Survey Research Methods (2023)
Vol. 17, No. 2, pp. 193-204
doi:10.18148/srm/2023.v17i2.8006
European Survey Research Association

Investigating direction effects in rating scales with five and seven points in a probability-based online panel

Jan Karem Höhne[1,2], Dagmar Krebs[3], and Steffen-M. Kühnel[4]
[1]University of Hannover, German Centre for Higher Education Research and Science Studies
[2]RECSM-University Pompeu Fabra
[3]University of Gießen
[4]University of Göttingen

Survey questions with rating scales are a common method in attitude measurement. Similar to other scale characteristics, scale direction and its effects on answer behavior and data quality deserves special attention. This particularly applies to scale direction effects across different scale lengths. In order to contribute to the current state of research, we investigate scale direction effects across scales with five and seven points by analyzing observed and latent answer distributions including composite reliabilities. We conducted an experiment in the probability-based German Internet Panel ($N = 4676$) using questions on achievement and job motivation that vary scale direction (i.e., decremental and incremental) and scale length (i.e., five and seven points). The results reveal differences between scales with five and seven points. Five-point scales are more robust against scale direction effects than their seven-point counterparts. In addition, decremental and incremental scales with five points are invariant. This does not apply to decremental and incremental scales with seven points. However, composite reliabilities are higher for scales with seven points than for scales with five points. This is irrespective of the scale direction.

Keywords: Composite reliabilities; Data quality; Latent means; Measurement invariance; Online survey; Scale direction effects; Survey-satisficing

1 Introduction and Background

Survey questions with rating scales are a frequently used method in social science research and many adjacent research fields for measuring respondents' attitudes and opinions. For example, major national and international social surveys, such as the American National Election Study (ANES) and the European Social Survey (ESS), regularly employ survey questions with rating scales. However, as shown by previous research, the design of rating scales can have a profound impact on respondents' answer behavior, potentially inducing systematic measurement error (DeCastellarnau, 2018; Krosnick & Presser, 2010; Schaeffer & Dykema, 2020; Schaeffer & Presser, 2003). When designing rating scales researchers must carefully decide about the inclusion of a scale midpoint (i.e., an even or uneven number of scale points), the length of the scale (i.e., the actual number of scale points), the labeling of te scale (i.e., complete or partial), the polarity of the scale (i.e., unipolar or bipolar), the inclusion of numeric values (i.e., numbers that accompany scale points), the alignment of the scale (i.e., horizontal or vertical), and the direction of the scale (i.e., decremental or incremental).

Compared to other scale design effects, effects of the scale direction (or response order) received relatively little attention so far (Menold & Bogner, 2015). Most studies investigated response order effects in lists of unordered options that, for example, captured qualities that might be important for a child to have (Krosnick & Alwin, 1987, p. 205). The occurrence of response order effects, however, is not restricted to unordered lists of options, but can also occur in ordered lists of options, as it is the case in rating scales. In this context, we usually speak of scale direction effects instead of response order effects.

As outlined by Sudman et al. (1996), scale direction effects can occur in two different ways. If respondents' answers are shifted toward the beginning of the rating scale, we speak of primacy effects. If respondents' answers are shifted toward the end of the rating scale, we speak of recency effects. However, as mentioned by Keusch and Yan (2018), the terms primacy and recency effects, including their theoretical frameworks, are usually associated with unordered rather

Contact information: Jan Karem Höhne, German Centre for Higher Education Research and Science Studies (DZHW), Research Area 4: Research Infrastructure and Methods, Lange Laube 12, 30159 Hannover, Germany (E-mail: hoehne@dzhw.eu)

than ordered options. Nevertheless, we adopt the terms primacy and recency effects to simply describe respondents' answering tendency—either to the beginning of the rating scale or to the end of the rating scale (for a similar practice see Sudman et al., 1996, p. 157).

Rating scales can have a decremental direction (e.g., "applies completely" to "applies not at all") or an incremental direction (e.g., "applies not at all" to "applies completely"). Overall, it seems that respondents' answers are more likely to shift toward the beginning of rating scales, resulting in primacy effects (see Galesic et al., 2008; Höhne & Lenzner, 2015; Höhne et al., 2018; Keusch & Yan, 2018; Krebs, 2012; Krebs & Bachner, 2018; Krebs & Hoffmeyer-Zlotnik, 2010; Krebs & Höhne, 2020, 2021; Mavletova, 2013; Rammstedt & Krebs, 2007; Sudman et al., 1996; Toepoel, 2008; Yan & Keusch, 2015). Interestingly, some studies find that primacy effects only occur in decremental rating scales that range from the high to the low end (Krebs & Hoffmeyer-Zlotnik, 2010; Krebs & Höhne, 2020), whereas some others find that primacy effects only occur in incremental rating scales that range from the low to the high end (Toepoel, 2008). There are even studies that find no effects at all (Keusch & Yan, 2018; Rammstedt & Krebs, 2007; Weng & Cheng, 2000).

