Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories influences extreme responding

Weijters, Bert; Millet, Kobe; Cabooter, Elke

published in
International Journal of Research in Marketing
2021

DOI (link to publisher)
10.1016/j.ijresmar.2020.04.002

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

Link to publication in VU Research Portal

citation for published version (APA)
Weijters, B., Millet, K., & Cabooter, E. (2021). Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories influences extreme responding. *International Journal of Research in Marketing*, *38*(1), 85-103. https://doi.org/10.1016/j.ijresmar.2020.04.002

International Journal of Research in Marketing 38 (2021) 85–103

Contents lists available at ScienceDirect

IJRM

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories influences extreme responding

Bert Weijters [a,*], Kobe Millet [b], Elke Cabooter [c,d]

[a] *Ghent University, Dunantlaan 2, B9000 Ghent, Belgium*
[b] *Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, the Netherlands*
[c] *IESEG School of Management, Rue de la Digue 3, F-59000 Lille, France*
[d] *LEM-CNRS 9221, Rue de la Digue 3, F-59000 Lille, France*

ARTICLE INFO

Article history:
First received on February 19, 2019 and was under review for 7 months
Available online 27 April 2020

Area Editor: Joel C. Huber

Keywords:
Scale format
Likert
Vertical
Horizontal
Extreme response style
Visual distance

In four survey experiments we show that people generally answer more extremely to survey items presented in vertical versus horizontal Likert formats. Our findings suggest that this effect may be at least partly driven by differences in the visual range spanned by the response scale (i.e. the visual distance between endpoint response categories is larger in horizontal than in a vertical format). In addition, compared to traditional horizontal Likert data, vertical Likert data contain more variance, which is mainly non-substantive. As a result, data obtained with scale formats that have different distances between response categories (as is typically the case for vertical vs. horizontal formats) may lead to differences in measurement model parameter estimates like residual terms, and in some cases factor loadings and construct correlations. Based on these results, we provide recommendations on the use of response scale formats in online surveys, bearing in mind that several online survey tool providers promote the use of vertical Likert formats and even automatically change traditional horizontal formats of Likert-type items to vertical Likert formats when viewed on small screens (e.g., on mobile phones).

1. Introduction

The past decade has witnessed a tremendous increase in internet usage which has brought along an explosive growth of online surveys (Ramsey, Thompson, McKenzie, & Rosenbaum, 2016; Zhang, Kuchinke, Woud, Velten, & Margraf, 2017). The shift from offline to online has opened up new possibilities and has resulted in changes in survey practices. One such change is that online surveys have led to a greater diversity in scale format types that are being used by researchers (Blasius, 2012; Couper, Traugott, & Lamias, 2001; Evans & Mathur, 2005; Funke, 2016), for varying reasons: online survey tool developers promote the use of different scale formats and make it increasingly easy to implement them (Qualtrics, 2016; SurveyGizmo, 2016). Survey methods researchers propose and evaluate new or alternative scale formats based on criteria like data quality (e.g., item non-response rates), succinctness or respondent convenience (Liu, 2017; Liu & Conrad, 2016). In some cases, researchers have been found to choose scale for-

[*] Corresponding author.
E-mail addresses: bert.weijters@ugent.be, (B. Weijters), kobe.millet@vu.nl, (K. Millet), e.cabooter@ieseg.fr. (E. Cabooter).

https://doi.org/10.1016/j.ijresmar.2020.04.002

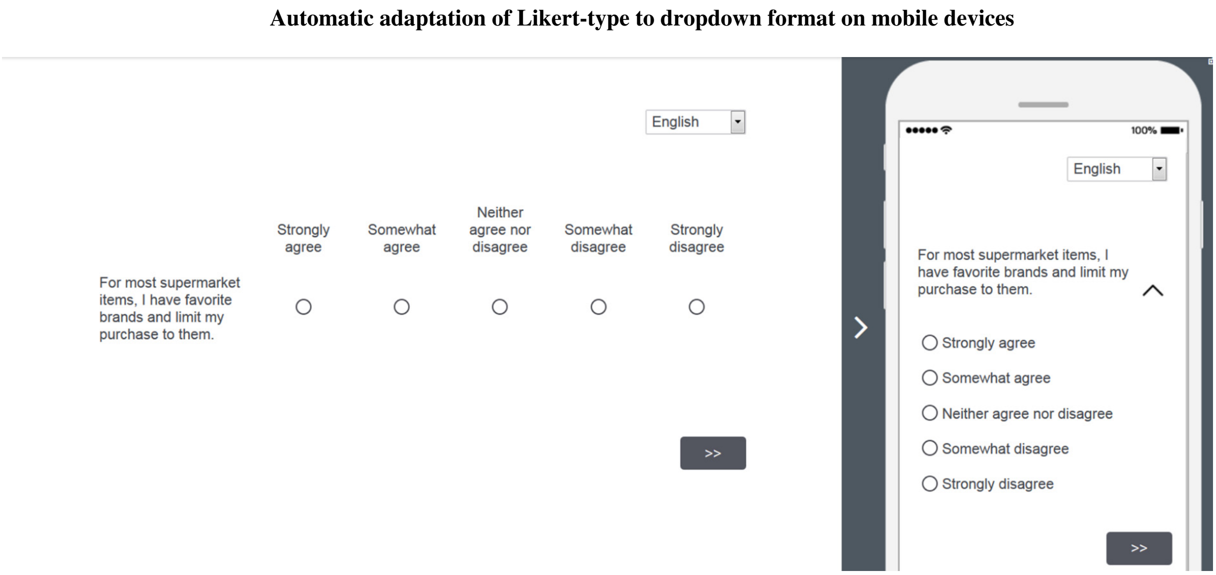
mats arbitrarily and without providing a clear rationale (Cabooter, Weijters, Geuens, & Vermeir, 2016). Finally, in other instances, researchers have been found to deliberately use a variety of different scale formats to prevent common method variance (Hulland, Baumgartner, & Smith, 2018; Rindfleisch, Malter, Ganesan, & Moorman, 2008).

One scale format that has gained prominence is the vertical Likert format (Couper, Antoun, & Mavletova, 2017; Gunn, 2002; Liu & Conrad, 2016). This scale format takes up less space and eliminates the need for horizontal scrolling which is convenient on small devices such as mobile phones (De Bruijne & Wijnant, 2014; Qualtrics, 2016). For this reason, most online survey packages automatically adapt the visual presentation of traditional horizontal Likert-type items to a vertical Likert format when viewed on smaller screens. To illustrate, Fig. 1 shows how Qualtrics automatically changes traditional (horizontal) Likert formats to a vertical 'accordion format' on smartphone screens at the time of the current study. Other popular survey tools apply similar automatic modifications for mobile users (SurveyGizmo, 2016).

There is reason to suspect, however, that responses to vertical Likert formats are not equivalent to responses to horizontal Likert formats. In the present set of experimental studies, we focus on the difference in Extreme Response Style (ERS) between horizontal vs. vertical Likert scale formats. ERS is defined as the tendency to give extreme responses (e.g., strongly disagree and strongly agree in a typical five point Likert-type scale format) regardless of item content. We argue that the difference in visual distance between response categories in the two scale formats contributes to the difference in ERS. More specifically, we expect that visual distance serves as a cue for the perceived level of extremeness of the scale endpoints. Compared to horizontal Likert scale formats, vertical Likert scale formats are typically more compact. As a consequence, vertical format endpoints tend to be perceived as less extreme and will therefore be more likely to be selected, because endpoint response categories are selected less often if they are perceived as more extreme. In line with this reasoning, we provide evidence that ERS is higher (a) in vertical Likert scale formats compared to horizontal Likert scale formats; (b) in more compact compared to less compact horizontal Likert type scales and (c) in more compact compared to less compact vertical Likert scales. Furthermore, we provide evidence that the additional ERS in vertical Likert formats often represents non-substantive variance, which leads to different estimates of important measurement model parameters (like residual terms and factor loadings) and different estimates of correlations between constructs in data obtained with different scale formats.

2. Conceptual framework

ERS has received considerable research attention in the last few years (Aichholzer, 2013; Cabooter, Millet, Weijters, & Pandelaere, 2016; de Jong, Steenkamp, Fox, & Baumgartner, 2008; de Langhe, Puntoni, Fernandes, & van Osselaer, 2011; Jin & Wang, 2014; Morren, Gelissen, & Vermunt, 2011; Naemi, Beal, & Payne, 2009; Wetzel, Carstensen, & Böhnke, 2013). A key reason is that ERS has been found to exert a pervasive and consistent influence on item responses across different scales included in the same questionnaire (Weijters, Geuens, & Schillewaert, 2010a; Wetzel et al., 2013; Zettler, Lang, Hülsheger, & Hilbig, 2015). ERS is influenced by dispositional variables like demographics (Weijters, Geuens, & Schillewaert, 2010b), personality (Naemi et al., 2009) and culture (de Jong, Steenkamp, Fox, & Baumgartner, 2008), but also by situational variables related to the measurement situation (Baumgartner & Steenkamp, 2001; Cabooter et al., 2016; Weijters, Cabooter, & Schillewaert, 2010). Our focus is on the latter type of antecedent (i.e. situational variables), more specifically response scale format characteristics.



Automatic adaptation of Likert-type to dropdown format on mobile devices

English

	Strongly agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Strongly disagree
For most supermarket items, I have favorite brands and limit my purchase to them.	○	○	○	○	○

>>

English

For most supermarket items, I have favorite brands and limit my purchase to them.

○ Strongly agree
○ Somewhat agree
○ Neither agree nor disagree
○ Somewhat disagree
○ Strongly disagree

>>

Fig. 1. Automatic adaptation of Likert-type to dropdown format on mobile devices.

An important determinant of the way respondents use response scales is the interpretation of the endpoint categories (de Langhe et al., 2011; Janiszewski, Silk, & Cooke, 2003; Tourangeau, Couper, & Conrad, 2004; Tourangeau, Couper, & Conrad, 2007; Weijters, Geuens, & Baumgartner, 2013). Respondents are less likely to endorse endpoint response categories that are perceived as more extreme as such extreme positions are typically seen as more exceptional (de Langhe et al., 2011; Weijters, Geuens, & Baumgartner, 2013). We posit that the visual distance between response categories of a scale format will affect the perceived extremity of its endpoints as it may signal to respondents the conceptual distance covered by the rating scale. Consequentially, we expect that respondents will be less likely to endorse endpoints when the visual distance between response categories is increased.

Previous survey methods research provides initial evidence demonstrating potential effects of response option spacing and position on responses. In an interesting experiment (experiment 3), Tourangeau et al. (2004) manipulate the rating scale format used for a question concerning respondents' expected chance of falling ill in the coming year. They compare a traditional, equally spaced rating scale format to an experimental (rather unusual) format with uneven spacing. Both formats use the same response options (certain, very likely, probable, even chance, possible, unlikely, impossible), but in the uneven format, the high-probability options, shown on the left hand side, are condensed, such that the visual midpoint (i.e., the geometric middle of the rating scale) no longer corresponds to the conceptual midpoint (i.e., the middle option ‘even chance’), and the low-probability options shift towards the visual middle. In support of their proposed “middle means typical” heuristic, they find that when “unlikely” was displaced towards the visual midpoint, it was selected more often, presumably because it seemed to represent a less extreme probability. Our aim with the current study, is to further explore how the distance of a response option from the midpoint affects responses and to show its effect in scale formats where the response options are symmetrically spaced around the midpoint (which is more common than the experimental unevenly spaced format just discussed).

