

# Do Different Devices Perform Equally Well with Different Numbers of Scale Points and Response Formats? A test of measurement invariance and reliability

Sociological Methods &amp; Research

2024, Vol. 53(2) 898–939

© The Author(s) 2022



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/00491241221077237

[journals.sagepub.com/home/smr](https://journals.sagepub.com/home/smr)

Natalja Menold <sup>1</sup>  
and Vera Toepoel <sup>2</sup>

## Abstract

Research on mixed devices in web surveys is in its infancy. Using a randomized experiment, we investigated device effects (desktop PC, tablet and mobile phone) for six response formats and four different numbers of scale points.  $N = 5,077$  members of an online access panel participated in the experiment. An exact test of measurement invariance and Composite Reliability were investigated. The results provided full data comparability for devices and formats, with the exception of continuous Visual Analog Scale (VAS), but limited comparability for different numbers of scale points. There were device effects on reliability when looking at the interactions with formats and number of scale points. VAS, use of mobile phones and five point scales consistently gained lower reliability. We suggest technically less demanding implementations as well as a unified design for mixed-device surveys.

<sup>1</sup>Technische Universität Dresden, Institute of Sociology, Technische Universität Dresden, D-01062 Dresden

<sup>2</sup>Utrecht University, Padualaan 14, 3584CH Utrecht, Netherlands

## Corresponding Author:

Natalja Menold, Technische Universität Dresden, Institute of Sociology, Technische Universität Dresden, D-01062 Dresden.

Email: [natalja.menold@tu-dresden.de](mailto:natalja.menold@tu-dresden.de)

**Keywords**

web surveys, mixed-device surveys, scale points, slider bars, visual analogue scales, response formats

**1. Introduction**

Surveys are important methods of collecting data in social science, sociological, psychological, behavior and related research, and online or web surveys are becoming increasingly popular (Dillman, Smith and Christian 2014). Online surveys are not only accessed on regular desktop PCs, but also on other devices such as tablets or mobile phones. Web survey designers can use many different response formats to program an online survey, and these may have an impact on respondents' behavior due to a wide variation of screen sizes between devices and different methods of navigation. Likert-type rating scales, in which respondents are asked to select an answer falling within a continuum (e.g., agree-disagree), are commonly used as a response format in surveys. Traditional PC-based web surveys usually gather data by means of rating scales made from Radio Buttons that are circles to click on. With the rise of mobile-friendly (responsive) design, tiles or so called Big Buttons are often used to increase the size of the clickable format. Alternatives are Visual Analogue Scales (VAS), which represent clickable lines. These are frequently used in the medical sector. Slider scales, which are often used in market research, use a slider to define the answer position on the line (Funke 2016).

It is well known that response formats (Jenkins and Dillman 1995; Couper et al. 2004; Toepoel and Dillman 2011; Toepoel, Das, and Van Soest 2009) and their visibility (Couper et al. 2004) influence respondents' answers. Visibility in particular could be more problematic on mobile phones, owing to the small size of their screens. Visibility may also be related to the length of the response format, in particular to the number of scale points. As Radio Buttons do not make efficient use of space, the number of scale points quickly affects the visibility and usability of answer scales on mobile phones (because of the small screen sizes). It is not clear whether the answers given on a mobile phone can be compared with answers given using a tablet or desktop, as using a mouse and a large desktop PC or using a finger on a small mobile phone screen are two very different ways of pointing to a desired answer category. This raises the question of how device, scale points and response formats interact, in order to see if they

provide equivalent measures. In addition, research is needed not only on how devices and different response format realizations affect respondents' behavior, but also on whether there is a response format that can be recommended for use with different devices.

In this paper we present results from an experiment conducted in the GfK Online Panel in the Netherlands. Respondents were randomly requested to complete the survey on either a regular desktop PC, a tablet, or a mobile phone. The response formats (Radio Buttons, Big Buttons, Slider Bars, or VAS and combinations of the latter two) and the number of scale points (5-, 7- 11, or 100-point scale) were also varied.