Following the survey-satisficing theory proposed by Krosnick (1991) and Krosnick and Alwin (1987), scale direction or primacy effects are a consequence of weak satisficing. Accordingly, respondents circumvent the effort that is necessary to provide decent and thoughtful answers to survey questions. In doing so, they draw on a variety of superficial answer strategies, such as selecting the first answer option that seems to constitute a reasonable answer (Krosnick, 1991, p. 213). The underlying mechanisms responsible for the occurrence of primacy effects are twofold: First, it is possible that respondents choose the first adequately appearing answer option avoiding reading the entire range of options. Second, it is possible that respondents process the first answer options more deeply so that these have a higher selection chance. Galesic et al. (2008) and Höhne and Lenzner (2015) provide supporting evidence for both answer strategies. The author groups use eye-tracking methodology and show that respondents look more frequently and longer at the top (first appearing) answer options shifting their answers into this direction. This kind of answer behavior applies to lists with ordered (Galesic et al., 2008; Höhne & Lenzner, 2015) and unordered answer options (Galesic et al., 2008; Galesic & Yan, 2011).

One special factor that affects the likelihood of survey-satisficing is task difficulty (Krosnick, 1991; Krosnick et al., 1996). Task difficulty is a feature of a survey question and depends on how much mental work is required to accomplish the task set out by a survey question (Anand, 2008, p. 798). The higher the task difficulty, the higher the likelihood of survey-satisficing. Particularly, the length of the rating scale can increase task difficulty potentially fostering the occurrence of survey-satisficing in the form of primacy effects. When including too many answer options, the meaning of the individual options becomes less clear, which in turn impedes the selection of an option (Menold & Bogner, 2015; Parducci, 1983). In line with this theoretical reasoning some studies indicate that scale direction effects are more common in longer than in shorter rating scales (Liu, 2017; Tourangeau et al., 2017; Yan et al., 2018). Tourangeau et al. (2017), for example, investigated direction effects in rating scales with five and seven points. While the authors found a significant shift towards the first answer options in seven-point scales (indicating a primacy effect), they did not find a shift towards the first answer options in five-point scales. This finding suggests that the seven-point scales increased task difficulty promoting survey-satisficing in the form of selecting the first reasonable answer option.

Considering the existing studies on direction effects in rating scales with different lengths it is to observe that their analyses of answer distributions mostly remain on the observational level. Studies investigating scale direction effects on the latent level are rather scarce (exceptions are studies by Chan, 1991; Liu, 2017; Salzberger & Koller, 2013). This particularly applies to studies that also vary scale length. In addition, there is a lack of studies investigating data quality of rating scales differing in terms of direction and length.

In this study, we address this research gap using experimental survey data that were collected in the probability-based German Internet Panel. We investigate scale direction effects across decremental and incremental rating scales with five and seven points. In doing so, we analyze observed and latent answer distributions. Only analyses on the latent level support the investigation of measurement invariance among these rating scale designs. Measurement invariance indicates that the relationship between test scores (or answers to questions) and latent attributes is not affected by measurement methods (Meredith, 1993; Millsap, 2007) that vary with respect to scale direction and length. Obtaining measurement invariance also facilitates the comparison of (latent) means. In addition to measurement invariance, we consider data quality in terms of composite reliabilities (Raykov & Marcoulides, 2015). The investigation of composite reliabilities also requires latent variable modeling. By analyzing scale direction effects on the latent level, our study stands out of previous studies contributing to the eminent survey literature on rating scale design.

2 Research Hypotheses

Following the notion that an increasing number of scale points increases task difficulty and thus the occurrence of survey-satisficing (Krosnick, 1991; Krosnick & Alwin, 1987), it is assumable that scale direction (or primacy) effects are more likely to occur in longer than shorter rating scales. In longer scales, respondents may have trouble to properly distinguish the answer options so that they either select the first adequately appearing option or process the first options more deeply than the later ones (Galesic et al., 2008; Galesic & Yan, 2011; Höhne & Lenzner, 2015). Both answer strategies result in a higher selection chance of the answer options at the rating scale beginning. In correspondence with this theoretical reasoning, we propose the following research hypothesis:

Research hypothesis 1: Decremental and incremental rating scales with seven points are associated with stronger answer shifts to the scale beginning than their five-point counterparts.

Studies on direction effects in rating scales of different lengths mostly remained on the observational level by analyzing answer distributions (see Tourangeau et al., 2017; Yan et al., 2018) Thus, so far, it is unclear whether and to what extent these effects manifest themselves on the latent level as well. By latent level, we refer to measurement properties in terms of measurement invariance and latent means. As indicated by previous research, rating scales with seven points are particularly prone to measurement error that is caused by the direction of the scale (Tourangeau et al., 2017). Measurement error has the potential to affect the attainment of measurement invariance (Steinmetz, 2013) and thus we propose the following research hypothesis:

Research hypothesis 2: Measurements with seven-point decremental and incremental rating scales are not invariant, whereas their five-point counterparts are invariant.