To resume, respondents tend to interpret the visual midpoint of a rating scale as the conceptual midpoint (Tourangeau et al., 2004). Adjacent response categories (e.g., slightly disagree and slightly agree in some scale formats) are visually and conceptually closer to the midpoint than are the endpoint categories (e.g., strongly disagree and strongly agree). Given this clear relation between visual distance and conceptual distance, it seems natural for respondents to infer that endpoint categories that are visually more distant from the midpoint are also conceptually more distant from it. In line with this, it is likely that a visually more expanded rating scale will lead respondents to perceive the endpoints as more extreme compared to the endpoints of a scale that is relatively compact, even when using the same response category labels. As a result, the endpoint categories are more likely to be selected if response scales are compact, as depicted in Fig. 2. We call this the *response category distance effect*.

The distance in between endpoint categories is usually different in traditional, horizontal Likert-type than in vertical Likert-type scale formats. A primary reason is that the minimum horizontal range covered by a response category label corresponds to the longest label (e.g., ‘strongly disagree’), whereas the minimal vertical range covered by a response category label spans only the height of one letter. Thus, a crucial difference between horizontal vs. vertical Likert-type scale formats is the visual distance typically covered by the rating scale: the endpoints in a horizontal Likert scale format are visually further apart from the midpoint compared to a vertical Likert scale format, where mere line breaks delineate different response categories. The resulting difference

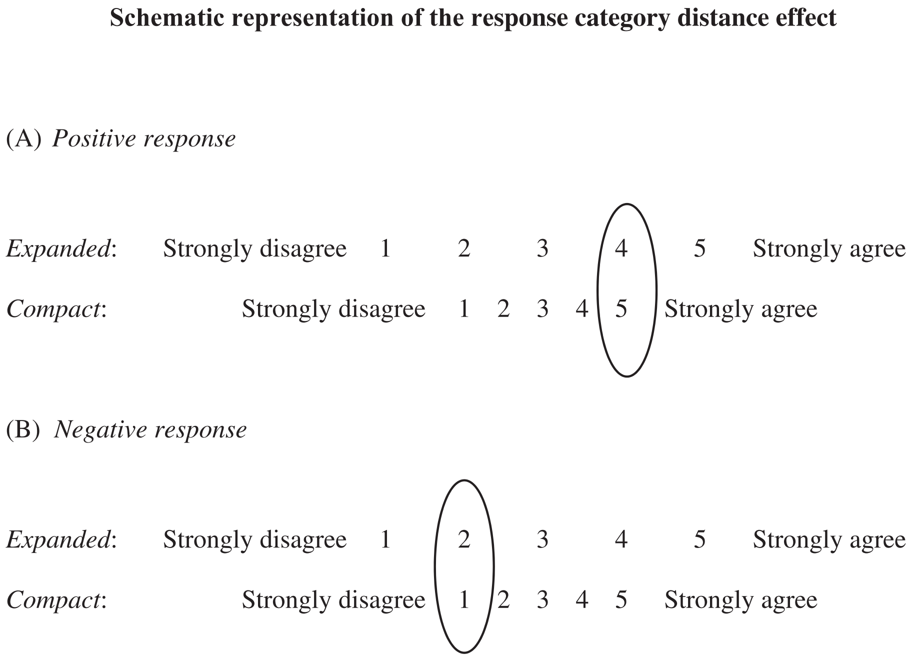

Schematic representation of the response category distance effect
(A) Positive response
Expanded: Strongly disagree 1 2 3 4 5 Strongly agree
Compact: Strongly disagree 1 2 3 4 5 Strongly agree
(B) Negative response
Expanded: Strongly disagree 1 2 3 4 5 Strongly agree
Compact: Strongly disagree 1 2 3 4 5 Strongly agree

Fig. 2. Schematic representation of the response category distance effect.

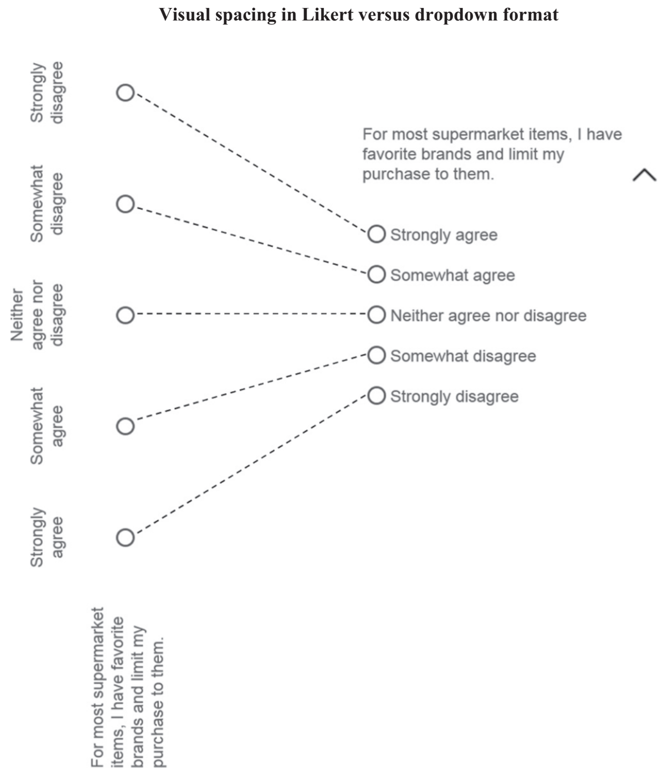

Visual spacing in Likert versus dropdown format
Strongly disagree
Somewhat disagree
Neither agree nor disagree
Somewhat agree
Strongly agree
For most supermarket items, I have favorite brands and limit my purchase to them.
For most supermarket items, I have favorite brands and limit my purchase to them.
Strongly agree
Somewhat agree
Neither agree nor disagree
Somewhat disagree
Strongly disagree

Fig. 3. Visual spacing in Likert versus Likert dropdown format. Note: Comparison of a horizontal Likert-type scale format, rotated 270° for comparability (left hand side) versus the same item shown in vertical format (right hand side, also called accordion view) that is automatically used in Qualtrics when taking the survey on a mobile device. Defaults settings were used for designing the item. The response categories in the vertical format can be collapsed and reopened by clicking the arrow above the response categories (i.e., dropdown format).

in visual distance between the endpoints tends to be remarkably large, as illustrated in Fig. 3. As a result of this difference in response category distance, an identically labeled response category (e.g., strongly disagree) may be perceived as more extreme in the horizontal Likert scale format than in the vertical dropdown Likert scale format. We predict that, in line with the response category distance effect, ERS will consequently be larger in vertical than in horizontal Likert-type scale formats.[1]

The resulting ERS difference is important because ERS can affect survey data by adding variance to item response data (Greenleaf, 1992a). The added variance may be substantive, i.e., variance due to individual differences related to the construct that the item purports to measure, but it may also be non-substantive, i.e. variance due to random error, item-specific variance and/or method variance (where method variance is defined as individual variation in the way rating scales are used irrespective of item content). Although this effect of ERS is well established, the results presented by Greenleaf (1992a) focus on between-respondents individual variation in ERS, and are limited to one response scale format (a six-category endpoint labeled Likert-type scale format). It remains an open question whether the additional variance we expect to find in vertical data will be substantive or non-substantive variance.

3. Overview of empirical studies

We present four experiments. Study 1 tests the general notion that presenting commonly used marketing scales in a vertical Likert scale format (vs. a horizontal Likert format) leads to higher ERS, and investigates how this affects measurement model estimates. Study 2a demonstrates the response category distance effect more directly by showing that the visual width of a horizontal Likert scale format influences ERS, which in turn affects residual terms. Study 2b replicates the effects found in study 2a across multiple marketing scales. Finally, study 3 compares a vertical Likert format with extra spacing between response categories to a regular vertical Likert format and a regular horizontal Likert format. This last study shows how increased spacing between response options reduces ERS in vertical Likert formats, which makes the resulting response more comparable to those obtained by means of a horizontal Likert scale format.

[1] Although distances between response categories are along the horizontal axis in traditional horizontal Likert-type items and along the vertical axis in the vertical Likert items, there is no decisive reason why this would neutralize the relation between visual and conceptual distance. After all, respondents mainly expect response categories to be ordered in a logical sequence, be it from left to right or from high to low (Tourangeau et al., 2004). Nevertheless, in our experiments we will counter the potential confounding effect of vertical vs. horizontal direction.

In all our experiments, we manipulate the scale format while keeping constant the device that participants use. This way, scale format is not automatically adapted by the survey software (Qualtrics) and the device that participants use is not confounded with the scale format that they see. We use a 5-point fully labeled scale as this format is commonly used and has been shown to be well-suited for all population types (Weijters, Cabooter, & Schillewaert, 2010). In addition, labeling all response categories reduces ambiguity in the interpretation of response categories (Moors, Kieruj, & Vermunt, 2014). Since all of the response categories are labeled, it seems likely that our findings are conservative, since response category distance might affect ERS even more if the response categories are unlabeled.

We measure ERS as the proportion of items for which a participant selects a response located at the (low or high) end of the scale (i.e., 'strongly disagree' or 'strongly agree') (Baumgartner & Steenkamp, 2001). To operationalize ERS it is important that content is controlled for. This way, individual differences in the proportion of endpoints cannot be attributed to item content (but to response styles). In experiments that focus on scale format effects, content can be controlled for simply by keeping it constant across conditions (Cabooter, Weijters, et al., 2016; Weijters, Cabooter, & Schillewaert, 2010). That is, if the proportion of extreme responses differ when participants respond to the same scale in randomly assigned alternative formats, this difference can only be attributed to scale format (and not to content). This is the approach taken in one of our studies (study 2a). To enhance generalizability, for the other three studies we based ERS measures on responses to a large number of items that are diverse in content (study 3), or by using items from several multi-item scales (study 1 and study 2b), where items from the same scale share content, but the different scales presumably have limited overlap in terms of content (Baumgartner & Steenkamp, 2001; de Jong, Steenkamp, Fox, & Baumgartner, 2008; Greenleaf, 1992a; Greenleaf, 1992b; Weijters, Schillewaert, & Geuens, 2008).

