Measurement Properties of Completely and End Labeled Unipolar and Bipolar Scales in Likert-type Questions on Income (In)equality

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

Dagmar Krebs
University of Gießen (Germany)

Steffen-M. Kühnel
University of Göttingen (Germany)

Abstract

The measurement of respondents' attitudes is key in social science research and many adjacent research fields. A common method to measure this information is to use Likert-type questions that consist of a statement that is evaluated with a rating scale. As shown by previous research, the scale design of Likert-type questions can have a profound impact on respondents' answer behavior. In this study, we therefore investigate the measurement properties of scales that systematically vary with respect to polarity (i.e., unipolar and bipolar) and labeling (i.e., completely and end). We conducted a survey experiment in a probability-based online panel (N = 4,851) and used questions on income (in)equality that were adopted from the European Social Survey (ESS). The results reveal considerable differences between the scales under investigation. They show that end labeled unipolar and bipolar scales accomplish the criteria of equidistance best. Completely labeled bipolar scales, in contrast, only show a poor performance in terms of equidistance. Completely labeled unipolar scales are somewhere in between. Overall, our findings suggest that researchers should be careful when using survey data measured with (slightly) different scales because the results might not be comparable.

Keywords: Latent thresholds, Likert-type questions; Measurement invariance, Online survey, Rating scale design

1. Introduction and Background

Measuring respondents' attitudes using survey questions following the notion of Likert (1932) is a very common method in social science research and adjacent research fields.[1] Likert-type questions usually consist of a request for an answer, which is followed by a statement and a rating scale for providing an answer (Saris & Gallhofer, 2014). Even though numerous studies

This document is a preprint and thus it may differ from the final version: Höhne, Jan K., Krebs, Dagmar, & Kühnel, Steffen-M. (2021). Measurement properties of completely and end labeled unipolar and bipolar scales in Likert-type questions on income (in)equality. Social Science Research, 97. DOI: 10.1016/j.ssresearch.2021.102544.

[1] It must be mentioned that Likert (1932) tested survey questions with five-point, completely labeled bipolar "approval/disapproval" rating scales that were horizontally aligned and that contained numeric values running from 1 to 5.

suggest that Likert-type questions are associated with serious methodological drawbacks, such as being cognitively demanding and prone to response bias (Carpenter & Just, 1975; Converse & Presser, 1986; Fowler, 1995; Fowler & Cosenza, 2008; Höhne, 2019; Höhne & Krebs, 2018; Höhne & Lenzner, 2018; Kuru & Pasek, 2016; Lelkes & Weiss, 2015; Liu, Lee, & Conrad, 2015; Saris, Revilla, Krosnick, & Shaeffer, 2010), they have some tempting practical advantageous (Revilla, Saris, & Krosnick, 2014; Saris et al., 2010): First and foremost, they allow researchers to ask about a variety of unrelated topics (e.g., political efficacy and job motivation) without changing the scale. Second, they streamline questionnaires and save space and time in self-administered surveys; this particularly applies when using matrix questions. Finally, since Likert-type questions represent a long-standing method this may encourage researchers to reuse established batteries of such questions instead of developing new ones that, for instance, employ item-specific questions (i.e., survey questions that express the content dimension of the question stem in the scale directly; Höhne & Lenzner, 2018).

When designing scales for Likert-type questions researchers must take several design aspects into account because these aspects can profoundly influence respondents' answer behavior (see DeCastellarnau, 2018; Schaeffer & Dykema, 2020). For instance, researchers must decide about an even or uneven number of scale points (i.e., including a middle option or not), the length of the scale (i.e., the number of scale points), the inclusion of numeric values (i.e., providing the answer options with or without numbers), the direction of the scale (i.e., decremental or incremental), the alignment of the scale (i.e., horizontal or vertical), the polarity of the scale (i.e. unipolar or bipolar), and the extent of scale labeling (e.g., completely or end).

In general, Likert-type questions can consist of unipolar or bipolar scales. In the case of unipolar scales, the answer options can proceed from the uppermost scale point (e.g., "agree strongly") to the lowermost scale point (e.g., "agree not at all"). In the case of bipolar scales, the answer options can proceed from the uppermost positive scale point (e.g., "agree strongly") through a "transition point" (Schaeffer & Presser, 2003; Schaeffer & Dykema, 2020) that is located in the middle of the scale to the opposite lowermost negative scale point (e.g., "disagree strongly"). The current state of research on the use of unipolar and bipolar scales for measuring respondents' attitudes lacks empirical evidence and there is no scientific consensus (see Alwin, 2007, 2010; DeCastellarnau, 2018; Höhne, Krebs, & Kühnel, 2020; Krosnick & Fabrigar, 1997; Menold, 2019; Menold & Raykov, 2015; Schaeffer & Presser, 2003; Thomas & Barlas, 2018).

Similar to scale polarity, scale labeling is also a somewhat controversial issue when designing scales (see DeCastellarnau, 2018; Krosnick & Presser, 2010; Menold & Bogner, 2015). The main reason is that verbal labels for all options (i.e., completely labeled) or only for the end options (i.e., end labeled) convey crucial information that respondents, as "cooperative communicators" (Schwarz, 1996), use to understand and answer survey questions meaningfully (Höhne, Lenzner, Neuert, & Yan, 2019; Höhne & Yan, 2019; Parducci, 1983; Sudman, Bradburn, & Schwarz, 1996; Toepoel & Dillman, 2011a, 2011b; Tourangeau, Couper, & Conrad, 2004, 2007). This particularly applies to scale polarity because verbal labels – besides numerical values – disclose the polarity of scales (O'Muircheartaigh, Gaskell, & Wright, 1995; Schaeffer & Dykema, 2020). Table 1 shows an example of a Likert-type question on social equality with completely and end labeled unipolar and bipolar scales.

Table 1. Example of a Likert-type question on social equality with completely and end labeled unipolar and bipolar scales

Question parts	Unipolar scales	Bipolar scales
Requests for an answer	To what extent do you agree with the following statement?	To what extent do you agree or disagree with the following statement?
Statements	Social benefits lead to more equality in society.	Social benefits lead to more equality in society.
Completely labeled scales	○ Agree strongly ○ Agree somewhat ○ Agree moderately ○ Agree hardly ○ Agree not at all	○ Agree strongly ○ Agree somewhat ○ Neither agree nor disagree ○ Disagree somewhat ○ Disagree strongly
End labeled scales	○ Agree strongly ○ ○ ○ ○ Agree not at all	○ Agree strongly ○ ○ ○ ○ Disagree strongly

Note. The survey question was adopted from the European Social Survey (2016).

