Using Only Numeric Labels Instead of Verbal Labels: Stripping Rating Scales to Their Bare Minimum in Web Surveys

Gummer, Tobias; Kunz, Tanja

Article

Social Science Computer Review
2021, Vol. 39(5) 1003-1029

DOI: 10.1177/0894439320951765
journals.sagepub.com/home/ssc

Using Only Numeric Labels Instead of Verbal Labels: Stripping Rating Scales to Their Bare Minimum in Web Surveys

Tobias Gummer[1] and Tanja Kunz[1]

Abstract

With the increasing use of smartphones in web surveys, considerable efforts have been devoted to reduce the amount of screen space taken up by questions. An emerging stream of research in this area is aimed at optimizing the design elements of rating scales. One suggestion that has been made is to completely abandon verbal labels and use only numeric labels instead. This approach deliberately shifts the task of scale interpretation to the respondents and reduces the information given to them with an intention to reduce their response burden while still preserving the scale meaning. Following prior research, and by drawing on the established model of the cognitive response process, we critically tested these assumptions. Based on a web survey experiment, we found that omitting verbal labels and using only numeric labels instead pushed respondents to focus their responses on the endpoints of a rating scale. Moreover, drawing on response time paradata, we showed that their response burden was not reduced when presented with only numeric labels; quite the opposite was the case, especially when respondents answered the scale with only numeric labels for the first time, which seemed to entail additional cognitive effort. Based on our findings, we advise against using only numeric labels for rating scales in web surveys.

Keywords

response styles, satisficing, data quality, attitude scales, paradata, web surveys

Previous research shows that the design of rating scale labels may substantially influence the way in which respondents answer survey questions. In general, the numeric and verbal labels assigned to each response option may help respondents interpret a rating scale and clarify the meaning of a survey question (Krosnick et al., 2005; O'Muircheartaigh et al., 1995; Schwarz & Hippler, 1995; Schwarz et al., 1991). For instance, Alwin and Krosnick (1991) and Krosnick (1991) report a higher reliability of answers to rating scales when fully labeled scales with verbal labels are used in comparison to scales with numeric labels with verbally labeled endpoints only. Obviously, the use

[1] GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author:
Tobias Gummer, GESIS Leibniz Institute for the Social Sciences, Mannheim, Germany.
Email: tobias.gummer@gesis.org

of numeric scale labels can result in diverging interpretations of scale meaning, which in turn leads to between-respondent differences in response behavior and quality.

As web surveys grew more important in survey research, experimental studies were conducted to identify the effects of rating scale design decisions on the cognitive response process. Regarding rating scale labels, these studies established heuristics on which respondents rely when interpreting scales (e.g., Tourangeau et al., 2004) and a hierarchy of cues that are of relevance in this context with verbal information taking precedence over numeric information (Toepoel & Dillman, 2011; Tourangeau et al., 2007). These studies suggest that the use of fully labeled scales with verbal labels eases the interpretation of scale meaning for respondents, so verbal labels should be preferred over numeric labels. Based on an eye-tracking study, Menold (2020) added further support to these findings by reporting that it is "more difficult for the respondents to map their responses onto the response categories with the numeric than with the verbal rating scales" (p. 22).

With the recent development of high Internet and smartphone penetration rates in many countries (Fuchs & Busse, 2009; Mohorko et al., 2013; Poushter, 2016), the use of smartphones to complete web surveys also has increased (e.g., Gummer et al., 2019). As a consequence, it has become increasingly important to design web surveys in such a way that device effects on response behavior and data quality are minimized (Antoun et al., 2017; Couper & Peterson, 2017; Keusch & Yan, 2017; Lugtig & Toepoel, 2016; Mavletova, 2013). In this context, the discussion about replacing verbal labels with numeric labels in rating scales has resurfaced as a way to reduce the space that questions occupy on the smartphone screen. Thomas and Barlas (2018, 2019) proposed stripping rating scales down by using only numeric labels for all response options and even omitting verbal endpoint labels. This approach rests on the assumption that respondents can adequately interpret the meaning of numeric labels and thus provide reliable and valid responses more quickly than when presented with scales with verbal labels. The use of only numeric labels can be considered a logical consequence and continuation of previous efforts to fit questions on smaller screens. Advocates of this perspective find that replacing verbal labels with only numeric labels is appealing mainly for two reasons. First, rating scales with only numeric labels take up less space on a screen, which makes it easier to display questions on a small screen, for instance on smartphones. Second, the use of only numeric labels instead of verbal labels reduces the amount of verbal information that a respondent must read to grasp the meaning of the rating scale, which is reasoned to make the cognitive response process faster and reduce the response burden.

However, based on previous research on rating scale designs (e.g., Schwarz & Hippler, 1995; Schwarz et al., 1991; Tourangeau et al., 2007), we remain skeptical whether abandoning verbal labels and instead using only numeric labels will not adversely affect response behavior and quality by making the cognitive response process more difficult for respondents. This is to be expected, especially when it involves more complex attitudinal items (compared to few-word items as in the examples of Thomas and Barlas, 2018, 2019).

To investigate whether it is feasible to use a rating scale with only numeric rather than verbal labels, we conducted a survey experiment in which we systematically varied the scale labeling and examined the effects with regard to the various indicators of response behavior and quality. In the following sections, we present our hypotheses, the data and methods we used, and our results. We then close with concluding remarks and suggestions for future research.

Hypotheses

The established model of the cognitive response process (e.g., Tourangeau et al., 2000) differentiates between four steps that a respondent has to process through when answering a survey question: (1) comprehending the question meaning, (2) retrieving the relevant information, (3) forming a judgment, and (4) mapping the answer to the response options.

With respect to rating scales with verbal labels, respondents can directly deduce the meaning of the verbally labeled response options and map their answers to one of these. In contrast, "numbered scale points have no inherent meaning" (Krosnick & Fabrigar, 1997, p. 149). Thus, when using only numeric labels, respondents must first create a verbal equivalent for each response option before they can match these "translations" to their mental judgment. This additional interpretation requirement complicates the cognitive processing of rating scales with only numeric labels and increases the required cognitive effort (Krosnick & Fabrigar, 1997; Krosnick et al., 2005).

Moreover, when respondents interpret the meaning of a rating scale that has no verbal labels, the endpoints are deemed an important source of information since the numeric values alone have no clear meaning without knowledge of their range. Thus, for respondents, endpoints are anchors or reference points for interpreting the intermediate response options and mapping answers accordingly (Tourangeau et al., 2004, 2007). Two effects can be expected. First, respondents' cognitive effort is increased since they first must relate the individual scale points to each other before making a meaningful interpretation of the entire scale. Second, it can be assumed that the endpoints are more prevalent in the interpretation process of respondents, so they can reduce the ambiguity of intermediate response options, and thus, the endpoints are more central in importance when a rating scale with only numeric labels is presented compared to a rating scale with verbal labels. Therefore, the reduction of the amount of verbal information in rating scales with only numeric labels may enable respondents to read a question faster (Menold, 2020; Menold et al., 2014; Thomas & Barlas, 2018, 2019). However, due to the lack of information in scales with only numeric labels, respondents' interpretation of the meaning of the scale will be more difficult and time-consuming and thus more burdensome.

According to satisficing theory (Krosnick, 1991, 1999), respondents try to reduce their response burden by using various cognitive shortcuts. By skipping one or more steps of the cognitive response process, they aim at a satisfactory answer without overdoing the cognitive effort. This finding is in line with the conclusion drawn by previous studies that suggest "that if no verbal or numerical labels are used, respondents become more susceptible to hints and thus more inclined to use other heuristics [. . .] to arrive at acceptable answers" (Moors et al., 2014, p. 374). Satisficing response behavior associated with lowered data quality as a consequence of skipping one or more steps of the cognitive response process can take various forms (Krosnick, 1991, 1999; Tourangeau et al., 2000).