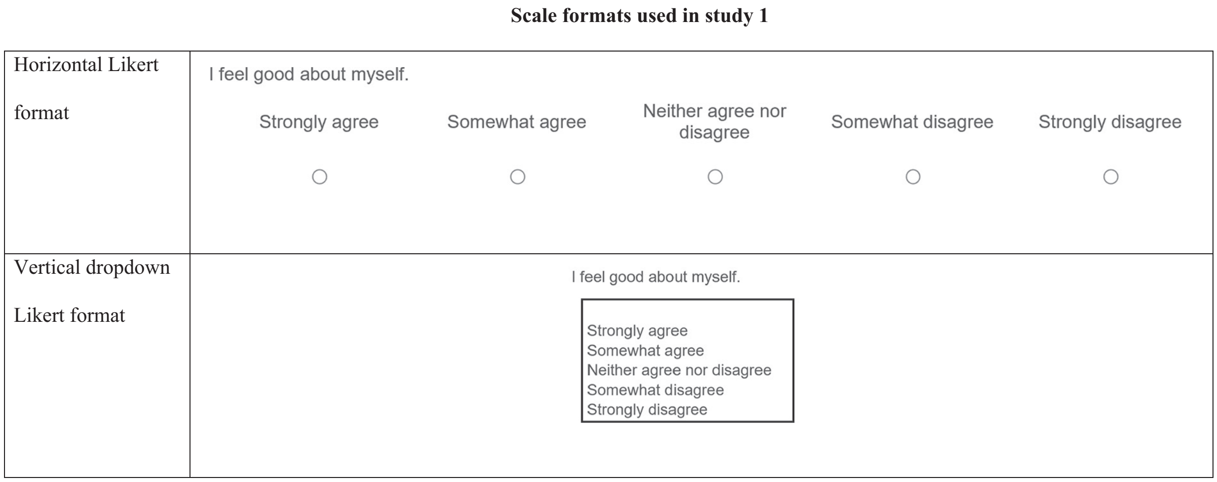
4. Study 1

In study 1 we test whether a vertical Likert format (vs. a horizontal Likert format) leads to higher ERS in commonly used marketing scales, and if so, whether the increased ERS in vertical Likert data affects estimates of measurement model parameters (factor loadings and residual terms).

4.1. Method

In the current study, to avoid confounding ERS with directional bias (acquiescence), we selected marketing scales with reversed items (Baumgartner & Steenkamp, 2001). Specifically, we selected 14 measurement scales that contain at least two reversed items and that are commonly used in marketing research. For each selected scale, we included two non-reversed items and two reversed items (i.e., for scales consisting of more than four items, we included only four items to ensure comparability across scales). The scale items are reported in Appendix A. Participants were randomly assigned to either vertical or horizontal Likert format scales (between subjects), both using the same five response category labels (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree). We showed one scale of four items on a screen (14 screens with 56 items in total of which 28 are reversed items), presenting the scales in randomized order (presenting each item as a separate question). As shown in Fig. 4, the horizontal format in study 1 is a radio button horizontal Likert format. The vertical format uses a dropdown list (select objects), comparable with the accordion format Qualtrics uses on mobile devices.

We collected data from the MTurk panel in the U.S. with the explicit instruction to use a desktop/laptop computer. From the initial 520 completed questionnaires we deleted 16 inattentive respondents (in the horizontal vs. vertical conditions, respectively


Scale formats used in study 1

Horizontal Likert format	I feel good about myself. Strongly agree ○ Somewhat agree ○ Neither agree nor disagree ○ Somewhat disagree ○ Strongly disagree ○
Vertical dropdown Likert format	I feel good about myself. Strongly agree Somewhat agree Neither agree nor disagree Somewhat disagree Strongly disagree

Fig. 4. Scale formats used in study 1. Note: The horizontal format in study 1 is a radio button horizontal Likert format. The vertical format uses a dropdown list (select objects), comparable with the accordion format Qualtrics uses on mobile devices.

eight and five responded negatively to the question 'did you seriously participate in this study?'; three and zero responded positively to the question 'were you in any way distracted during the survey?'). In the resultant sample, $N = 504$, age ranged from 18 to 77 (M = 37.2, SD = 12.8) and 44.4% were women. As per instructions, all respondents were desktop/laptop users.

4.2. Results

We compute ERS as the proportion of items (out of 56 items) for which an endpoint response was endorsed (i.e., strongly agree or strongly disagree). As hypothesized, ERS is higher in the vertical format ($M_{vertical} = 0.362$, SE = 0.011) than the horizontal format ($M_{horizontal} = 0.311$, SE = 0.013). An independent samples *t*-test shows that the difference between the vertical vs. horizontal format is statistically significant ($t(502) = 3.050$, $p = .002$), with a small to medium effect size (Cohen's D = 0.272).

To assess the extent to which this ERS difference can bias substantive findings, we perform an additional analysis in which we investigate the nature of the increased item variance due to the higher ERS in vertical formats. In particular, we compare measurement parameter estimates in a confirmatory factor analysis (CFA) model, which we run for each scale separately. The model includes a residual covariance between the two negatively worded items to account for method effects (Marsh, 1996). In a CFA framework, item variance can be decomposed into the following components (Baumgartner & Weijters, 2017a; Baumgartner & Weijters, 2017b): (a) Substantive variance represents variance due to the construct of interest that an item purports to measure; if substantive variance is higher in the vertical format, absolute average factor loadings should be higher in this condition. (b) Unique variance related to an item represents item-specific variance that is not related to the substantive construct of interest, including random error; if unique variance is higher in the vertical format, the residual variance terms of the items should be higher in this condition. (c) Method variance, in the current context, refers to variance that is shared by items that are worded in the same direction (i.e., reversed items); if method variance is higher in the vertical format, the residual covariance between the two reversed items in each scale should be higher in this condition. (d) Composite reliability or C.R. is an indicator of internal consistency and is higher if items measuring the same latent construct contain a large proportion of substantive variance (i.e., high absolute factor loadings) relative to their unique variance. C.R. is a generalization of coefficient alpha to a situation in which items can have different loadings on the underlying factor and is therefore the preferred indicator of internal consistency (Baumgartner & Weijters, 2017a).

We run the same CFA model for each scale, using Mplus 8.0, and store the parameter estimates. Based on the results, scale 6 (the involvement facet of the spousal conflict arousal scale; see Appendix A) is omitted from further analysis at this point due to problematic estimates (more precisely, a negative residual variance estimate). The CFA model fits the data well for the remaining 13 scales, with average $\text{chi}^2(2) = 2.81$, $p = .245$, RMSEA = 0.027, CFI = 0.998, TLI = 0.991, SRMR = 0.008. Using the 13 scales as the unit of analysis, vertical Likert vs. horizontal Likert format as the independent variable, and the parameter estimates corresponding to the aforementioned variance components as dependent variables, we run paired samples *t*-tests for (a) the average absolute factor loadings (averaged across the four items per scale), (b) the residual variance terms (also averaged across the four items per scale), (c) the residual covariance terms, and (d) composite reliabilities. As shown in Table 1, the average absolute factor loadings are not significantly different (and neither are the residual covariance terms), whereas average residual variances are significantly higher in the vertical Likert than the horizontal Likert format. As a result of this, composite reliabilities are significantly lower in the vertical Likert scale format than in the horizontal Likert scale format. These results show that estimates of important measurement model parameters may be affected by the use of vertical formats and that internal consistency may suffer as a consequence. If distinct subgroups within a sample (e.g., different nationalities, different age groups,…) use different scale formats, this can consequently result in artificial differences in measurement results (including differences in internal consistency).

5. Study 2a

The results of study 1 show that the use of a vertical format leads to higher ERS compared to a horizontal Likert format. Respondents are more likely to select endpoint responses in vertical formats, regardless of what is being measured. We also demonstrated downstream effects of the ERS difference on measurement parameter estimates in a CFA model. Yet, it remains unclear what causes the increased ERS. The aim of study 2 is to explore if the response category distance effect is one of the potential drivers of this effect. Presuming that the effect is driven by visual distance, we should be able to see a similar effect when simply

Table 1
Measurement parameters in vertical vs. horizontal Likert data (study 1).

	$M_{horizontal}$	$M_{vertical}$	Difference	
			t	*p*
Absolute loading	0.688	0.658	−1.543	.149
Residual variance	0.592	0.717	4.791	<.001
Residual covariance	0.245	0.266	0.321	.753
Composite reliability	0.762	0.730	−3.051	.010

Note: The marketing scales in Appendix A (except scale 6) are the unit of analysis. The reported loadings and residual variance estimates are based on the means across the four items of each scale. The inter-item correlations by condition are reported in the web appendix.

changing distance between response labels in horizontal scales. The current study tests exactly this idea. In line with our reasoning, we will first demonstrate in an experimental study that the visual width of a horizontal Likert scale format influences ERS, which in turn affects measurement parameter estimates. Study 2a tests this proposition on data for a single scale in which visual distance was manipulated by enlarging the response options in the expanded condition. Study 2b tries to extrapolate the results to a larger number of scales that are commonly used in marketing research.

5.1. Method

We recruited U.S. participants on MTurk with the explicit instruction to use a desktop/laptop computer. From the initial sample of $N = 403$ completed questionnaires, we deleted 18 inattentive respondents (who answered negatively to the question 'Did you seriously participate in this study?' and/or positively to the question 'Were you in any way distracted during the survey?'; there were respectively eight and ten inattentive respondents in the expanded vs. compact conditions), and three respondents using a mobile device (as identified in the paradata concerning the operating system; we dropped one mobile respondent from the expanded and two from the compact condition). In the resulting sample, $N = 382$, age ranged from 19 to 75 years (M = 36.4, SD = 11.9) and 42.7% were women.

In the questionnaire, participants answered Likert items measuring current mood valence, adapted from Weijters, Baumgartner, and Schillewaert (2013); half the items are reverse worded: 'I feel contented right now,' 'I feel troubled at this moment,' 'I feel pleased at present,' 'I currently feel annoyed,' 'At the moment, I feel happy,' 'Presently, I feel sad'. Scale format width was experimentally manipulated (compact vs. expanded): the graphical width of the response category boxes was varied, as depicted in Fig. 5, while keeping the response categories' labels constant ('strongly disagree,' 'disagree,' 'neither agree nor disagree,' 'agree,' and 'strongly agree'). All questions appeared on the same screen, although each item was presented as a separate question.

5.2. Results & discussion

We compute ERS as the proportion of items for which an endpoint response was endorsed (i.e., strongly agree or strongly disagree). As hypothesized, ERS is higher in the compact format ($M_{compact} = 0.335$, SE = 0.025) than the expanded format ($M_{expanded} = 0.251$, SE = 0.023). An independent samples *t*-test shows that the difference between the compact vs. expanded format is statistically significant ($t(380) = 2.481$, $p = .014$), with a small to medium effect size (Cohen's D = 0.254).

To further investigate how the increased ERS in the compact format affects measurement quality, we apply the same rationale as in study 1 and specify a CFA model using scale format (expanded vs. compact Likert scale format) as the grouping variable, and including a residual covariance between the two reversed items (Marsh, 1996). Two items (contented and troubled) were omitted due to problematic fit (i.e., RMSEA > 0.100; including these items led to similar conclusions, i.e., higher residual variance in the compact format data). The resulting model fits the data well (chi^2(3) = 3.757, $p = .2890$, RMSEA = 0.036, CFI = 0.999, TLI = 0.997, SRMR = 0.021). Using the model constraint functionality in Mplus 8.0, we test for between-group differences in (1) average absolute factor loadings, (2) average residual variance terms, (3) residual covariance, and (4) composite reliability. The results are reported in Table 2. While the average absolute factor loadings are not significantly different, the compact format leads to significantly higher residual variance and residual covariance terms. Composite reliability is significantly lower in the compact format as a result. These results are in line with the hypothesized *response category distance effect*, which states that respondents are less likely to use the endpoints of a rating scale if the endpoints are visually further apart.