Considering the unipolar and bipolar scales in Table 1 it is to see that the verbal differences between both scales are more pronounced for the completely labeled versions in which the middle options (i.e., “agree moderately” vs. “neither agree nor disagree”) and the second lowest options (i.e., “agree hardly” vs. “disagree somewhat”) additionally differ from each other. While the middle option in the unipolar scale indicates a moderate position towards the object under investigation, the middle option in the bipolar scale can have two different meanings. More specifically, “neither/nor” formulations indicate a neutral position or no position at all towards the object under investigation depending on the interpretation of respondents (Höhne et al., 2020; Krosnick & Fabrigar, 1997; Menold, 2019; Schaeffer & Dykema, 2020; Sturgis, Roberts, & Smith, 2014; Wang & Krosnick, 2020). This ambiguity might push respondents to one side of the two opposing parts of bipolar scales (Höhne et al., 2020). The formulation disparities regarding the middle options, coupled with the formulation disparities regarding the second lowest and the lowest options, have the potential to alter the evaluative scale character as well as the position of each verbal label (or answer option) on the respective scale continuum (Höhne et al., 2020; Mohler, Smith, & Harkness, 1998; Rohrmann, 1978). In contrast, in end labeled scales the differences between unipolar and bipolar scales are limited to the lowermost options (i.e., “agree not at all” vs. “disagree strongly”), which substantially reduces disparities between unipolar and bipolar scales.

It is also to mention that the middle options of completely labeled unipolar and bipolar scales frequently do not match with the actual polarity of the scale (Menold, 2019). For instance, the German version of the International Social Survey Program (ISSP; 2012) uses Likert-type questions with unipolar agreement scales that contain a bipolar “neither/nor” middle option (see also Scholz & Jutz, 2014). Furthermore, the English source questionnaire of the ISSP (2012) uses a bipolar agreement/disagreement scale, but the German questionnaire uses an inconsistently developed unipolar agreement scale. This points to the fact that, in practice, unipolar and bipolar scales are inconsistently and interchangeably used. This may have serious consequences for attitude measurement in general and cross-cultural and cross-national surveys

in particular. As shown by Höhne et al. (2020), answers to completely labeled unipolar and bipolar scales differ regarding measurement properties, such as answer distributions, measurement invariance, and equidistance or equality of intervals between scale points (see Stevens, 1946).

Even though completely labeled scales require a higher cognitive effort than their end labeled counterparts (respondents need to read and process more information), they usually better clarify the meaning of different scale points and decrease ambiguity (Alwin, 2007; DeCastellarnau, 2018; Krosnick & Presser, 2010). Some research also suggests that respondents express higher satisfaction with completely than end labeled scales (see Krosnick & Presser, 2010; Menold & Bogner, 2015). Moreover, previous research indicates that completely labeled scales, compared to end labeled scales, result in better data quality, such as reliability and validity (Alwin, 2007; DeCastellarnau, 2018; Menold, Kaczmirek, Lenzner, & Neusar, 2014; Krosnick & Fabrigar, 1997; Revilla et al., 2014). However, completely and end labeled unipolar and bipolar scales can vary with respect to their literal meaning and the cognitive and communicative processes they initiate (Schwarz, 1996; Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). While the meaning of the middle part of the completely labeled unipolar and bipolar scales highly depends on the verbal labels and the adverbial intensifiers (Rohrmann, 1978; Mohler et al., 1998), the meaning of the middle part of the end labeled unipolar and bipolar scales conveys the impression of equal intervals.

In line with this reasoning, it is to assume that completely and end labeled unipolar and bipolar scales may affect question understanding differently. This may impede the comparability of answers to identical Likert-type questions that employ (slightly) different scales. Comparability is evaluated in terms of answer distributions, measurement invariance, and equidistance between scale points (or answer options).

In this article, we build on the study by Höhne et al. (2020) and extend it by additionally comparing end labeled unipolar and bipolar scales. Thus, we investigate the impact of completely and end labeled unipolar and bipolar scales on answer behavior by analyzing the observed and latent answer distributions and comparing latent thresholds of answer options. By investigating completely and end labeled unipolar and bipolar scales at this level of analysis, our study stands out of previous studies contributing to the eminent survey literature on scale design.

2. Research Hypotheses

Höhne et al. (2020) compared Likert-type questions with completely labeled unipolar and bipolar scales (using the verbal labels in Table 1). In line with previous research, the authors found significantly different answer distributions in the form of more positive agree answers (particularly, “agree somewhat”) in bipolar scales and more middle answers (i.e., “agree moderately”) in unipolar scales. One reason for the significant differences between unipolar and bipolar scales may lie in the complete labeling of the scales. More specifically, the disparities associated with the middle and the second lowest and lowest options may have changed the evaluative character of the scales and the meaning of the verbal labels (or answer options). In end labeled unipolar and bipolar scales, in contrast, the verbal differences are limited to the lowermost options (i.e., “agree not at all” vs. “disagree strongly”) decreasing the disparities between both scales. In addition, the unlabeled answer options between the scale

endpoints convey the impression of equally distanced intervals so that the two scales appear virtually equivalent. We therefore expect no significantly different answer distributions between end labeled unipolar and bipolar scales (Hypothesis 1a).

We also investigate respondents' answer behavior with respect to completely and end labeled unipolar and bipolar scales. Compared to end labeled scales, completely labeled scales provide more information on how to interpret and understand the scale (Alwin, 2007; DeCastellarnau, 2018; Krosnick & Presser, 2010). However, following the notion of "cooperative communicators" (Schwarz, 1996), this additional information may affect respondents cognitive and communicative processes, which, in turn, can have an impact on respondents' answer behavior. This similarly applies to unipolar and bipolar scales. Thus, we expect that the answer distributions of completely and end labeled unipolar scales significantly differ from each other (Hypothesis 1b). Furthermore, we expect significant differences between the answer distributions of completely and end labeled bipolar scales (Hypothesis 1c).

In addition to our hypotheses on the observational level, we also compare completely and end labeled unipolar and bipolar scales on the latent level. More specifically, we test for measurement invariance using multigroup confirmatory factor analysis (MG-CFA). Consistent with our previous hypotheses on the observational level, we expect to obtain measurement invariance between end labeled unipolar and bipolar scales (Hypothesis 2a). Since scale labeling has the great potential to influence respondents' answer behavior, we expect to obtain measurement non-invariance between completely and end labeled unipolar scales (Hypothesis 2b). In addition, we expect completely and end labeled bipolar scales not being invariant (Hypothesis 2c).

In a next step, we follow the strategy described in Höhne et al. (2020) and switch from the multiple-indicator factor level to the survey question level focusing on answer options and their latent thresholds. Following an Item Response Theory (IRT) approach, we account for each answer option of the completely and end labeled unipolar and bipolar scales using threshold parameters. A threshold parameter marks the point on a latent continuum where an answer to an answer option x is more likely than an answer to an answer option x – 1 (Wetzel & Carstensen, 2014, p. 766). In general, it is assumed that answer options and their latent thresholds are equidistantly and normally distributed (Rost, 1988). However, as suggested by Höhne et al. (2020) and Rohrmann (1978), verbal labels may have an impact on the equidistance between answer options (or scale points), which can apply to the observational as well as the latent level. As a consequence, scale polarity and scale labeling have the potential to influence latent thresholds. In line with our previous hypotheses and the findings reported by Höhne et al. (2020), we expect the latent thresholds of end labeled unipolar and bipolar scales to be similarly equidistantly distributed (Hypothesis 3a). We also expect that the latent thresholds of completely labeled unipolar scales are less equidistantly distributed than the latent thresholds of their end labeled counterparts (Hypothesis 3b). Finally, we expect that the latent thresholds of completely labeled bipolar scales are less equidistantly distributed than those of end labeled bipolar scales (Hypothesis 3c).

