M2) zeigen sich signifikante Unterschiede zwischen den latenten Varianzen, es gibt aber keine signifikanten Unterschiede bei den Mittelwerten der latenten Antwortverteilungen. Allerdings führt diese Gleichsetzung, wie bereits bei der Frage nach dem Erreichen der Lebensziele, zu einem unzureichenden Modellfit (M2 $\chi^2(4) = 24.833$, $p < .001$; RMSEA = .077), weswegen wir auch dieses Modell nicht graphisch darstellen.

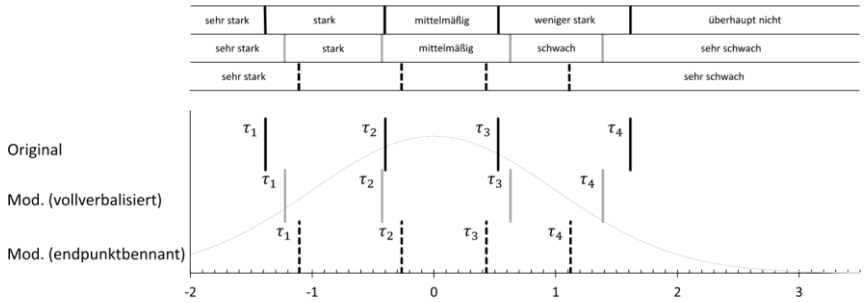

Abbildung 5. Schwellenwerte und latente Antwortverteilung bei der Frage nach dem politischen Interesse (M1)

Im Unterschied zu den beiden zuvor betrachteten Fragen stimmen hier die ersten drei Antwortkategorien der Originalversion und der modifizierten vollverbalisierten Version überein, sodass theoretisch impliziert werden kann, dass sich die Abstände zwischen den verbalen Labels der beiden ersten Antwortkategorien gleichen. Entsprechend sollten sich der erste und der zweite Schwellenwert zwischen den Gruppen nicht unterscheiden. Bei unterschiedlichen Schwellenwerten in allen drei Gruppen zwischen den letzten beiden

Antwortkategorien und zusätzlicher Freigabe des vorletzten Schwellenwerts in der modifizierten vollverbalisierten Vorgabe (Abbildung 6) wird ein hinreichender Modellfit erreicht (M5: $\chi^2(1) = 1.134$, $p = .287$; RMSEA = .012). Der vierte Schwellenwert unterscheidet sich dabei deutlich zwischen allen drei Gruppen und kann auch zwischen den beiden modifizierten Antwortskalen nicht gleichgesetzt werden, ohne dass sich der Modellfit signifikant verschlechtert (Differenzentest: $\chi^2(1) = 7.434$, $p = .006$). Entgegen unserer Erwartungen zeigt sich in diesem Modell, dass der vierte Schwellenwert der modifizierten vollverbalisierten Version deutlich weiter vom dritten Schwellenwert entfernt ist als in der ursprünglichen Version, in welcher die Schwellenwerte gleichmäßiger verteilt sind.

Zwischen der endpunktbenannten und der ursprünglichen Skalenversion unterscheidet sich lediglich der dritte Schwellenwert. Allgemein zeigen sich bei diesen beiden Versionen jedoch gleichmäßig verteilte Schwellenwerte. Äquidistanz scheint hier eher gegeben zu sein als bei der modifizierten vollverbalisierten Antwortskala. Dies spricht gegen unsere Annahme der Verbesserung der Äquidistanz durch unsere Modifikation. Da sich in dem Modell auch die latenten Verteilungen zwischen den drei Gruppen unterscheiden, zeigt sich Unterstützung für H2; dagegen wird H3 widerlegt.