6. Study 2b

In study 2a, we found that a more compact Likert format results in more ERS, as compared to a more expanded horizontal Likert format, using responses to a single mood scale. In study 2b, we aim to test whether this effect generalizes to several often used marketing scales. In addition, we manipulated visual distance differently (compared to study 2a) namely by adding

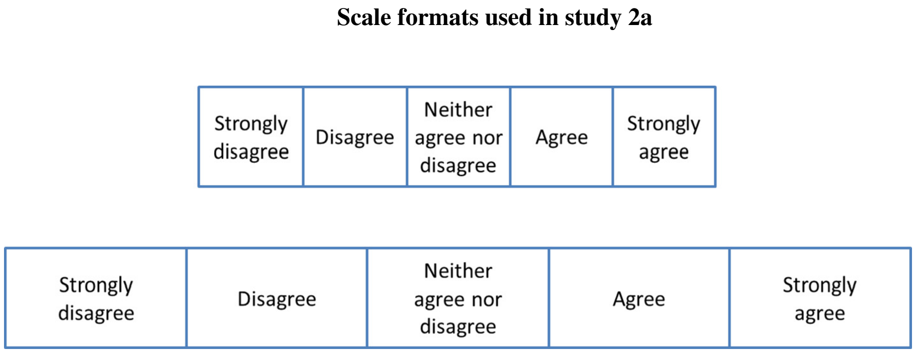


Fig. 5. Scale formats used in study 2a. Note: Study 2a uses two horizontal scale formats with select boxes of varying width: Expanded (bottom) vs. compact (top) horizontal Likert scale format, with the compact format having a width of just under 60% of the expanded format.

Table 2
Measurement parameters in expanded vs. compact Likert format (study 2a).

	$M_{expanded}$	$M_{compact}$	Difference	
			t	*p*
Absolute loading	0.830	0.785	−0.663	.507
Residual variance	0.298	0.416	3.246	.001
Residual covariance	0.119	0.336	3.035	.002
Composite reliability	0.701	0.605	−2.160	.031

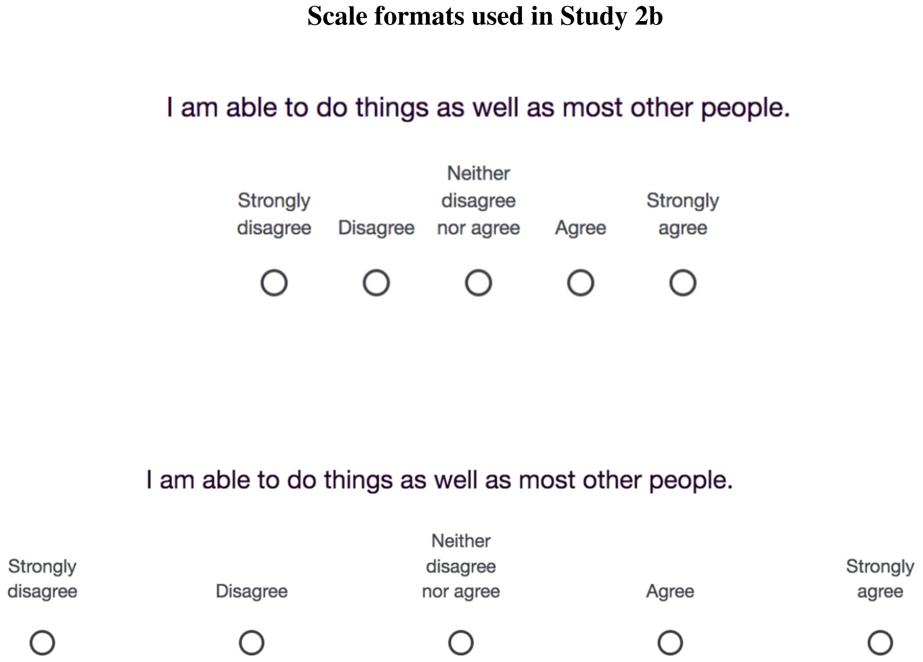
Note: The reported loadings and residual variance estimates are based on the means across the scale items. The inter-item correlations by condition are reported in the web appendix.

extra space between the response options (see Fig. 6 vs. Fig. 5). In the current survey experiment, we compare responses to several marketing scales in a compact vs. expanded horizontal Likert format to once more test the hypothesis that reducing the visual distance between response categories leads to more extreme responses. In addition, we again investigate whether and how this affects measurement model parameter estimates (loadings and residual terms). Finally, we explore whether using the compact vs. expanded format may also influence estimates of correlations between constructs.

6.1. Method

We recruited UK residents on Prolific with the explicit instruction to use a desktop/laptop computer. Of the $N = 819$ completed surveys, we deleted five respondents (two from the compact and three from the expanded condition) who answered negatively to the question “Did you maximize your browser when asked for in the beginning of the study?” at the end of the survey, leaving us with $N = 814$ usable responses (no mobile users were detected). In this sample, 541 (66.5%) are women and age ranges from 18 to 74 (M = 36.5, SD = 13.4).

The survey contains the short versions of 15 commonly used marketing scales, each consisting of two regular and two reversed items, for a total of 60 items. Appendix A reports the items and scales. Respondents were randomly assigned to either a compact or an expanded version of the horizontal Likert scale format, as depicted in Fig. 6, and as in our previous studies we kept the response category labels constant (‘strongly disagree,’ ‘disagree,’ ‘neither agree nor disagree,’ ‘agree,’ and ‘strongly agree’). We presented each item as a separate question and displayed 4 items per screen (all 60 items were presented in fully randomized order).



Scale formats used in Study 2b

I am able to do things as well as most other people.

I am able to do things as well as most other people.

Fig. 6. Scale formats used in study 2b. Note: Study 2b uses two horizontal scale formats of varying width with radio buttons: Expanded (bottom) vs. compact (top) horizontal Likert scale format, with the compact format having a width of just under 60% of the expanded format.

6.2. Results

We run three sets of analyses, first testing the main hypothesis that ERS is higher in the compact vs. expanded format, then investigating the effects thereof on factor loadings and residual terms, and finally exploring the effects on factor correlations.

First, we test our main hypothesis, which states that extreme responding is higher in a compact vs. expanded horizontal Likert scale format. We compute ERS as the proportion of endpoint responses over the 60 items in the questionnaire. ERS is higher in the compact format ($M_{compact} = 0.176$, SD = 0.144) than in the expanded format ($M_{expanded} = 0.153$, SD = 0.125), and this difference is statistically significant ($t(812) = 2.487$, $p = .013$), with small effect size (Cohen's D = 0.174).

Second, we replicate the analysis of CFA parameter estimates in previous studies to investigate the effect of the ERS difference between formats on estimates of measurement model parameters of interest. We specifically compare item loadings, item residual variances and the residual covariance between the two reversed items in each scale. To do so, we specify a one-factor CFA model. In the model, the four items of a scale are the indicators of a latent construct. The two reversed items have a residual covariance to account for shared method variance. We estimate this model for each scale separately, using experimental condition (compact vs expanded Likert format) as the grouping variable, and save the model fit indices and parameter estimates. Overall, the model shows good fit to each scale, with average $chi^2(2) = 2.94$, TLI = 0.995, CFI = 0.998, RMSEA = 0.026, SRMR = 0.007. The nostalgia scale (scale 8 in Appendix A) has a negative residual variance estimate and is therefore excluded from further analyses.

Using the 14 remaining scales as the unit of analysis, compact vs. expanded Likert format as the independent variable, and the parameter estimates as dependent variables, we run paired samples *t*-tests for (1) the average absolute factor loadings (averaged across the four items per scale), (2) the residual variance terms (also averaged across the four items per scale), (3) the residual covariance terms, and (4) composite reliabilities. As shown in Table 3, the average absolute factor loadings are significantly higher in the compact vs. expanded format ($p = .02$). But at the same time, the residual variance terms are also significantly higher in compact vs. expanded format ($p < .001$). Higher loadings contribute positively to composite reliability (C.R.), but higher residuals contribute negatively to C.R., and the net result here is that composite reliabilities are not significantly different across conditions. Finally, the residual covariance terms are nearly identical (and not significantly different). To sum up, these results indicate that item responses in the compact format contain more unique variance but also some more substantive variance (while there is no difference in the amount of method variance, i.e., the residual covariance between reversed items).

Third, to explore whether and how the difference in ERS due to the effect of inter-category distance affects factor correlations, we estimate a CFA model that simultaneously incorporates all 14 scales (i.e., excluding the nostalgia scale, which was dropped in the previous step) as latent factors with four indicators each. Experimental condition (compact vs. expanded format) is used as the grouping variable. To account for reversed item method bias in a parsimonious way, we drop the residual covariances between the reversed item pairs in each scale and instead specify a method factor on which all items have unit loadings, regardless of their wording direction (Maydeu-Olivares & Coffman, 2006; Weijters, Baumgartner, & Schillewaert, 2013).

The resulting model can be thought of as a large nomological network including many commonly studied marketing variables. The model includes more constructs than is common in this type of model, and probably largely because of this, the model shows only borderline acceptable fit to the data (see below). But to avoid overfitting and to limit researchers' degrees of freedom, we refrain from ad hoc model adaptations and focus instead on investigating the effect of scale format. To do so, we run a sequence of model comparisons to evaluate invariance of the factor loadings (i.e., metric invariance), item intercepts (i.e., scalar invariance) and residual variance terms (i.e., residual invariance) across the two experimental conditions, as reported in Table 4.

In line with the results in Table 3, the factor loadings are non-invariant at $p < .05$ (see the row labeled metric invariance in Table 4; $D(chi^2)$ represents the deterioration in model fit when imposing factor loading equality across the two conditions) and the residual variance terms are non-invariant with $p < .01$ (see the row labeled residual invariance in Table 4). This measurement non-invariance between the compact vs. expanded scale format implies that in instances where different groups of participants use differently spaced formats, results may become essentially incomparable, since measurement invariance is a requirement for parameter estimates to be comparable (Steenkamp & Baumgartner, 1998). The non-invariance occurs despite the fact that the internal consistency of the scales (C.R.) is not significantly or notably different across conditions, which might erroneously reassure researchers who only check internal consistencies (without performing measurement invariance tests).

To evaluate how findings might be biased by the change in scale format if researchers are not aware of the invariance issue, we compare the factor correlations in both experimental groups. To do so, we focus on the parameter estimates in the scalar invariant model (since residual variance invariance is clearly rejected and is usually not imposed in marketing research). In particular, using the model constraint functionality in Mplus 8.0, we compute the factor correlations and test the between-group difference of each of the factor correlations (i.e., all correlations between the fourteen constructs in our model, for a total of $N = 91$ correlations;

Table 3
Measurement parameters in expanded vs. compact horizontal Likert format (study 2b).

	$M_{expanded}$	$M_{compact}$	$t(13)$	p
Absolute loading	0.610	0.640	−2.649	.020
Residual variance	0.509	0.546	−5.307	<.001
Residual covariance	0.163	0.156	0.532	.603
Composite reliability	0.433	0.441	−0.961	.354

Note: The inter-item correlations by condition are reported in the web appendix.

