Survey Research Methods (2025)
Vol. 19, No. 2, pp. 175-186
https://doi.org/10.18148/srm/2025.v19i2.8384
European Survey Research Association

The Impact of Scale Direction on Data Quality

Ting Yan[1] · Alexandru Cernat[2] · Florian Keusch[3]
[1]University of Chicago, NORC
[2]University of Manchester
[3]University of Mannheim

Survey literature has documented a tendency for respondents to select scale options closer to the start of a scale, resulting in scale direction effects. This paper builds upon the literature and expands it in two important ways. First, we experimentally manipulated scale length, scale labeling, and scale type, and fully crossed them with scale direction. This experimental setup allows us to examine the moderating effects of other scale features on scale direction effects. Second, previous literature largely examined the impact of scale direction on distributional properties of answers. In this paper, we conducted a systematic assessment of reliability and validity of answers as well as measurement equivalence by scale direction, allowing us to gain a deeper understanding and a more complete picture of the impact of scale direction on resultant answers. The findings will have important implications for question writers and will provide practical guidance on the direction of scales.

Keywords: scale direction effects; reliability; validity; measurement equivalence; data quality

1 Introduction

Scale direction effects refer to respondents' tendency to select scale points closer to the beginning of a response scale, holding other scale features constant (Yan & Keusch, 2015). Unlike response order effects observed for unordered response categories, which are manifested as primacy effects when the communication channel is visual (e.g., a web or paper survey) and recency effects when the response option list is read to respondents (e.g., a telephone survey), scale direction effects are not contingent upon the mode of data collection. Scale direction is empirically shown to produce primacy effects across a variety of modes of data collection such as telephone surveys (e.g., Yan & Keusch, 2015), face-to-face surveys (e.g., Carp, 1974), web surveys (e.g., Garbarski, Schaeffer, & Dykema, 2015, 2019; Höhne, Krebs, & Kühnel, 2023; Keusch & Yan, 2018; Tourangeau et al., 2017), mobile web surveys (e.g., Keusch & Yan, 2017; Tourangeau et al., 2017), and paper questionnaires (e.g., Höhne & Krebs, 2018; Israel, 2006).

Two mechanisms have been proposed to account for scale direction effects. According to the satisficing account (Krosnick, 1991, 1999), respondents unwilling or unable to exert the cognitive effort required to provide an optimal response satisfice by taking a cognitive short-cut and selecting the first acceptable or satisfactory scale point, leading to scale direction effects. Researchers resorting to the satisficing notion to explain scale direction effects essentially treat scale direction effects as a special case of primacy effects (e.g., Krosnick & Presser, 2010). The second possible mechanism is the anchoring-and-adjustment heuristic proposed by Tversky and Kahneman (1974). According to this account, respondents who are provided a response scale use the beginning scale point as an anchor and then make adjustments to that anchor until they reach a satisfactory scale point. Yan and Keusch (2015) demonstrated empirically that the anchoring-and-adjustment heuristic is at work for scale direction effects. Because both mechanisms predict primacy effects, it is hard to pinpoint which mechanism is at work under what circumstances. Conceptually speaking, the two mechanisms differ with respect to moderators of the scale direction effects. The satisficing account argues for a stronger scale direction effect for difficult tasks and

Supplementary Information The online version of this article (https://doi.org/10.18148/srm/2025.v19i2.8384) contains supplementary information.

Corresponding author: Ting Yan, University of Chicago, NORC, Chicago, Illinois, USA (Email: tyanuconn@gmail.com)

among respondents with limited cognitive capacity and decreased motivation (Krosnick, 1991, 1999). However, empirical research showed that satisficing cannot entirely account for scale direction effects as these tend to be observed across the board among respondents who were at a high risk of satisficing and those who were not (e.g., Keusch & Yan, 2018) and under conditions that were conducive to satisficing and conditions that were not (e.g., Mingay & Greenwell, 1989; Carp, 1974). By contrast, conditions facilitating the use of anchoring-and-adjustment heuristic include relevant anchors (Mussweiler & Strack, 1999), plausible anchors (Wegener & Petty, 1995), respondents with a high need for cognition (Epley & Gilovich, 2006), respondents with high agreeableness (Eroglu & Croxton, 2010), respondents without required knowledge (Wilson et al., 1996), and respondents attentive to the anchor (Wilson et al., 1996). So far, only Yan and Keusch (2015) demonstrated empirically a stronger scale direction effect among respondents without the necessary knowledge, consistent with the predictions of the anchoring-and-adjustment process.

There are a few additional gaps in the literature. First, most of the research on scale direction effects focused on rating scales with attitudinal items. Example rating scales examined in earlier research include agreement scales (e.g., Höhne & Krebs, 2018; Leon et al., 2022; Yan & Keusch, 2018), satisfaction scales (e.g., Smyth et al., 2019), and evaluative scales (e.g., Garbarski et al., 2015, 2019). By contrast, only three studies investigated scale direction effects for questions using frequency scales (Carp, 1974; Keusch & Yan, 2019; Tourangeau et al., 2017). Carp (1974) examined 10 questions on frequency of trips using an 8-point fully labeled frequency scale but failed to find evidence indicating that scale direction affected answers to these behavioral questions. Keusch and Yan (2019) varied scale direction, scale alignment, and verbal labeling of a 5-point fully labeled unipolar frequency scale to 10 survey items. Again, they did not find any significant effect of scale direction on resultant answers. Tourangeau and colleagues (2017) assessed scale direction effects to six behavior items using frequency scales and found significant scale direction effects only when the frequency scale had seven scale points but not when a 5-point scale was used. To advance scale direction literature, this paper examines the impact of scale direction on answers to questions that measure the same constructs but use either agreement or frequency scales.

Second, most studies on scale direction effects did not attempt to take into consideration potential confounding or moderating impact of other characteristics of the question and the scale. They tend to use scales as they are, without experimentally manipulating other features of the question and the scale. As a result, scale direction effects are observed in existing empirical research on both end-labeled scales (e.g., Keusch & Yan, 2018) and fully-labeled scales (e.g., Leon et al., 2022), on 5-point scales (e.g., Garbarski et al., 2015, 2019) and longer scales (Krebs and Hoffmeyer-Zlotnik, 2010), on scales vertically aligned (e.g., Christian et al., 2009; Höhne and Lenzner, 2015) and horizontally aligned (e.g., Keusch & Yan, 2018), and on bipolar scales (e.g., Hofmans et al., 2007) and unipolar scales (e.g., Höhne et al., 2023). But it is not clear what conditions are more prone to scale direction effects.