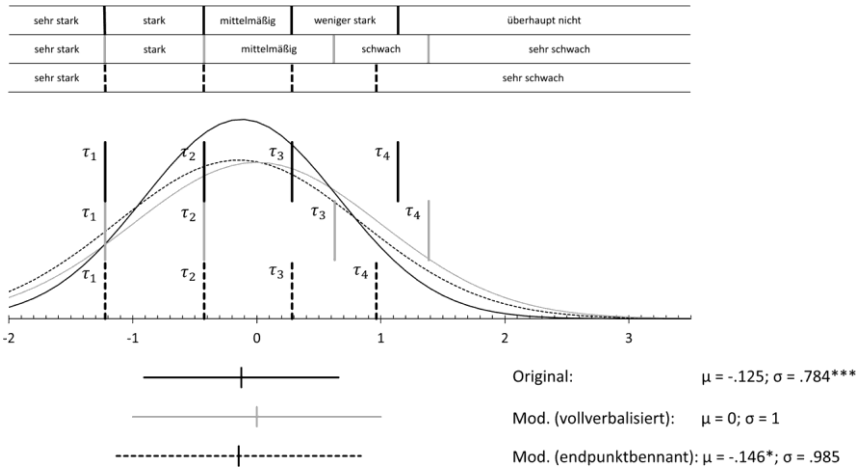

Abbildung 6. Schwellenwerte und latente Antwortverteilungen im angepassten Probitmodell (M5) bei der Frage zum politischen Interesse(M5)
Anmerkungen: *: $p < .05$, **: $p < .01$, ***: $p < .001$ im Vergleich zu modifiziert (vollverbalisiert)

Wird wieder mit Chi-Quadrat-Differenzentests zwischen jeweils zwei Gruppen bei gleichen Schwellenwerten (M2a-c) die Gleichheit der latenten Antwortverteilungen (M3a-c) geprüft (Tabelle 10), zeigt die endpunktbenannte Antwortskala jeweils signifikante Unterschiede zu den beiden vollverbalisierten Skalen (jeweils $p < .001$). Beim Test der beiden vollverbalisierten Versionen zeigen sich dagegen keine signifikanten Unterschiede bei den latenten Mittelwerten und Signifikanzen ($p = .109$). Allerdings passen bereits die Modelle ohne Gleichheitsrestriktion über die latenten Verteilungen nicht zu den Daten (M2b: $p < .001$).

Interessant ist auch hier der Vergleich zwischen den Gruppen mit modifizierten vollverbalisierten und mit endpunktbenannten Antwortskalen, wenn bei gleichgesetzten Antwortverteilungen entweder alle Schwellenwerte (M3a: $\chi^2(4) = 36.682$, $p < .001$) oder

lediglich der erste und der letzte Schwellenwert und die latenten Verteilungen gleichgesetzt werden (M4a: $\chi^2(2) = 14.565$, $p < .001$). Im Falle der beiden vierstufigen Antwortkategorien zu den Fragen nach der Lebenszufriedenheit und der Erreichung der Lebensziele hatte sich gezeigt, dass sich die beiden Skalen nicht signifikant unterscheiden. Bei der gleichen Prüfung für die fünfstufige Skala zum politischen Interesse zeigt der Chi-Quadrat-Differenzentest ein signifikantes Ergebnis (Diff. M3a-M4a: $\chi^2(2) = 22.117$, $p < .001$). Daraus schließen wir, dass die modifizierte vollverbalisierte und die modifizierte endpunktbenannte Skala nicht gleichgesetzt werden können.