First, *acquiescence* is a respondent's tendency to agree with statements independently of their content (e.g., Paulhus, 1991; Rammstedt et al., 2017). Among other behaviors, acquiescence is considered especially likely with respect to issues about which respondents are uncertain (Baumgartner & Steenkamp, 2001), or when the items are ambiguous and vague in content (Weijters et al., 2013). Moreover, acquiescence increases with the cognitive load; this is in case respondents are required to divide their attention (Cabooter, 2010; Knowles & Condon, 1999).

Second, *extreme responding* is a respondent's tendency to choose one of the extreme endpoints of the rating scale regardless of the content of the statements (Paulhus, 1991). Among other influencing factors, an extreme response is less likely with verbally labeled scales, which provide respondents with a higher perceived meaning and salience of all response options (Weijters et al., 2010).

Third, *midpoint responding* refers to the respondent's tendency to use the midpoint of a rating scale regardless of item content. Similar to extremity, midpoint responding is less likely when all response options are verbally labeled and thus easier for respondents to interpret (Weijters et al., 2010). Midpoint responding is negatively correlated with extreme responding (Baumgartner & Steenkamp, 2001; He et al., 2014).

Fourth, *item nonresponse* occurs when responses to one or more items are missing, which may be due to respondents' deliberate decisions to reduce the extent of their response burden (Beatty & Herrmann, 2002; Krosnick et al., 2002). These respondents skip parts of their cognitive response

process or at least the step of mapping their answer to the response scale by not providing an answer. Item nonresponse is especially likely for items that are considered difficult due to the question content or the complex question format (Dixon & Tucker, 2010).

Fifth, *nondifferentiation* is a respondent's tendency to first select the answer that seems satisfactory and then anchor all subsequent responses to that first answer, which results in the selection of the same or nearly the same response options to answer a set of rating scale items, rather than making use of the full range of response options (cf. McCarty & Shrum, 2000; Roßmann et al., 2018). Often, only a few response options are selected that are very close to each other (e.g., "fully agree" and "rather agree"), although nondifferentiation also can mean limiting the selection to a few response options that are at a maximum distance from each other (e.g., "fully agree" and "not at all agree").

Based on the model of the cognitive response process, we hypothesized:

Hypothesis 1 (H1): Rating scales with only numeric labels are associated with a higher likelihood of cognitive shortcuts than rating scales with verbal labels due to the greater cognitive effort involved with answering scales with only numeric labels.

With respect to the different aspects of response behavior previously outlined, we put forward the following specific hypotheses:

Hypothesis 1.1 (H1.1): Using only numeric labels instead of verbal labels will increase the likelihood of acquiescent responding.

Hypothesis 1.2 (H1.2): Using only numeric labels instead of verbal labels will increase the likelihood of extreme responding.

Hypothesis 1.3 (H1.3): Using only numeric labels instead of verbal labels will increase the likelihood of midpoint responding.

Hypothesis 1.4 (H1.4): Using only numeric labels instead of verbal labels will increase the likelihood of item nonresponse.

Hypothesis 1.5 (H1.5): Using only numeric labels instead of verbal labels will increase the likelihood of nondifferentiated responding so that fewer of the available response options are used.

Previously, we argued that we expect that using rating scales with only numeric labels will lead to an increased response burden. Response times frequently are used as a measure for how burdensome answering a question is for respondents and how much effort they devote to the answering process (Greszki et al., 2015; Zhang & Conrad, 2014). We concur that it would be faster to read the digits in only numeric labels compared to the text in verbally labeled scales. However, as indicated previously, we assume that using a scale with only numeric labels will increase the complexity of a respondent's interpretation task. Accordingly, the potential time savings due to less verbal information are eliminated by the additional "translation effort" and the need to interpret the individual response options in relation to each other in scales with only numeric labels. Consequently, we hypothesized:

Hypothesis 2 (H2): Rating scales with only numeric labels will not lead to time savings compared to rating scales with verbal labels.

To evaluate the robustness of the potential effects of different scale labels on response behavior and quality, we varied other relevant scale characteristics that had been shown to impact response behavior in previous research. Our aim was to test whether effects occurred not only in one specific

scale design in which we varied the use of numeric and verbal labels but also across different designs of the same scale. We varied two additional scale characteristics—scale specification and scale orientation—which both can influence the interpretation of response scales and thus may interact with the effects of scale labeling.

Scale specification is a design decision that determines whether the question stem and response options are generic (agree–disagree) or tailored (construct-specific) for each item of a rating scale (e.g., Höhne et al., 2018; Saris et al., 2010). For example, in a generic case that asks about the importance of tasks at work, one might present a statement about the importance of a task and provide a scale ranging from *agree* to *disagree*. When using construct-specific scales, one might ask about the importance of a task and provide a scale ranging from *important* to *unimportant*. Concerning response quality, previous research has shown that responses to construct-specific scales are of comparable or even higher quality than responses to items with agree–disagree scales (Hanson, 2015; Höhne et al., 2018; Lelkes & Weiss, 2015; Liu et al., 2015). Moreover, construct-specific scales take longer to complete, which suggest a more conscientious cognitive processing compared to that involved with agree–disagree scales (Höhne et al., 2018). Similarly, based on an eye-tracking study, Höhne and Lenzner (2018) showed that respondents expend more cognitive effort in processing response options when a construct-specific scale is used. With regard to previous research, we expect respondents to invest more effort when answering construct-specific compared to agree–disagree scales. Especially with respect to scales with only numeric labels that are construct-specific, we assume that respondents would invest high cognitive effort when answering without shortcutting the response process.

Scale orientation is a design decision that determines the order in which the response options are presented in a rating scale (e.g., Krebs & Bachner, 2018; Yan & Keusch, 2015). For instance, rating scales might be ordered in incremental or decremental order: low to high, high to low, disagree to agree, agree to disagree, and so on. Previous research has hinted at differences in response behavior depending on the particular order of response options (Hofmans et al., 2007; Krebs & Bachner, 2018; Yan & Keusch, 2015). As argued previously, we assume that respondents draw on information provided by the response scale. With respect to scales with verbal labels, the information provided easily reveals which scale endpoint resembles the "high" or "low" end of a scale (i.e., start and endpoint). With respect to scales with only numeric labels, we assume that the interpretation task is more complex. In contrast to scales with verbal information, the numeric value of a scale endpoint does not provide information regarding whether this response option is the start or end of the scale. When only numeric labels are provided, respondents must compare the numeric values to each other to deduce their meaning. Consequently, we assume that scale orientation affects the interpretation of scales, presumably differently when using only numeric labels compared to using verbal labels.

Data and Method

Sample

We used data from an experiment embedded in a web survey that was conducted in October 2018 with quota-sampled panelists from a large German online access panel. Quotas on age, sex, and education were set to resemble distributions of the last German census. 5,563 panelists were invited of which 824 were screened out and 238 broke off, which resulted in a final sample of 4,371 cases with a participation rate of 92% (American Association for Public Opinion Research, 2016) and a break-off rate of 5% (Callegaro & DiSogra, 2008). Our questionnaire featured a variety of questions regarding political behavior and attitudes and took an average of 21.7 min to complete ($Mdn = 18.7$).

Table 1. Rating Scale Characteristics in the 2 × 2 × 2 Between-Subjects Design (Example Item).