The hypotheses on completely labeled unipolar and bipolar scales are discussed and reported in Höhne et al. (2020).

3. Method

3.1 Study Design

To investigate the effects of scale polarity (i.e., unipolar and bipolar) and scale labeling (i.e., completely and end) on respondents' answer behavior we conducted a survey experiment and randomly assigned respondents to one out of four experimental groups. Table 2 describes the four experimental groups.

Table 2. Experimental design defined by scale polarity and scale labeling

Experimental group	Scale polarity	Scale labeling	Group size
1	Unipolar	Completely	1,214
2	Unipolar	End	1,214
3	Bipolar	Completely	1,207
4	Bipolar	End	1,216

To evaluate the effectiveness of random assignment to the four experimental groups, we conducted chi-square tests. The results showed no statistically significant differences regarding age, gender, and education.

3.2 Questions

In total, this study consisted of five Likert-type questions, which were adopted from the German versions of the European Social Survey (2002, 2016) dealing with different aspects of income (in)equality. We excluded one survey question from the analyses because it did not load on the same latent factor as the other four. For each survey question adopted from the European Social Survey, we developed completely and end labeled unipolar and bipolar scales. To limit question order effects, we randomized the order of the survey questions. Each survey question was presented on a separate page using five-point, vertically aligned scales with radio buttons for selecting an answer option (see Appendix A for English translations and Appendix B for an illustration of the question and scale design).

3.3 Study Procedure

Data were collected in the German Internet Panel, which is part of the Collaborative Research Center 884 "Political Economy of Reforms" at the University of Mannheim. The German Internet Panel is based on an initial recruitment in 2012 and two refresher recruitments in 2014 and 2018. While the recruitments in 2012 and 2014 are based on a three-stage stratified probability sample, the recruitment in 2018 is based on a two-stage stratified probability sample of the German population aged 16 to 75 years. For a detailed methodological description of the German Internet Panel, we refer interested readers to Blom, Gathmann, and Krieger (2015).

The German Internet Panel invites all panelists every two months to participate in a self-administered online survey that deals with a variety of economic, political, and social topics. Each online survey lasts about 20 minutes. For their participation in each wave, respondents receive a compensation of 4 Euros.

At the beginning of each wave, panelists are directed to a short welcome page announcing the approximate length of the online survey and informing them that the compensation for their participation will be credited to their study account after survey completion. The survey

questions used in this study were included at the beginning of the online survey limiting carry-over effects from other panel modules.

3.4 Sample Characteristics

In this article, we use data that were collected in wave 40 of the German Internet Panel. This wave ran from March 1 to March 31, 2019, with a total of 4,890 respondents (response rate: 64.1%). Out of those, 4,851 respondents took part in the present study (participation rate: 99.2%). The mean age of these respondents is 49.7 (SD = 15.8), and 48.4% of them are female. In terms of education, 13.3% graduated from a lower secondary school, 30.6% from an intermediate secondary school, and 52.1% from a college preparatory secondary school or university. Furthermore, 1.4% were still attending school or left school without a diploma and 2.6% reported having a different degree from those mentioned above.

3.5 Analyses

In a first step, we investigate the answer distributions of all four Likert-type questions used in this study and conduct chi-square tests to test our hypotheses on the observational level (1a, 1b, and 1c). We additionally conduct several directed Z-tests to investigate potential differences between specific answer options.

In order to test our hypotheses on measurement invariance between completely and end labeled unipolar and bipolar scales (2a, 2b, and 2c) we start with confirmatory factor analyses (CFAs). In a next step, we conduct multigroup confirmatory factor analyses (MG-CFAs) for completely and end labeled unipolar and bipolar scales to test for configural invariance (equality of dimensional structure), metric invariance (additional equality of factor loadings), and scalar invariance (additional equality of intercepts).

As criteria for accepting measurement invariance, we use non-significant differences between chi-square values (Bryant & Satorra, 2012; Byrne, 2012) between the hierarchically ordered (configural, metric, and scalar) models. In addition, we look at the differences between CFIs (Comparative Fit Index) and RMSEAs (Root Mean Square Error of Approximation) between the three hierarchically ordered models. In line with Cheung and Rensvold (2002), these differences should be lower than 0.01. If these criteria cannot be obtained, we assume measurement non-invariance. Due to the fact that all indicators (Likert-type questions) were measured with five-point scales, we assume a continuous scale level (Rhemtulla, Brosseau-Liard, & Savalei, 2012) and use the MLR (Robust Maximum Likelihood) discrepancy function.

For testing our hypotheses on the equidistance of latent thresholds of answer options (3a, 3b, and 3c), we follow the analysis strategy by Höhne et al. (2020) and consider answer options (and their latent thresholds) as approximate ordinal measures of a continuous latent variable. First, we compute unrestricted univariate probit models for each Likert-type question using the WLS (Weighted Least Squares) discrepancy function. To get information on the sequential order of latent thresholds of answer options we regress these model-estimated values on ascending integers 1 to 4 and inspect explained variances (R^2 values). This is done for completely and end labeled unipolar and bipolar scales.

Second, for each Likert-type question, we again compute univariate models and constrain the latent thresholds to equal distances. For these models, we inspect the model fits (RMSEAs) for completely and end labeled unipolar and bipolar scales. Note that higher R^2 values indicate

a better sequential representation of latent thresholds and lower RMSEA values indicate more equidistantly distributed latent thresholds of answer options.

The chi-square and Z-tests and the linear regressions are conducted using SPSS version 24. The analyses on measurement invariance and the latent thresholds are conducted using Mplus version 6.12. Appendix C contains Mplus commands to track the analyses of measurement invariance and latent thresholds.

4. Results

4.1 Answer Distributions

In order to test our hypotheses on the answer distributions of completely and end labeled unipolar and bipolar scales (1a, 1b, and 1c), we conducted chi-square tests. Table 3 reports all answer distributions and test results. In line with Hypothesis 1a, the answer distributions of the Likert-type questions with end labeled unipolar and bipolar scales do not differ significantly at the 5%-level, except for question 3. Nonetheless, we count these findings as supporting evidence for our hypothesis.