Tabelle 10. Chi-Quadrat-Differenzentests zur Überprüfung der Invarianz der latenten Verteilungen zwischen den Modellen M2(a-c) und M3 (politisches Interesse)

	M2(a-c) gleichgesetzte Schwellenwerte		M3(a-c) M2 + gleichgesetzte lat. Verteilung		M3-M2(a-c)		
	χ^2	df	χ^2	df	χ^2	df	p
a) Mod. (vollverbalisiert) vs. Mod. (endpunktbenannt)	15.472	2	36.682	4	21.21	2	< .001
b) Original vs. Mod. (vollverbalisiert)	17.131	2	21.564	4	4.433	2	.109
c) Original vs. Mod. (endpunktbenannt)	5.249	2	51.683	4	46.434	2	< .001

6.4 Gesundheit

In Tabelle 11 sind die beobachteten Antwortverteilungen für die Frage zur Gesundheit wiedergegeben. Im Gegensatz zu den anderen Fragen stimmen im Fall der beiden vollverbalisierten Versionen vier Label überein und die Skala ist in ihrer ursprünglichen Version lediglich durch eine weitere extrem positiv formulierte Antwortkategorie ergänzt, während sie in der modifizierten vollverbalisierten Form durch eine weitere extrem negativ formulierte Antwortkategorie ergänzt ist.

Durch die Modifizierung erwarten wir eine bessere Übereinstimmung zwischen dem inhaltlich-verbalisierten und dem visuellen Mittelpunkt der Skala. In den beiden Gruppen, in denen eine Vollverbalisierung der Antwortskala vorgegeben wurde, gibt es ähnlich hohe Anteile an Befragungspersonen, die ihre Gesundheit (eher) positiv (86.7% und 81.9%) bzw. positiv oder mittelmäßig (98.7% und 96.2%) einschätzen. Die negativste Antwortkategorie wählt jeweils nur ein sehr kleiner Teil der Befragungspersonen, in der endpunktbenannten Version sogar fast niemand. Bei der modifizierten vollverbalisierten Version kann jedoch beobachtet werden, dass ein leicht höherer Anteil der Befragungspersonen eine negative Kategorie auswählt als in der ursprünglichen Version. Unterschiede zwischen den beobachteten Verteilungen sind zwischen allen drei Versionen signifikant, was wiederum H1a unterstützt. Zudem sind die Unterschiede zwischen den beiden vollverbalisierten Versionen größer als zwischen der modifizierten vollverbalisierten und der endpunktbenannten Version, was ebenfalls Hypothese H1b unterstützt.

Können die Schwellenwerte zwischen den Gruppen bei gleicher latenter Verteilung frei variieren (M1), zeigen sich deutliche Unterschiede zwischen den Schwellenwerten (Abbildung 7), sodass eine Gleichsetzung aller Schwellenwerte (M2) keinen Erfolg

verspricht. Aufgrund der gleichen Antwortformulierungen sollten bei beiden vollverbalisierten Versionen Schwellenwerte zwischen Antwortkategorien mit gleichem Label gleichgesetzt werden können, da wir nur eine Verschiebung der Schwellenwerte für die einzelnen Antwortkategorien erwarten. Daher sind die Schwellenwerte τ_2, τ_3 und τ_4 der Originalversion und τ_1, τ_2 und τ_3 der modifizierten Versionen gleichgesetzt. Bei den beiden modifizierten Versionen sind alle Schwellenwerte gleichgesetzt. Dies führt jedoch zu keinem adäquaten Fit (M6: $\chi^2(3) = 27.905$, $p < .001$; RMSEA = .097), weshalb wir auch hier auf eine graphische Abbildung verzichten.

Tabelle 11. Verteilungen und kumulierte Verteilungen zu den Fragen hinsichtlich Gesundheit

	Alles in allem betrachtet, würden Sie sagen, Ihre Gesundheit ist …				
Original	ausgezeichnet	sehr gut	gut	mittelmäßig	schlecht
% (kum. %)	13.0	41.5 (54.6)	32.2 (86.7)	12.0 (98.7)	1.3 (100)
Modifiziert (vollverbalisiert)	sehr gut	gut	mittelmäßig	schlecht	sehr schlecht
% (kum. %)	28.1	53.9 (81.9)	14.3 (96.2)	2.9 (99.1)	0.9 (100)
Modifiziert (endpunktbenannt)	sehr gut				sehr schlecht
% (kum. %)	29.3	44.3 (73.6)	20.7 (94.3)	5.6 (99.9)	0.1 (100)