Agree–disagree	It is very important for me to decide for myself how I do my work.	
Verbal	Positive–negative	Negative–positive
	○ Fully agree	○ Not at all agree
	○ Rather agree	○ Rather not agree
	○ Reasonably agree	○ Reasonably agree
	○ Rather not agree	○ Rather agree
	○ Not at all agree	○ Fully agree
Numeric	High–low	Low–high
	○ +4	○ 0
	○ +3	○ +1
	○ +2	○ +2
	○ +1	○ +3
	○ 0	○ +4
Construct-specific	How important is it for you to decide for yourself how you do your work?	
Verbal	Positive–negative	Negative–positive
	○ Very important	○ Not at all important
	○ Rather important	○ Rather not important
	○ Reasonably important	○ Reasonably important
	○ Rather not important	○ Rather important
	○ Not at all important	○ Very important
Numeric	High–low	Low–high
	○ +4	○ 0
	○ +3	○ +1
	○ +2	○ +2
	○ +1	○ +3
	○ 0	○ +4

Experimental Design

Our experimental question was a 10-item battery comprising items adapted from the German version of the Achievement Motivation Inventory designed by Schuler et al. (2004). More information on the items is provided in Appendix A. We presented the items in an item-by-item format (i.e., all items were presented individually on the same screen, and the response options for each item were repeated). We vertically aligned the 5-point unipolar rating scales. We randomly assigned respondents to experimental groups by three factors with two dimensions each: verbal versus only numeric labels, agree–disagree versus construct-specific response options, and positive to negative versus negative to positive/high to low versus low to high response option order. Table 1 provides examples for all experimental groups. This full-factorial 2 × 2 × 2 between-subjects design resulted in eight experimental groups, each having a sample size of approximately 545 respondents.

Response Behavior and Quality Indicators

With respect to H1, we computed a set of indicators that described different aspects of response behavior and quality. We calculated all the indicators based on the 10-item battery for each respondent individually.

Acquiescence. We computed the share of agreeing answers to the 10-item battery (i.e., *fully agree*, +4). This indicator takes values between [0; 1]. The interpretation of the acquiescence indicator as a

share is straightforward. For instance, a value of .5 indicates that 50% of the answers made by a respondent were in agreement with the statements of the 10-item battery.

Extremity. Again, we relied on the share of responses on the endpoint response options of the 10-item battery per respondent (i.e., *fully/not at all agree*, +4/0). This indicator takes values between [0; 1] and can be interpreted as illustrated above.

Midpoint responding. We calculated the share of answers on the midpoint response options of the 10-item battery per respondent (i.e., *reasonably agree*, +2). Again, the indicator takes values between [0; 1] and can be interpreted accordingly.

Item nonresponse. We did not force respondents to answer, and they were free to skip items. As before, this indicator denotes the share of missing answers in the 10-item battery in the range of [0; 1] per respondent.

Nondifferentiation. We applied two different indicators to account for nondifferentiation. First, we examined the *coefficient of variation* measured as the distance between the scale points used in the 10-item battery (McCarty & Shrum, 2000). This indicator is defined as the standard deviation of a respondent's answers to an item battery and takes values in the range [0; 2] in our 10-item, 5-point rating scale. A value of 0 indicates straightlining (i.e., perfect nondifferentiation by giving the same answer to all items in a battery), whereas higher values indicate a greater variation between responses to a rating scale. Second, we calculated ρ which measures how many different scale points a respondent selected when answering a set of rating scale items (Krosnick & Alwin, 1988; McCarty & Shrum, 2000). ρ takes values between [0; 1]. Again, a value of 0 indicates straightlining, whereas higher values indicate that a respondent selected a higher number of different response options to answer a rating scale.

With respect to H2, we were interested in gauging the response burden operationalized as the time respondents spent on comprehending and answering the 10-item battery. Therefore, we relied on client-side time stamps captured by the Embedded Client Side Paradata script (Schlosser & Höhne, 2018). We computed two different response time indicators as proxy for the response burden.

Response time. To gauge the cognitive response process as a whole, we relied on the total time respondents spent on answering the 10-item battery. Time stamps used in our study were captured in milliseconds, but in the analyses, the response times were shown in seconds. To exclude outliers from our analyses, we coded all response times t_i to missing, which were above the commonly used criterion of $\bar{t} + 2SD(t)$. For analytical purposes, we used the logarithm of the cleaned response times, which reduced the skewness to −.14.

Time until the first click. We used the time until the first click to answer the 10-item battery, which we also measured in milliseconds. This indicator describes the time that respondents spent reading the question, comprehending its meaning, and mapping their answer to the response options for the first time. In our view, this time is indicative of the differences in the effort necessary to interpret the meaning of the scale labels as a prerequisite for matching respondents' interpretations with their mental judgments. Again, we omitted outliers with the same procedure as indicated previously and used the logarithm of the cleaned response times (in seconds) to reduce the skewness to −.04.

Method

For each of the eight indicators, we fitted separate ordinary least squares regressions with the respective indicator as a dependent variable. Regressions were based on the full sample ($N = 4{,}371$), and we added dummy variables for all eight experimental groups (each around $n = 545$). Regression outputs are provided in Appendix B, Table B1. We computed Wald tests to gauge whether experimental groups differed with respect to specific response behavior indicators (i.e., we tested the regression parameters for significant differences). Per model, we compared each of the eight experimental groups to the remaining seven groups (i.e., 28 Wald tests per model). We report selected tests in the "Results" section, and we present all tests in Appendix B, Tables B2–B9.

Since we employed a $2 \times 2 \times 2$ experimental design to control the robustness of our findings, we compared the effect of using a rating scale with only numeric labels instead of verbal labels in four instances (i.e., across the two additional experimental factors). Accordingly, to test our hypotheses, we tested regression parameters for differences per response behavior indicator 4 times. We assumed an effect was robust if it was significant across all four tests. For example, when assessing H1.1, we tested for the significance of differences in acquiescence between:

- Only numeric labels versus verbal labels, both with *agree–disagree* response options and a positive to negative (high to low) response option order;
- Only numeric labels versus verbal labels, both with *agree–disagree* response options and a negative to positive (low to high) response option order;
- Only numeric labels versus verbal labels, both with *construct-specific* response options and a positive to negative (high to low) response option order;
- Only numeric labels versus verbal labels, both with *construct-specific* response options and a negative to positive (low to high) response option order.

Then, we utilized the remaining Wald tests to investigate the potential interaction effects of scale labeling with scale orientation and scale specification.

To ease the interpretation and comparability of our findings, we present plots with predicted values and confidence intervals. We estimated the predicted values based on the regression models for each response behavior indicator, which can be interpreted accordingly. For instance, a predicted value of .5 in extremity for an experimental group means that the respondents in this group have an average share of 50% answers based on the scale endpoints. Note that confidence intervals of the predicted values do not necessarily represent significant differences between parameters and should not be interpreted in this way (e.g., Knol et al., 2011; Schenker & Gentleman, 2001). With respect to hypothesis testing, we refer readers to the tests discussed above.

Results

H1 stated that the use of rating scales with only numeric instead of verbal labels would foster the likelihood of cognitive shortcuts. Figure 1 presents the predicted values for acquiescent and extreme responding by experimental group. When comparing the rating scales with numeric and verbal labels, we did not find a consistent pattern regarding acquiescence. The difference between scales with only numeric versus verbal labels was not significant for agree–disagree scales (both Wald tests $p > .05$), whereas for construct-specific scales, we found respondents to be more likely to acquiesce when using only numeric instead of verbal labels (both Wald tests $p < .05$), irrespective of the response option order. Based on these findings, we rejected H1.1.