The comparability of data between different devices, formats or modes can be assessed by two main methods (Hox et al. 2015): i) comparison of distributions or means and ii) evaluation of comparability of variance, covariance and mean structures. The first method was used by Toepoel and Funke (2018), who found strong differences in mean scores and item nonresponse between devices and response formats with regards to this experiment. To investigate comparability of data more thoroughly, we apply the second method in this paper. To evaluate comparability of data between devices, with different formats and different number of scale points, we conducted exact measurement invariance tests (Meredith 1993). In addition, we compared measurement quality by means of reliability scores among different groups using a Latent Variable Modeling (LVM) approach (Muthén 2002; Raykov and Marcoulides 2011).

## **2. Background**

### ***2.1. Mixed-Device Surveys***

Respondents can complete an online survey on a regular PC, a tablet or a mobile phone. Use of these devices within one online-survey has been referred to as mixed-device surveys (Toepoel and Lugtig 2015; Toepoel and Funke 2018). In this sense, mixed-device surveys are similar to mixed-mode designs where a combination of different survey modes, e.g. face-to-face, telephone, paper-and-pencil, or web, is used. Research on mixed-mode designs shows that respondents' participation behavior and measurement effects can make it difficult to compare results obtained in different modes (e.g., Hox et al. 2015; Schouten et al. 2013). Respondents chose a device for survey completion at their own preference. Haan, Lugtig and Toepoel (2019) show that device ownership does not predict device use. Rather, it is the attitude of respondents towards the use of mobile devices that drives the intention to use them and which in turn explains the use of mobile devices in surveys.

There has been an increasing amount of research on mixed-mode design in recent decades (Couper 2011; DeLeeuw 2018; Dillman, Smyth, and Christian 2014). Although the literature on mixed-device surveys is still in its infancy, it is reasonable to expect similar problems in mixed-device surveys.

More and more people access online surveys via mobile devices (cf. Toeпоel and Funke, 2018). Johnson (2015) shows that around 25% of respondents to an online survey use a mobile device (tablet or phone). De Bruijne and Wijnant (2014) report around 15% use in the Dutch probability-based Center and LISS Panel. Poggio, Bosnjak and Weyandt (2015) report about 18% use of a mobile device in the German Social Science Infrastructure Services (GESIS) Panel. Lugtig, Toeпоel, and Amin (2016) show that about 30% of respondents sometimes complete surveys on a mobile device and about 12% always use a mobile device in the American Life Panel. For different German election studies, Gummer, Quoß and Roßmann (2019) report an increase in the use of mobile phones to approx. 20% and in the use of tablets to approx. 10%. This suggests that mobile-friendly web survey design is growing in importance (see, e.g., Revilla, Toninelli, Ochoa and Loewe 2016).

With growing mobile Internet use, people increasingly expect surveys to be adapted for and work well on mobile devices. This has an impact on the way we design online surveys. Tourangeau, Couper, and Conrad (2013) show that the position of an item on a screen has a systematic (even if not always large) effect on responses. They argue that screen position effects may jeopardize comparisons if items have different screen positions. De Bruijne and Wijnant (2013) found no significant differences in answer distributions in a regular PC layout and a mobile web layout using responsive design (where software detects the respondent's browser type based on the automatically logged user agent string as soon as a respondent accesses the survey). Toeпоel and Funke (2018) found lower item non-response for desktop PC than mobile devices. In their study, respondents evaluated the survey more negatively on smartphones than on other devices. With small screen sizes on mobile phones, and the fact that visibility is a powerful indicator for response endorsements (Couper et al. 2004), there is a need to study the functionality of different realizations of answer formats or their length across devices in order to find out which particular format is desirable in an era of multi-device completion.

## 2.2. Number of Scale Points

It is common practice to use a scale length of between five and eleven scale points for Likert-type rating scales (De Beuckelaer, Toonen and Davidov

2013). 11-point scales may be beneficial because they allow a higher level of precision, but they also put a high cognitive burden on respondents. This can result in higher levels of measurement error. De Beuckelaer et al. (2013) demonstrate that an 11-point scale is beneficial in that the analytical operations to be performed result in more consistent scoring (i.e., higher reliability). However, their results also indicate that a seven-point scale is a reasonable alternative. A further reduction to five-point scales was troublesome and produced a relatively high level of inconsistency in answer scores. Choudhury and Bhattacharjee (2014) noted an increase in reliability with a larger number of scale points although such increments were found to be insignificant. These authors conclude that a 5-point scale may be easier to implement and consequently preferable. However, 5-point scales elicited higher non-response rates than longer answer scales and 11 points gained more positive evaluation of the questionnaire in a mixed-device context (Toepoel and Funke 2018). Preston and Colman (2000) conclude that response scales with seven or more points perform better on indices of reliability, validity, and discriminating power. The test-retest reliability of scales with more than ten response categories tended to decrease while respondent preferences were highest for 10-point scales, closely followed by 7- and 9-point scales. 7-point scales seem to be an optimal solution as far as the number of scale points is concerned. Preston and Colman (2000) argue that the superiority of scales with around seven response categories is in line with Miller's (1956) theoretical analysis of human information-processing. The preference for 7-point scales is confirmed in a literature review by Maitland (2009). In addition, a review by Krosnick and Fabrigar (1997) found a curvilinear pattern in which scales of 5 to 7 points were more reliable than scales with either fewer than 5 or more than 7 points.