Anmerkungen: $\chi^2(8) = 205.9$, $p < .01$; Original n = 867, Modifiziert (vollverbalisiert) n = 897, Modifiziert (endpunktbenannt) n = 871; Original vs. Modifiziert (vollverbalisiert) $\chi^2(4) = 173.7$, $p < .001$; Modifiziert (vollverbalisiert) vs. Modifiziert (endpunktbenannt) $\chi^2(4) = 31.7$, $p < .001$; Original vs. Modifiziert (endpunktbenannt) $\chi^2(4) = 105.2$, $p < .001$

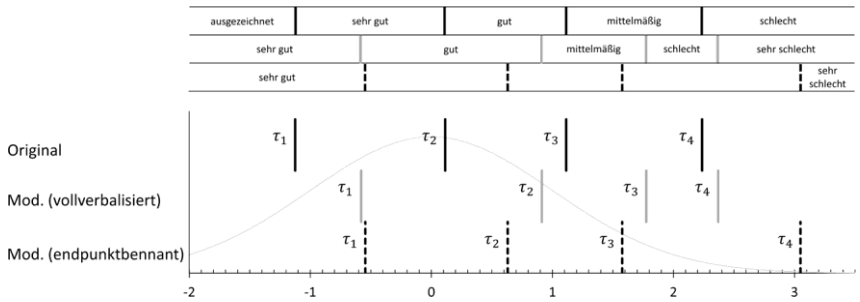

Abbildung 7. Schwellenwerte und Antwortverteilung bei den Fragen zur Gesundheit (M1)

Ein besserer Modellfit ergibt sich erst, wenn nur der zweite und der dritte Schwellenwert der originalen Antwortskala mit dem ersten und dem zweiten Schwellenwert der beiden modifizierten Antwortskalen gleichgesetzt werden und bei den modifizierten Skalen zusätzlich der vierte Schwellenwert freigegeben wird (M7: $\chi^2(1) = 6.706$, $p < .010$; RMSEA = .081). Allerdings ist auch dieser Fit-Wert nach gängigen Faustregeln, die gemeinhin einen RMSEA von < .05 verlangen, noch zu hoch, sodass Zweifel an einem akzeptablen Modell bleiben.

Gemäß unserer Erwartung ist in der Antwortskala der Originalversion eine deutliche Verschiebung des ersten Schwellenwertes nach links zu erkennen (Abbildung 8). Der vierte Schwellenwert (zwischen den Kategorien „mittelmäßig“ und „schlecht“) in der Originalversion ist gegenüber dem dritten Schwellenwert (ebenfalls zwischen den Kategorien „mittelmäßig“ und „schlecht“) der modifizierten vollverbalisierten Version deutlich nach

rechts verschoben. Dies weist darauf hin, dass nicht nur die Formulierung der Antwortkategorien einen Einfluss auf die Interpretation einer Skala haben, sondern auch ihre relative Position. Ebenfalls zeigt sich ein Unterschied im Schwellenwert zwischen den beiden letzten Antwortkategorien bei der modifizierten vollverbalisierten und der endpunktbenannten Antwortskala, die in der vollverbalisierten Version mit „schlecht“ und „sehr schlecht“ benannt sind. Die letzte Kategorie wird in der endpunktbenannten Version, dies zeigten auch schon die beobachteten Verteilungen, deutlich seltener ausgewählt. Korrespondierend ist der Schwellenwert in der endpunktbenannten Version auch deutlich nach rechts verschoben. Dies bestätigt die Vermutung, dass nicht nur verschiedene vollverbalisierte Versionen einer Antwortskala Unterschiede in den Schwellenwerten aufweisen können, sondern auch, dass endpunktbenannte Skalen nicht als grundsätzlich äquidistant wahrgenommen werden. Die unterschiedlichen Schwellenwerte unterstützen Hypothese H2. Da sich in dem Modell zudem signifikante Unterschiede zwischen den latenten Antwortverteilungen zeigen, wird H3 widerlegt. Dies zeigt sich auch, wenn wieder im Vergleich von jeweils zwei Antwortskalen mit gleichgesetzten Schwellenwerten die Gleichheit der latenten Antwortverteilungen getestet wird.