In case of measurement invariance (in the form of scalar invariance; see 3.5 Analytical Strategies), we propose the following research hypothesis on latent means:

Research hypothesis 3: Seven-point decremental and incremental rating scales show larger latent mean differences than their five-point counterparts.

In order to draw conclusions about data quality of rating scales differing in terms of direction and length we investigate composite reliabilities of multi-item measurement instruments. In doing so, we follow an approach suggested by Raykov and Marcoulides (2015) that is based on latent variable modeling. It assumes one-dimensionality of the components (items or questions) of a measurement instrument and allows point and interval estimations of group differences in composite reliabilities. We address the following final research hypothesis:

Research hypothesis 4: Five-point decremental and incremental rating scales result in higher data quality in terms of composite reliabilities than their seven-point counterparts.

3 Method

3.1 Data Collection and Study Procedure

Data were collected in the German Internet Panel, which is part of the Collaborative Research Center 884 "Political Economy of Reforms". 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 from 16 to 75 years. For a detailed methodological description of the German Internet Panel, we refer interested readers to Blom et al. (2015).

The German Internet Panel invites all panel members every two months to participate in a self-administered online survey that deals with a variety of economic, political, and social topics. At the beginning of each wave, panelists are directed to a short welcome page announcing the approximate length of the online survey (about 20 minutes) and informing them that the compensation for their participation (in the amount of 4€) will be credited to their study account after survey completion.

3.2 Sample Characteristics

We use data that were collected in wave 42 of the probability-based German Internet Panel (see Blom et al., 2020). This wave ran from July 1 to July 31, 2019, with a total of 4714 respondents. Of these respondents, 38 broke off before being asked any study-relevant questions. This leaves us with 4676 respondents for statistical analyses. The median birth year category is "1965 to 1969", and 48% of them are female. In terms of education, 14% graduated from a lower secondary school, 31% from an intermediate secondary school, and 51% from a college preparatory secondary school or university. Furthermore, 4% were still attending school, left school without a diploma, or reported having a different degree from those mentioned above.

3.3 Experimental Design

In order to investigate the effects of scale direction (i.e., decremental and incremental) across different scale lengths (i.e., five and seven points) we conducted a split-ballot experiment and randomly assigned respondents to one out of four experimental groups. Table 1 describes the four experimental groups.

3.4 Survey Questions on Achievement and Job Motivation

In this study, we adopted 12 survey questions from the "Cross Cultural Survey for Work and Gender Attitudes"

Table 1

Experimental design defined by scale direction and scale length

Experimental group	Scale direction	Scale length	Group size
1	Decremental	Five points	1173
2	Incremental	Five points	1171
3	Decremental	Seven points	1167
4	Incremental	Seven points	1165

The four experimental groups do not differ significantly with respect to age, gender, and education.

(2010). Five of these questions deal with achievement motivation, four deal with intrinsic job motivation, and three deal with extrinsic job motivation (see also footnote 2). All questions were presented individually (i.e., one question per online survey page). The rating scales were end-labeled and vertically aligned with no numeric values.[1] All survey questions including rating scales were in German (see Appendix for English translations).

3.5 Analytical Strategies

Research hypothesis 1

In order to investigate whether there is a shift of answers towards the scale beginning, we initially recode all rating scales to identical values running from "applies not at all" to "applies completely". We then compute dummy-variables with the value 1 for "applying" answer options (last two for five- and last three for seven-point scales) as well as dummy-variables with the value 1 for "non-applying" answer options (first two for five- and first three for seven-point scales). In doing so, we follow a similar analytical strategy as Tourangeau et al. (2017). Subsequently, we calculate the average proportions for the questions on achievement motivation, intrinsic job motivation, and extrinsic job motivation, respectively, and compare these proportions by conducting Z tests to determine significant differences.

Research hypothesis 2

In order to investigate scale direction effects on the latent level, we first conduct confirmatory factor analyses (CFAs) with three latent variables (achievement motivation and intrinsic and extrinsic job motivation) and 12 indicators for decremental and incremental rating scales with five and seven points.[2] We then conduct multigroup confirmatory factor analyses (MG-CFAs) to test for configural invariance; using an identical dimensional structure across decremental and incremental scales with five, and seven points. Subsequently, we successively impose increasing equality constraints on the parameters (Byrne, 2008; Davidov et al., 2014): first, by constraining the factor loadings to be equal (metric invariance). Second, by constraining the intercepts to be equal (scalar invariance). While metric invariance allows the comparison of correlations, scalar invariance allows the comparison of (latent) means.