It is possible to make an almost unlimited increase in the number of scale points for web surveys. For example, every pixel could potentially be a response option. Relatively little is known about how these continuous scales (VAS) perform in terms of reliability.

Visibility is affected by the number of scale points. Couper et al. (2004) show that options that are initially visible (in a dropdown list) are endorsed more often than response options that are not initially visible. The visibility of answer options is influenced by the size of the screen and personal settings. De Bruijne and Wijnant (2014) demonstrate that the visibility of answer options differed significantly on mobile phones: an entire 5-point scale was visible on mobile phones for 99 percent of respondents, while an 11-point scale was only visible for 59 percent of respondents.

### 2.3. Response Formats

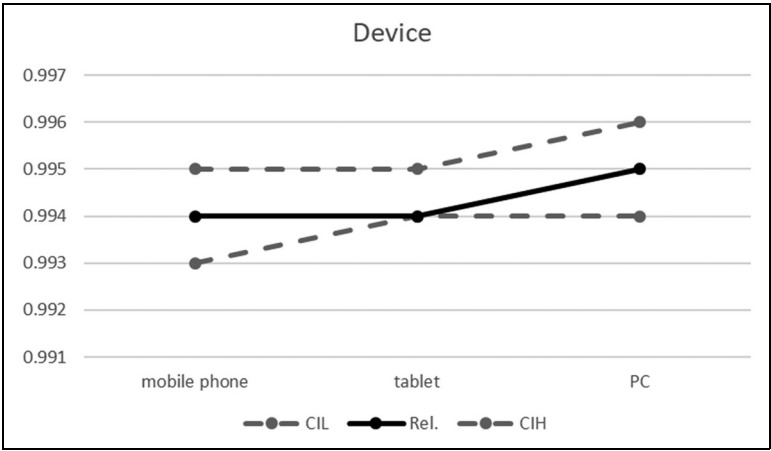
Apart from scale points, response formats may interact with the device being used. Survey software vendors are already trying to adapt to more mobile survey completion by applying relatively large buttons (tiles) that are preferable in a touchscreen layout as they are easier to pinpoint than small Radio Buttons. In addition, bars (slider scales or VAS) might become more popular response formats because they require less space on a screen than Radio Buttons. The choice for a particular format is often based on the designer's preferences, with little consideration given to measurement error. However, it is a well-known fact that response formats affect respondents' answers (Smith 1995; Jenkins and Dillman 1995; Couper et al. 2004; Toepoel and Dillman 2011; Toepoel, Das, and Van Soest 2009). In the following, we provide descriptions of response formats relevant to online and mixed-device surveys, according to Toepoel and Funke (2018: 114-115).

*Radio buttons* (see Appendix A, Figure 1) are round circles which a respondent clicks to provide an answer. Radio buttons use standard HTML and work with every browser. They are a low-tech response format and all respondents know how to use them. A problem with Radio Buttons is that they are not very efficient in the use of space on a screen because they are not scalable.