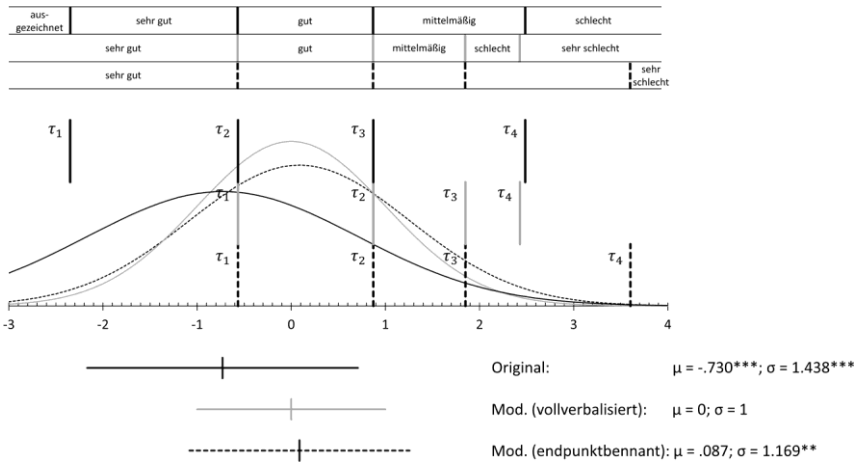

Abbildung 8. Vergleich der latenten Mittelwerte und Streuung bei teilweise gleichgesetzten Schwellenwerten zu den Fragen zu Gesundheit (M7)
Anmerkungen: **: p < .01, ***: p < .001 im Vergleich zu modifiziert (vollverbalisiert)

Tabelle 12. Chi-Quadrat-Differenzentests zur Überprüfung der Invarianz der latenten Verteilungen zwischen den Modellen (M3 und M4) (Gesundheit)

	M2(a-c) gleichgesetzte Schwellenwerte		M3(a-c)4 M3 + gleichgesetzte lat. Verteilung		M3-M2(a-c)		
	χ^2	df	χ^2	df	χ^2	df	p
Mod. (vollverbalisiert) vs. Mod. (endpunktbenannt)	16.737	2	28.895	4	12.158	2	.002
Original vs. Mod. (vollverbalisiert)	22.023	2	177.212	4	155.189	2	< .001
Original vs. Mod. (endpunktbenannt)	1.581	2	106.993	4	105.412	2	< .001

Wird schließlich wie bei den übrigen Fragen getestet, ob bei gleichen latenten

Antwortverteilungen bei den modifizierten voll- und endverbalisierten Antwortkategorien alle Schwellenwerte (M3a: $\chi^2(4) = 28.895$, $p < .001$) oder nur der erste und letzte Schwellenwert (M4a: $\chi^2(2) = 4.554$, $p = .103$) gleichgesetzt werden können, ergibt die Gleichsetzung aller Schwellenwerte wie beim politischen Interesse einen signifikant schlechteren Fit (Diff. M3a-M4a: $\chi^2(2) = 24.341$, $p < .001$). Dies deutet darauf hin, dass nicht nur unterschiedliche Verbalisierungen zu unterschiedlichen Messmodellen führen können, sondern auch endpunktbenannte und vollverbalisierte Skalen nicht ohne weiteres als äquivalent wahrgenommen werden. Dabei legen unsere Ergebnisse nahe, dass das Risiko der Nichtäquivalenz zwischen vollverbalisierten und endpunktbenannten Antwortskalen möglicherweise mit steigender Skalenlänge zunimmt. Dies kann aber erst durch weitere Untersuchungen geklärt werden. Alternativ ist auch denkbar, dass die Unterschiede Folge der unterschiedlichen Bewertungsdimensionen sind, die mit den jeweiligen Fragen erfasst werden sollen.