A secondary data analysis demonstrates stronger scale direction effects for longer scales, questions with both subjective and behavioral components, and survey items appearing earlier in a questionnaire (Yan et al., 2018). In addition, the moderating impact of question type, question location, and scale length on scale direction effects is more pronounced for items administered via Computer-Assisted Personal Interviewing (CAPI) than in self-administration.

Two studies experimentally varied scale length in addition to scale direction (Höhne et al., 2023; Tourangeau et al., 2017). Both found scale direction effects for 7-point scales only, but not for 5-point scales. Three studies experimentally varied scale alignment but failed to find significant interaction effects between scale direction and scale alignment (Tourangeau et al., 2017; Keusch and Yan, 2019; Garbarski, Schaeffer, & Dykema, 2019). Tourangeau and colleagues (2017) also varied scale labeling so that half of respondents received fully-labeled scales and the other half end-labeled scales. They did not find a significant interaction between scale direction and scale labeling. Keusch and Yan (2019) manipulated verbal labels of frequency scales; frequency scales were labeled with quantifiers only (e.g., never, a little of the time, some of the time, most of the time, all of the time), precise frequency labels (e.g., zero days, one or two days, three or four days, five or six days, seven days), or a combination of both (e.g., never [zero days], a little of the time [one or two days], some of the time [three or four days], most of the time [five or six days], all of the time [seven days]). Although verbal labeling had a significant main effect on resultant answers, it did not interact with scale direction.

This paper reports findings from an experiment fully crossing scale direction with three other scale features: scale length (5- or 7-point), scale labeling (end-labeled or fully-labeled), and scale type (agreement or frequency). This experiment allows us to clearly tease apart the moderating effects of scale features on scale direction effects.

Third, most studies (especially the earlier ones) examined scale direction effects in terms of respondents' selection of response options closer to the beginning of a scale (e.g., Israel, 2006; Tourangeau et al., 2017) and means (e.g., Yan & Keusch, 2015; Garbarski et al., 2015). Only a few studies used latent variable models in their analysis of scale direction effects. Three studies showed measurement invari-

ance by scale direction (Höhne et al., 2018; 2021; Krebs and Hoffmeyer-Zlotnik, 2010). However, Höhne and colleagues (2023) found that measurement invariance was achieved for 5-point scales but not for 7-point scales. Shifts in latent means were found for agreement scales (Höhne & Krebs, 2018) and 7-point scales (Höhne et al., 2023), but were not found in the Höhne & Krebs, 2018 study, which used 7-point end-labeled scales. Liu and Keusch (2017) showed that the latent content factor did not differ by scale direction. Two studies examined reliability or validity. Höhne and colleagues (2023) found that the composite reliability did not differ for 7-point scales, but was higher for the 5-point scale running from a high (or positive) end to a low (or negative) end than for the 5-point scale progressing from a low (or negative) end to a high (or positive) end. Saris and Gallhofer (2007) showed that reliability and validity estimates did not significantly differ by scale direction in a meta-analysis of MultiTrait-MultiMethod (MTMM) experiments. In terms of indirect indicators of data quality, Liu and Keusch (2017) found that scale direction affected acquiescence but Yan and Keusch (2018) did not find evidence that scale direction affected acquiescence, mid-point response style, straightlining, and internal consistency in four surveys conducted face-to-face and online.

In this paper, we conducted a comprehensive analysis of scale direction effects by examining the impact of scale direction on means, validity, reliability, and other indicators of data quality including acquiescence, straightlining, extreme answers, and midpoint answers.

This paper uses data from two waves of a web survey collected about a month apart. In the first wave of the web survey, a 2 * 2 * 2 * 2 experiment was implemented on a set of 15 survey questions that fully crosses scale direction (ascending direction progressing from the low/negative end to the high/positive end vs. descending from the high/positive end to the low/negative end), scale length (5-point vs. 7-point), scale labeling (fully labeled vs. end labeled), and scale type (agreement scale vs. frequency scale). In the second wave, half of respondents were randomly assigned to receive the same scale direction assignment as in the first wave whereas the other half received a different scale direction. Taking advantage of the between-subject and within-subject design, we answer three research questions:

RQ1 How does scale direction affect the distribution of answers in terms of means? How do other scale features (scale length, scale labeling, scale type, using same direction in both waves) moderate the impact of scale direction on means?

Both satisficing and anchoring-and-adjustment predict a primacy effect. As a result, we hypothesize smaller means for ascending scales than for descending scales since ascending scales start with a low/negative end and descending scales begin from a high/positive end.

Based on the existing empirical literature, we expect a stronger scale direction effect for 7-point scales, fully-labeled scales, and attitudinal scales than for 5-point scales, end-labeled scales, and frequency scales.

RQ2 How does scale direction affect reliability, validity, and measurement invariance?

Based on the current literature, we don't expect scale direction to affect reliability, validity, and measurement variance under the satisficing account.

RQ3 How does scale direction affect proxy indicators of data quality in terms of straightlining, acquiescence, extreme responses, and midpoint answers? How do other scale features (scale length, scale labeling, scale type, using same direction in both waves) moderate the impact of scale direction on data quality?

We examine these quality indicators to help identify a combination of scale features that yields the best data quality. As a result, we generally do not have a priori hypothesis on the impact of scale direction on straightlining, acquiescence, extreme responses, and midpoint answers. We also do not have a priori hypothesis on the moderating impact of other scale features on the impact of scale direction on these proxy indicators of data quality. However, we expect a higher prevalence of acquiescing answers for descending agreement scales starting with agreement options because acquiescing answers are conflated with scale direction effects.