Criteria for testing measurement invariance between models with increasing equality constraints are non-significant differences between chi-square values (Bryant & Satorra, 2012; Byrne, 2012) and differences between comparative fit indices (CFIs) and root mean square errors of approximation (RMSEAs) above 0.01 (Cheung & Rensvold, 2002). Opposing results imply measurement non-invariance. Since the indicators of the latent variables are measured with five- and seven-point scales, we assume a continuous scale level (see Rhemtulla et al., 2012) and use the robust maximum likelihood (MLR) discrepancy function.

Research hypothesis 3

After testing for measurement invariance between decremental and incremental rating scales with five and seven points, respectively, we analyze shifts in latent means. This is only done for groups being invariant.

Research hypothesis 4

In order to shed light on data quality we compute composite reliabilities that are based on unidimensional multigroup confirmatory factor analyses (MG-CFAs) for decremental and incremental scales with five and seven points. This is separately done for—the multi-item instruments—achievement motivation, intrinsic job motivation, and extrinsic job motivation. This procedure results in both point and interval estimations of composite reliabilities for decremental and incremental scales with five and seven points.

[1] Vertical scale alignment is the default setting in the German Internet Panel.

[2] According to the results of the confirmatory factor analyses (CFAs), the indicator of intrinsic job motivation on autonomy was moved to extrinsic job motivation. Thus, in the analyses, intrinsic job motivation consists of three indicators and extrinsic job motivation consists of four indicators (see Appendix for the survey questions and the online replication files).

The descriptive statistics including Z tests are conducted with SPSS version 27 and the multigroup confirmatory factor analyses (MG-CFAs), latent mean comparisons, and composite reliabilities are computed with Mplus version 6.12 (replication files are published online).

4 Results

4.1 Research Hypothesis 1

Considering Table 2 it is to observe that there is no statistically significant difference between decremental and incremental scales with five points. This applies to the questions on achievement motivation, intrinsic job motivation, and extrinsic job motivation. Thus, there is no supporting evidence for the occurrence of primacy effects in five-point scales.

When looking at Table 3 it is to observe that two out of six comparisons—between decremental and incremental scales with seven points—result in significant differences. More specifically, for the questions on extrinsic job motivation, respondents' answers are significantly shifted toward the beginning of the scales. In these cases, respondents selected significantly more often "applying" options on decremental scales and significantly more often "non-applying" options on incremental scales. This indicates that primacy effects are at work, corroborating findings reported by Tourangeau et al. (2017). For the questions on achievement motivation and intrinsic job motivation, in contrast, no significant differences can be observed. However, for achievement motivation the differences are much more pronounced than for intrinsic job motivation and only slightly miss the p-level of 5%. Overall, these findings provide some supporting evidence for our first research hypothesis.

4.2 Research Hypothesis 2

We initially computed separate but identical confirmatory factor analyses (CFAs) baseline models for each scale direction (i.e., decremental and incremental) within each scale length (i.e., five and seven points). Each of these four baseline models contained three latent variables with 12 indicators. We admitted one error covariance between two indicators of achievement motivation. All baseline models had satisfactory goodness-of-fit statistics.

Next, we conducted multigroup confirmatory factor analyses (MG-CFAs). To this end, we first tested configural invariance by simultaneously analyzing the baseline model for the two scale directions within each scale length. Table 4 reports the statistical results. Given CFI values above 0.95 and RMSEA values ≤ 0.05, configural invariance was accepted for rating scale directions within both scale lengths. In order to test metric invariance, factor loadings were constrained to equality between decremental and incremental scale directions within scale lengths. The chi-square difference tests between the metric and configural models were not significant and, thus, we accepted metric invariance. Finally, to compare latent means, scalar invariance was tested by imposing equality constraints on the intercepts. The results are mixed. While we obtain scalar invariance for decremental and incremental scales with five points, we do not obtain scalar invariance for decremental and incremental scales with seven points.

The two criteria for measurement invariance between models with increasing equality constraints—non-significant differences between (mean-adjusted) chi-square values (Byrne, 2012) and differences between CFI and RMSEA values below 0.01 (Cheung & Rensvold, 2002)—hold for the model of five-point scales. The model of seven-point scales does not meet these criteria. This provides strong evidence for our second hypothesis.

4.3 Research Hypothesis 3

In line with the results on measurement invariance, we only compare latent means for rating scales with five points. Differences in latent means between answers across rating scale directions are tested using the decremental direction as reference group. Since the answers for both scale directions were coded from 1 "applies not at all" to 5 "applies completely" estimates with negative signs indicate that answers on the incremental scales are slightly more negative (i.e., the answers have lower values) than those on the decremental scales. Table 5 reports the statistical results. Fit of the mean comparing model: $\chi^2(118) = 422.52(1.36)$; RMSEA = 0.047; CFI = 0.961. The shifts in latent means between decremental and incremental scales with five points are negligibly small and non-significant. Thus, there is no supporting evidence for scale direction effects in five-point scales. The findings on the latent level correspond to those on the observational level.