7 Diskussion

In unseren Analysen haben wir untersucht, ob unterschiedliche Verbalisierungen von Antwortskalen nicht nur zu Unterschieden in den beobachteten Antwortverteilungen führen, sondern auch, wie sich die Beziehungen zwischen der latenten kontinuierlichen Dimension und der beobachteten ordinalen Antwortskala ändern und ob sich auch Verteilungen auf der latenten Dimension unterscheiden. Aufgrund unseres experimentellen Designs können Veränderungen der Verteilung auf der latenten Dimension nur Folge der unterschiedlichen Antwortvorgaben sein, aber nicht Folge unterschiedlicher Positionen der Befragungspersonen. Eine Veränderung der Verteilung auf der latenten Dimension ist sehr problematisch, da es dann beispielsweise nicht möglich ist, verschiedenen Wellen einer Panelbefragung oder Ergebnisse unterschiedlicher Querschnittsbefragungen zu vergleichen, wenn sich zwar nicht die Fragen, aber die Antwortvorgaben unterscheiden.

Für die Analyse wurden verschiedene Fragen aus sozialwissenschaftlichen Befragungen ausgewählt, bei denen wir vermutet haben, dass das Kriterium der Äquidistanz der dazugehörigen Antwortskalen nicht erfüllt ist. Die Antwortskalen wurden dann mit dem Ziel überarbeitet, gleichabständige Skalen zu generieren. Die Probitmodelle zeigen, dass dieses Ziel schwer zu erreichen ist. Dies bedeutet aber auch, dass die Berechnung von Mittelwerten auf der Basis der ordinalen Antwortskala problematisch sein kann. Noch problematischer ist allerdings, dass wir mit unseren Experimentaldaten unter Anwendung ordinaler Probitmodelle zeigen können, dass bei allen vier Items die Verbalisierung (zumindest beim Vergleich der ursprünglichen mit den modifizierten Antwortvorgaben) auch einen Einfluss auf die latenten Antwortverteilungen haben kann (Widerlegung von H3). Interessant ist zudem, dass bei einigen Fragen Unterschiede auch zwischen endpunktbenannten und vollverbalisierten Skalen mit identischen Endpunkten auftreten.

Unsere experimentellen Ergebnisse legen somit eine Interaktion zwischen den verschiedenen Antwortvorgaben und den zu messenden Eigenschaften nahe. Vergleiche zwischen Personen oder Gruppen, denen die gleichen Fragen mit unterschiedlichen Antwortskalen vorgelegt wurden, sind daher offenbar selbst dann nicht möglich, wenn die Schwellenwerte gleichgesetzt werden können und damit formal skalare Messäquivalenz der True Scores gegeben ist. Unklar bleibt allerdings in dieser Analyse, in welchem Maße Unterschiede zwischen Personen tatsächlich substantiell sind. Schon die klassischen Studien

über Frage- und Antwortformulierungen von Sudman et al. (1996) zeigen, dass unterschiedliche Skalen zu unterschiedlichen beobachteten Antwortverhalten führen. Unsere Analysen zeigen, dass dies auch für die latenten True Scores gilt.