2 Data and Method

This paper uses data from two web surveys conducted as part of the LISS (Longitudinal Internet Studies for the Social Sciences) panel administered by CentERdata (Tilburg University, the Netherlands). The LISS panel builds on a probability sample of households drawn from the Dutch population register and aims to represent the Dutch-speaking population permanently residing in the Netherlands. Recruited households without Internet access are provided with a computer and Internet connection. LISS panel members are invited to complete online questionnaires every month (see Scherpenzeel & Das (2010) and Scherpenzeel (2011) for more details on LISS). For the purpose of this study, the same set of 15 survey items were asked twice about one month apart in the LISS panel. Six items ask about mindfulness and nine about political efficacy. The

exact question wording and response options are displayed in the Supplementary Materials (Appendix A).

The first wave of data was collected in September 2014, achieving a total of 3007 completes at a response rate of 83% using AAPOR RR1 formula. Wave 2 was conducted in October 2014 with a total of 2740 completes and a response rate of 75% again using AAPOR RR1. Both response rates do not take into consideration panel recruitment response rates. In addition, panelists were allowed to use their own device to complete the two web surveys. Only about 5% of respondents completed the web survey on a mobile device. We pooled data across devices for the analysis but controlled for device use in regression models.

We experimentally manipulated scale direction so that half of the respondents received scales in an ascending order, that is, the scale begins with the low (i.e., never) or the negative end (i.e., totally disagree) and progresses to the high (i.e., always) or the positive end (i.e., totally agree), and the other half scales in a descending order (i.e., starting with the high or positive end). The 15 target items either used a bipolar agreement scale (ranging either from totally agree to totally disagree or from totally disagree to totally agree) or a unipolar frequency scale (from never to always or from always to never). Scales had either five or seven points. Either all scale points had a verbal label (fully-labeled conditions) or only the two endpoints had a verbal label (end-labeled conditions). All items were presented individually, that is, one item per screen, with response options shown in a vertical line.

All four experimental factors are fully crossed and respondents were randomly assigned to one of the sixteen cells at Wave 1 and Wave 2 separately. The assignment of scale type, scale length, and scale labeling was kept the same across waves. However, for scale direction assignment, a random half of respondents were given the same scale direction across waves, but the other half received scales of different directions (for instance, if they were assigned scales in ascending order at Wave 1, they would get scales in descending order at Wave 2). All survey data collected as part of the experiment is available at https://doi.org/10.17026/dans-z3f-jc65.

To answer RQ1, we re-scaled answers to all questions on a 0–1 scale where 1 represents more agreement or higher frequency. We then compared the **means** of the re-scaled answers to every item by scale direction. We restructured the data to the long format and ran a multilevel model to test the impact of scale direction on the rescaled means. As shown in the formula below, we accounted for the nested structure of the data by allowing for a random intercept at the individual level. We also estimated separate multilevel models testing interactions between the scale direction and each of the three moderating factors (scale type, scale length, and scale labeling).

The formula used is (Snijders & Bosker, 2012):

$$Y_{i,j} = \gamma_0 + \Sigma\gamma_h x_{hij} + u_{0j} + \varepsilon_{ij}$$

Where $Y_{i,j}$ is the rescaled response to question i by individual j, γ_0 is the intercept of the regression, and $\Sigma\gamma_h x_{hij}$ are a set of h predictors of the outcomes (e.g., experimental group assignment). The variance is decomposed in the individual level variation (u_{0j}) and the residual (ε_{ij}).

To address RQ2 on **reliability** by scale direction, we utilized the within-person design of the study and estimated reliability as the correlation between Wave 1 and Wave 2 measures for each question. We ran the correlations separately by scale direction as well as by scale direction and four moderator variables: scale type, scale length, scale labeling, and an indicator of whether scale direction in Wave 2 was changed from Wave 1. To test whether the difference is significant we first estimate reliability separately for each combination of scale direction and the three moderators (scale labeling, scale length and scale type). We then pooled all reliability estimates into a dataset consisting of 240 reliability estimates and used t-test to compare reliability by scale direction.

To address RQ2 on **validity** by scale direction, we developed a MultiTrait-MultiError model (Cernat & Oberski, 2019, 2022, 2023) in which scale direction is considered as a potential source of systematic bias. Due to the within-subject experimental setup, two measurements are available from each individual. Furthermore, a random half of respondents received a different scale direction at Wave 2. Consequently, we can estimate the following model in the Structural Equation Modeling framework (Bollen, 1989):

$$y^*_{td} = \lambda^{(T)^*}_{td} T_t + \lambda^{(D)^*}_{td} D + \varepsilon_{td}$$

where y^*_{td} is the observed variable measuring a particular trait or topic, t, using a particular scale direction, d. We decompose the observed variance into three sources of variation: T, measuring the trait variance, D, measuring the scale direction variance, and an item specific random error, ε_{td}. The trait variance represents the valid source of variation that measures the concept of interest. The direction variance is systematic measurement error as it represents consistent answering patterns due to the format of the response scale and not the content. The random error represents noise in the data that can bias confidence intervals and multivariate analyses. Fig. 1 visually represents the model for the six mindfulness items as an illustration.

To address RQ2 on **measurement equivalence** by scale direction, we ran a Confirmatory Factor Analysis (CFA) in a sequence of models on Wave 1 and Wave 2 data separately:

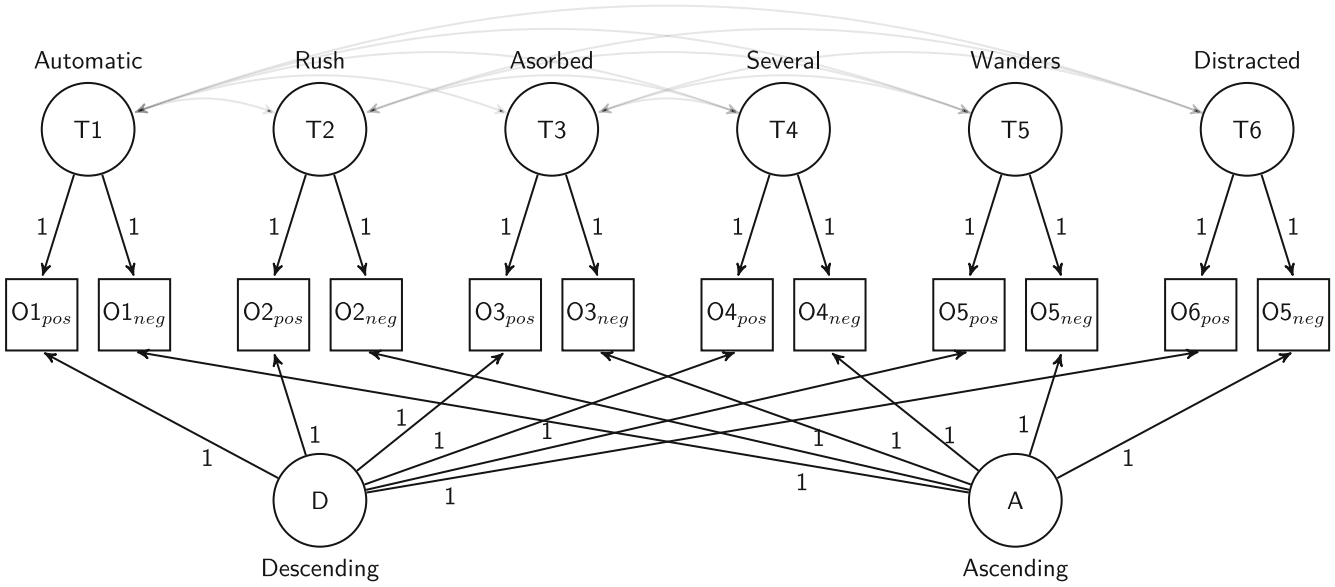

Fig. 1

MTME model for the mindfulness scale, where circles represent latent variables while squares are observed variables. The T latent variables measure the concept of interest while D and A are systematic variance due to the scale direction. Residuals, estimates of random error, are not shown for ease of reading

1. a model without any constraints across groups (known as configural model),
2. adding constraints that all the loadings to be equal across groups (known as the metric model),
3. adding constraints that all the intercepts are equal across groups (known as the scalar model), and
4. adding constraints that the means of the latent variables are equal across groups.

This sequence of models allows us to identify potential causes for differences across groups. If the final model is not significantly worse we can conclude that the measurement model is the same regardless of the scale direction. We used the Comparative Fit Index (CFI) to assess whether or not restrictions make the models significantly worse. A decrease of more than 0.01 in CFI was considered an indicator of decrease of fit (Chen, 2007).

To answer RQ3, we calculated a number of proxy data quality indicators at the respondent level and investigated how they are affected by scale direction. **Acquiescence** was evaluated on questions using agreement scales and was calculated as the percentage of times respondents selected "totally agree." For **straightlining** we created a binary indicator for each respondent to indicate whether or not they provided the same answer to either battery of items. **Extreme response style** was calculated as the proportion of times one of the most extreme categories was chosen. **Middle response style** was calculated as the proportion of times the middle category was chosen.

To test whether differences due to scale direction are statistically significant, we ran OLS regression models (or logistic regression models if the outcome was dichotomous) with the experimental factors as predictors in addition to conducting t-tests and chi squared tests. We then specified separate models testing interactions between scale direction and each of the three moderator variables. We ran all the models for Wave 1 data and Wave 2 data separately. For models on Wave 2 data, we also included, as another possible moderator variable, whether respondents received the same or different scale direction. All regression models used to address RQ3 are at the respondent level, and are not multilevel models.

All data was cleaned and analyzed in R 4.3.2. Respondent-level regression models were run using the lme4 package (Bates et al., 2015). The multilevel models were run using the lme4 package (Bates et al., 2015) while the equivalence testing and the MTME were estimated using lavaan (Rosseel, 2012).

3 Results

3.1 RQ1. Scale Direction Effects On Means

We first examined the impact of scale direction on the means to answer RQ1. The averages of answers to survey items are displayed in Table A1 in the Appendix separately for each scale direction. The trend is that a descending scale starting with the high/positive end tends to elicit higher means (more positive attitudes or higher frequency) than an ascending scale starting with the low/negative end, regardless of whether questions were worded positively or negatively. The multilevel model fit on the long dataset shows that this difference by scale direction is statistically significant (Table A2 in the Appendix), demonstrating the presence of scale direction effects despite the small overall difference (descending = 0.567 vs. ascending = 0.551). (We reran the multilevel model including an indicator for question battery to account for the nesting of individual items within question battery. Conclusions remain the same.)

The moderation effects of scale length, scale type, and scale labeling are plotted in Fig. 2, which shows the predicted values from the multilevel model (Table A2 in the Appendix). Scale direction effects are moderated by scale length. In particular, a significant scale direction effect was observed for 7-point scales but not for 5-point scales (see Table A2 in the Appendix). Neither scale labeling nor scale type moderated the impact of scale direction on means.

The means of answers to the 15 survey items at Wave 2 by scale direction are displayed in Table S1 in the Supplemental Materials. Significant scale direction effects are also found in Wave 2 data (see Table S2 in the Supplemental Materials) but we did not find significant moderating effects of scale length, scale type, scale labeling, and whether or not the same scale direction was used in Wave 2 (Table S2 in the Supplemental Materials).

3.2 RQ2: Scale Direction Effects on Reliability

Fig. 3 presents the reliability of answers to each of the 15 survey items by scale direction. Scale direction does not seem to have a consistent impact on reliability. The overall difference in reliability by scale direction is small and not statistically significant ($p > 0.05$). The average reliability is 0.56 for the ascending order and 0.57 for the descending order (t = –0.58, $p = 0.55$). Scale length, labeling, and type did not moderate the impact of scale direction on reliability (see Table A3 in the Appendix). The interaction term between scale direction and whether or not scale direction was changed from Wave 1 to Wave 2 is also not statistically significant (Table A3 in the Appendix).