4.4 Research Hypothesis 4

Since we obtained metric invariance for both scale directions within both scale lengths we are able to conduct comparisons on the correlational level. Therefore, we now investigate data quality in terms of composite reliabilities of decremental and incremental rating scales with five and seven points. In doing so, we follow an approach suggested by Raykov and Marcoulides (2015) and computed multigroup confirmatory factor analysis (MG-CFA) models with equality constraints on factor loadings and intercepts. This was done for each of the three latent variables (achievement motivation, intrinsic job motivation, and extrinsic job motivation). We then computed composite reliabilities, including 95% confidence intervals. Tables 6 and 7 report the statistical results.

Considering the composite reliability coefficients, we observe that they have values above 0.80. This similarly applies to decremental and incremental scales with five and seven

Table 2

Average proportions of respondents selecting applying and non-applying options of decremental and incremental scales with five points

	Achievement motivation	Intrinsic job motivation	Extrinsic job motivation
	Applying answers (%)	Applying answers (%)	Applying answers (%)
Scale direction			
Decremental	46	83	61
Incremental	46	85	61
Z	0.42	−0.90	−0.26
	Non-applying answers (%)	Non-applying answers (%)	Non-applying answers (%)
Scale direction			
Decremental	22	3	23
Incremental	20	2	23
Z	1.58	0.40	0.00

Five questions for achievement motivation, three questions for intrinsic job motivation, and four questions for extrinsic job motivation (see footnote 2).

Table 3

Average proportions of respondents selecting applying and non-applying options of decremental and incremental scales with seven points

	Achievement motivation	Intrinsic job motivation	Extrinsic job motivation
	Applying answers (%)	Applying answers (%)	Applying answers (%)
Scale direction			
Decremental	55	88	73
Incremental	51	8867	61
Z	1.87	0.06	3.281***
	Non-applying answers (%)	Non-applying answers (%)	Non-applying answers (%)
Scale direction			
Decremental	26	4	25
Incremental	29	5	30
Z	−1.70	−0.51	−2.75**

Five questions for achievement motivation, three questions for intrinsic job motivation, and four questions for extrinsic job motivation (see footnote 2).
** $p < 0.01$ *** $p < 0.001$

points. The only exception is the incremental scale with five points for extrinsic job motivation with a coefficient of 0.776. For five-point scales, we also find a significant difference between reliability coefficients for achievement motivation. Specifically, composite reliability is significantly higher for the decremental scale than for the incremental scale. This indicates that the incremental scale direction can decrease data quality. For seven-point scales, in contrast, we neither find reliability coefficients below 0.80 nor significant differences between scale directions. Composite reliabilities of decremental and incremental scales with seven points are slightly but consistently higher than composite reliabilities of decremental and incremental scales with five points. The only exception is the decremental scale for achievement motivation (coefficients = 0.837 and 0.836). Furthermore, the inspection of the confidence intervals shows that reliability coefficients of decremental and incremental scales with seven points are more similar than those of decremental and incremental scales with five points. Overall, the results on composite reliabilities contradict our fourth research hypothesis.

5 Discussion and Conclusion

The goal of this study was to investigate the occurrence of scale direction effects across scales that differ in terms of

Table 4

Testing measurement invariance between decremental and incremental rating scales with five and seven points

Invariance level	Chi-square value	Scale correction factors	df	RMSEA	CFI	Chi-square difference test
Five-point scales						
Configural	394.10	1.40	100	0.050	0.962	
Metric	402.35	1.38	109	0.048	0.962	1.97
Scalar	425.92	1.35	121	0.047	0.961	18.33
Seven-point scales						
Configural	416.43	1.36	100	0.052	0.964	
Metric	431.65	1.35	109	0.051	0.963	15.22
Scalar	463.74	1.31	121	0.050	0.961	26.17**

The results are based on MLR estimation. Scale correction factors for model comparisons. Five questions for achievement motivation (latent factor 1), three questions for intrinsic job motivation (latent factor 2), and four questions for extrinsic job motivation (latent factor 3). For "metric" the increase in df is 9 (because of the free estimation of three factor variances), whereas for "scalar" the increase in df is 12 (corresponding to the number of indicators).

** $p < 0.01$

Table 5

Latent mean differences between decremental and incremental rating scales with five points (unstandardized results)

	Estimate	Standard error	Critical ratio	p-values
Achievement motivation	−0.038	0.047	−0.811	0.417
Intrinsic job motivation	−0.032	0.044	−0.724	0.469
Extrinsic job motivation	−0.006	0.048	−0.134	0.894

Answers to decremental and incremental rating scales with five points were coded to values ranging from 1 "applies not at all" to 5 "applies completely". The reference group is the decremental scale direction. Five questions for achievement motivation (latent factor 1), three questions for intrinsic job motivation (latent factor 2), and four questions for extrinsic job motivation (latent factor 3).

length. For this purpose, we conducted a survey experiment in the probability-based German Internet Panel using questions on achievement and intrinsic and extrinsic job motivation. The four experimental groups differed with respect to scale direction (i.e., decremental or incremental) and scale length (i.e., five or seven points). In a first step, we compared the answer distributions of decremental and incremental scales with five and seven points. Then, we investigated measurement invariance, latent means, and composite reliabilities. Our findings revealed differences in scale direction effects and data quality.