Considering Table 3, it is also to see that there is supporting evidence for Hypothesis 1b on completely and end labeled unipolar scales. Corresponding with our expectation, completely and end labeled unipolar scales produce significantly different answer distributions. This applies to all four questions. Interestingly, compared to completely labeled unipolar scales, answers in end labeled unipolar scales pile up at the endpoints (i.e., “agree strongly” and “agree not at all”). To investigate these differences, we aggregated the percentages of the first and last answer options of the completely and of the end labeled unipolar scales and conducted one-sided directed Z-tests. We tested the following survey questions for significant differences (end labeled > completely labeled): Question 1 ($Z = 3.24$, $p < 0.001$), Question 2 ($Z = 3.77$, $p < 0.001$), Question 3 ($Z = 5.37$, $p < 0.001$), and Question 4 ($Z = 5.69$, $p < 0.001$).

With respect to Hypothesis 1c on completely and end labeled bipolar scales the results of the chi-square tests reveal significantly different answer distributions. This applies to all four questions, supporting our hypothesis. Overall, Table 3 shows that completely labeled bipolar scales lead to more positive agree answers than their end labeled counterparts. We therefore aggregated the percentages of the first two agree answer options of the completely and of the end labeled bipolar scales and conducted one-sided directed Z-tests (completely labeled > end labeled): Question 1 ($Z = 3.96$, $p < 0.001$), Question 2 ($Z = 4.43$, $p < 0.001$), Question 3 ($Z = 3.94$ $p < 0.001$), and Question 4 ($Z = 4.96$, $p < 0.001$). Even though the results show that respondents tend to prefer positive agree options of completely labeled bipolar scales, this only applies to the second positive answer option (i.e., “agree somewhat”). For end labeled bipolar scales, in contrast, the results show that respondents tend to prefer the middle answer option. The results of one-sided directed Z-tests provide supporting evidence (percentages of middle answers: end labeled > completely labeled): Question 1 ($Z = 5.68$, $p < 0.001$), Question 2 ($Z = 3.22$, $p < 0.001$), Question 3 ($Z = 3.14$ $p < 0.001$), and Question 4 ($Z = 4.01$, $p < 0.001$).

Table 3. Answer distributions in percentages

		End labeled scales		Unipolar scales		Bipolar scales	
		Unipolar	Bipolar	Completely labeled	End labeled	Completely labeled	End labeled
Question 1		$\chi^2(4) = 0.64$, $p = 0.958$		$\chi^2(4) = 22.59$, $p < 0.001$		$\chi^2(4) = 47.31$, $p < 0.001$	
Scale points	1	14	14	11	14	12	14
	2	36	37	36	36	47	37
	3	29	30	36	29	20	30
	4	15	15	14	15	18	15
	5	6	5	4	6	4	5
Question 2		$\chi^2(4) = 8.98$, $p = 0.062$		$\chi^2(4) = 18.46$, $p < 0.001$		$\chi^2(4) = 54.84$, $p < 0.001$	
Scale points	1	16	17	13	16	12	17
	2	27	29	33	27	43	29
	3	32	33	34	32	27	33
	4	19	18	18	19	16	18
	5	6	3	3	6	2	3
Question 3		$\chi^2(4) = 16.99$, $p < 0.002$		$\chi^2(4) = 46.27$, $p < 0.001$		$\chi^2(4) = 44.58$, $p < 0.001$	
Scale points	1	13	14	9	13	10	14
	2	29	32	37	29	44	32
	3	35	36	33	35	30	36
	4	17	15	18	17	15	15
	5	7	3	3	7	2	3
Question 4		$\chi^2(4) = 5.01$, $p = 0.286$		$\chi^2(4) = 42.26$, $p < 0.001$		$\chi^2(4) = 67.87$, $p < 0.001$	
Scale points	1	23	23	17	23	18	23
	2	26	28	32	26	43	28
	3	29	28	31	29	21	28
	4	15	16	17	15	15	16
	5	7	5	3	7	3	5

Note. Due to rounding, the percentages may not add up to 100%. Verbal labels of completely and end (points 1 and 5 only) labeled unipolar scales: 1 “agree strongly”, 2 “agree somewhat”, 3 “agree moderately”, 4 “agree hardly”, and 5 “agree not at all”. Verbal labels of completely and end (points 1 and 5 only) labeled bipolar scales: 1 “agree strongly”, 2 “agree somewhat”, 3 “neither agree nor disagree”, 4 “disagree somewhat”, and 5 “disagree strongly”. See Appendix A for the statements of the Likert-type questions. The test results for completely labeled unipolar and bipolar scales are reported in Höhne et al. (2020).

4.2 Measurement Invariance

We initially computed the same confirmatory factor analysis (CFA) baseline models for completely and end labeled unipolar and bipolar scales. All baseline models included one latent variable with four indicators (Likert-type questions). We then conducted multigroup confirmatory factor analyses (MG-CFAs) and tested for configural, metric, and scalar invariance between the following scale conditions: end labeled unipolar and bipolar scales (Hypothesis 2a), completely and end labeled unipolar scales (Hypothesis 2b), and completely and end labeled bipolar scales (Hypothesis 2c). Table 4 reports the results on measurement invariance.

In line with Hypothesis 2a, measurement invariance can be accepted for end labeled unipolar and bipolar scales. This is indicated by the non-significant result of the chi-square difference test and implies that both scales are comparable. However, in contrast to Hypothesis 2b, we find that measurement invariance is also established for completely and end labeled unipolar scales. Again, this finding is supported by the non-significant result of the chi-square difference test. Finally, we find supporting evidence for Hypothesis 2c on measurement non-invariance between completely and end labeled bipolar scales. More specifically, we obtain metric invariance but not scalar invariance, which is suggested by the significant result of the chi-square difference test. Consequently, completely and end labeled bipolar scales are not comparable.

Table 4. Testing for measurement invariance

Invariance levels	χ^2 values	Df	RMSEA	CFI	χ^2 difference test
End labeled unipolar and bipolar scales					
Configural	2.36 (1.43)	2	0.012	1	
Metric	4.79 (1.30)	5	0	1	2.35
Scalar	11.52 (1.19)	8	0.019	0.997	7.43
Completely and end labeled unipolar scales					
Configural	0.97 (1.38)	2	0	1	
Metric	1.90 (1.23)	5	0	1	0.88
Scalar	4.61 (1.14)	8	0	1	2.95
Completely and end labeled bipolar scales					
Configural	1.92 (1.37)	2	0	1	
Metric	5.22 (1.26)	5	0.006	1	3.33
Scalar	16.90 (1.16)	8	0.030	0.992	13.11**

Note. **p < 0.01. The results are based on the MLR discrepancy function. Scale correction factors are in parentheses. The test results for completely labeled unipolar and bipolar scales are reported in Höhne et al. (2020).

4.3 Latent Thresholds

Next, we investigate the equidistance between the latent thresholds of answer options. Thus, we now change our focus from the multiple indicator factor level to the question level following an Item Response Theory (IRT) approach.