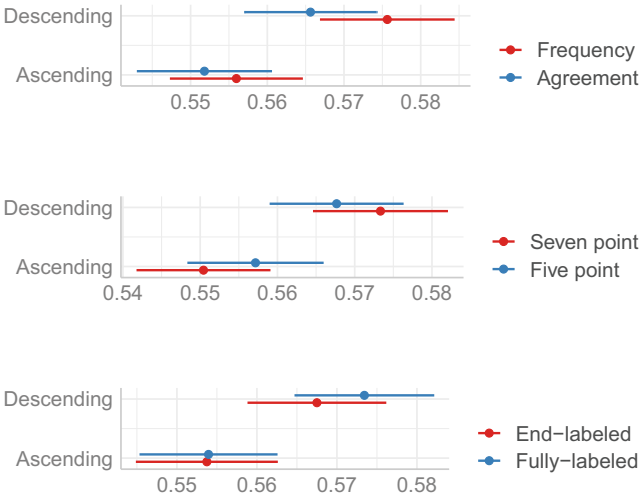

Fig. 2

Wave 1: Predicted means with 95% confidence intervals for scale direction and moderating factors based on multilevel models (Table A2 in the Appendix)

3.3 RQ2: Scale Direction Effects on Validity

We decomposed the variance of the responses to mindfulness questions using the MTME approach and plotted in Fig. 4 the estimated validity, the systematic variance due to the scale direction, and the random error. Overall, the validity for the mindfulness questions is relatively low, at around 50%. The systematic variation due to the scale direction is around 9% while the remaining variance is due to random error (Fig. 4). Validity did not seem to differ by scale direction though. We carried out variance decomposition for each of the six items measuring mindfulness. As shown in Figure A4 in Appendix, validity varies considerably with the lowest validity for the "absorbed" item, which is worded in the opposite direction to the rest of the mindfulness items. However, the effect of the scale direction seems relatively consistent across items.

By comparison, validity for the nine political efficacy questions is larger than 0.50, and the effect of scale direction is smaller (around 3%, Fig. 5). Validity differs by item but the effect of scale direction is rather consistent across the nine items on political efficacy (Figure A5 in Appendix).

3.4 RQ2: Scale Direction Effects on Measurement Equivalence

We next run the increasingly restrictive models for both sets of questions to investigate measurement equivalence across scale direction. The initial models had moderate to low fit with CFI values of 0.85 and 0.62 and RMSE val-

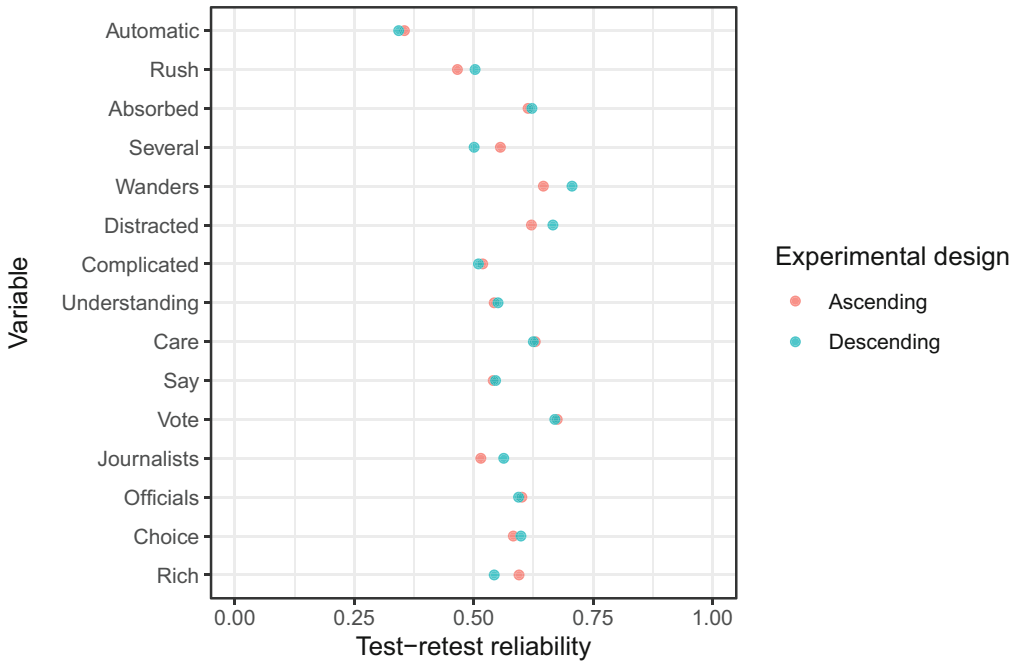

Fig. 3

Reliability by item and scale direction

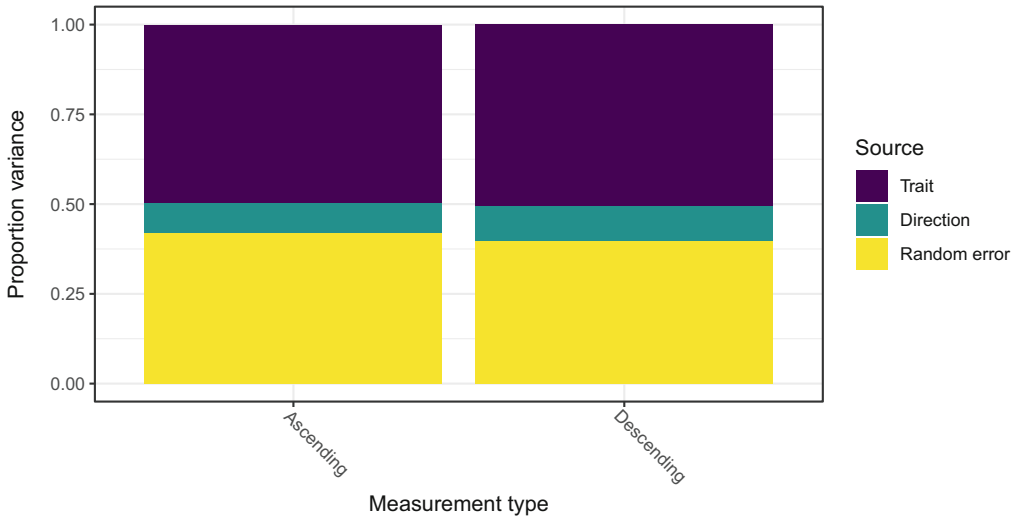

Fig. 4

Pooled variance decomposition based on the MTME model for mindfulness questions by scale direction

ues of 0.15 and 0.18. However, our focus here is on relative fit after invariance restrictions are added. As shown in Table 1, increasingly restrictive models are not significantly worse than the less restrictive models. For instance, the metric model restricting all factor loadings to be equal across scale direction is not significantly worse than the configural model with no restrictions, supporting metric invariance of mindfulness and political efficacy questions across scale direction. The scalar model is not significantly worse than the metric model and the final model not significantly worse than the scalar model, supporting equivalence of intercepts and means of latent variables across scale direction. There-