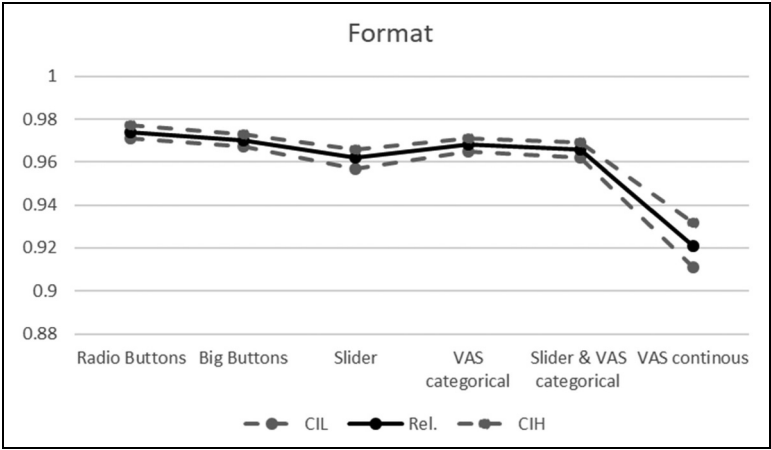
*Tiles or Big Buttons* (see Appendix A, Figure 5), in which entire cells can be clicked, instead of a small circle in traditional Radio Buttons, are becoming more and more apparent in surveys because they are easier to handle on touch screen devices with small screen sizes.

*Slider scales* (see Appendix A, Figure 2) consist of a line. Sliders work on a drag-and-drop principle: respondents drag a "handle", such as the circle on the left hand side of Figure 2 (Appendix A), to the desired answer point on the line. Slider scales suffer from the problem of the handle's starting position. Funke, Reips, and Thomas (2011) have demonstrated that the initial position of the slider leads to a different distribution of answer scores in comparison to Radio Buttons. Sliders can be realized with HTML5 or client-side technology like JavaScript and are efficient in the use of space (see VAS).

VASs (see Appendix A, Figure 3) are similar to sliders in that they also consist of a plain, mostly horizontal line but, unlike sliders, they do not use a handle. Respondents give a rating by placing a click on the line. VASs are very efficient in the use of space. A VAS with about 50 response options can take about the same space as a 3-point scale made from Radio Buttons (Funke 2016). VAS can be operationalized as a continuous (a

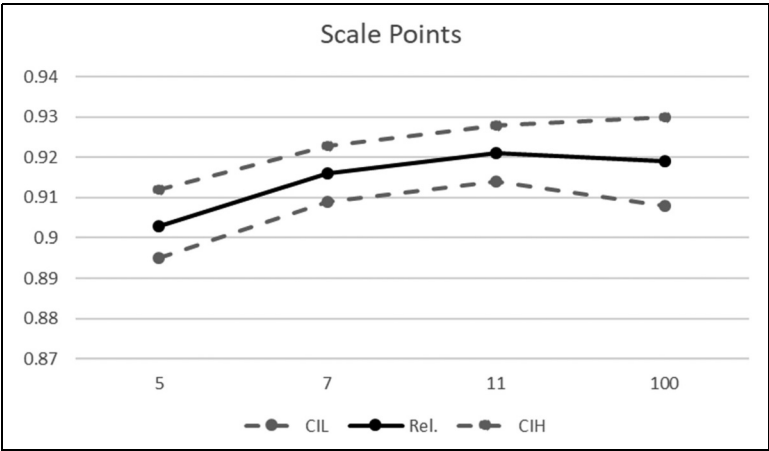


**Figure 1.** Reliability coefficients by device.  
Note. Rel.: Composite Reliability; CIL and CIH: lower and higher borders of the 95% Confidence Interval.



**Figure 2.** Reliability coefficients by format.  
Note. Rel.: Composite Reliability; CIL and CIH: lower and higher borders of the 95% Confidence Interval.

pixel serves as a possible rating) or as a discrete n-point scale. VASs require client-side technology (e.g. JavaScript) and hence are more high-tech than Radio Buttons.



**Figure 3.** Reliability coefficients by scale points.  
Note. Rel.: Composite Reliability; CIL and CIH: lower and higher borders of the 95% Confidence Interval.

Bars, implemented as either slider bars that apply the drag-and-drop principle or categorical VAS with a point-and-click mechanism, save space on the screen and may for this reason be preferred to Radio Buttons on mobile devices.

Slider scales can have negative effects on data quality, such as response rate, sample composition, distribution of values, item non-response and response times, as compared to VAS and Radio Buttons (Funke 2016; Toepoel and Funke 2018). The similarity of VAS and Radio Buttons is confirmed in different studies (Funke and Reips 2012; Funke 2016). Further studies are needed to test the differences in response formats between online devices in experimental settings to advance our understanding of why and when differences between online devices in response formats are likely to occur.