In order to test our hypotheses on the equidistance of latent thresholds of answer options (3a, 3b, and 3c) we initially investigated the sequential order of unrestricted latent thresholds. For this purpose, we computed a univariate probit model estimating the unrestricted latent thresholds for each Likert-type question. Then, for each Likert-type question, we conducted a

linear regression with these estimated unrestricted threshold values, as dependent variable, and ascending integers from 1 to 4, as independent variable. We use the R^2 values of these linear regressions to determine whether the sequential order of latent thresholds is appropriately represented. Higher R^2 values signify a better sequential representation of the latent thresholds of answer options. Table 5 reports the results of the regressions.

The R^2 values reveal that the order of unrestricted latent thresholds is slightly better represented for the end labeled bipolar scales than for their unipolar counterparts, except for the second question. However, the differences are negligibly small and, thus, provide supporting evidence for Hypothesis 3a. We only find partial evidence for Hypothesis 3b on completely and end labeled unipolar scales. The sequential order of the unrestricted latent thresholds is somewhat better for end than completely labeled unipolar scales with two exceptions. First, for the first question we observe a higher R^2 value for the completely labeled unipolar scale. Second, for the second question we observe equally sized R^2 values for completely and end labeled unipolar scales. Finally, we find substantially higher R^2 values for end than completely labeled bipolar scales. This similarly applies to all four Likert-type questions with no exception. We take these findings as supporting evidence for Hypothesis 3c.

To investigate the equidistance of latent thresholds, we again computed a univariate model for each Likert-type question and constrained the latent thresholds to equal distances. We now use RMSEA values to evaluate the performance of the models. Lower RMSEA values signify more equidistantly distributed latent thresholds of answer options. Table 6 reports the results. Taking a closer look at Table 6 the results of the models with equality constraints for latent thresholds correspond to the results of the regressions with unrestricted latent threshold (see Table 5).

Appendix D contains graphical illustrations of the results on the unrestricted latent thresholds. Figures D1 to D3 show the distances between the unrestricted latent thresholds.

Table 5. R^2 and adjusted R^2 values of linear regressions of estimated unrestricted latent thresholds (Y) on ascending integers (X = 1 to 4)

	Unipolar scales				Bipolar scales			
	Completely labeled		End labeled		Completely labeled		End labeled	
Questions	R^2	Adjusted R^2	R^2	Adjusted R^2	R^2	Adjusted R^2	R^2	Adjusted R^2
1	0.996	0.995	0.993	0.989	0.975	0.962	0.993	0.990
2	1	0.999	1	0.999	0.993	0.989	0.999	0.998
3	0.997	0.995	0.998	0.997	0.994	0.990	1	1
4	0.998	0.997	0.999	0.999	0.988	0.982	1	1

Note. See Appendix A for the statements and answer options of the Likert-type questions. The results for completely labeled unipolar and bipolar scales are also reported in Höhne et al. (2020).

Table 6. Fit indices of univariate models with latent thresholds constrained to equal distances

	Unipolar scales						Bipolar scales					
	Completely labeled			End labeled			Completely labeled			End labeled		
Questions	χ^2 value	Df	RMSEA	χ^2 value	Df	RMSEA	χ^2 value	Df	RMSEA	χ^2 value	Df	RMSEA
1	11.50	2	0.063	25.99	2	0.100	172.61	2	0.266	25.69	2	0.099
2	3.27	2	0.023	1.44	2	0.000	56.08	2	0.150	5.46	2	0.038
3	23.14	2	0.097	7.32	2	0.047	49.70	2	0.141	0.93	2	0.000
4	4.90	2	0.035	2.96	2	0.020	83.84	2	0.185	0.60	2	0.000

Note. The results are based on the WLS discrepancy function for categorical data with THETA parameterization. No CFIs were reported because they are not defined in univariate models. See Appendix A for the statements and answer options of the Likert-type questions. The results for completely labeled unipolar and bipolar scales are also reported in Höhne et al. (2020).

5. Discussion and Conclusion

The aim of this experimental study was to extend the study by Höhne et al. (2020) and to investigate the effect of completely and end labeled unipolar and bipolar scales on respondents' answer behavior. For this reason, we analyzed observed and latent answer distributions and additionally compared latent thresholds of answer options. Table 7 provides a summary of our empirical findings in relation to the research hypotheses.

Table 7. Summary of the empirical findings

Hypotheses	Empirical findings
Answer distributions	
1a: Comparable	Supporting evidence
1b: Incomparable	Supporting evidence
1c: Incomparable	Supporting evidence
Measurement invariance	
2a: Comparable	Supporting evidence
2b: Incomparable	No evidence
2c: Incomparable	Supporting evidence
Latent thresholds	
3a: Comparable	Supporting evidence
3b: Incomparable	Partial evidence
3c: Incomparable	Supporting evidence

Note. All "a" hypotheses deal with end labeled unipolar and bipolar scales, all "b" hypotheses deal with completely and end labeled unipolar scales, and all "c" hypotheses deal with completely and end labeled bipolar scales.

We found supporting evidence for all hypotheses on end labeled unipolar and bipolar scales (1a, 2a, and 3a). This implies that end labeled unipolar and bipolar scales are comparable with respect to answer distributions, measurement invariance, and latent thresholds. The main reason for this comparability is that end labeled unipolar and bipolar scales only differ with respect to the verbal labels of the lowermost answer options (i.e., "agree not at all" vs. "disagree strongly"). Furthermore, their middle parts (i.e., the answer options between the endpoints) are not labeled, fostering the idea of equally distanced intervals. Our findings point to the fact that respondents seem to perceive and treat end labeled unipolar and bipolar scales equivalently.

The results on completely and end labeled unipolar scales are somewhat mixed and not entirely in line with our hypotheses (1b, 2b, and 3b). On the observational level, end labeled unipolar scales, compared to their completely labeled counterparts, seem to be associated with an extreme response style (see van Vaerenbergh & Thomas, 2013). Nevertheless, the results of the multigroup confirmatory factor analysis (MG-CFA) indicate scalar invariance for both unipolar scales. The latent thresholds also show a good sequential representation and equidistant distribution. Overall, these results support the comparability of answers to completely and end labeled unipolar scales.

We also found supporting evidence for the hypotheses on completely and end labeled bipolar scales (1c, 2c, and 3c). The results indicate that both scales are not comparable. This is supported by significantly different answer distributions, pointing to a positivity bias (see Tourangeau, Rips, & Rasinski, 2000) in completely labeled bipolar scales and concurrently to a middle response style (see van Vaerenbergh & Thomas, 2013) in end labeled bipolar scales.

Even though metric invariance can be obtained for completely and end labeled bipolar scales, both scales are not invariant because of differing intercepts (lacking scalar invariance). This points to the presence of systematic measurement error. Possible sources are the positivity bias in completely labeled bipolar scales and/or the middle response style in end labeled bipolar scales. This is only an attempted explanation and, thus, further refined research on completely and end labeled bipolar scales is necessary. Finally, the latent thresholds of answer options show a much better sequential representation and a more equidistant distribution for end labeled than for completely labeled bipolar scales.