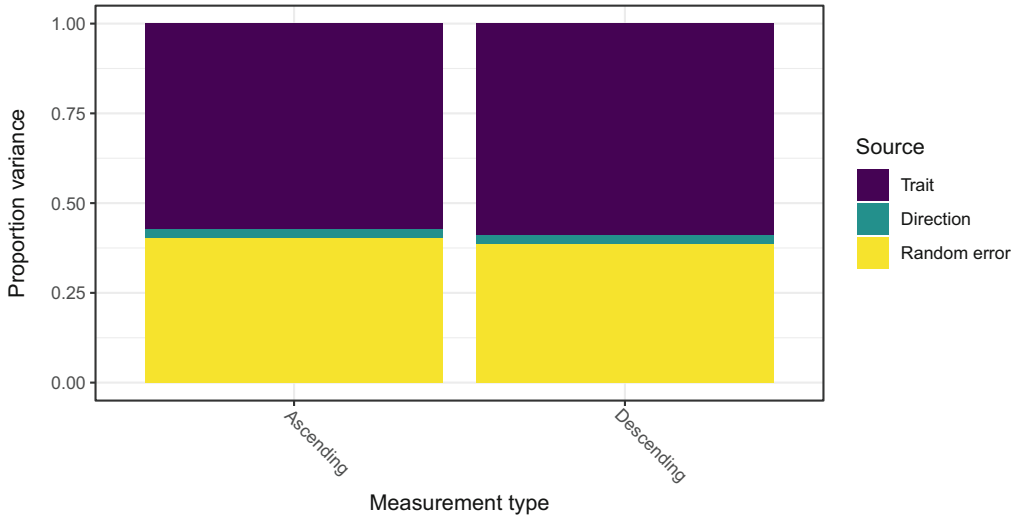

Fig. 5

Pooled variance decomposition based on the MTME model for political efficacy questions by scale direction

Table 1

Equivalence testing for Wave 1 data

Model	Chisq	Df	CFI	RMSEA	AIC	BIC
Mindfulness Questions						
Configural	661.93	18	0.85	0.15	–3906.1	–3689.8
Metric	661.93	18	0.85	0.15	–3906.1	–3689.8
Scalar	661.93	18	0.85	0.15	–3906.1	–3689.8
Means	661.93	18	0.85	0.15	–3906.1	–3689.8
Political Efficacy Questions						
Configural	2612.2	54	0.62	0.18	–5678.8	–5354.5
Metric	2612.2	54	0.62	0.18	–5678.8	–5354.5
Scalar	2612.2	54	0.62	0.18	–5678.8	–5354.5
Means	2612.2	54	0.62	0.18	–5678.8	–5354.5

fore, both mindfulness and political efficacy questions are found to be invariant across scale direction. We ran the same models on Wave 2 data and, again, found that the two scales are equivalent across scale direction (see Table S4 in Supplemental Materials).

3.5 RQ3: Scale Direction Effects on Proxy Quality Indicators

To examine scale direction effect on acquiescence, we examined the percentage of times respondents selected "totally agree" to questions using the agreement scale. We found that respondents reported "totally agree" to more items at Wave 1 when the agreement scale starts with it (that is, the descending order) than when the agreement scale is in the ascending order starting with "totally disagree" (3% vs. 12%, t = –1.81, df = 1474, p = 0.07, Fig. 6). Scale direction effect on acquiescence does not seem to be moderated by any of the other scale features (see table A6 in Appendix). At Wave 2 the differences in how often respondents selected "totally agree" by scale direction are not significant (see Figure S4 and Table S6 in Supplemental Materials).

To examine scale direction effect on straightlining, we examined the percent of respondents who provided the same answers to either questions on mindfulness or questions on political efficacy by scale direction. We did not find significant differences by scale direction in respondents' propensity to straightline in Wave 1 (5% for the descending order vs. 4% for the ascending order, X-squared = 0.14469, df = 1, p-value = 0.7037, Fig. 6). We also did not find evidence of moderating effects of other scale features (see table A7 in the Appendix). Wave 2 results are consistent with Wave 1 findings (Figure S5 in Supplemental Materials); scale direction did not influence straightlining in Wave 2 and no other scale feature moderated scale direction effect on straightlining (Table S7 in Supplemental Materials).