Table 4
Measurement invariance expanded vs. compact horizontal Likert format (study 2b).

Model	Chi^2	DF	D(Chi^2)	DF	P	CFI	TLI	RMSEA	SRMR
Unconstrained	4922.731	2784				0.877	0.864	0.043	0.063
Metric invariance	4998.633	2840	75.902	56	0.0395	0.876	0.866	0.043	0.066
Scalar invariance	5048.878	2896	50.245	56	0.6915	0.876	0.869	0.043	0.067
Residual invariance	5141.067	2952	92.189	56	0.0017	0.874	0.869	0.043	0.069

since this analysis merely aims to illustrate possible outcomes without testing any specific hypotheses, we did not implement a family-wise error correction). As we found in the previous analysis, higher ERS in the compact format shows up as part substantive variance (stronger factor loadings), part noise (greater residual variance terms; see above), so it is not easy to predict how correlations will be affected. The averaged difference between factor correlations in the compact vs. expanded format is near zero (d = 0.01). Yet, this average hides a large variety in differences between correlations. Fig. 7 shows a histogram of between-format differences in correlations (i.e., each observation is the difference in correlation $r_{compact} - r_{expanded}$ for a given pair of factors). Indeed, results show that some correlations are higher whereas others are lower: four correlations are significantly lower at $p < .05$ and another two at $p < .10$ in expanded vs. compact format, whereas three correlations are significantly higher at $p < .05$ and another five at $p < .10$ in expanded vs. compact format. To illustrate the extent to which ERS can affect specific correlations, consider the correlation between the Self-Monitoring factor and the Life Orientation Test factor which equals $r_{compact} = 0.331$ ($p < .001$) in the compact format versus $r_{expanded} = 0.559$ ($p < .001$) in the expanded format (this difference is statistically significant, $p = .002$). In other words, the relation between self-monitoring and optimism is significantly and substantially suppressed in the compact condition. By contrast, the correlation between Need for Precision and Resisting Requests for Compliance is $r_{compact} = 0.422$ ($p < .001$) in the compact format, but only $r_{expanded} = 0.161$ ($p = .017$) in the expanded format (this difference is significant as well, $p = .005$, albeit in the other direction). That is, here the relation between Need for Precision and Resisting Requests for Compliance is significantly inflated in the compact condition relative to the expanded condition. In still other examples, a factor correlation may be statistically significant in one format, but not in the other (e.g., the correlation between the Enjoyment facet of Coupon antecedents and Self-Monitoring, $r_{compact} = -0.023$, $p = .712$, $r_{expanded} = 171$, $p = .003$, difference $p = .021$).

6.3. Discussion study 2b

The results of the current study provide further support for the category distance effect: reducing the distance between response categories in a horizontal Likert format increases the proportion of extreme responses. In a CFA model, we find that the

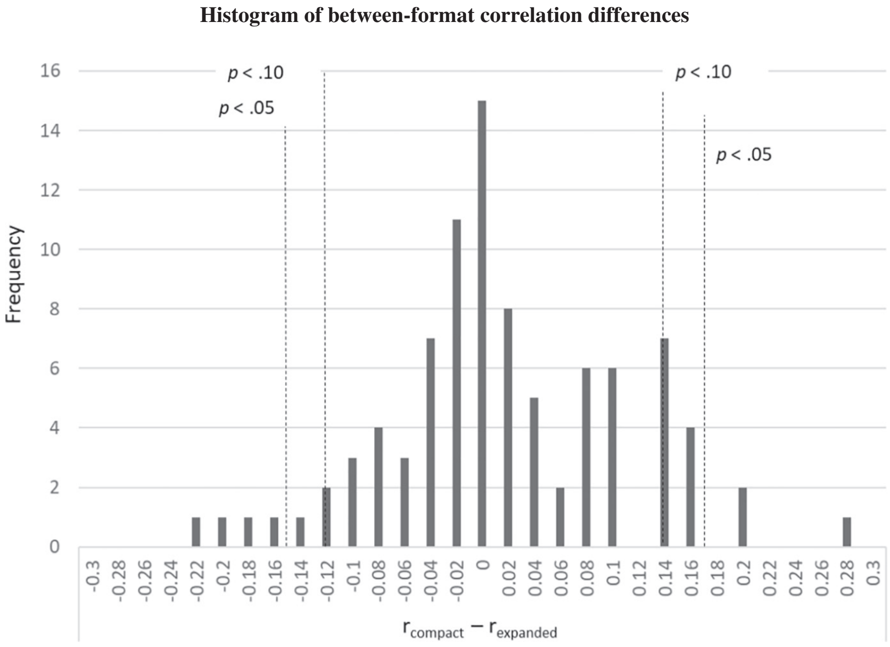


Fig. 7. Histogram of between-format correlation differences. Note: Histogram of between-format differences in correlations (i.e., each observation is the difference in correlation $r_{compact} - r_{expanded}$ for a given pair of factors). Note that the X-axis labels depict the upper boundary of the histogram bins (hence the asymmetry in critical values on the left vs. right hand side).

increase of ERS in the compact format again leads to higher residual item variance terms (which is typically not desirable, even though factor loadings are also slightly higher). Finally, some factor correlations are affected by response category distances, resulting in correlations that may either be smaller or greater if the visual distance between response categories is lower. Since the added variance due to increased ERS in compact formats results in more non-substantive but also somewhat more substantive variance, correlations may go up or may go down as a consequence, depending on a complex interplay between scale format, content and ERS. If such differences occur, it is difficult to say which of the two correlation estimates is most 'correct', since the constructs under investigation are latent variables (and their scaling is consequently arbitrary to some extent). Because correlations can be affected by scale format, spurious moderation effects could result if different groups (e.g., demographic groups, like youngsters using mobile devices) are presented with different scale formats. In summary, graphically expanding inter-category horizontal Likert formats decrease extreme responding, which reduces both desirable and less desirable types of item variance and which may render correlations incomparable across groups using different formats, unless measurement equivalence can first be established.

7. Study 3

In studies 2a and 2b, we found that a more compact Likert format results in more ERS, as compared to a more expanded horizontal Likert format. In study 3 we test whether this distance effect can be generalized to vertical Likert formats. More specifically, we test whether the difference in ERS between vertical Likert and horizontal Likert format data can be reduced by increasing the spacing between the vertical response categories. To test this, we compare a horizontal Likert format to a regular vertical Likert format and an expanded vertical Likert format with increased spacing between the response categories.

7.1. Method

We recruited U.S. participants on MTurk with the explicit instruction to use a desktop/laptop computer. From the initial sample of 508 completed questionnaires we deleted 21 inattentive respondents (who answered negatively to the question 'Did you seriously participate in this study?' and/or positively to the question 'Were you in any way distracted during the survey?'; there were 13 inattentive respondents in the compact and eight in the expanded vertical format conditions), and three respondents using a mobile device (as identified through the operating system identified in the paradata; all three mobile users were in the compact vertical format condition). In the resulting sample, $n = 484$, age ranged from 18 to 74 years (M = 34.7 SD = 11.4) and 44.6% were women.

We used a two-factor mixed experimental design, where all participants responded to both horizontal and vertical Likert tasks (24 items each), but whether the vertical task was regular or expanded was manipulated between subjects. The vertical scale format came in two versions: the regular format vs. an expanded format with one line of white space in between every two response category labels. The scale formats used in this study are shown in Fig. 8. The 48 items were sampled from many different existing scales, covering many diverse and unrelated topics (e.g., 'I like to serve unusual dinners,' 'I certainly feel useless at times'), and two sets of 24 items each were randomly assigned to either the horizontal or the vertical format for each participant. Each item was presented as a separate question and 12 items were presented per screen. In line with the heterogeneous content, the average

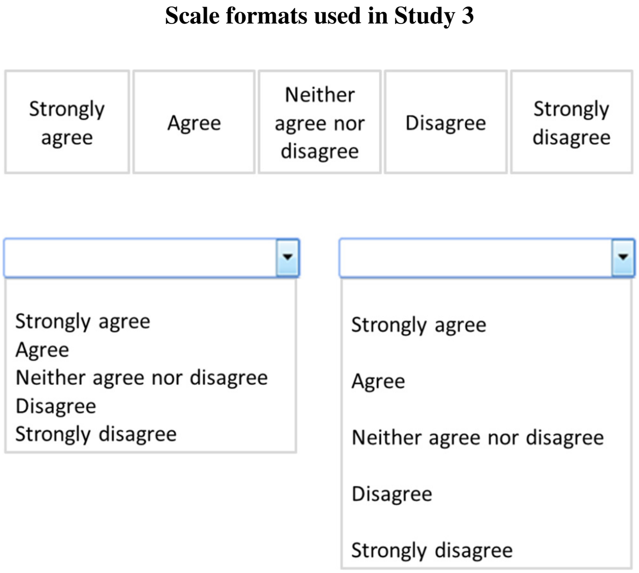


Fig. 8. Scale formats used in study 3. Note: Study 3 uses three formats: A horizontal Likert-scale format with select boxes (top), vertical dropdown format with select items (bottom left); and an expanded version of the latter that uses double spacing between categories (bottom right).

absolute inter-item correlation was $r = 0.098$. This way, we can study response patterns regardless of item content (Baumgartner & Steenkamp, 2001; Greenleaf, 1992a; Weijters et al., 2008).

7.2. Results & discussion

To simultaneously test all the mean differences of interest while accounting for the partially nested design and while allowing for heterogeneity of variances, we specify a basic two group path model with ERS $_{\text{Horizontal}}$ and ERS $_{\text{Vertical}}$ as freely correlating variables (horizontal vs. vertical Likert was manipulated within-subject) and using the default vs. expanded vertical format condition as the grouping variable (as vertical format was manipulated between-subjects). We set the mean of ERS $_{\text{Horizontal}}$ to equality across groups (fit results indicate that this assumption is acceptable: $\text{chi}^2(1) = 0.396$, $p = .529$); preliminary tests indicated that neither set content nor set order significantly affected ERS, with all p's > 0.10, so these variables were not included in the path model (Bernerth & Aguinis, 2016). Using the model constraint procedure in Mplus 8.0, we compare mean ERS levels between the horizontal format and the two vertical formats (default vs. expanded).

First, as before, ERS is higher in the regular vertical format ($M_{\text{regular vertical}} = 0.269$, SE = 0.011) compared to the horizontal format ($M_{\text{Horizontal}} = 0.237$, SE = 0.009), and this difference is statistically significant (Est. = 0.031, SE = 0.008, $t = 3.702$, $p < .001$; Cohen's D = 0.164). Second, compared to the regular vertical format, ERS is lower in the expanded vertical format ($M_{\text{expanded vertical}} = 0.247$, SE = 0.010), and this difference too is statistically significant (Est. = 0.021, SE = 0.011, $p = .048$; Cohen's D = 0.115). Finally, the difference between ERS in the horizontal format and ERS in the expanded vertical format is not statistically significant (Est. = 0.010, SE = 0.007, $t = 1.375$, $p = .169$).

To summarize, the regular vertical format shows higher ERS than the horizontal format and higher ERS than the expanded vertical format in which the visual distance between response categories has been increased, in line with the response category distance effect. Moreover, also in line with expectations, the spacing manipulation alleviates the ERS difference between horizontal and vertical Likert data.