As indicated by Schwarz (1996), the design of rating scales conveys information that can affect respondents' cognitive and communicative processes when answering survey questions. In our study, for instance, we show that unipolar and bipolar scales affect respondents' answer processes in such a way that the scales result in significantly different answer distributions. As a consequence, theory testing may not only depend on proper theoretical frameworks but also on the design of rating scales used for survey data collection. There are legitimate reasons to apprehend that two empirical studies on the same theoretical phenomenon may come to different conclusions because they used different rating scale designs. At first glance, this may sound far-fetched, but when considering our results, a proper rating scale design is key for a sound theory testing. This especially applies to cross-cultural and cross-national surveys in which, for instance, scale polarity is frequently not taken into consideration (see the German and English questionnaires of the ISSP 2012). We therefore call for attention when designing rating scales for collecting respondents' attitudes.

This study has some limitations that may provide perspectives for future studies on the scale design of Likert-type questions. First, even though we investigated completely and end labeled unipolar and bipolar scales, we did not consider other important scale aspects. For instance, previous research has shown that especially numeric values may have a powerful effect on answer behavior (see DeCastellarnau, 2018; Schwarz et al., 1991). We therefore suggest that future research additionally investigates the effect of numeric values when comparing unipolar and bipolar scales in terms of answer distributions, measurement invariance, and equidistance between scale points (or answer options). Second, we only focused on Likert-type questions dealing with income (in)equality. However, the specialty of these questions is that numerous unrelated topics (e.g., political efficacy and job motivation) can be measured without the necessity of changing the scales (see Revilla et al., 2014; Saris et al., 2010). For this reason, it would be worthwhile to cover a variety of question topics and to investigate measurement properties across different topics. Third, we adopted questions from the European Social Survey (ESS) and investigated the effects of completely and end labeled unipolar and bipolar scales on answer behavior in Germany. However, the ESS is conducted in a variety of countries and, thus, we advocate for a cross-cultural or cross-national comparison. In our opinion, this is a necessity to properly evaluate the comparability of scales that differ with respect to polarity and labeling. Fourth, in this study, we only looked at measurement properties, such as measurement invariance and latent thresholds of answer options. However, in order to make informed decisions about scale design it is crucial to look at data quality, such as reliability and validity. We therefore suggest that future studies provide further evidence on these aspects of data quality. In addition, it might be worthwhile to investigate respondents'

satisfaction with differently designed scales (see Krosnick & Presser, 2010; Menold & Bogner, 2015).

Despite its limitations, this study provides some important insights on the impact of scale polarity and labeling in Likert-type questions. Taking a closer look at our results it seems best to avoid completely labeled bipolar scales because they violate the criteria of equidistance (see Stevens, 1946). In addition, they are not comparable to end labeled bipolar scales in terms of answer distributions and measurement invariance. Even though completely labeled unipolar scales perform quite well, our results suggest that end labeled unipolar and bipolar scales are superior (higher applicability) and, thus, they should be preferred when measuring different aspects of income (in)equality. The main reason is that the two scales result in very similar (observed and latent) answer distributions. Both scales also show a good sequential representation and equidistance of latent thresholds of answer options.

References

Alwin, D. F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley and Sons.

Alwin, D. F. (2010). How good is survey measurement? Assessing the reliability and validity of survey measures. In P. V. Marsden & J. Wright (Eds.), *Handbook of Survey Research,* (pp. 405–434). London, UK: Emerald Group Publishing.

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. Field Methods, 27, 391–408.

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling, 19,* 372–398.

Byrne, B. M. (2012). *Structural equation modeling with Mplus. Basic concepts, applications, and programming*. New York, NY: Routledge.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review, 82,* 45–73.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Converse, J. M., & Presser S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: Sage.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity, 52,* 1523–1559.

Fowler, F. (1995). *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage.

Fowler, F., & Cosenza, C. (2008). Writing effective questions. In E. de Leeuw, J.J. Hox, & D.A. Dillman (Eds.), *International Handbook of Survey Methodology* (pp. 136–160). New York, NY: Taylor & Francis.

Höhne, J. K. (2019). Eye-tracking methodology: Exploring the processing of question formats in web surveys. *International Journal of Social Research Methodology, 22,* 199-206.

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,* 91–103.

Höhne, J. K., Krebs, D., Kühnel, S.-M. (2020). Measuring income (in)equality: Comparing survey questions with unipolar and bipolar scales in a probability-based online panel. *Social Science Computer Review*. DOI: 10.1177/0894439320902461

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., Lenzner, T., Neuert, C. E., & Yan, T. (2019). Re-examining the middle means typical and the left and top means first heuristics using eye-tracking methodology. *Journal of Survey Statistics and Methodology*. DOI: 10.1093/jssam/smz028

Höhne, J. K., & Yan, T. (2020). Investigating the impact of violations of the "left and top means first" heuristic on response behavior and data quality. *International Journal of Social Research Methodology, 23*, 347–353.

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, NY: John Wiley and Sons.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*, (pp. 263–313). Bingley: Emerald.

Kuru, O., & Pasek, J. (2016). Improving social media measurement in surveys: Avoiding acquiescence bias in facebook research. *Computers in Human Behavior, 57*, 82–92.

Lelkes, Y., & Weiss, R. (2015). Much ado about acquiescence: The relative validity and reliability of construct-specific and agree-disagree questions. *Research and Politics*. DOI: 10.1177/2053168015604173

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

Liu, M., Lee, S., & Conrad, F.G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales. *Public Opinion Quarterly, 79*, 952–975.

Menold, N. (2019). Response bias and reliability in verbal agreement rating scales: does polarity and verbalization of the middle category matter? *Social Science Computer Review*. DOI: 10.1177/0894439319847672

Menold, N., & Bogner, K. (2015). Gestaltung von Ratingskalen in Fragebögen. Mannheim (Germany): GESIS – Leibniz-Institute for the Social Sciences (SDM Survey Guidelines).

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., & Raykov, T. (2015). Can reliability of multiple component measuring instruments depend on response option presentation mode? *Educational and Psychological Measurement, 76*, 454–469.

Mohler, P. P., Smith, T. W. & Harkness, J. A. (1998). Respondents' ratings of expressions from response scales: A two country, two language investigation on equivalence and translation. In J. A. Harkness (Ed.), *Cross-Cultural Survey Equivalence*, (pp. 159–184). Mannheim: Zentrum für Umfragen, Methoden und Analysen.

O'Muircheartaigh, C., Gaskell, G., & Wright, D. B. (1995). Weighing anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics, 11*, 295–308.

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 (Eds.), *Modern issues in perception,* (pp. 262–82). Berlin: VEB Deutscher Verlag der Wissenschaften.

Revilla, M., Saris, W., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods and Research, 43*, 73–97.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*, 354–373.