2.4. Measurement Invariance

The comparability of data between different devices, formats or modes can be assessed by two main methods (Hox et al. 2015): i) comparison of distributions or means and ii) evaluation of the comparability of variance, covariance and mean structures. The second evaluation has been referred to as

measurement invariance analysis (Jöreskog 1971; Meredith 1993). Scheuch (1993) refers to it as “functional equivalence”, by which is meant not only comparability of question wording or visual design, but in particular comparable suitability of data for the analysis of interest. Measurement invariance analysis, however, is only applicable to multi-item measures (scalar questions) that have a known one-dimensional or multi-dimensional factor structure. This means that indicators (single items or questions; manifest variables) represent either a single concept (unidimensional case) or a number of related sub-concepts (multidimensional case). The concepts are referred to as latent variables. Using the language of factor analysis, on which the analysis of measurement invariance is based, the items have to represent one factor or multi-factorial structure for a concept. The relationship between an indicator and latent variable can be either linear or non-linear. Assuming a linear case (that is of relevance for our study), this relationship is mainly described by two metrics: the intercept and the slope of the linear function. The latter is referred to as factor loading of an indicator on the latent factor.

Measurement invariance analysis is typically conducted by a sequence of subsequent steps within the frame of Multi-Group Confirmatory Factor Analysis (MGCFA). The following increasing degrees of measurement invariance are differentiated, with each subsequent one including the preceding, as introduced by Meredith (1993; see also Kline 2016; Mayerl 2016).

- (i) The first step is to evaluate configural invariance. Configural invariance occurs when the same latent factor structure underlies a given set of manifest variables in each group. Establishing configural invariance, however, does not allow for statistical comparisons of latent variables using structural equation modeling (SEM) or simple sum scores.
- (ii) The second step is to evaluate metric or weak invariance, which implies that the covariances between the manifest and latent variables are comparable among the groups. Metric invariance is evaluated by restricting corresponding factor loadings between the groups to be equal in the configural model. Equality of factor loadings is proven if the introduced restriction does not significantly decrease model fit. If factor loadings are found to be equal between the groups, metric invariance is given. Establishing metric invariance allows for comparison of correlations between the groups (those of latent variables using SEM or simple sum scores).
- (iii) The final step is to test for scalar or strong invariance, which means that the manifest variables have comparable metrics among groups

and are tapping comparable parts of the latent means. Scalar invariance is evaluated by restricting the intercepts of the manifest variables to make them equal between the groups. Again, this restriction should not significantly decrease model fit to assume the model to be scalar invariant. Establishing scalar invariance allows for comparisons of either latent mean scores or those of summarized scores.

- (iv) If the scalar invariance is evident in the data, latent means or summarized mean scores can be compared between the groups.

Hox et al. (2015) stress how powerful the testing of measurement invariance is, compared with simple inspections of distributions or means. The analysis of measurement invariance allows for the disentangling of the non-systematic and systematic measurement error associated with group membership (i.e. participation using a certain device, a special format or a certain number of scale points). If configural, metric, or scalar invariance is not given, the implications are severe as this shows that respondents understand questions or response options differently depending on the modes, devices etc. that they use. Consequently, non-systematic measurement error is evident and a given set of indicators cannot be assumed to measure a concept in a comparable manner in each of the groups. If scalar invariance is established, latent means or summarized mean scores (step iv) can be compared. In the case of a randomized experiment, however, the difference of latent means demonstrates an effect of experimental manipulation on the measurement, which is associated with a systematic measurement error. For the comparison of modes, Hox et al. (2015:3) stipulate: "This fourth step tests if the latent mean or sum scores in different modes are equal. If not, we may have measurement equivalence, but the different modes still result in a response shift, with some modes reporting higher scores than other modes. This response shift points toward either a systematic bias in one of the modes or different systematic biases across modes."

Hox et al. (2015) provide an overview of past research that evaluates mode effects when using measurement invariance tests. Considering their results as well, previous research provides a mixed picture, as some researchers have found scalar invariance for modes with and without the interviewer (e.g., Revilla 2013; Heerwegh and Loosveldt 2011), but others did not (Hox et al. 2015; De Leeuw 1992; Klausch, Hox and Schouten 2013). In one study, scalar measurement invariance between self-administered modes, that is mail and web surveys, could be supported (Klausch et al. 2013). While some research is cited in which measurement invariance across different modes is evaluated, the authors are not aware of such research for

different devices. Rather, for mixed-device surveys, response behavior and potential mean differences have been addressed. Toepoel and Funkte (2018) found mean differences among different rating scale formats when different devices were used. However, mean differences can only be interpreted as a bias if there has been measurement invariance among the investigated groups, which we will address in the present study.