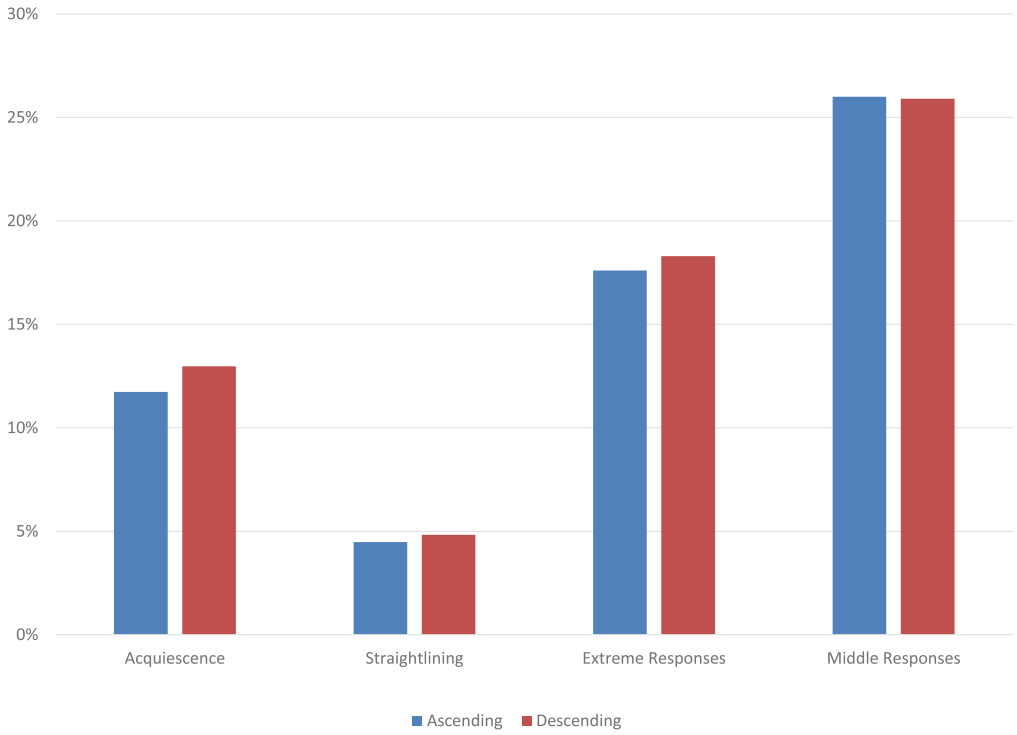

Fig. 6

Scale direction effects on proxy quality indicators at Wave 1

We found no significant differences in respondents' likelihood to select extreme responses (18% vs. 18%, t = –1.129, df = 2997.6, *p*-value = 0.259) and middle responses by scale direction at both Wave 1 (26% vs. 26%, t = 0.13981, df = 2985.3, *p*-value = 0.8888, Fig. 6) and Wave 2 (Figure S5 in Supplemental Materials). Furthermore, there was no moderating effects by any of the three scale features on extreme responses and middle responses (Tables A8 and A9 in Appendix for Wave 1 results, and tables S8 and S9 in Supplemental Materials for Wave 2 results).

4 Discussion

We conducted an experiment in two waves of a web survey collected one month apart in the LISS panel, a probability panel of the adult population in the Netherlands. The experiment fully crossed the manipulation of four scale features (scale direction, scale type, scale length, and scale labeling) on 15 items at both waves. In addition, at Wave 2, the experiment further varied whether or not respondents received scales of the same direction as Wave 1 or scales of the opposite direction from Wave 1. We then conducted comprehensive analyses to examine scale direction effects on multiple aspects of data quality and summarized results in Table 2.

Following the suit of the majority of earlier empirical research on scale direction effects, we examined the impact of scale direction on means of resultant answers. Consistent with the literature, we found a significant scale direction effect on means for both Wave 1 and Wave 2 data; means are larger for descending scales starting with the high/positive end than for ascending scales beginning with the low/negative end. For Wave 1 data, only one scale feature significantly moderated the impact of scale direction on means. Scale direction effect was found for seven-point scales but not for five-point scales consistent with the literature (e.g., Höhne et al., 2023; Tourangeau et al., 2017; Yan et al., 2018). The literature is mixed on whether scale direction affects answers to frequency scales. Two studies found no impact of scale direction for frequency scales (Carp, 1974; Keusch & Yan, 2019) and one study found a significant scale direction effect on 7-point frequency scales, but not 5-point scales (Tourangeau et al., 2017). We did not find a significant interaction effect of scale direction and scale type in this study.

We further found that scale direction had no significant impact on test-retest reliability and that none of the other scale features varied in our experiment moderated the effect of scale direction on reliability, different from the findings

Table 2

Summary of Findings

Outcomes Evaluated	Scale Direction Effects	Moderators
RQ1: Means	Yes	Wave 1: scale length Wave 2: no
RQ2: Reliability	No	–
RQ2: Validity	No	Not examined
RQ2: Measurement Equivalence	No	Not examined
RQ3: Acquiescence	No	No
RQ3: Straightlining	No	No
RQ3: Extreme Response Style	No	No
RQ3: Middle Response Style	No	No

on composite reliability (Höhne et al., 2023) but consistent with reliability estimates from MTMM models (Saris and Gallhofer, 2007). Furthermore, scale direction was found to have no impact on validity estimated from MTMM models, consistent with earlier research by Saris and Gallhofer (2007). We found evidence of measurement equivalence by scale direction, consistent with Höhne and colleagues (2023) and Liu and Keusch (2017).

Consistent with Yan and Keusch (2018), we found no evidence of scale direction affecting straightlining, extreme responses, and middle responses. However, scale direction did marginally affect the proportion of times respondents selected "strongly agree" at Wave 1, but not at Wave 2.

Our findings have important practical implications. Survey researchers and practitioners have been concerned about scale direction effects and searching for evidence-based practical guidelines on which scale direction to use (see Discussion in Yan and Keusch, 2015). Our experimental findings provide good news for researchers and practitioners who are concerned with reliability, validity, measurement equivalence, straightlining, extreme responses, and middle responses. However, if survey researchers and practitioners are to use means of resultant answers for classification and comparison purposes, they should decide on one scale direction and use that direction consistently throughout the questionnaire and across different waves of a panel study. At the same time, users should be mindful of scale direction when making comparisons cross surveys and/or waves.

The survey literature recommends researchers and practitioners to avoid using agreement scales and to use construct or item specific scales instead (Saris, Revilla, Krosnick, & Schaeffer, 2010). We did not find evidence supporting worse performance of the agreement scales than frequency scales. However, if an agreement scale has to be used, we suggest using an ascending order starting with disagree options to reduce inflated acquiescing answers due to scale direction effects.

We attempted to understand mechanisms accounting for observed scale direction effects but due to lack of informative moderators (such as perceived relevant of anchors) we could not draw conclusions on whether the observed scale direction effects were due to satisficing or the use of anchoring-and-adjustment heuristics. We interpreted the presence of a moderating impact of scale length as a piece of evidence supporting the satisficing notion as the mechanism accounting for scale direction effects. This is because longer scales are cognitively harder to process than shorter scales but the anchors (that is, the scale endpoints) are the same regardless of scale length. We interpret the absence of scale direction effects on validity, reliability, and measurement equivalence as a support for the anchoring-and-adjustment heuristics as the working mechanism because satisficing, by definition, induces data of lower quality. However, studies are needed that systematically vary the moderators described in the Introduction section in order to tease apart the two mechanisms. We recommend continuing the research on establishing mechanisms accounting for scale direction effects and on uncovering circumstances under which quality of answers differ by scale direction.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bollen, K. A. (1989). Structural equations with latent variables. John Wiley & Sons. https://doi.org/10.1002/9781118619179.