In contrast, looking at the extent of extremity, we found a consistent effect of using numeric instead of verbal labels. Respondents answering rating scales with only numeric labels were more likely to choose one of the endpoints compared to those who were presented with scales with verbal

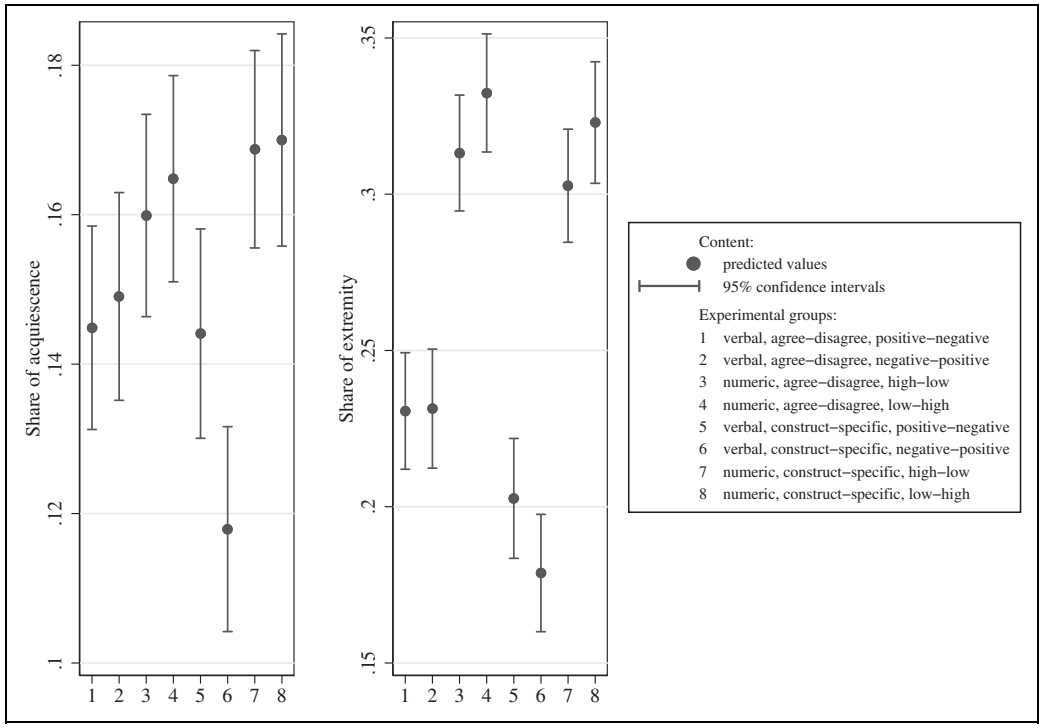

Figure 1. Predicted values of acquiescence and extremity across experimental groups.

labels (four of four Wald tests $p < .05$). This effect was robust with respect to the other scale characteristics that we varied in our experiment, and thus, we interpret our findings to support H1.2. In combination with our finding that respondents were not consistently more likely to acquiesce regardless of whether the rating scale used only numeric or verbal labels, we found that respondents increasingly used both endpoints of the rating scale when only numeric labels were presented.

Related to midpoint responding, we found that respondents who answered scales with only numeric labels were less likely to use the middle response option compared to those who answered scales with verbal labels (Figure 2). Differences in midpoint responding were robust and appeared independent from scale specification and scale orientation (four of four Wald tests $p < .05$). Although these findings were opposite to H1.3, given previous research that showed a negative correlation between midpoint and extreme responding (Baumgartner & Steenkamp, 2001; He et al., 2014), these findings were consistent with our earlier findings that respondents focused their answers more on the endpoints of the response scales. It seems that endpoints were a more prevalent anchor than midpoints when respondents were presented with scales with only numeric labels.

With respect to item nonresponse, we did not find consistent significant differences between respondents' use of scales with only numeric versus verbal labels (three of four Wald tests $p > .05$). We found a significant difference only when respondents answered construct-specific scales with a response option order from positive to negative/high to low ($F = 4.255$, $p < .05$). Given this finding, we rejected H1.4.

With respect to the two indicators of nondifferentiation (Figure 3), the coefficient of variation consistently indicated more differentiated responses to rating scales with only numeric compared to verbal labels (four of four Wald tests $p < .05$). This finding can be explained easily by the more

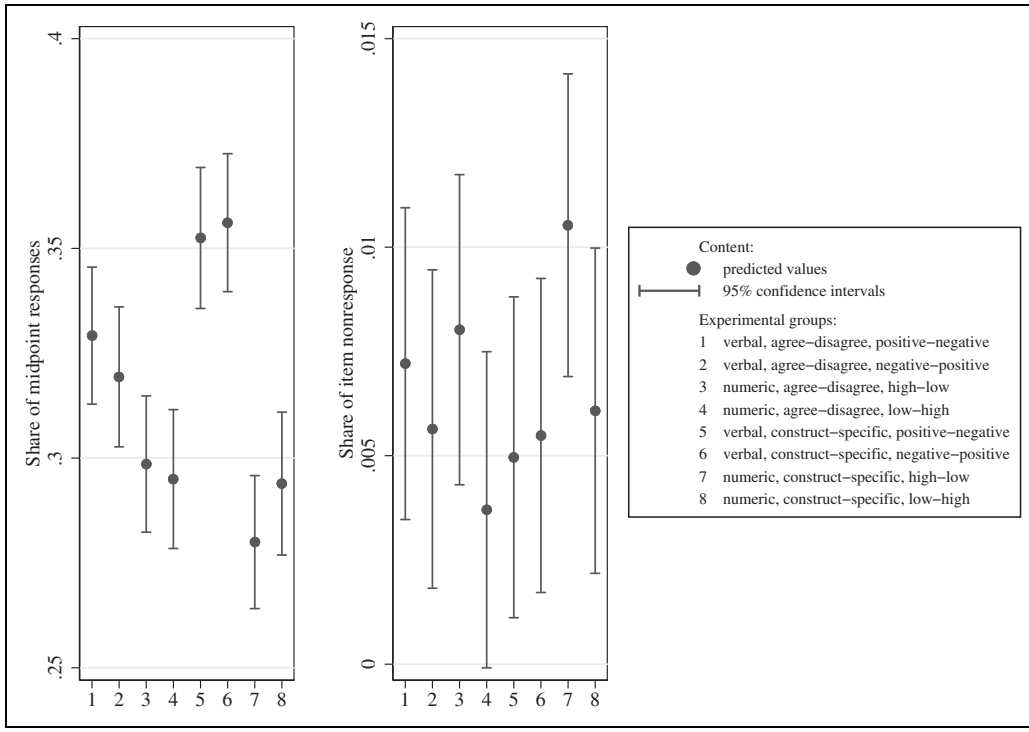

Figure 2. Predicted values of midpoint responding and item nonresponse across experimental groups.

frequent use of the two endpoints of the rating scale when only numeric labels were provided, which maximized the distance between responses and thus the coefficient of variation. We found this pattern to be independent of scale specification and scale orientation. Yet, Wald tests showed that the coefficient of variation was generally lower if construct-specific scales were used (four of four Wald tests $p < .05$). With regard to our second measure of nondifferentiation ρ, which reflects the extent to which different response options were used to answer the rating scale, we did not find consistent differences between scales with only numeric and verbal labels (three of four Wald tests $p > .05$). We found a significant difference ($F = 4.809$, $p < .05$) only when respondents answered agree–disagree scales ranging from positive to negative/high to low. Considering both indicators of nondifferentiation, we rejected H1.5. Respondents who answered a rating scale with only numeric labels were more likely to select response options that were as far apart as possible (i.e., endpoints). However, they were not more inclined to use the full range of response options compared to respondents who answered a scale with verbal labels.

Overall, based on the six indicators of response behavior described previously, the results with respect to H1 were mixed.[1] Two findings that were consistent across all experimental groups showed that respondents who answered the 10-item battery based on a rating scale with only numeric labels were more likely to rely on the endpoints of the scale and less likely to select the middle response option compared to those who answered scales with verbal labels. We did not find clear differences in nondifferentiation and item nonresponse. For acquiescence, we found an interaction with scale specification. With respect to construct-specific scales, acquiescent responding differed according to whether these scales were labeled with only numeric or verbal labels.