Die Auswirkung unterschiedlicher Formen der Verbalisierung scheint zudem sehr fragespezifisch zu sein. Liao (2014) kam zu dem Ergebnis, dass nicht-äquidistante Skalen möglicherweise von Vorteil sind, um im Falle einer sehr schiefen Verteilung der zu messenden Eigenschaft besser zwischen Befragungspersonen differenzieren zu können, weil nicht äquidistante Antwortkategorien zu einer gleichmäßigeren Verteilung über alle Antwortkategorien führen. Auf Basis unserer Ergebnisse ist diese Vermutung schwer zu prüfen. Bei den in unseren Analysen auftretenden Interaktionen von Antwortskala und zu messender Dimension ist nämlich unklar, an welchem Kriterium eine gute Messung festzumachen ist. Es fehlen zudem empirische Analysen zur Übertragbarkeit von Ergebnissen auf verschiedene Umfragemodi sowie verschiedene Gruppen von Befragungspersonen.

Insgesamt halten wir es für notwendig, weitere Untersuchungen zum Zusammenspiel zwischen verbalen und visuellen Attributen von Antwortskalen durchzuführen, da sowohl die visuelle Positionierung als auch die Verbalisierung einen Einfluss auf das Antwortverhalten haben kann. Bis auf einige wenige Arbeiten (siehe Rohrmann 1978) fehlen weiterhin systematische Untersuchungen unterschiedlicher Verbalisierungen. Dabei sollte auch die Ordinalität der Antwortskalen berücksichtigt werden. Wenn ordinale Kategorien als metrisch aufgefasst werden, sollten zumindest die Antwortvorgaben äquidistant sein, sich die Schwellenwerte also gleichmäßig über den Wertebereich verteilen. Unsere Ergebnisse weisen darauf hin, dass diese Bedingung meist nicht gegeben ist.

Als ein Problem sehen wir auch die schwierige Unterscheidung zwischen „äquidistant“ und „balanciert“. Während in dieser Arbeit der Begriff der Äquidistanz gewählt wurde, um Divergenzen in der Verbalisierung der Labels zu beschreiben, gibt es andere Studien, die in dieser Hinsicht von Balancierung sprechen (siehe Friedman et al. 1981; Liao 2014). Allerdings können die Begriffe der Äquidistanz und der Balancierung unserer Einschätzung nach nicht immer trennscharf voneinander abgegrenzt werden. Die Verwendung unterschiedlich vieler positiver/negativer Antwortkategorien kann entweder zur kognitiven Verschiebung des inhaltlichen Mittelpunkts bei Befragungspersonen führen – dies wäre ein Problem der Balancierung – andererseits können in diesem Fall auch Abstände zwischen Antwortkategorien unterschiedlich groß wahrgenommen werden; dies wäre dann ein Problem der Äquidistanz.

Unsere Analysen zeigen die Notwendigkeit weiterer Studien zur Wahrnehmung von Antwortkategorien, bei denen visuelle und verbale Charakteristika von Antwortskalen berücksichtigt werden. Hier erwarten wir auch gerade von Eyetracking-Studien (siehe Höhne und Lenzner 2018; Komoen et al. 2011) und der Analyse von Paradaten in Onlinebefragungen (siehe Höhne, Schlosser und Krebs 2017) neue Erkenntnisse. Ebenfalls interessant ist die Frage, ob sich im Verlauf von Panelbefragungen das Verständnis von Antwortkategorien verändert, wenn Befragungspersonen sich stärker an die Verwendung bestimmter Formen von Antwortskalen gewöhnt haben.

8 Literatur

Coromina, L., & Coenders, G. (2006). Reliability and validity of egocentered network data collected via web. *Social Networks* 28(3), 209–231.

DeCastellarnau, A. (2017). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*. doi: 10.1007/s11135-017-0533-4

Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviorally anchored rating and mixed standard scale formats. *Journal of Applied Psychology*, 65(2), 147-154.

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

Friedman, H., Wilamowsky, Y., & Friedman, L. (1981). A comparison of balanced and unbalanced rating scales. *The Mid-Atlantic Journal of Business* 19, 1–7.

Geiser, C., Eid, M. (2010). Item-Response-Theorie. In C. Wolf, & H. Best (Hrsg.): *Handbuch der sozialwissenschaftlichen Datenanalyse*, (S.311-332). Wiesbaden: VS Verlag für Sozialwissenschaften.