8. General discussion

Although many marketing scales were initially constructed and validated using horizontal Likert-type scale formats (Weijters, Cabooter, & Schillewaert, 2010), it is quite common to use alternative scale formats when including these scales in online questionnaires, even more so on mobile devices. Vertical Likert formats are an often used scale format in such instances, among other reasons because they are relatively compact (Couper et al., 2017; Gunn, 2002; Liu & Conrad, 2016). Moreover, popular online survey tools automatically adapt horizontal Likert-type formats to vertical Likert-type formats when viewed on a small screen, as illustrated in Fig. 1.

We provide evidence that people give more extreme responses in vertical than horizontal Likert scale formats. The increase in extreme responding does not affect substantive content variance, but does increase residual variance, potentially resulting in lower internal consistency of marketing scales shown in vertical Likert format (study 1). In studies 2a and 2b we provide evidence for the response category distance effect: our findings suggest that categories that are visually closer are treated as if they are less extreme (despite them being identically labeled) which leads to increased ERS (see Fig. 2). Again this increased ERS seems to

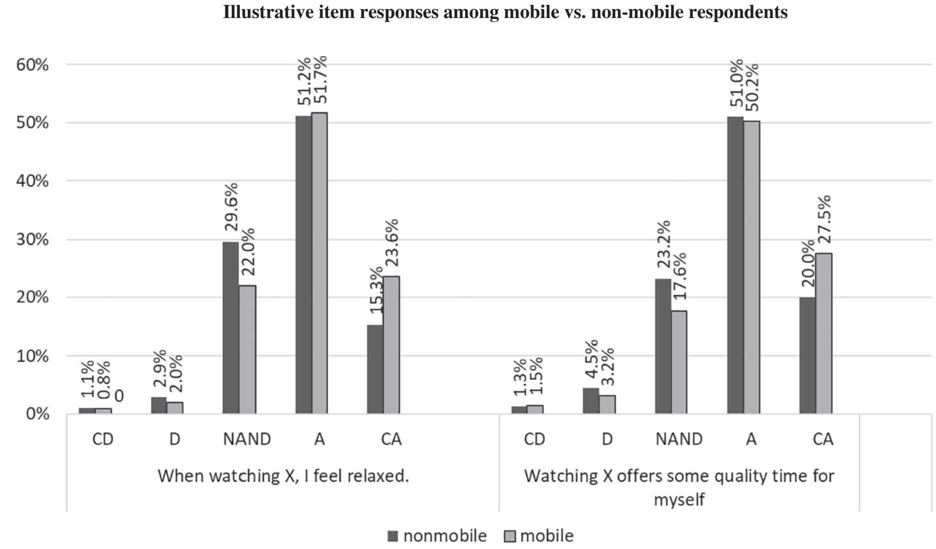


Fig. 9. Illustrative item responses among mobile vs. non-mobile respondents. Note: CD = Completely disagree, D = disagree, NAND = Neither agree nor disagree, A = Agree, CA = Completely agree.

increase non-substantive variance as exemplified by increased residual variance (studies 2a and 2b) and different estimates of correlations between constructs (study 2b). In line with the response category distance effect, the difference between horizontal Likert and vertical Likert formats in terms of ERS is mitigated when spacing between dropdown response categories is increased (study 3).

In our empirical studies, we used validated marketing scales that contain equal numbers of reversed and regularly formulated items; that is, the scales are balanced. We used balanced scales because this is generally recommended but also with the aim of obtaining measures of ERS that are not confounded with acquiescent responding. Perhaps partly because of this conservative testing approach, our results did not show evidence of substantial systematic differences in scale means across differently spaced (and/or oriented) scale formats. Furthermore, the response distributions of the items in our empirical studies showed only limited skewness and typically item means were rather close to the rating scale midpoint (none of the item means were close to the endpoints). To illustrate what can happen with data based on less well behaved item response distributions, consider the following real-life case from an applied marketing survey by a media company among users of its online video content platform ($N =$ 5569; 45.2% women; with age M = 42.5, SD = 17.2). Importantly, 63.05% of respondents took the survey using a mobile device. The survey contained several Likert items using five response categories ('completely disagree,' 'disagree,' 'neutral,' 'agree,' 'completely agree') and the response categories switched from a horizontal Likert format to a vertical Likert format for devices identified as mobile devices by the software. Fig. 9 displays the item response distribution for two items of interest ("When watching X, I feel relaxed" and "Watching X offers some quality time for myself") as a function of mobile device use. As shown in the figure, mobile users (vertical Likert format) more frequently endorse the 'completely agree' option. Both items have higher means among mobile vs. non-mobile respondents ($M_{\text{nonmobile}} = 3.77$, SE = 0.017, $M_{\text{mobile}} = 3.95$, SE = 0.013, $t(5567) = 8.556$, $p < .001$; and $M_{\text{nonmobile}} = 3.84$, SE = 0.018, $M_{\text{mobile}} = 3.99$, SE = 0.014, $t(5567) = 6.527$, $p < .001$, respectively). At first glance, these results might be interpreted substantively, as indicating that mobile users may be more at ease with an online video platform, for instance because they may feel more comfortable using mobile devices in general. However, the results we presented in this paper suggest an alternative explanation, where the mean difference between mobile vs. non-mobile respondents is due (at least in part) to the different response scale format they used when responding to the items. However, in absence of a useful ERS measure (that is unrelated to the content of the survey), researchers confronted with this type of result cannot be sure what is really driving the difference in item means between the two groups.

8.1. Theoretical implications

Our findings are especially relevant given that scale formats are sometimes chosen arbitrarily (Cabooter, Weijters, et al., 2016), are deliberately varied to prevent common method variance (Hulland et al., 2018; Rindfleisch et al., 2008) or are simply automatically adapted by survey software in function of the screen size on which they are displayed (Qualtrics, 2016). The observation that changing the scale format from horizontal Likert to vertical Likert leads to increased non-substantive variance can be especially troublesome when changes in scale format are related to other individual difference variables of interest, for instance when scale format changes in function of device use and device use is related to demographic profile. Future research needs to test this possibility and might even consider to reassess demographic effects in surveys that use device-adaptive scale formats.

Studies 2a and 2b do not only provide evidence for one interesting driver that may explain how vertical scale formats result in increased ERS compared to horizontal Likert-type items, but also demonstrate the more general point that the distance between response categories affects extreme responding. Standard survey software packages typically adapt the distance between response categories depending on the available space. If the screen is wider, the distance between response categories is often increased accordingly. Our findings suggest that ERS is influenced because of these changes in distance between response categories and as such adds to the literature on ERS by highlighting visual distance as an antecedent. Future research may therefore focus on differences in ERS between additional alternative types of scale formats that differ in visual distance and test if differences in ERS arise in line with the response category distance effect. A specific instance where such effects may be problematic is when translations of response category labels in different languages vary in width.

In study 2b, reduced inter-category spacing led to higher ERS, which translated not only in increased residual variance estimates, but also slightly higher factor loadings. This effect of response category distance on factor loadings in study 2b is relatively small and less significant than the effect of response category spacing on residual variance estimates. Thus, a first possible explanation that comes to mind is sampling variation between studies. However, since study 2b uses large samples of both items and respondents, it seems plausible that this study actually has better statistical power and may actually be better suited than the other studies to provide a valid estimate of the loading effect. Therefore, it seems more likely that the increased variance that results from the increase in ERS is mostly non-substantive (cf. the residual variance estimates), but also partly substantive (as shown in the increased factor loadings). This finding is actually in line with earlier research: based on a study of survey data containing a large number of attitude scales, Greenleaf (1992a) concludes that individuals' response range (which is closely related to ERS) has both "attitude information and bias components". In line with this, it has indeed been suggested that factor loadings tend to be higher among respondents who respond more extremely (Cheung & Rensvold, 2000; Little, 2000; Weijters, Puntoni, & Baumgartner, 2017). ERS can boost factor loadings if it leads respondents to more strongly differentiate their responses in line with the construct that is being measured. As an exaggerated example, if respondents disregard the endpoints completely, a five point scale reduces to a three point scale, which will suppress correlations (and hence factor loadings) because three categories cannot carry as much information as five (Bollen & Barb, 1981).

8.2. Practical implications

There are several situations where our findings are of relevance. As a first example, consider a market researcher who surveys customers of a brand to study the relation between age and satisfaction. If older customers are less likely to use a mobile device for filling out the questionnaire, and if the satisfaction items are shown in horizontal Likert format on PCs, but in vertical format on mobile devices, results may show a spurious negative correlation between age and satisfaction if satisfaction items are scored above the midpoint on average (since ERS in such instances will primarily increase the number of positive endpoint responses (Baumgartner & Steenkamp, 2001)). As another example, imagine a researcher who is testing measurement invariance of a marketing scale across two countries, with one country's residents being more likely to use mobile devices (on which the items of the scale are shown in vertical Likert format). Results may indicate that measurement invariance is violated, as the sample that used the vertical format is likely to show more residual variance. But the response category distance effect can also affect results in situations where all respondents use a horizontal Likert format, for instance if the survey software adapts the width of the scale to the screen that respondents use or if different researchers using the same scale (e.g., in a cross-national cooperation) apply different layout templates when designing their questionnaire.

In sum, our results suggest that scale format standardization is crucial as validity may suffer when trying to increase user friendliness, for instance. Ideally, scale format should be kept constant, and ideally, respondents should use similar devices. It is commendable to pretest online surveys with multiple browsers and screen settings to be aware of potential scale format changes, and it is also commendable to collect paradata, including data on screen resolution, type of browser and operating system (McClain et al., 2019). If researchers cannot implement any of the measures to avoid scale format differences, they might want to use items that have means close to the midpoint, since differences in ERS minimally affect such midpoint-centered items