H2 stated that the use of rating scales with only numeric labels would not lead to time savings compared to using rating scales with verbal labels. As depicted in Figure 4, we did not find

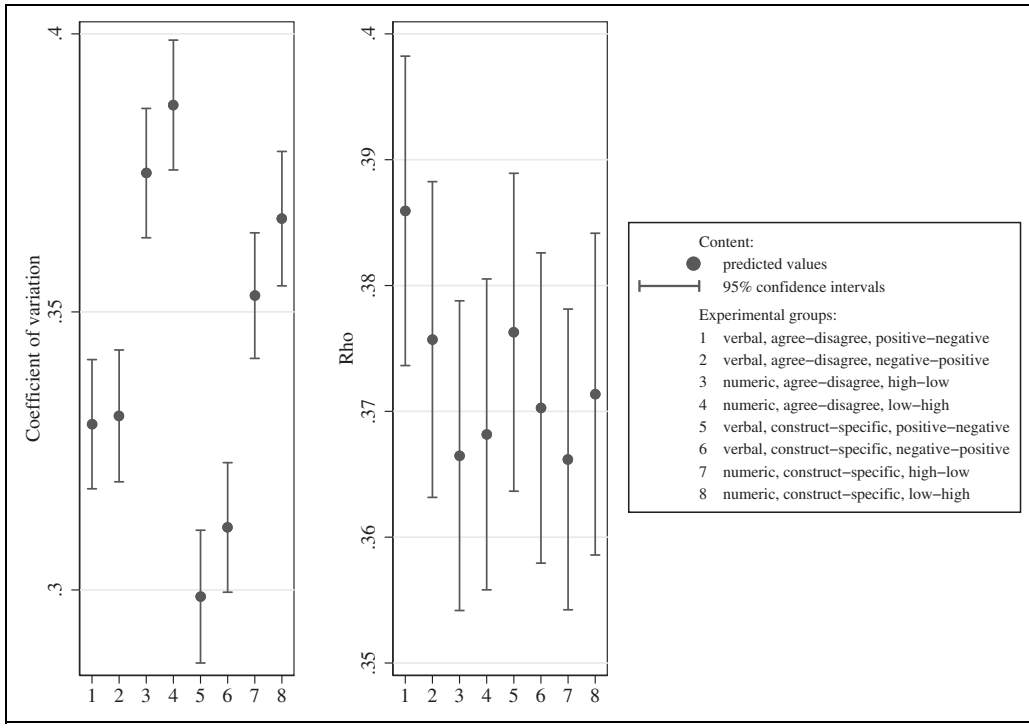

Figure 3. Predicted values of nondifferentiation across experimental groups.

consistent significant differences in the overall time that respondents required to answer the 10-item battery (three of four Wald tests $p > .05$). This finding is line with our expectation that, on the one hand, reducing the amount of verbal information in rating scales likely would reduce the time needed to read the question, whereas on the other hand, answering rating scales with only numeric labels would require more interpretation by respondents and thus take more time. This additional cognitive task associated with rating scales with only numeric labels should increase response time, which cancels out any time savings due to quicker reading times and, thus, supports H2. However, we would like to add that we found an effect of using a rating scale with only numeric labels instead of verbal labels when construct-specific response options were provided that used a low to high order ($F = 4.959$, $p < .05$).

In addition, we compared the time until the first click was made by respondents who answered a rating scale with only numeric labels and respondents who answered a rating scale with verbal labels. We argued that the time until the first click captured the respondents' cognitive response process of interpreting the meaning of the scale labels and then matching this interpretation with their mental judgments for the first time. In line with our theoretical reasoning, we found that when using only numeric labels, the time until the first click was longer compared to when using verbal labels. Differences were significant across different scale specifications and orientations (four of four Wald tests $p < .05$), which hints at the robustness of this finding. However, we also found an interaction between scale labeling and scale specification: For rating scales with only numeric labels, the time until the first click was significantly higher when respondents answered construct-specific scales compared to agree–disagree scales. We found this relationship independent of whether the scale was presented from high to low ($F = 12.400$, $p < .05$) or from low to high ($F = 4.473$, $p < .05$). For rating scales with verbal labels, we found a significantly longer time until the

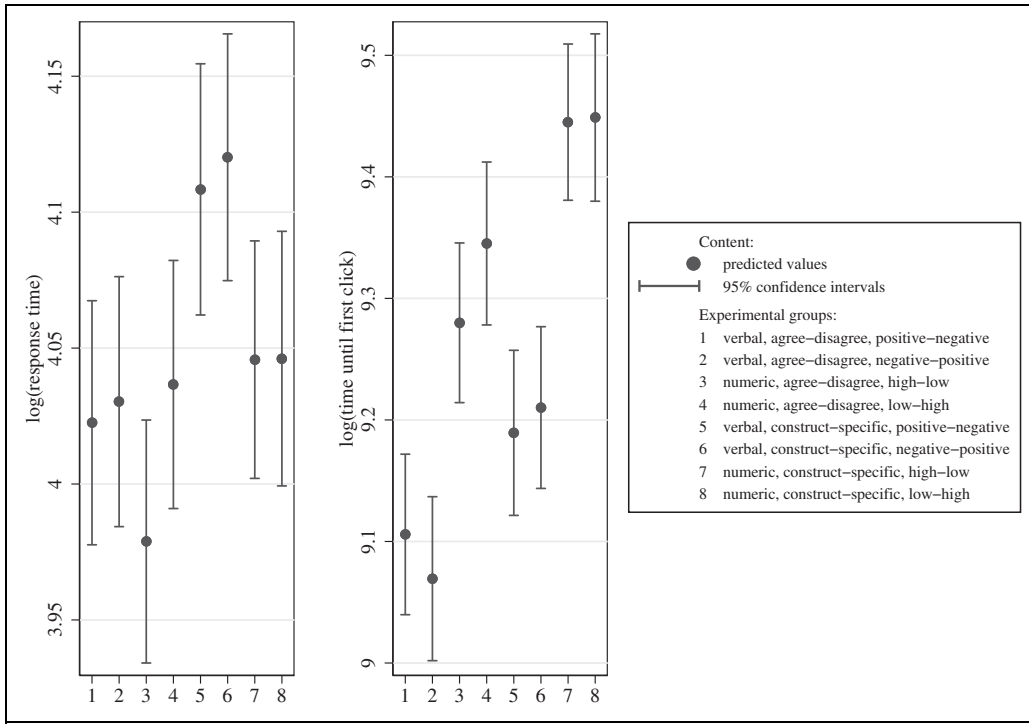

Figure 4. Predicted values of response time indicators across experimental groups.

first click for construct-specific compared to agree–disagree scales, only when the scale ranged from negative to positive ($F = 8.488$, $p < .05$). The use of only numeric labels provides little information—compared to verbal labels that ease interpretation and presumably help to link a construct-specific question stem to the response options—and thus makes a response to this condition even more demanding.

Our findings on the time until the first click indicate a more cumbersome cognitive response process when respondents answer a rating scale with only numeric labels, which is, in our view, the result of an increased interpretation task due to the lack of verbal information on the meaning of the response options.

Conclusion

In this study, we set out to investigate the effects of using a rating scale with only numeric labels instead of verbal labels. With respect to the increasing use of smartphones to complete web surveys, survey researchers and practitioners are searching for ways to reduce the screen space occupied by survey questions. The goal behind these approaches is to fit questions on smaller screens, enhance the survey experience, lessen the response burden, and, thus, increase data quality. In our study, we investigated a rather extreme solution to this problem that has emerged recently in the academic debate, which proposes to completely omit verbal labels in rating scales and rely only on numeric labels. At first glance, this appears to be a consistent continuation of the idea of reducing the amount of information presented to respondents to a minimum. Given the consequences that previous research on rating scale designs led us to expect, empirical tests on the effects of this strategy on response behavior seemed necessary. We have presented experimental evidence that stripping down

a rating scale to its bare minimum and presenting only numeric labels will result in respondents heavily focusing their answers on the endpoints of a rating scale. Apparently, the meaning of a rating scale with only numeric labels becomes clear only when the endpoints of the scale are interpreted, which suggests that the endpoints serve as an anchor or reference point for respondents to interpret the intermediate response options and make their judgments (Tourangeau et al., 2004, 2007). Following this reasoning, we argued that endpoint information becomes the most relevant information in rating scales with only numeric labels, and thus, respondents select one of the most extreme response options—instead of making use of the full range of response options—to cope with the ambiguity and complexity of the response task. Moreover, we found that the response burden with respect to the cognitive response process was not lessened by presenting respondents only with numeric labels. Quite the contrary, our results showed that more time was required to perform additional interpretation tasks of the rating scale. This finding is in line with previous research showing that numeric labels require increased interpretation by respondents (Schwarz & Hippler, 1995; Schwarz et al., 1991).