Höhne, J. K., & Krebs, D. (2018). Scale direction effects in agree/disagree and item-specific questions: A comparison of question formats. *International Journal of Social Research Methodology*, 21(1), 91–103.

Höhne, J. K., & Lenzner, T. (2018). New insights on the cognitive processing of agree/disagree and item-specific questions. *Journal of Survey Statistics and Methodology*, 6, 401-417.

Höhne, J. K., Schlosser, S., & Krebs, D. (2017). Investigating cognitive effort and response quality of question formats in web surveys using paradata. *Field Methods*, 29(4), 365–382.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles. *Journal of Cross-Cultural Psychology* 36, 264–277.

Jöreskog, K. G. (1994). Structural equation modeling with ordinal variables. *Lecture Notes-Monograph Series* 24, 297-310.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Processes* 48, 355–385.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement* 25, 85–96.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science* 37, 941-946.

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 (Hrsg.), *Survey measurement and process quality* (S. 141–164). Hoboken: John Wiley & Sons.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Hrsg.), *Handbook of survey research* (S. 263–313). San Diego: Elsevier.

Kühnel, S. M. (1993). Lassen sich ordinale Daten mit linearen Strukturgleichungsmodellen analysieren? *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, 33, 29-51.

Lantz, B. (2013). Equidistance of Likert-type scales and validation of inferential methods using experiments and simulations. *The Electronic Journal of Business Research Methods* 11, 16–28.

Liao, P.-S. (2014). More happy or less unhappy? Comparison of the balanced and unbalanced designs for the response scale of general happiness. *Journal of Happiness Studies* 15, 1407–1423.

Menold, N. (2017). Rating-scale labeling in online surveys: An experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. *Sociological Methods & Research*. doi: 10.1177/0049124117729694

Menold, N., & Bogner, K. (2015). Gestaltung von Ratingskalen in Fragebögen (SDM - Survey Guidelines). Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften. doi: 10.15465/sdm-sg_015

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

Menold, N., & Kemper, C. (2015). The impact of frequency rating scale formats on the measurement of latent variables in web surveys - an experimental investigation using a measure of affectivity as an example. *Psihologija* 48, 431–449.

Menold, N., & Tausch, A. (2016). Measurement of latent variables with different rating scales. *Sociological Methods & Research* 45, 678–699.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research* 39, 479–515.

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology* 44, 369–399.

Moosbrugger, H. (2012). Klassische Testtheorie (KTT). In: H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 103–117). Berlin, Heidelberg: Springer.

Mullahy, J. (1990). Weighted least squares estimation of the linear probability model, revisited. *Economics Letters* 32, 35–41.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 49, 115–132.

Muthén, L. K, & Muthén, B. O. (1998-2012). Mplus User's Guide. Siebte Auflage. Los Angeles, Ca: Muthén & Muthén.

Ostrom, T. M., & Gannon, K. M. (1996). Exemplar generation: Assessing how respondents give meaning to rating scales. In N. Schwarz & S. Sudman (Hrsg.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (S. 293–318). San Francisco: Jossey-Bass.

Parducci, A. (1983). Category ratings and the relational character of judgment. In H.-G. Geissler, H.F.J.M. Buffort, E.J. Leeuwenberg, & V. Sarris (Hrsg.), *Advances in Psychology. Modern Issues in Perception* (S. 262–282). Amsterdam: North-Holland Publishing Company.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie* 9, 222–245.

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

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

Publishers.

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

van de Vijver, F. J. R. (2003). Bias and equivalence: cross-cultural perspectives. In J. A. Harkness, F. J. R. van de Vijver, & P. Ph. Mohler (Hrsg.) *Cross-cultural survey methods*, S.143-156. Hoboken: Wiley.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement* 64, 956–972.

Zaller, J. R. (Ed.) 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.