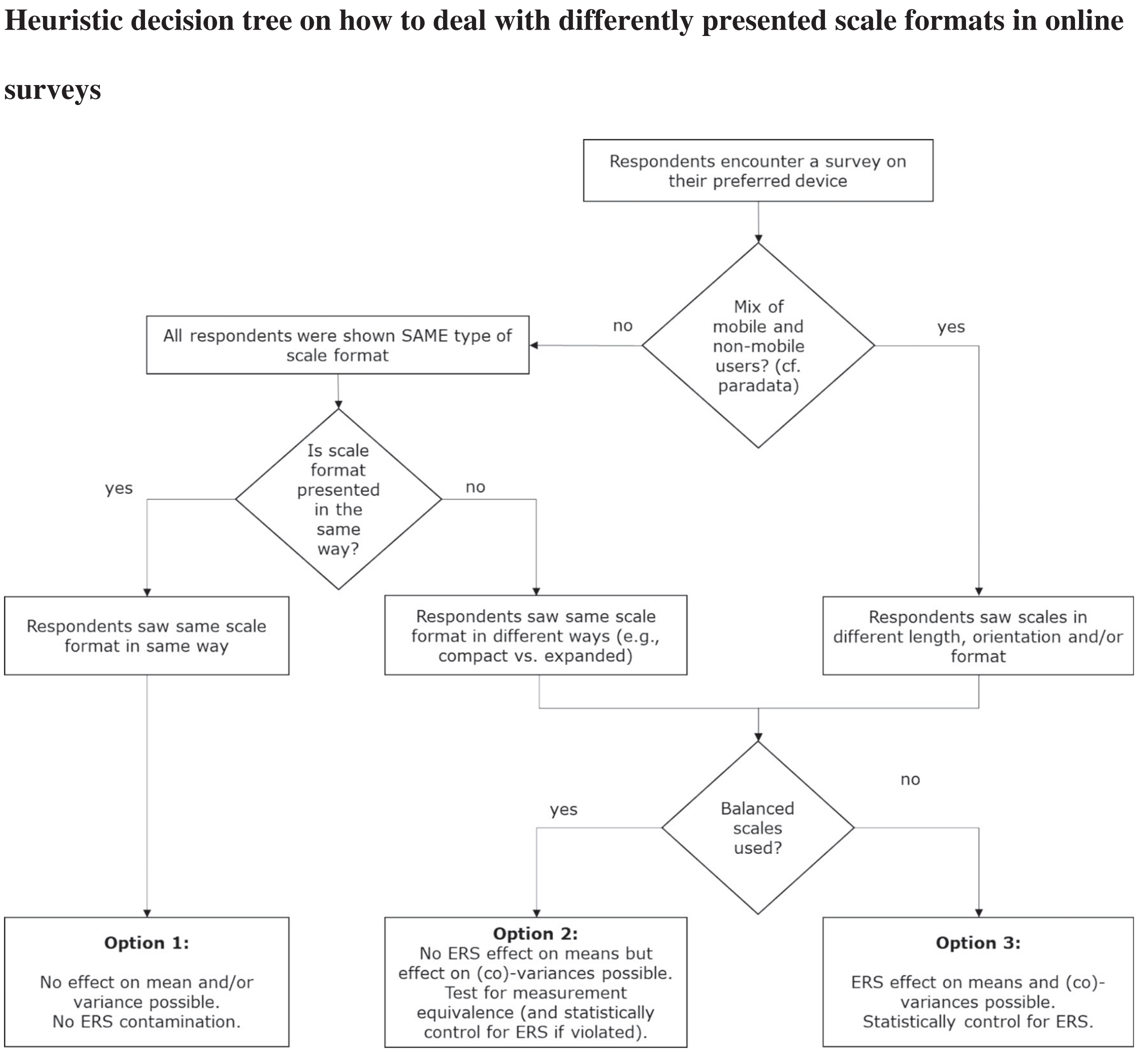

Heuristic decision tree on how to deal with differently presented scale formats in online surveys
Respondents encounter a survey on their preferred device
Mix of mobile and non-mobile users? (cf. paradata)
no
yes
All respondents were shown SAME type of scale format
Is scale format presented in the same way?
yes
no
Respondents saw same scale format in same way
Respondents saw same scale format in different ways (e.g., compact vs. expanded)
Respondents saw scales in different length, orientation and/or format
Balanced scales used?
yes
no
Option 1:
No effect on mean and/or variance possible.
No ERS contamination.
Option 2:
No ERS effect on means but effect on (co)-variances possible.
Test for measurement equivalence (and statistically control for ERS if violated).
Option 3:
ERS effect on means and (co)-variances possible.
Statistically control for ERS.

Fig. 10. Heuristic decision tree on how to deal with differently presented scale formats in online surveys.

(Baumgartner & Steenkamp, 2001), and use scale formats with labels for all response options as these formats are least ambiguous in terms of response category interpretation (Weijters, Cabooter, & Schillewaert, 2010). Finally, if differences in ERS might still emerge during data collection, researchers should test for measurement equivalence (Steenkamp & Baumgartner, 1998) or statistically control for ERS afterwards using one of the methods developed with this aim (Baumgartner & Weijters, 2015).

The flowchart presented in Fig. 10 serves as a heuristic decision tree to decide what can be expected in terms of ERS contamination and how to deal with it. More specifically, when balanced scales are used ERS is only to be expected to affect variances whereas if no balanced scales are used and respondents saw the scale differently, ERS might affect both variances and means. The ERS effects can be even more profound in case of endpoint labeled scales as the labeled endpoints will attract the respondents more (Weijters, Cabooter, & Schillewaert, 2010). Although fully labeled scales might reduce the effect sizes, our set of studies provides evidence for these effects even with fully labeled scales.

8.3. Limitations and suggestions for future research

In our studies we identified a response category distance effect with commonly used agreement ratings (i.e. in all studies our questions were labeled with labels ranging from strongly disagree to strongly agree). Even though we did not leave much room for interpretation given the labeling of each response option, we still believe that the strength of this effect may depend on the specific types of ratings and/or labels. That is, the difference between 'agree' and 'strongly agree' may be less obvious in a respondent's mind than giving 4 or 5 stars as a product evaluation. As such, we suggest that the response category distance effect may be attenuated for rating scales with labels that clearly belong to distinct categories, but may be stronger for scales with only the endpoints labeled (thereby leaving more room for interpretation). Therefore, an interesting avenue for further research is to explore how the distance between response categories may affect responses depending on how the scale is labeled. For instance, using numeric anchors for response categories may affect the response category distance effect. Also, potential interaction effects with other scale format characteristics may be studied.

The current research used survey experiments to assess the impact of response category spacing (partly in function of the vertical vs horizontal orientation of a scale format) on responses. As in any research, we were faced with tradeoffs, one of which pertains to the question of which formats to use. We decided to use commonly used formats, focusing on horizontal vs. vertical Likert formats that are part of the default templates in a popular survey software (Qualtrics). By doing so we increased the ecological validity of our work. Still, further research would benefit from replicating the effects in a more controlled (i.e., lab) environment where one would have complete control of the scale presentation and the visual angle between adjacent response categories, while keeping constant question type and just varying orientation and spacing. In a similar vein, it would be helpful to run a replication study with the presentation of only one item per page to further minimize noise, although we do not expect major differences as previous research did not find significant differences in response behavior between a single item per screen format versus multiple items per screen format (Toepoel, Das, & Van Soest, 2009). For now, our potentially noisier but more ecologically valid study procedures already provide consistent evidence for a response category distance effect which seems to be of importance to understand differences in ERS between traditional horizontal Likert-type items and vertical Likert formats.

As it is, even respondents using the same type of device (e.g., desktop computers) may have larger or smaller monitors, which may be running at higher or lower resolution. All this may affect the absolute and/or relative visual distance between response categories. The results presented in the current set of studies may not depend on visual distance per se, but rather visual distance relative to some frame. Whether this is the screen, the window or something else would need to be determined by further research.

As is common in online surveys, in our empirical studies, respondents provided a response on the displayed scale. Although this approach is relevant since it has high external validity, it necessarily confounds display and control response properties. Consequently, our results cannot unambiguously be attributed to either purely perceptual or purely motoric mechanisms. It is in other words possible that respondents are more likely to endorse endpoints because they are perceptually closer when looking at the response options, but it is also possible that respondents are more likely to endorse endpoints because they are physically closer when clicking one of the options. We have no conclusive evidence in support of either mechanism, but it seems plausible that the two mechanisms are not mutually exclusive, and that response options that are more closely spaced simultaneously look and feel conceptually closer. An alternative account where respondents are more likely to endorse more endpoint options by accident (i.e., due to problematic eye-hand coordination while selecting an option) seems less plausible for several reasons. First, in our empirical studies we used commonly used scale formats, as visualized in Figs. 4–6 and 8. None of these formats seem likely to be challenging for respondents, especially experienced users from online panels. Second, to select a response option, respondents need to click an area of the screen that immediately surrounds the radio button, box or text that represents the response option to be selected. This area is not systematically smaller or larger for endpoints relative to other options; also, if respondents click an area outside the endpoint area, no option is selected. Thus, there is no clear reason why motor error would specifically increase the proportion of endpoint responses. Finally, the surveys in our empirical studies typically contain an open question towards the end of the survey, asking respondents whether they encountered any issues or whether they have remarks. None of the respondents in our studies mentioned that it was hard to hit the right response option in any of the experimental conditions. Nevertheless, when buttons are small or closely spaced, inadvertent errors are more likely (Chen, Savage, Chourasia, Wiegmann, & Sesto, 2013), and the fact that respondents did not mention that they made response errors is not conclusive evidence that they indeed did not (in that, given the nature of the controls when closely spaced, they may not even be aware if they make an error). In sum, to further disentangle the relative contribution of the purely visual distance between response options from the contribution of the motoric distance, further research is needed. An important challenge in this context would be to strictly disentangle perceptual vs. motoric effects, both conceptually (since motoric feedback may influence perception) and procedurally (since the nature of the task become qualitatively different when trying to disentangle perception and movement).

The focus of the current research was on the distance between response options, which is sometimes (but not always) caused by vertically versus horizontally presenting response scales. We have decided to disentangle vertical vs. horizontal position from inter-category distance (cf. study 2a, study 2b and study 3). However, early psychophysical research has shown that vertical and horizontal distances are not similarly perceived (Avery & Day, 1969; Finger & Spelt, 1947). Remarkably, research also suggests differential effects of vertical vs. horizontal orientation on distance for the perceptual vs. visuomotor systems (Servos, Carnahan, & Fedwick, 2000). Future research could use these insights to further investigate the moderating effect of vertical vs. horizontal orientation on the response category distance effect.

Finally, it would be relevant to investigate the response category distance effect (and its potential implications) in different non-survey settings that involve rating scales, like for instance customer reviews or job performance evaluations in professional settings. Interestingly, the personal involvement or care in filling out the evaluation at hand (which may depend on the particular evaluative setting) can be included as a potentially relevant moderator to get more insight into the psychological mechanism behind the response category distance effect.

In conclusion, we emphasize the need for more research into the effects and the generalizability of the response category distance effect over a range of scale formats. Our set of studies at least provides evidence that vertical labeled 5-point Likert scale formats lead to more ERS than horizontal scales and that distance between response categories is of importance.

Appendix A. Multi-item scales used in study 1 and study 2b

Scales 1 through 14 were used in study 1. For study 2b, we dropped scale 6 (because of problematic estimates in study 1) and added scales 15 and 16 instead.