While the increasing need to specifically cater to smartphone respondents when designing web surveys seems widely accepted, survey researchers and practitioners should carefully consider the consequences of these design decisions. In this study, by drawing on the previous research on the cognitive response process, we highlighted the role of interpretation tasks in answering rating scales. By providing verbal labels, researchers try to facilitate the interpretation of the meaning of a rating scale to ensure that a respondent's and their own interpretations align. From this perspective, it does not seem reasonable to leave the interpretation process entirely to the respondent by providing only numeric labels. If verbal labels are omitted completely, we should think about other ways of facilitating the interpretation of rating scales with the aim of reducing the complexity of the cognitive task delegated to respondents. Possible solutions might include color coding or other visual design features of questions that give meaning to scale points (e.g., "green" for positive or "red" for negative, "thumps up" for agreement and "thumps down" for disagreement) and provide clues on how scale points relate to each other (e.g., by tuning down color by 50%). Research in this area is budding (Stange et al., 2018; Toepoel & Dillman, 2010; Toepoel & Funke, 2018; Toepoel et al., 2019; Tourangeau et al., 2007), although recent studies on the use of smiley faces were not encouraging in terms of data quality and ease of cognitive response processing (Cernat & Liu, 2019; Gummer et al., 2020). Unless convincing progress is made to make rating scales more compact and less screen size space consuming, for now, we recommend complying with the classical advice on using fully labeled rating scales with verbal labels (e.g., Krosnick & Fabrigar, 1997).

As always, our study is not without limitations and thus yields opportunities for future research. First, we decided to study the effects of using rating scales with only numeric labels in comparison to verbal labels in a sample that allowed respondents to answer on desktop PCs, laptops, tablets, and smartphones. Our rationale behind this approach was that even now, general purpose web surveys will likely be answered on a mix of devices, and we argue that methodological experimentation should be done in a setting as close as possible to real-world applications. However, we would like to acknowledge that the idea of using only numeric labels is strongly focused on smartphones. Therefore, we encourage future studies that investigate the device-specific effects of using only numeric labels by experimentally assigning respondents to specific devices.

Second, our study made use of a sample drawn from an online access panel, which might impair its generalizability to other samples. Having said that, we assume that the effects of leaving the interpretation of a rating scale to respondents (i.e., when presenting only numeric labels) would be even more pronounced in general population samples composed of respondents who are less used to answering survey questions compared to the more experienced respondents of online access panels.

Third, we compared a fully verbally labeled scale to a scale only labeled 0–4. Using this range of numbers to label scale points seems to be common practice in survey research, and we made an

implicit assumption that these numbers correspond to verbal labels. A future study investigating this topic in more detail could draw on a rather simplistic experimental design: One experimental group could receive fully verbally labeled scales, while the other experimental groups receive scales with the same number of response options labeled with different ranges of numbers, for instance, a first group gets "0, 1, 2, 3, 4," a second "0, 10, 20, 30, 40," etc. A comparison of response behaviors between these groups could be used to identify which numbers best correspond to verbal labels. Moreover, such an experimental study could be used to test the replicability and generalizability of our findings.

Finally, this study focused on the effects of stripping down a rating scale to its bare minimum by using rating scales with only numeric labels and comparing them to the use of rating scales fully labeled with verbal labels. Certainly, the degrees in-between these two ways of labeling rating scales also warrant investigation. Again, experimentation might be a viable way of advancing our study in this direction.

Appendix A

Our experimental question was a 10-item battery comprising items adapted from the German version of the Achievement Motivation Inventory (AMI) designed by Schuler et al. (2004). The AMI is an inventory capturing 17 dimensions of achievement motivation by using 10 items to measure each dimension.

For the purpose of our study, we adapted the following 10 items:

- Item 5: dimension fearlessness;
- Item 17: dimension goal setting;
- Item 21: dimension flow;
- Item 31: dimension dominance;
- Item 87: dimension flexibility;
- Item 91: dimension pride in productivity;
- Item 123: dimension status orientation;
- Item 129: dimension preference for difficult tasks;
- Item 135: dimension independence;
- Item 160: dimension competitiveness.

Appendix B

Table B1. Regression Models on Response Behavior and Quality Indicators.

Regression Parameters	Acquiescence Coef. (*SE*)	Extremity Coef. (*SE*)	Midpoints Coef. (*SE*)	Item Nonresponse Coef. (*SE*)	CV Coef. (*SE*)	ρ Coef. (*SE*)	Response Time Coef. (*SE*)	Time Until the First Click Coef. (*SE*)
Experimental groups								
(1) Verbal, agree–disagree, positive–negative	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
(2) Verbal, agree–disagree, negative–positive	.004 (.010)	.001 (.014)	−.010 (.012)	−.002 (.003)	.001 (.008)	−.010 (.009)	.008 (.033)	−.036 (.048)
(3) Numeric, agree–disagree, high–low	.015 (.010)	.083*** (.013)	−.031** (.012)	.001 (.003)	.045*** (.008)	−.019* (.009)	−.044 (.032)	.174*** (.048)
(4) Numeric, agree–disagree, low–high	.020* (.010)	.102*** (.014)	−.034** (.012)	−.004 (.003)	.057*** (.008)	−.018* (.009)	.014 (.033)	.239*** (.048)
(5) Verbal, construct-specific, positive–negative	−.001 (.010)	−.028* (.014)	.023 (.012)	−.002 (.003)	−.031*** (.008)	−.010 (.009)	.086** (.033)	.084 (.048)
(6) Verbal, construct-specific, negative–positive	−.027** (.010)	−.052*** (.013)	.027* (.012)	−.002 (.003)	−.019* (.008)	−.016 (.009)	.098** (.033)	.104* (.048)
(7) Numeric, construct-specific, high–low	.024* (.010)	.072*** (.013)	−.049*** (.012)	.003 (.003)	.023** (.008)	−.020* (.009)	.023 (.032)	.339*** (.047)
(8) Numeric, construct-specific, low–high	.025* (.010)	.092*** (.014)	−.035** (.012)	−.001 (.003)	.037*** (.009)	−.015 (.009)	0.024 (.033)	.343*** (.049)
Intercept	.145*** (.007)	.231*** (.010)	.329*** (.008)	.007*** (.002)	.330*** (.006)	.386*** (.006)	4.023*** (.023)	9.106*** (.034)
N	4,358	4,358	4,358	4,358	4,165	4,165	4,336	4,348
R^2	.008	.057	.017	.000	.042	.000	.005	.026

Note. Coef = regression coefficient; *SE* = standard error; CV = coefficient of variation.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table B2. Wald Test on Difference Between Regression Parameters for Acquiescence.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	0.179 (.673)	—						
Numeric, agree–disagree, high–low	**2.355 (.125)**	1.198 (.274)	—					
Numeric, agree–disagree, low–high	4.072 (.044)	**2.486 (.115)**	0.249 (.618)	—				
Verbal, construct-specific, positive–negative	0.006 (.938)	0.244 (.621)	2.531 (.112)	4.272 (.039)	—			
Verbal, construct-specific, negative–positive	7.479 (.006)	9.778 (.002)	18.242 (.000)	22.341 (.000)	6.850 (.009)	—		
Numeric, construct-specific, high–low	6.099 (.014)	4.055 (.044)	0.845 (.358)	0.164 (.686)	**6.312 (.012)**	27.406 (.000)	—	
Numeric, construct-specific, low–high	6.276 (.012)	4.268 (.039)	1.020 (.313)	0.264 (.608)	6.489 (.011)	**26.763 (.000)**	0.016 (.900)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. Tests used for hypothesis testing in bold. $F = F$ statistic, $p = p$ value.