With regard to the effects of rating scales Menold and Tausch (2016) and Menold and Kemper (2015) compared rating scales that use five and seven categories with different degrees of verbalization using multi-item sets for different concepts (opinions on the European Union, studying effort, affectivity). They employed either university students' samples and paper-and-pencil mode (Menold and Tausch 2016) or heterogeneous quota samples of adults and web surveys (Menold and Kemper 2015). The generalizable finding was that there was no metric and scalar measurement invariance between different rating scale realizations. Variation in points of rating scales might therefore be a crucial factor that negatively impacts comparability in mixed-device surveys.

## **2.5. Reliability**

Besides the comparability of devices when using different formats and numbers of scale points, measurement quality is also an issue. Previous research on survey and questionnaire design has mainly focused on response sets, such as acquiescent, middle or extreme responding, item-non response or on differences in means and distributions (e.g. Toepoel and Funke 2018, see also the research overview by Schaeffer and Dykema 2011). However, analysis of reliability and validity allows more direct metrics of measurement quality than rather indirect evaluations of response sets or distributions. In general, when looking at the research on mode effects and other features of questionnaire design, little is known about how these factors affect the reliability or validity of measurements (Schaeffer and Dykema 2011:912).

In the present study, we focus on reliability scores, as reliability is a prerequisite for valid measures. Reliability means that the variation of data is mainly due to the true variation, and not to non-systematic measurement error (e.g., Raykov and Marcoulides 2011; Rammstedt et al. 2015; Groves et al. 2009). Non-reliable measures make the true relationships between concepts or changes in values difficult to identify, e.g., as a result of interventions.

As the research on mixed-devices is still largely in its infancy, the authors are not aware of studies that compare reliability metrics among different devices. Little is known about the effects of different formats either.

The relationship between the number of scale points and reliability has been an issue for many decades, however (see section 2.2). In general, the research can be divided into i) experimental studies with between-subject-design (for overviews see Krosnick and Fabrigar 1997; Maitland 2009) and ii) nonrandomized quasi-experimental studies (Alwin 2007; Saris and Gallhofer 2007; Churchill and Peter 1984). The implications of these different lines of research differ. Whereas the first show that five to seven categories are associated with the highest level of reliability, the results of the latter are mixed and suggest that there is no relationship between number of categories and reliability (Churchill and Peter 1984), maximum reliability for four (Alwin 2007), seven to nine (Alwin and Krosnick 1991) or eleven categories (Saris and Gallhofer 2007). Cronbach's Alpha (Cronbach 1951) has been the preferred method for analyzing reliability (cf., Menold and Tausch 2016). Composite reliability is superior to Cronbach's Alpha in survey setting, because the latter requires indicators with equal loadings on the latent variable and an absence of correlated error terms (see, e.g., Raykov and Marcoulides 2011). Although testable in terms of Confirmatory Factor Analysis (CFA), these prerequisites were not evaluated in previous studies that used Cronbach's Alpha. The method of Composite Reliability, within the frame of LVM, is based on the sole assumption that manifest variables represent one latent variable and relaxes other requirements for the data. It is also possible to assess a single reliability metric while taking account of multi-factor structure (so called general structure, Raykov 2012). Composite Reliability can also be compared between groups and differences can be tested for significance by means of MGCFA (Menold and Raykov 2016). Menold and colleagues (Menold 2020; Menold and Kemper 2015; Menold and Tausch 2016) compared five and seven point scales with respect to Composite Reliability (e.g., Raykov and Marcoulides 2011). Hence, they emphasized the verbalization of rating scales, in particular, which is not the focus of this paper.

Our analyses investigate the effects of device, scale points and formats on reliability and assess these by applying the Composite Reliability method.

### 3. Method

Data were collected in the GfK Online Panel on the Dutch population with regards to age (15+), gender and education. This panel is ISO certified.

The panel is a nonprobability online access panel and recruitment is multi-sourced. Panel members that own a PC, and or tablet, or mobile phone were used as a sampling frame for our study.