Carp, F.M. (1974). Position effects on interview response. *Journal of Gerontology*, *29*, 581–587.

Cernat, A., & Oberski, D.L. (2019). Extending the within-persons experimental design: the Multitrait-Multierror (MTME) approach. In P. Lavrakas, M. Traugott, C. Kennedy, A. Holbrook, E. de Leeuw & B. West (Eds.), *Experimental methods in survey research* (1st edn., pp. 481–500). New York: Wiley. https://doi.org/10.1002/9781119083771.ch24.

Cernat, A., & Oberski, D.L. (2022). Estimating stochastic survey response errors using the Multitrait-Multierror model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(1), 134–155. https://doi.org/10.1111/rssa.12733.

Cernat, A., & Oberski, D.L. (2023). Estimating measurement error in longitudinal data using the longitu-

dinal MultiTrait MultiError approach. *Structural Equation Modeling: A Multidisciplinary Journal.*, *30*(4), 592–603. https://doi.org/10.1080/10705511.2022.2145961.

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal.*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834.

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

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic. Why the adjustments are insufficient. *Psychological Science*, *17*, 311–318.

Eroglu, C., & Croxton, K.L. (2010). Biases in judgmental adjustments of statistical forecasts: the role of individual differences. *International Journal of Forecasting*, *26*, 116–133. https://doi.org/10.1016/j.ijforecast.2009.02.005.

Garbarski, D., Schaeffer, N.C., & Dykema, J. (2015). The effects of response option order and question order on self-rated health. *Qual Life Res*, *24*(6), 1443–1453. https://doi.org/10.1007/s11136-014-0861-y.

Garbarski, D., Schaeffer, N.C., & Dykema, J. (2019). The effects of features of survey measurement on self-rated health: response option order and scale orientation. *Applied Research in Quality of Life*, *14*(2), 545–560.

Hofmans, J., Theuns, P., Baekelandt, S., et al. (2007). Bias and change in perceived intensity of verbal qualifiers effected by scale orientation. *Survey Research Methods*, *1*, 97–108.

Höhne, J.K., & Krebs, D. (2018). Scale direction effects in agree/disagree and item-specific questions: a comparison of question formats. *International Journal of Social Research Methodology*, *21*(1), 91–103. https://doi.org/10.1080/13645579.2017.1325566.

Höhne, J.K., & Lenzner, T. (2015). Investigating response order effects in web surveys using eye tracking. *Psiholgija*, *48*, 361–377.

Höhne, J.K., Krebs, D., & Kühnel, S.M. (2023). Investigating direction effects in rating scales with five and seven points in a probability-based online panel. *Survey Research Methods*, *17*(2), 193–204. https://doi.org/10.18148/srm/2023.v17i2.8006.

Israel, G.D. (2006). *Visual cues and response format effects in mail surveys*. Revised version of the paper presented at the Annual Meeting of the Southern Rural Sociological Association, Orlando, FL (7 February).

Keusch, F., & Yan, T. (2017). Web versus mobile web: an experimental study of device effects and self-selection effects. *Social Science Computer Review*, *35*(6), 751–769. https://doi.org/10.1177/0894439316675566.

Keusch, F., & Yan, T. (2018). Is satisficing responsible for response order effects in rating scale questions? *Survey Research Methods*, *12*(3), 259–270. https://doi.org/10.18148/srm/2018.v12i3.7263.

Keusch, F., & Yan, T. (2019). Impact of response scale features on survey responses to factual/behavioral questions. In P.J. Lavrakas, M.W. Traugott, C. Kennedy, A. Holbrook, E. de Leeuw & B.T. West (Eds.), *Experimental methods in survey research: techniques that combine random sampling with random assignment* (pp. 131–150). Hoboken: John Wiley & Sons, Inc..

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

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

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537–567.

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 Group Publishing.

Leon, C.M., Aizpurua, E., & van der Valk, S. (2022). Agree or disagree: does it matter which comes first? An examination of scale direction effects in a multi-device online survey. *Field Methods*, *34*(2), 125–142. https://doi.org/10.1177/1525822X211012259.

Liu, M., & Keusch, F. (2017). Effects of scale direction on response style of ordinal rating scales. *Journal of Official Statistics*, *33*, 137–154.

Mingay, D.J., & Greenwell, M.T. (1989). Memory bias and response-order effects. *Journal of Official Statistics*, *5*, 253–263.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, *35*, 136–164.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02.

Saris, W.E., & Gallhofer, I.N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken: John Wiley & Sons, Inc.. https://doi.org/10.1002/9780470165195.

Saris, W.E., 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*(1), 61–79. https://doi.org/10.18148/srm/2010.v4i1.2682.

Scherpenzeel, A. (2011). Data collection in a probability-based internet panel: how the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *5661*, 109. https://doi.org/10.1177/0759106310387713.

Scherpenzeel, A., & Das, J. (2010). Ttrue longitudinal and probability-based internet panels—research portal. In J. Das, P. Ester & L. Kaczmirek (Eds.), *Social and behavioral research and the internet* (pp. 77–103). Boca Raton: Taylor & Francis.

Smyth, J.D., Israel, G.D., Newberry, M.G., & Hull, R.G. (2019). Effects of stem and response order on response patterns in satisfaction ratings. *Field Methods*, *31*(3), 260–276. https://doi.org/10.1177/1525822X19860648.

Snijders, T.A.B., & Bosker, R.J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling* (2nd edn.). London: SAGE.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*, 1124–1131.

Wegener, D.T., & Petty, R.E. (1995). Flexible correction processes in social judgment: the role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology*, *68*, 36–51.

Wilson, T.D., Houston, C., Etling, K.M., & Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology*, *125*, 387–402.

Yan, T., & Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public Opinion Quarterly*, *79*, 145–165.

Yan, T., & Keusch, F. (2018). *Direction of agree-disagree rating scales and data quality*. Paper presented at the Annual Conference of the American Association for Public Opinion Research.

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