Table B3. Wald Test on Difference Between Regression Parameters for Extremity.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	0.003 (.955)	—						
Numeric, agree–disagree, high–low	**37.905 (.000)**	36.418 (.000)	—					
Numeric, agree–disagree, low–high	56.511 (.000)	**54.507 (.000)**	2.025 (.155)	—				
Verbal, construct-specific, positive–negative	4.200 (.040)	4.340 (.037)	65.963 (.000)	89.220 (.000)	—			
Verbal, construct-specific, negative–positive	14.755 (.000)	14.872 (.000)	99.713 (.000)	127.812 (.000)	3.042 (.081)	—		
Numeric, construct-specific, high–low	29.597 (.000)	28.345 (.000)	0.628 (.428)	4.950 (.026)	**55.322 (.000)**	86.814 (.000)	—	
Numeric, construct-specific, low–high	45.141 (.000)	43.500 (.000)	0.506 (.477)	0.468 (.494)	74.516 (.000)	**109.309 (.000)**	2.228 (.136)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. $F = F$ statistic, $p = p$ value.

Table B4. Wald Test on Difference Between Regression Parameters for Midpoints.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	0.681 (.409)	—						
Numeric, agree–disagree, high–low	**6.790 (.009)**	3.064 (.080)	—					
Numeric, agree–disagree, low–high	8.307 (.004)	**4.130 (.042)**	0.091 (.763)	—				
Verbal, construct-specific, positive–negative	3.797 (.051)	7.519 (.006)	20.444 (.000)	22.817 (.000)	—			
Verbal, construct-specific, negative–positive	5.190 (.023)	9.465 (.002)	23.819 (.000)	26.363 (.000)	0.092 (.761)	—		
Numeric, construct-specific, high–low	17.978 (.000)	11.265 (.001)	2.583 (.108)	1.653 (.199)	**37.864 (.000)**	42.715 (.000)	—	
Numeric, construct-specific, low–high	8.584 (.003)	4.376 (.037)	0.150 (.698)	0.008 (.929)	23.014 (.000)	**26.517 (.000)**	1.382 (.240)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. $F = F$ statistic, $p = p$ value.

Table B5. Wald Test on Difference Between Regression Parameters for Item Nonresponse.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	0.331 (.565)	—						
Numeric, agree–disagree, high–low	**0.092 (.762)**	0.768 (.381)	—					
Numeric, agree–disagree, low–high	1.665 (.197)	**0.497 (.481)**	2.542 (.111)	—				
Verbal, construct-specific, positive–negative	0.673 (.412)	0.060 (.806)	1.257 (.262)	0.209 (.648)	—			
Verbal, construct-specific, negative–positive	0.405 (.524)	0.003 (.955)	0.883 (.347)	0.427 (.513)	0.036 (.849)	—		
Numeric, construct-specific, high–low	1.560 (.212)	3.309 (.069)	0.894 (.345)	6.499 (.011)	**4.255 (.039)**	3.573 (.059)	—	
Numeric, construct-specific, low–high	0.168 (.682)	0.025 (.875)	0.500 (.480)	0.733 (.392)	0.160 (.689)	**0.046 (.830)**	2.680 (.102)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. $F = F$ statistic, $p = p$ value.

Table B6. Wald Test on Difference Between Regression Parameters for Coefficient of Variation.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	0.031 (.860)	—						
Numeric, agree–disagree, high–low	**29.006 (.000)**	26.569 (.000)	—					
Numeric, agree–disagree, low–high	46.687 (.000)	**43.394 (.000)**	2.122 (.145)	—				
Verbal, construct-specific, positive–negative	13.314 (.000)	14.340 (.000)	80.268 (.000)	107.747 (.000)	—			
Verbal, construct-specific, negative–positive	4.892 (.027)	5.597 (.018)	57.591 (.000)	81.536 (.000)	2.138 (.144)	—		
Numeric, construct-specific, high–low	7.832 (.005)	6.709 (.010)	7.100 (.008)	17.117 (.000)	**41.707 (.000)**	25.375 (.000)	—	
Numeric, construct-specific, low–high	18.692 (.000)	16.875 (.000)	0.917 (.338)	5.685 (.017)	61.541 (.000)	**42.062 (.000)**	2.692 (.101)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. $F = F$ statistic, $p = p$ value.

Table B7. Wald Test on Difference Between Regression Parameters for ρ.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	1.303 (.254)	—						
Numeric, agree–disagree, high–low	**4.809 (.028)**	1.061 (.303)	—					
Numeric, agree–disagree, low–high	3.995 (.046)	**0.705 (.401)**	0.036 (.849)	—				
Verbal, construct-specific, positive–negative	1.151 (.283)	0.004 (.949)	1.189 (.276)	.812 (.368)	—			
Verbal, construct-specific, negative–positive	3.113 (.078)	0.368 (.544)	0.182 (.670)	.056 (.814)	0.447 (.504)	—		
Numeric, construct-specific, high–low	5.104 (.024)	1.163 (.281)	0.001 (.973)	.051 (.820)	**1.299 (.254)**	0.218 (.641)	—	
Numeric, construct-specific, low–high	2.591 (.108)	0.225 (.635)	0.292 (.589)	.125 (.724)	0.288 (.592)	**0.015 (.903)**	0.338 (.561)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. $F = F$ statistic, $p = p$ value.

Table B8. Wald Test on Difference Between Regression Parameters for Response Time.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	0.056 (.813)	—						
Numeric, agree–disagree, high–low	**1.828 (.176)**	2.475 (.116)	—					
Numeric, agree–disagree, low–high	0.186 (.666)	**0.036 (.849)**	3.143 (.076)	—				
Verbal, construct-specific, positive–negative	6.827 (.009)	5.516 (.019)	15.611 (.000)	4.696 (.030)	—			
Verbal, construct-specific, negative–positive	8.998 (.003)	7.443 (.006)	18.932 (.000)	6.484 (.011)	0.128 (.721)	—		
Numeric, construct-specific, high–low	0.528 (.467)	0.228 (.633)	4.406 (.036)	0.081 (.777)	**3.730 (.054)**	5.369 (.021)	—	
Numeric, construct-specific, low–high	0.508 (.476)	0.224 (.636)	4.152 (.042)	0.081 (.775)	3.443 (.064)	**4.959 (.026)**	0.000 (.991)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. $F = F$ statistic, $p = p$ value.

Table B9. Wald Test on Difference Between Regression Parameters for Time Until First Click.

Experimental Groups	Verbal, Agree–Disagree, Positive–Negative *F* (*p*)	Verbal, Agree–Disagree, Negative–Positive *F* (*p*)	Numeric, Agree–Disagree, High–Low *F* (*p*)	Numeric, Agree–Disagree, Low–High *F* (*p*)	Verbal, Construct-Specific, Positive–Negative *F* (*p*)	Verbal, Construct-Specific, Negative–Positive *F* (*p*)	Numeric, Construct-Specific, High–Low *F* (*p*)	Numeric, Construct-Specific, Low–High *F* (*p*)
Verbal, agree–disagree, positive–negative	—							
Verbal, agree–disagree, negative–positive	0.571 (.450)	—						
Numeric, agree–disagree, high–low	**13.435 (.000)**	19.219 (.000)	—					
Numeric, agree–disagree, low–high	24.877 (.000)	**32.327 (.000)**	1.857 (.173)	—				
Verbal, construct-specific, positive–negative	2.993 (.084)	6.040 (.014)	3.532 (.060)	10.247 (.001)	—			
Verbal, construct-specific, negative–positive	4.767 (.029)	8.488 (.004)	2.140 (.144)	7.853 (.005)	0.184 (.668)	—		
Numeric, construct-specific, high–low	52.080 (.000)	62.466 (.000)	12.400 (.000)	4.444 (.035)	**28.731 (.000)**	24.770 (.000)	—	
Numeric, construct-specific, low–high	49.696 (.000)	59.565 (.000)	12.106 (.001)	4.473 (.034)	27.667 (.000)	**23.881 (.000)**	0.006 (.937)	—

Note. Wald tests on significance of differences between regression parameters of model presented in Appendix Table B1. $F = F$ statistic, $p = p$ value.

Data Availability

The data set generated and analyzed during this study is available on request from the corresponding author. Email: tobias.gummer@gesis.org.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Software Information

All analyses in this study were conducted using Stata Version 15.1.