The experiment was not pre-registered. The questionnaire consisted of three blocks of questions: three questions about attitudes towards surveys, 16 items about vacation experiences that served as our test items, and seven questionnaire evaluation questions. The 16 vacation items varied in the number of scale points and response format. The 16 items are intended to measure four sub-factors according to Pine and Gillmore's (2011) book on the Experience Economy. These items are designed to measure the four realms of a vacation experience, which can be sorted into four broad categories: (i) entertaining quality of vacation experience, including having fun and positive emotions from vacation activities; (ii) educational effect of vacation including new experiences and learning new matters; (iii) escapist quality, meaning escaping and recuperating from everyday routines, and (iv) aesthetic quality, meaning enjoying beauty or other similar experiences. The questionnaire took about five minutes to complete (see Appendix B for information on items and their wording).

We used a three-factor design with factors Device, Format and Number of Scale Points in which we randomly assigned respondents to the following conditions (Table 1):

1. Desktop PC vs. tablet vs. mobile phone (Factor Device)
2. Radio Buttons vs. Big Buttons vs. Slider vs. categorical VAS vs. a combination of Slider/VAS vs. continuous VAS (Factor Formats)
3. 5- vs. 7- vs. 11- vs. 100 (continuous VAS) point scale (Factor Scale Points)

Every format was made identical (e.g., length, color) except for functionality (e.g., point-and-click vs. drag-and-drop) and answer type (e.g., bar vs. buttons); see Appendix A. This was done in order to avoid the problem of confounding scale length and the number of scale points. In order to be able to differentiate between nonresponse and responses of a substantial response category the initial place of the handle in the slider scales was placed outside the line with valid ratings, to the left (see Appendix A, Figure 2). In the combination Slider/categorical VAS condition, respondents can use both the drag-and-drop and point-and-click operation. The continuous VAS with 100 points was used to compare it with other formats and with shorter rating scales.

**Table 1.** Overview of Experimental Design over Factors Device, Format and Number of Scale Points (5, 7, 11, 100).

Device Format	Desktop PC			Tablet			Mobile Phone		
Radio Buttons	5	7	11	5	7	11	5	7	11
Big Buttons	5	7	11	5	7	11	5	7	11
Slider	5	7	11	5	7	11	5	7	11
Categorical VAS	5	7	11	5	7	11	5	7	11
Slider/VAS	5	7	11	5	7	11	5	7	11
VAS	100			100			100		

We used a horizontal orientation of rating scales, as response order effects may occur with vertical orientation in web surveys as well (Hu 2020; Toepoel, Das and van Soest 2009). Although verbal over numerical labels have been recommended (e.g., Krosnick and Fabrigar 1997; Menold 2020), their use is restricted by the number of scale points. It is difficult to find an appropriate verbalization for 11 points and not possible to use full verbalization for continuous VAS with 100 scale points. To avoid confounding the number of scale points and verbalization, we implemented all rating scales with numerical labels.

Fieldwork was done on April 8–16, 2014. The response percentage was 30%. Nonresponse was 34%. After the first invitation on April 8, 99.7% of respondents responded, who were assigned to the desktop PC-group, while of the assignments to the mobile phone and tablet 29% and 50% responded respectively. We therefore sent an additional invitation on April 11 for tablet and mobile respondents. We used a quota for the number of respondents on each device; 32% stopped due to quota fulfillment. The dropout rate was 4%. The total number of responses was 1,709 on desktop IPC, 1,702 on tablet, and 1,666 on mobile, with a total of 5,077 responses.

48% of participants were men and 52% women; the mean age was 46 years (SD = 16 years); 28% had higher education, 31% had lower education. The use of quotas to make random assignment to the devices meant that a target person could participate if he/she owned at least the required device. According to the self-reports in our sample, 58% of respondents possessed all three devices and 33% of respondents had access to two devices (11% PC and mobile phone; 9% PC and tablet; 13% tablet and mobile phone). 6% had only a PC and 3.4% one of the mobile devices. We also evaluated respondents' experience in the use of a device for the surveys. 83% reported

having considerable experience in using desktop PC, 50% had experience using tablets and 29% using smartphones.

In terms of browser, Safari was used by the majority of respondents who participated with mobile devices (mobile: 74%, tablet 89%). Desktop PC users used Chrome (34.4%), Mozilla or Netscape (31.3%), or Firefox (15%). Desktop PC participants primarily used Windows (92.3%). iPhones were used by 42% in the mobile phone group and by 72% in the tablet group. Nokia was also used among mobile devices (cell phones: 58%, tablets: 28%).