With respect to our first research hypothesis on shifts of answers to the scale beginning we partially replicated findings reported by Tourangeau et al. (2017). In line with the authors, we found significant differences for seven-point scales, but no significant differences for five-point scales. More specifically, for scales with seven points, we found answer shifts toward the beginning of decremental and incremental scales. Following the survey-satisficing theory (Krosnick, 1991), we argue that seven-point rating scales increase task difficulty fostering the occurrence of scale direction effects. Including too many answer options may blur the meaning of individual options. This makes answer option selection more difficult. Consequently, scale direction effects seem to be more common in longer than in shorter scales.

In order to investigate our second research hypothesis, we tested for measurement invariance between decremental and incremental scales within five- and seven-point scales, respectively. Scalar invariance could only be obtained for the scales with five points. For the scales with seven points, in contrast, only metric invariance could be obtained. This implies that for seven-point scales, the intercepts differ between

Table 6

Model-based composite reliabilities for decremental and incremental scales with five points

				95% C.I.			
Scale direction	Coefficients	Differences	p-value	lower	upper	RMSEA	CFI
Achievement motivation							
Decremental	0.837	0.030	0.031	0.819	0.855	0.032	0.993
Incremental	0.808			0.786	0.829		
Intrinsic job motivation							
Decremental	0.869	0.016	0.169	0.853	0.884	0.000	1
Incremental	0.852			0.834	0.870		
Extrinsic job motivation							
Decremental	0.813	0.037	0.054	0.790	0.836	0.040	0.985
Incremental	0.776			0.746	0.806		

Five questions for achievement motivation (latent factor 1), three questions for intrinsic job motivation (latent factor 2), and four questions for extrinsic job motivation (latent factor 3).

Table 7

Model-based composite reliabilities for decremental and incremental scales with seven points

				95% C.I.			
Scale direction	Coefficients	Differences	p-value	lower	upper	RMSEA	CFI
Achievement motivation							
Decremental	0.836	0.006	0.630	0.818	0.855	0.042	0.989
Incremental	0.830			0.812	0.849		
Intrinsic job motivation							
Decremental	0.873	−0.014	0.187	0.857	0.889	0.000	1
Incremental	0.886			0.873	0.900		
Extrinsic job motivation							
Decremental	0.825	0.002	0.890	0.799	0.850	0.041	0.985
Incremental	0.822			0.798	0.847		

Five questions for achievement motivation (latent factor 1), three questions for intrinsic job motivation (latent factor 2), and four questions for extrinsic job motivation (latent factor 3).

decremental and incremental scales. As suggested by Cox (1980), a higher number of answer options can increase discrepancies between the true (latent) and the observed scores. This potentially introduces systematic measurement error in the form of scale direction effects that are caused by superficial answer strategies when processing decremental and incremental scales with seven points. We see survey-satisficing responsible for the lack of scalar invariance. Importantly, the absence of measurement invariance precludes the comparison of (latent) means between the differently directed seven-point scales.

With respect to our third research hypothesis we investigated latent mean differences. Since scalar invariance is a substantial prerequisite for comparing latent means (Steinmetz, 2013) we only compared the latent means for scales with five points. In line with the results on the observational level, we did not find significant shifts in latent means. This similarly applies to the questions on achievement motivation as well as to the questions on intrinsic and extrinsic job motivation. This additionally indicates the robustness of five-point scales against scale direction (or primacy) effects.

In order to investigate our fourth research hypothesis on data quality, we computed composite reliabilities of the multi-item instruments achievement motivation, intrinsic job motivation, and extrinsic job motivation. We found one single significant difference between decremental and incremental scales with five points. This indicates that—depending on the question topic—scale direction does not affect data quality in terms of reliability. For seven-point scales, composite reliabilities were systematically higher than for five-point scales. The explanation for this finding can be twofold. First, seven-point scales indeed result in higher reliability than their five-point counterparts. Second, compared to five-point scales, reliability of seven-point scales might be artificially inflated by survey-satisficing in terms of primacy effects. As shown by Knowles and Condon (1999), measurement error in the form of survey-satisficing can increase reliability. Overall, it seems that scale direction does not impact reliability, but scale length does.

In this study, we followed the survey-satisficing theory (Krosnick, 1991) and explained our empirical findings by drawing conclusions from respondents' answers. However, in order to draw more robust conclusions about the mechanisms underlying scale direction effects across five- and seven-point decremental and incremental scales it might be worthwhile to employ more direct measures. For example, it is possible to ask respondents to evaluate the difficulty of rating scales varying in length. Another way would be the inclusion of eye-tracking methodology. Eye-tracking measures in the form of fixation count and time may provide valuable insights on scale processing and task difficulty (Galesic et al., 2008; Galesic & Yan, 2011; Höhne & Lenzner, 2015).