Supplemental Material

Supplemental material for this article is available online.

Note

1. To test the robustness of our findings, we ran item-specific regressions for our response behavior indicators (i.e., 10 regressions per indicator): acquiescence, extremity, midpoint responding, and item nonresponse. For acquiescence, in 10 of 10 models, we also would have rejected H1.1. For extremity, we found support for H1.2 based on seven of 10 models. For midpoint responding, we also would have rejected H1.3 based on 10 of 10 models. Finally, for item nonresponse, we also would have rejected H1.4 based on 10 of 10 models. We interpret these findings to support our overall conclusions.

References

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, *20*(1), 139–181.

American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. The American Association for Public Opinion Research.

Antoun, C., Couper, M. P., & Conrad, F. G. (2017). Effects of mobile versus pc web on survey response quality: A crossover experiment in a probability web panel. *Public Opinion Quarterly*, *81*(S1), 280–306.

Baumgartner, H., & Steenkamp, J. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156.

Beatty, P. C., & Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item nonresponse. In R. M. Groves, D. A. Dillmann, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresposne* (pp. 71–86). Wiley.

Cabooter, E. F. (2010). *The impact of situational and dispositional variables on response styles with respect to attitude measures* [PhD thesis, Universiteit Gent]. http://lib.ugent.be/catalog/rug01:001809333

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, *72*(5), 1008–1032.

Cernat, A., & Liu, M. (2019). Radio buttons in web surveys: Searching for alternatives. *International Journal of Market Research*, *61*(3), 266–286.

Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, *35*(3), 357–377.

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (pp. 593–630). Emerald.

Fuchs, M., & Busse, B. (2009). The coverage bias of mobile web surveys across European countries. *International Journal of Internet Science*, *4*(1), 21–33.

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing 'too fast' responses and respondents from web surveys. *Public Opinion Quarterly*, *79*(2), 471–503.

Gummer, T., Quoß, F., & Roßmann, J. (2019). Does increasing mobile device coverage reduce heterogeneity in completing web surveys on smartphones? *Social Science Computer Review*, *37*(3), 371–384.

Gummer, T., Vogel, V., Kunz, T., & Roßmann, J. (2020). Let's put a smile on that scale: Findings from three web survey experiments. *International Journal of Market Research*, *62*(1), 18–26.

Hanson, T. (2015). Comparing agreement and item-specific response scales: Results from an experiment. *Social Research Practice*, *1*, 17–25.

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology*, *45*(7), 1029–1045.

Hofmans, J., Theuns, P., Baekelandt, S., Mairesse, O., Schillewaert, N., & Cools, W. (2007). Bias and changes in perceived intensity of verbal qualifiers effected by scale orientation. *Survey Research Methods*, *1*(2), 97–108.

Höhne, J. K., & Lenzner, T. (2018). New insights on the cognitive processing of agree/disagree and item-specific questions. *Journal of Survey Statistics and Methodology*, *6*(3), 401–417.

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

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

Knol, M. J., Pestman, W. R., & Grobbee, D. E. (2011). The (mis)use of overlap of confidence intervals to assess effect modification. *European Journal of Epidemiology*, *26*, 253–254.

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., & Bachner, Y. G. (2018). Effects of rating scale direction under the condition of different reading direction. *Methods, Data, Analyses*, *12*(1), 105–126.

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., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*(4), 526–538.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141–164). Wiley.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., & Conaway, M. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, *66*(3), 371–403.

Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 21–76). Lawrence Erlbaum Associates.

Lelkes, Y., & Weiss, R. (2015, July–September). Much ado about acquiescence: The relative validity and reliability of construct-specific and agree–disagree questions. *Research and Politics*, 1–8. https://doi.org/10.1177/2053168015604173

Liu, M., Lee, S., & Conrad, F. G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales. *Public Opinion Quarterly*, *79*(4), 952–975. https://doi.org/10.1093/poq/nfv034

Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey. *Social Science Computer Review*, *34*(1), 78–94. https://doi.org/10.1177/0894439315574248

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

McCarty, J. A., & Shrum, L. J. (2000). The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly*, *64*(3), 271–298.

Menold, N. (2020). Rating-scale labeling in online surveys: An experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. *Sociological Methods & Research*, *49*(1), 79–107.

Menold, N., Kaczmirek, L., Lenzner, T., & Neusar, A. (2014). How do respondents attend to verbal labels in rating scales? *Field Methods*, *26*(1), 21–39. https://doi.org/10.1177/1525822X13508270

Mohorko, A., de Leeuw, E., & Hox, J. (2013). Internet coverage and coverage bias in Europe: Developments across countries and over time. *Journal of Official Statistics*, *29*(4), 609–622. https://doi.org/10.2478/jos-2013-0042

Moors, G., Kieruj, N. D., & Vermund, J. N. (2014). The effect of labelling and numbering of response scales on the likelihood response bias. *Sociological Methodology*, *44*(1), 369–399.

O'Muircheartaigh, C., Gaskell, G., & Wright, D. B. (1995). Weighing anchors: Verbal and numeric labels of rating scales. *Journal of Official Statistics*, *11*(3), 295–307.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press.

Poushter, J. (2016). *Smartphone ownership and internet usage continues to climb in emerging economies*. Pew Research Center.

Rammstedt, B., Danner, D., & Bosnjak, M. (2017). Acquiescence response styles: A multilevel model explaining individual-level and country-level differences. *Personality and Individual Differences*, *107*(1), 190–194.

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*, *6*(3), 376–400.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, *4*(1), 61–79.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–186.

Schlosser, S., & Höhne, J. K. (2018). *ECSP—Embedded Client Side Paradata*. https://zenodo.org/record/1218941

Schuler, H., Thornton, G. C., III, Frintrup, A., & Mueller-Hanson, R. (2004). *AMI: Achievement Motivation Inventory. Technical and user's manual*. Hogrefe & Huber.

Schwarz, N., & Hippler, H.-J. (1995). The numeric values of rating scales: A comparison of their impact in mail surveys and telephone interviews. *International Journal of Public Opinion Research*, *7*(1), 72–74.

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*(4), 570–582.

Stange, M., Barry, A., Smyth, J., & Olson, K. M. (2018). Effects of smiley face scales on visual processing of satisfaction questions in web surveys. *Social Science Computer Review*, *36*(6), 756–766. https://doi.org/10.1177/0894439316674166

Thomas, R. K., & Barlas, F. M. (2018). *We've got your number: Can numeric labels replace semantic labels in scales?* [Paper presentation]. 19th General Online Research Conference, Cologne, Germany.

Thomas, R. K., & Barlas, F. M. (2019). *New mobile-friendly labels: Using numbers as labels* [Paper presentation]. 8th Conference of the European Survey Research Conference, Zagreb, Croatia.

Toepoel, V., & Dillman, D. A. (2010). How visual design affects the interpretability of survey questions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet* (pp. 165–190). Routledge.

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

Toepoel, V., & Funke, F. (2018). Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies*, *25*(2), 112–122.

Toepoel, V., Vermeeren, B., & Metin, B. (2019). Smileys, stars, hearts, buttons, tiles or grids: Influence of response format on substantive response, questionnaire experience and response time. *Bulletin of Sociological Methodology*, *142*, 57–74.

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368–393. https://doi.org/10.1093/po1/nfh035

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, *71*(1), 91–112. https://doi.org/10.1093/poq/nfl046

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

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*(3), 320–334. https://doi.org/10.1037/a0032121

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*(3), 236–247.

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.

Zhang, C., & Conrad, F. G. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135.

Author Biographies

Tobias Gummer is a senior researcher and team leader at GESIS—Leibniz Institute for the Social Sciences, Germany. His methodological research interests include survey design, data quality, nonresponse, and correction methods for biases.

Tanja Kunz is a senior researcher at GESIS—Leibniz Institute for the Social Sciences, Germany. Her current research interests include issues of visual design and data quality in web surveys, paradata, and questionnaire design.