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

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12*, 397–409.

Saris, W., & Gallhofer, I. N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: John Wiley & Sons.

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

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual review of sociology, 46*, 10.1–10.24.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual review of sociology, 29*, 65–88.

Scholz, E., & Jutz, R. (2014). ISSP 2012 Germany: Family and Gender Roles IV. GESIS Report on the German Study. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-402638

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation*. New York: Psychology Press.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570–582.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.

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*, 15–38.

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

Thomas, R. K., & Barlas, F. M. (2018). We've got your number: Can numeric labels replace semantic labels in scales. *Paper presented at the General Online Research (GOR) conference in Cologne (Germany)*.

Toepoel, V., & Dillman, D. A. (2011a). How visual design affects the interpretability of survey questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and The Internet. Advances in Applied Methods and Research Strategies,* (pp. 165–190). New York: Routledge.

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

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.

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

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

van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*, 195–217.

Wang, R., & Krosnick, J. A. (2020). Middle alternatives and measurement validity: A recommendation for survey researchers. *International Journal of Social Research Methodology, 23*, 169–184.

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment, 21*, 765–774.

Appendix A

Table A1. Likert-type questions used in this study

Question parts	Unipolar scales	Bipolar scales
Requests for an answer	To what extent do you agree with the following statement?	To what extent do you agree or disagree with the following statement?
Statement 1	Employees need strong unions to protect their working conditions and wages.	
Statement 2	Large income differences are acceptable to adequately acknowledge different talents and achievements.	
Statement 3	To ensure fair society differences in people's living standards should be small.	
Statement 4	Social benefits lead to more equality in society.	
Completely labeled scales	○ Agree strongly ○ Agree somewhat ○ Agree moderately ○ Agree hardly ○ Agree not at all	○ Agree strongly ○ Agree somewhat ○ Neither agree nor disagree ○ Disagree somewhat ○ Disagree strongly
End labeled scales	○ Agree strongly ○ ○ ○ ○ Agree not at all	○ Agree strongly ○ ○ ○ ○ Disagree strongly

Note. The four Likert-type questions were adapted from the European Social Survey (2002, 2016). The order of the questions was randomized to limit question order effects. All questions were presented on a separate online survey page using five-point, vertically aligned scales with radio buttons for selecting an answer option. All questions included a brief instruction stating that only one answer option can be selected. The original German wordings of the statements and answer options of the Likert-type questions are available from the first author on request.

Appendix B

Screenshots illustrating the question and scale design presented on a PC (bipolar version only).

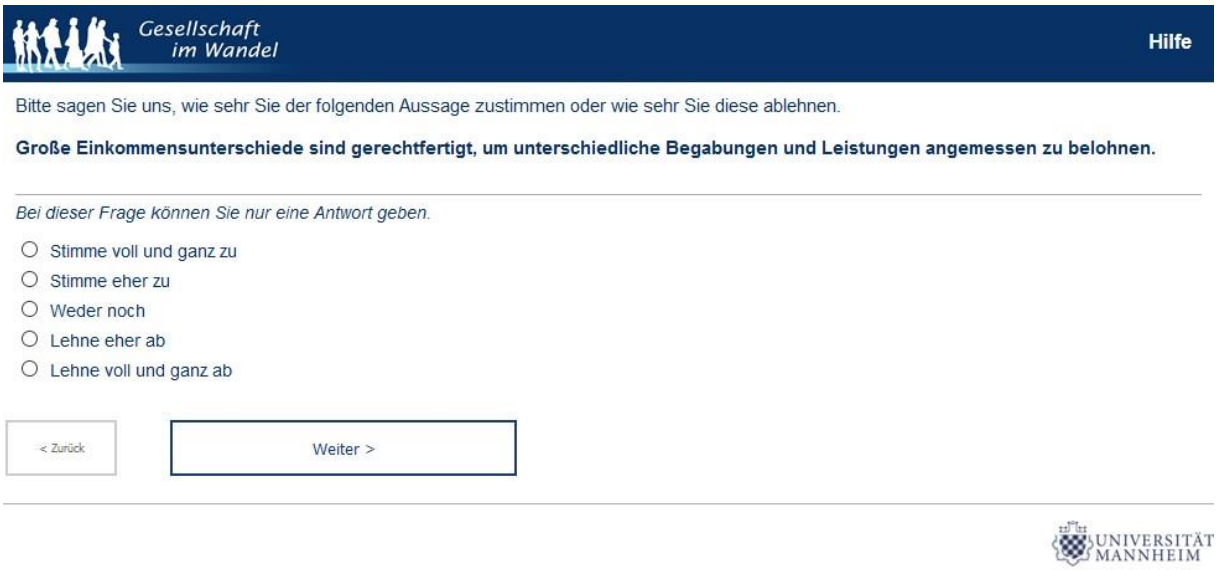

Figure B1. Screenshot illustrating completely labeled scales used in this study.

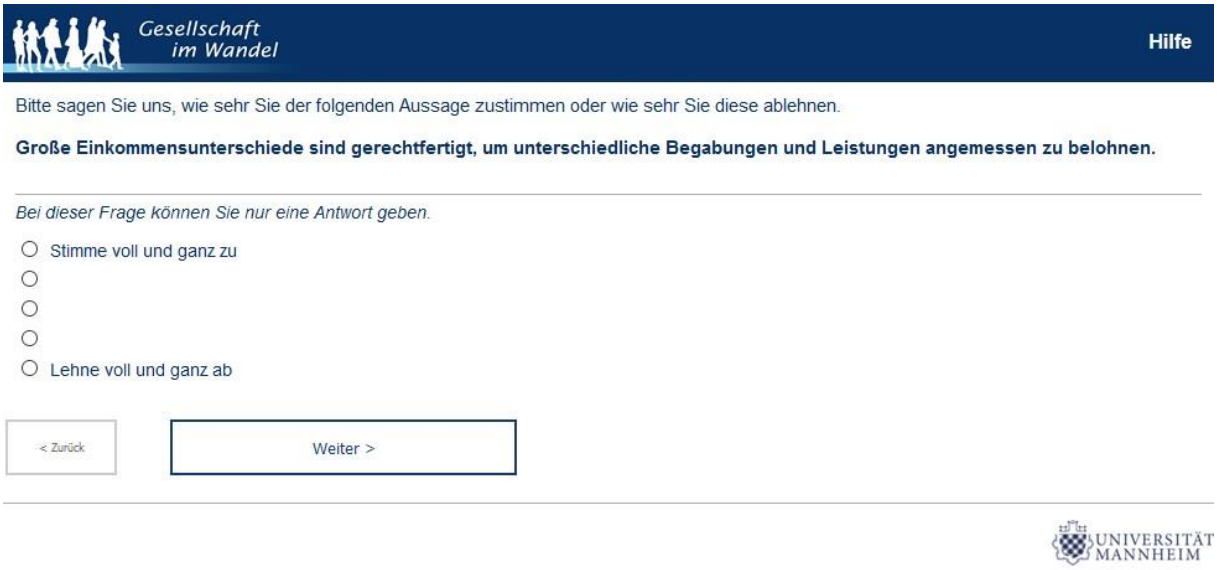

Figure B2. Screenshot illustrating end labeled scales used in this study.