This study has some limitations that provide avenues for future research. First, we only investigated scale direction effects in end-labeled, vertically aligned five- and seven-point scales with no numeric values. However, there are numerous design aspects that can be varied when investigating scale direction effects. In relation to this point, it would be interesting to compare completely labeled scales that potentially affect the processing of survey questions with decremental and incremental scales. Second, we conducted our study in one country (Germany). It remains unclear whether our findings hold in a cross-national or cross-cultural comparison because linguistic differences may also have an impact on the measurement properties and data quality of decremental and incremental scales. We therefore suggest that future research goes a step further and compares the rating scales in a cross-national or cross-cultural setting. Third, we only employed survey questions on achievement and intrinsic and extrinsic job motivation. Future research could employ survey questions on a variety of topics, such as income (in)equality or political efficacy, to provide further evidence for the effects of scale direction and scale length on answer behavior. Finally, in this study, we focused on data quality in terms of reliability. However, it would be worthwhile to additionally investigate validity. For this purpose, future studies could include measures (or questions) that are theoretically relevant to the experimentally manipulated target questions and estimate criterion validity (see Höhne & Yan, 2020; Yeager & Krosnick, 2012). This would allow to draw more robust conclusions about data quality of rating scales differing in terms of direction and length.

Our findings revealed that rating scales with five points are less prone to scale direction effects than rating scales with seven points. This is supported by the analyses on the observational and latent level. The comparability of answer distributions of differently directed seven-point scales is limited. The higher reliability associated with seven-point scales may be a methodological artefact and needs further investigation. For now, we recommend that survey researchers and practitioners go with five-point scales when measuring achievement and job motivation. This particularly applies when comparability is of main interest for the purposes of the study.

References

Anand, S. (2008). Satisficing. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research* (pp. 797–799). Sage.

Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 "Political Economy Of Reforms", Universität Mannheim. (2020). German Internet Panel, Welle 42 (Juli 2019) [GESIS Data Archive, Cologne (Germany)]. https://doi.org/10.4232/1.13465

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822x15574494

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference Chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 372–398. https://doi.org/10.1080/10705511.2012.687671

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*(4), 872–882. https://pubmed.ncbi.nlm.nih.gov/18940097/

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge. https://doi.org/10.4324/9780203807644

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement*, *51*(3), 531–540. https://doi.org/10.1177/0013164491513002

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/s15328007sem0902_5

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*(4), 407. https://doi.org/10.2307/3150495

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*, *52*(4), 1523–1559. https://doi.org/10.1007/s11135-017-0533-4

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, *72*(5), 892–913. https://doi.org/10.1093/poq/nfn059

Galesic, M., & Yan, T. (2011). Use of eye tracking for studying survey response processes. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 349–370). Routledge. https://doi.org/10.4324/9780203844922

Höhne, J. K., & Lenzner, T. (2015). Investigating response order effects in web surveys using eye tracking. *Psihologija*, *48*(4), 361–377. https://doi.org/10.2298/psi1504361h

Höhne, J. K., Revilla, M., & Lenzner, T. (2018). Comparing the performance of agree/disagree and item-specific questions across PCs and smartphones. *Methodology*, *14*(3), 109–118. https://doi.org/10.1027/1614-2241/a000151

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*(3), 347–353. https://doi.org/10.1080/13645579.2019.1696087

Keusch, F., & Yan, T. (2018). Is satisficing responsible for response order effects in rating scale question? *Survey Research Methods*, *12*(3), 259–270. https://doi.org/10.18148/SRM/2018.V12I3.7263

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379–386. https://doi.org/10.1037/0022-3514.77.2.379

Krebs, D. (2012). The impact of response format on attitude measurement. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, theories, and empirical applications in the social sciences* (pp. 105–113). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-18898-0_14

Krebs, D., & Bachner, Y. G. (2018). Effects of rating scale direction under the condition of different reading direction. *Methods, Data, Analyses*, *12*(1), 105–126. https://doi.org/10.12758/MDA.2017.08

Krebs, D., & Hoffmeyer-Zlotnik, J. H. P. (2010). Positive first or negative first? Effects of the order of answering categories on response behavior. *Methodology*, *6*(3), 118–127. https://doi.org/10.1027/1614-2241/a000013

Krebs, D., & Höhne, J. K. (2020). Antwortskalenrichtung und Umfragemodus. In A. Mays, A. Dingelstedt, V. Hambauer, S. Schlosser, F. Berens, J. Leibold, & J. K. Höhne (Eds.), *Grundlagen—Methoden—Anwendungen in den Sozialwissenschaften* (pp. 231–246). Springer VS. https://doi.org/10.1007/978-3-658-15629-9_12

Krebs, D., & Höhne, J. K. (2021). Exploring scale direction effects and response behavior across PC and smartphone surveys. *Journal of Survey Statistics and Methodology*, *9*(3), 477–495. https://doi.org/10.1093/jssam/smz058

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219. https://doi.org/10.1086/269029

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29–44.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd, pp. 63–313). Emerald.