Factor; scale (reference)	Items
1. Ability to modify self-presentation; self-monitoring (Lennox & Wolfe, 1984)	1. I have found that I can adjust my behavior to meet the requirements of any situation I find myself in. 2. Once I know what the situation calls for, it's easy for me to regulate my actions accordingly. 3. I have trouble changing my behavior to suit different people and different situations.* 4. Even when it might be to my advantage, I have difficulty putting up a good front.*
2. General self-esteem; Rosenberg Self-Esteem Scale (Marsh, 1996)	1. I feel good about myself. 2. I am able to do things as well as most other people. 3. At times I think I am no good at all.* 4. I feel I do not have much to be proud of.*
3. Life Orientation Test-Revised (Scheier, Carver, & Bridges, 1994)	1. In uncertain times, I usually expect the best. 2. I'm always optimistic about my future. 3. If something can go wrong for me, it will.* 4. I rarely count on good things happening to me.*
4. Preference for predictability; need for closure (Neuberg, Nicole Judice, & West, 1997)	1. I dislike unpredictable situations. 2. I don't like to go into a situation without knowing what to expect from it. 3. I enjoy the uncertainty of going into a new situation without knowing what might happen.* 4. I think it is fun to change my plans at the last minute.*
5. Resisting requests for compliance; consumer interaction styles (Richins, 1983)	1. I have no trouble getting off the phone when called by a person selling something I don't want. 2. If a salesperson comes to my door selling something I don't want, I have no trouble ending the conversation. 3. I really don't know how to deal with aggressive salespeople.* 4. If a salesperson has gone to a lot of trouble to find an item for me, I would be embarrassed not to buy it even if it isn't exactly right.*
6. Involvement; spousal conflict arousal (Seymour & Lessne, 1984)	1. The greater the number of cars considered when buying a car the better the results. 2. If information could be seen in terms of dollars, it is reasonable to spend a great deal of money for information before making a car purchase. 3. Because buying a car can be a fairly complex decision, a satisfactory solution is better than taking a great deal of energy to try and find the "best" solution.* 4. With the exception of price, there is not much difference between one car and another.*
7. Verbal; style of processing (Childers & Houston, 1984)	1. I like to think of synonyms for words. 2. I like learning new words.

(continued)

Factor; scale (reference)	Items
	3. I think I often use words in the wrong way.* 4. I spend very little time attempting to increase my vocabulary.*
8. Nostalgia (Holbrook, 1993)	1. Things used to be better in the good old days. 2. We are experiencing a decline in the quality of life. 3. Technological change will insure a brighter future.* 4. History involves a steady improvement in human welfare.*
9. Hedonic; shopping value (Babin, Darden, & Griffin, 1994)	1. My last shopping trip was truly a joy. 2. The last time I went shopping, I was able to forget my problems. 3. My last shopping trip was not a very nice time out.* 4. I felt really unlucky during my last shopping trip.*
10. Centrality; material values scale (Richins, 2004)	1. I usually buy only the things I need. 2. I try to keep my life simple, as far as possessions are concerned. 3. I enjoy spending money on things that aren't practical.* 4. Buying things gives me a lot of pleasure.*
11. Independence from government; psychographics social welfare (Ahmed & Jackson, 1979)	1. People should solve their own problems and not have to depend on government help. 2. People should not need the government to help them. 3. Retraining unemployed people is an important responsibility of the government.* 4. The government should see to it that every citizen enjoys the basic necessities of life.*
12. Enjoyment; coupon antecedents (Mittal, 1994)	1. I quite enjoy clipping, organizing, and using coupons. 2. Clipping, organizing, and using coupons has become a habit with me that I have come to like. 3. Clipping, organizing, and using coupons is no fun.* 4. It is a hassle to cut out, maintain, and redeem coupons.*
13. Product category knowledge; choice goal attainment and decision and consumption satisfaction (Heitmann, Lehmann, & Herrmann, 2007)	1. I know pretty much about computers. 2. Among my circle of friends, I am one of the "experts" on computers. 3. Compared to most other people, I know less about computers.* 4. I do not feel very knowledgeable about computers.*
14. Consumption satisfaction; choice goal attainment and decision and consumption satisfaction (Heitmann et al., 2007)	My choice to buy my current computer was a wise one. I have truly enjoyed my current computer. Sometimes I have mixed feelings about keeping my current computer.* If I could do it over again, I'd buy a different computer than my current computer.*
15. Need for precision (Viswanatian, 1997)	I like to use the precise information that is available to make decisions. I like putting things into exact categories as much as possible. I dislike tasks that require me to be exact.* I don't like tasks which require me to look for small differences between things. *
16. Need for cognition (Cacioppo & Petty, 1982)	I really enjoy a task that involves coming up with new solutions to problems. I prefer my life to be filled with puzzles that I must solve. I would prefer simple to complex problems.* I prefer watching entertainment to educational programs.*

* = reverse coded.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijresmar.2020.04.002.

References

Ahmed, S. A., & Jackson, D. N. (1979). Psychographics for social policy decisions: Welfare assistance. *Journal of Consumer Research*, *5*(4), 229–239.

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, *42*, 957–970.

Avery, G. C., & Day, R. H. (1969). Basis of the horizontal-vertical illusion. *Journal of Experimental Psychology*, *81*(2), 376.

Babin, B. J., Darden, W. R., & Griffin, M. (1994). Work and/or fun: Measuring hedonic and utilitarian shopping value. *Journal of Consumer Research*, *20*(4), 644–656.

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

Baumgartner, H., & Weijters, B. (2015). Response biases in cross-cultural measurement. *Handbook of culture and consumer behavior. 150–80.*.

Baumgartner, H., & Weijters, B. (2017a). Measurement models for marketing constructs. In B. Wierenga, & R. van der Lans (Eds.), *Springer handbook of marketing decision models*.

Baumgartner, H., & Weijters, B. (2017b). Structural equation modeling. In B. Wierenga, & R. van der Lans (Eds.), *Springer handbook of marketing decision models. Vol. 2.*. New York: Springer.

Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, *69*(1), 229–283.

Blasius, J. (2012). Comparing ranking techniques in web surveys. *Field Methods*, *24*(4), 382–398.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review*, *232–39*.

Cabooter, E., Millet, K., Weijters, B., & Pandelaere, M. (2016). The "I" in extreme responding. *Journal of Consumer Psychology*, *26*(4), 510–523.

Cabooter, E., Weijters, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research*, *69*(7), 2574–2584.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116.

Chen, K. B., Savage, A. B., Chourasia, A. O., Wiegmann, D. A., & Sesto, M. E. (2013). Touch screen performance by individuals with and without motor control disabilities. *Applied Ergonomics*, *44*(2), 297–302.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187–212.

Childers, T. L., & Houston, M. J. (1984). Conditions for a picture-superiority effect on consumer memory. *Journal of Consumer Research*, *11*(2), 643–654.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. *Total survey error in practice* (pp. 133–154).

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*(2), 230–253.

De Bruijne, M., & Wijnant, A. (2014). Improving response rates and questionnaire design for mobile web surveys. *Public Opinion Quarterly*, *78*(4), 951–962.

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*, *15*(2), 195–219.

Finger, F. W., & Spelt, D. K. (1947). The illustration of the horizontal-vertical illusion. *Journal of Experimental Psychology*, *37*(3), 243.

Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, *34*(2), 244–254.

Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*(2), 176–188.

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, *56*(3), 328–350.

Gunn, H. (2002). Web-based surveys: Changing the survey process. *First Monday*, *7*(12).

Heitmann, M., Lehmann, D. R., & Herrmann, A. (2007). Choice goal attainment and decision and consumption satisfaction. *Journal of Marketing Research*, *44*(2), 234–250.

Holbrook, M. B. (1993). Nostalgia and consumption preferences: Some emerging patterns of consumer tastes. *Journal of Consumer Research*, *20*(2), 245–256.

Hulland, J., Baumgartner, H., & Smith, K. M. (2018). Marketing survey research best practices: Evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science*, *46*(1), 92–108.

Janiszewski, C., Silk, T., & Cooke, A. D. J. (2003). Different scales for different frames: The role of subjective scales and experience in explaining attribute-framing effects. *Journal of Consumer Research*, *30*(3), 311–325.

Jin, K. -Y., & Wang, W. -C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138.

de Jong, M. G., Steenkamp, J. -B. E. M., Fox, J. -P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(February), 104–115.

de Langhe, B., Puntoni, S., Fernandes, D., & van Osselaer, S. M. J. (2011). The anchor contraction effect in international marketing research. *Journal of Marketing Research*, *48*(2), 366–380.

Lennox, R. D., & Wolfe, R. N. (1984). Revision of the self-monitoring scale. *Journal of Personality and Social Psychology*, *46*(6), 1349–1364.

Little, T. D. (2000). On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. *Journal of Cross-Cultural Psychology*, *31*(2), 213–219.

Liu, M. (2017). Web survey experiments on matrix questions. *Computers in Human Behavior*, *67*, 61–72.

Liu, M., & Conrad, F. G. (2016). An experiment testing six formats of 101-point rating scales. *Computers in Human Behavior*, *55*, 364–371.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–362.

McClain, C. A., Couper, M. P., Hupp, A. L., Keusch, F., Peterson, G., Piskorowski, A. D., & West, B. T. (2019). A typology of web survey paradata for assessing total survey error. *Social Science Computer Review*, *37*(2), 196–213.

Mittal, B. (1994). An integrated framework for relating diverse consumer characteristics to supermarket coupon redemption. *Journal of Marketing Research*, *31*(4), 533–544.

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

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. *Sociological Methodology*, *41*, 13–47.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261–286.

Neuberg, S. L., Nicole Judice, T., & West, S. G. (1997). What the need for closure scale measures and what it does not: Toward differentiating among related epistemic motives. *Journal of Personality and Social Psychology*, *72*(6), 1396.

Qualtrics (2016). Mobile survey optimization. available at https://www.qualtrics.com/support/survey-platform/survey-module/more-survey-module/mobile-survey-optimization/, Accessed date: 26 November 2016.

Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, *58*, 354–360.

Richins, M. L. (2004). The material values scale: Measurement properties and development of a short form. *Journal of Consumer Research*, *31*(1), 209–219.

Richins, M. L. (1983). An analysis of consumer interaction styles in the marketplace. *Journal of Consumer Research*, *10*(1), 73–82.

Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research*, *45*(3), 261–279.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem) — A reevaluation of the life orientation test. *Journal of Personality and Social Psychology*, *67*(6), 1063–1078.

Servos, P., Carnahan, H., & Fedwick, J. (2000). The visuomotor system resists the horizontal-vertical illusion. *Journal of Motor Behavior*, *32*(4), 400–404.

Seymour, D., & Lessne, G. (1984). Spousal conflict arousal: Scale development. *Journal of Consumer Research*, *11*(3), 810–821.

Steenkamp, J. -B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(June), 78–90.

SurveyGizmo (2016). Building mobile-friendly surveys. available at https://help.surveygizmo.com/help/mobile-friendly-surveys#layout, Accessed date: 4 January 2017.

Toepoel, V., Das, M., & Van Soest, A. (2009). Design of web questionnaires: The effects of the number of items per screen. *Field Methods*, *21*(2), 200–213.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order. Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368–393.

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, *71*(1), 91–112.

Viswanatian, M. (1997). Individual differences in need for precision. *Personality and Social Psychology Bulletin*, *23*(7), 717–735.

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

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*, 236–247.

Weijters, B., Geuens, M., & Baumgartner, H. (2013). The effect of familiarity with the response category labels on item response to Likert scales. *Journal of Consumer Research*, *40*(2), 368–381.

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*(2), 105–121.

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*(1), 96–110.

Weijters, B., Puntoni, S., & Baumgartner, H. (2017). Methodological issues in cross-linguistic and multilingual advertising research. *Journal of Advertising*, *46*(1), 115–128.

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*(3), 409–422.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178–189.

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2015). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self-and observer reports. *Journal of Personality*, *84*(4), 461–472.

Zhang, X. C., Kuchinke, L., Woud, M. L., Velten, J., & Margraf, J. (2017). Survey method matters: Online/offline questionnaires and face-to-face or telephone interviews differ. *Computers in Human Behavior*, *71*, 172–180.