Appendix C

Mplus commands to track the analyses of measurement invariance and latent thresholds.

MEASUREMENT INVARIANCE

```
VARIABLE:
  NAMES ARE scale v1 v2 v3 v4;
  USEVARIABLES ARE scale v1 v2 v3 v4;
  GROUPING scale (1 = unipolar 2 = bipolar);
ANALYSIS:
  ESTIMATOR IS MLR;
```

```
MODEL:
  F1 BY v1 v2 v3 v4;
  v1 WITH v3;
  [F1@0];
Model bipolar_end
```

LATENT THRESHOLDS

```
VARIABLE:
  NAMES ARE scale v1;
  CATEGORICAL IS v1;
  USEVARIABLES ARE v1;
  USEOBSERVATIONS ARE scale EQ 1;
ANALYSIS:
  ESTIMATOR IS WLS;
  PARAMETERIZATION IS THETA;
MODEL:
  v1@1;
  [v1$1] (t1);
  [v1$2] (t2);
  [v1$3] (t3);
  [v1$4] (t4);
  F1 BY v1@1; [F1@0]; F1@0;
MODEL CONSTRAINT:
  NEW (d*1.0);
  t2=t1+d;
  t3=t2+d;
  t4=t3+d;
```

Appendix D

Graphical illustrations of the distances between the unrestricted latent thresholds (question 4 only).

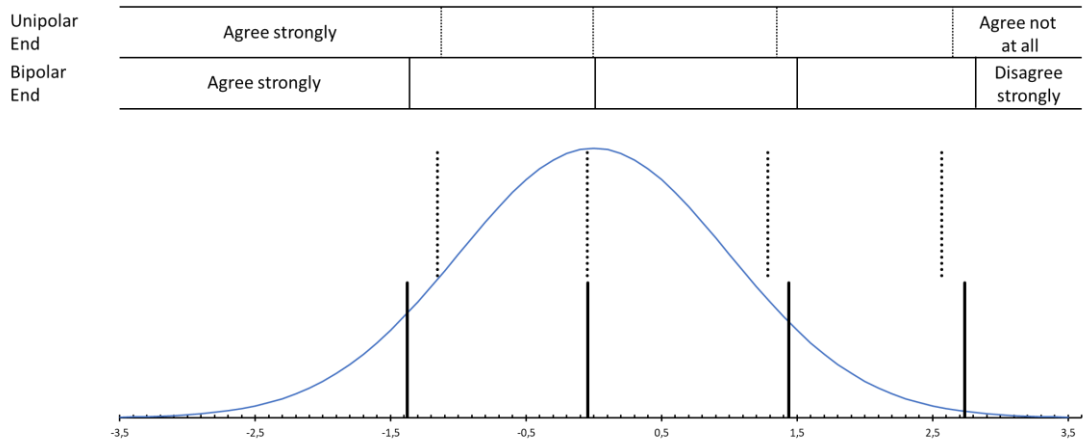

Figure D1. Distances between the unrestricted latent thresholds of end labeled unipolar and bipolar scales.

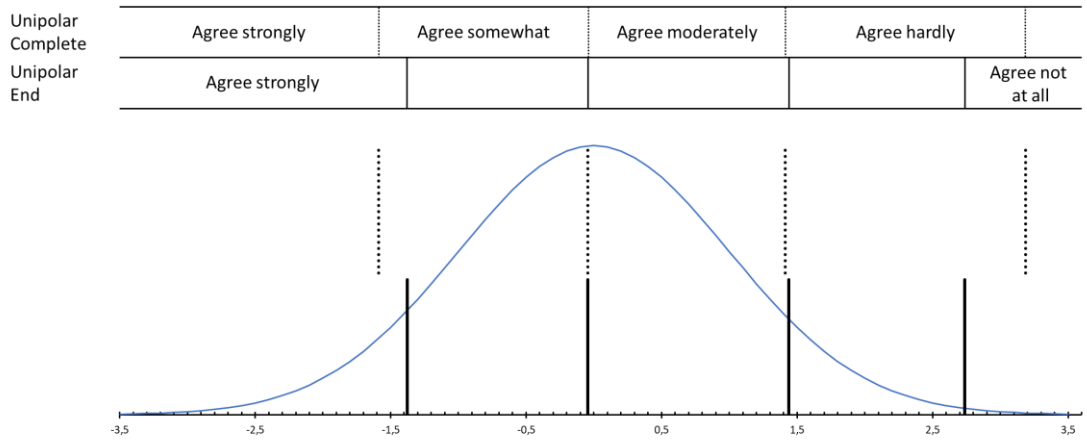

Figure D2. Distances between the unrestricted latent thresholds of completely and end labeled unipolar scales.

Note. The last verbal label of the completely labeled unipolar scale ("agree not at all") is not displayed above because of space limitations.

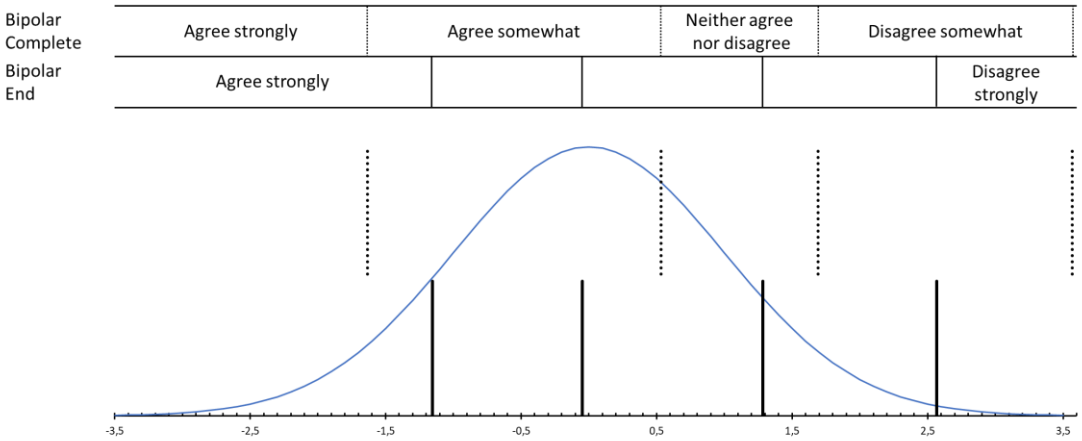

Figure D3. Distances between the unrestricted latent thresholds of completely and end labeled bipolar scales.

Note. The last verbal label of the completely labeled bipolar scale ("disagree strongly") is not displayed above because of space limitations.