Liu, M. (2017). Labelling and direction of slider questions: Results from web survey experiments. *International Journal of Market Research*, *59*(5), 601–624.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*(6), 725–743. https://doi.org/10.1177/0894439313485201

Menold, N., & Bogner, K. (2015). Gestaltung von Ratingskalen in Fragebögen [SDM-Survey Guidelines (GESIS—Leibniz Institute for the Social Sciences)].

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/bf02294825

Millsap, R. E. (2007). Invariance in measurement and prediction evisited. *Psychometrika*, *72*(4), 461–473. https://doi.org/10.1007/s11336-007-9039-7

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. 262–282). VEB Deutscher Verlag der Wissenschaften. https://doi.org/10.1016/s0166-4115(08)62067-1

Rammstedt, B., & Krebs, D. (2007). Does response scale format affect the answering of personality scales? *European Journal of Psychological Assessment*, *23*(1), 32–38. https://doi.org/10.1027/1015-5759.23.1.32

Raykov, T., & Marcoulides, G. A. (2015). Scale reliability evaluation with heterogeneous populations. *Educational and Psychological Measurement*, *75*(5), 875–892. https://doi.org/10.1177/0013164414558587

Rhemtulla, M., Brosseau-Liard, P. É., & 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*(3), 354–373. https://doi.org/10.1037/a0029315

Salzberger, T., & Koller, M. (2013). Towards a new paradigm of measurement in marketing. *Journal of Business Research*, *66*(9), 1307–1317.

Schaeffer, N. C., & Dykema, J. (2020). Advances in the science of asking questions. *Annual Review of Sociology*, *46*(1), 37–60. https://doi.org/10.1146/annurev-soc-121919-054544

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, *29*(1), 65–88. https://doi.org/10.1146/annurev.soc.29.110702.110112

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology*, *9*(1), 1–12. https://doi.org/10.1027/1614-2241/a000049

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

Toepoel, V. (2008). *A closer look at web questionnaire design* (Doctoral dissertation). Tilburg University (CentER, Center for Economic Research). https://research.tilburguniversity.edu/en/publications/a-closer-look-at-web-questionnaire-design

Tourangeau, R., Maitland, A., Rivero, G., Sun, H., Williams, D., & Yan, T. (2017). Web surveys by smartphone and tablets. *Public Opinion Quarterly*, *81*(4), 896–929. https://doi.org/10.1093/poq/nfx035

Weng, L.-J., & Cheng, C.-P. (2000). Effects of response order on Likert-type scales. *Educational and Psychological Measurement*, *60*(6), 908–924. https://doi.org/10.1177/00131640021970989

Yan, T., & Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public Opinion Quarterly*, *79*(1), 145–165. https://doi.org/10.1093/poq/nfu062

Yan, T., Keusch, F., & He, L. (2018). The impact of question and scale characteristics on scale direction effects. *Survey Practice*, *11*(2), 1–10. https://doi.org/10.29115/sp-2018-0008

Yeager, D. S., & Krosnick, J. A. (2012). Does mentioning "some people" and "other people" in an opinion question improve measurement quality? *Public Opinion Quarterly*, *76*(1), 131–141. https://doi.org/10.1093/poq/nfr066

(see Appendix on the next page)

Appendix

English translations of the survey questions on achievement motivation as well as intrinsic and extrinsic job motivation.

Questions on achievement motivation:
I like being in competition with other people.
It is satisfying when I achieve better results than other people.
I am always trying to perform better than other people.
I try harder when I am in competition with other people.
It is important for me to be the best at a task.
Answer options (decremental): "applies completely" to "applies not at all"
Answer options (incremental): "applies not at all" to "applies completely"

Questions on intrinsic job motivation:
A job that allows to make use of my skills and talents is important for me.
A job where I have responsibilities for specific tasks is important for me.
A job that allows me to develop my own ideas is important for me.
Answer options (decremental): "applies completely" to "applies not at all"
Answer options (incremental): "applies not at all" to "applies completely"

Questions on extrinsic job motivation:
A job with a high income is important for me.
A job with good promotion prospects is important for me.
A job with clear career perspectives is important for me.
A job that I can work autonomously on is important for me. (according to the results of the confirmatory factor analysis)
Answer options (decremental): "applies completely" to "applies not at all"
Answer options (incremental): "applies not at all" to "applies completely"

Note. The survey questions were adopted from the "Cross Cultural Survey for Work and Gender Attitudes" (2010). All questions were presented individually (i.e., one question per online survey page). The rating scales were end-labeled and vertically aligned without numeric labels. The original German wordings of the questions including answer options are available in the online questionnaire of wave 42 (July 2019) of the German Internet Panel.