There were significant differences between three devices with respect to gender ( $\chi^2_{(2,5077)} = 45.17, p < .001$ ), age ( $F_{(2, 5074)} = 235.44, p < .001$ ), education ( $\chi^2_{(4,5077)} = 102.68, p < .001$ ), device ownership ( $\chi^2_{(12,5077)} = 1227.54, p < .001$ ), and experience in the use of each of the devices for participation in surveys (Cramer's-V between .14 and .31,  $p < .001$ ). 58% of those in the tablet group, 52% of those in the mobile phone and 47% of those in the desktop PC group were women. There were fewer people who had not had higher education in the smartphone (22%) than in the tablet (32%) and desktop PC (37%) groups. The respondents in the smartphone group were younger ( $M = 39; SD = 14$ ) than in the tablet ( $M = 50, SD = 14$ ) and desktop/PC groups ( $M = 49; SD = 17$ ). In the mobile phone group, 70% of respondents assessed all three devices, whereas in the tablet group 63% and in the desktop/PC group 41% had access to both other devices. Respondents who participated using a device tended to have experience in the use of the corresponding device for surveys (desktop PC: 98%; tablet: 91% mobile phone: 69%).

As two other experimental factors - format and number of rating scale points - are considered, there were no significant differences in sample composition with respect to gender, age, education, device ownership and device experience between the different formats and scale points groups ( $p > .08$ ).

MGCFA's to evaluate Measurement Invariance and Composite Reliability were conducted with the software Mplus 8.2. We evaluated differences between the three devices, six formats and four groups with a different number of scale points. We then compared different formats when they were used by a device as well as different number of scale points per device. We waived the use of a full factorial comparison (device by format by scale points) as this would require the comparison of 45 groups.

The model fit of MGCFA's was evaluated using the chi-square ( $\chi^2$ ) test, the Root-Mean-Square Error of Approximation (*RMSEA*), and the Comparative Fit Index (*CFI*) (Beauducel and Wittmann 2005). The CFI should be 0.95

or higher, while an RMSEA of 0.08 or less indicates an acceptable fit (Hu and Bentler 1999). Robust Maximum Likelihood estimator (MLR) was used due to the non-normality of data in each experimental group (Muthén and Muthén 2014). Concerning the exact measurement invariance, a significant change of chi-square ( $\chi^2$ ) (Meredith 1993) or a change of  $\Delta CFI \geq .01$  and  $\Delta RMSEA \geq .015$  indicate significant differences (Chen 2007).

The analyses of measurement invariance were conducted both with and without consideration of socio-demographic variables to control for their individual effect, particularly if device effects were of interest. The socio-demographic variables gender, age, education, possession of a device and experience in the use of a device for surveys were included in the MGCFA by using MGSEM (Multi-Group Structural Equation Modeling). We modelled the regression paths of all socio-demographic variables on every manifest indicator (method suggested by Hox et al. 2015).

The Composite Reliability coefficient ( $\hat{\rho}$ ) is estimated within the framework of LVM as follows (Raykov and Marcoulides 2011:161):

$$\hat{\rho} = \frac{(\hat{b}_1 + \dots + \hat{b}_p)^2}{(\hat{b}_1 + \dots + \hat{b}_p)^2 + \hat{\theta}_1 + \dots + \hat{\theta}_p}$$

where  $b_1, \dots, b_p$  are the factor loadings and  $\theta_1, \dots, \theta_p$  the error variances, obtained from the MGCFA (see Menold and Raykov 2016).

Reliability was high and ranged between .993 to .995 in different sociodemographic groups when looking at gender, age, education, possession of device and device experience. There were no significant differences in reliability across the groups of any variable. The vacation items were found to be equally reliable in all different socio-demographic groups and differences in sample composition between devices reported above can be excluded as a possible alternative explanation for the effects of the experimental manipulation of devices on reliability coefficients.

In order to cross-validate the results for topic sensitivity with respect to the potential effect of a device on Measurement Invariance and reliability, we also conducted some analyses on the items on survey evaluation (QE, see Appendix C for wording). Respondents evaluated the survey using seven items in randomly distributed different device groups. The QE-items were administered on an 11 point Radio Buttons rating scale, so that scale formats and scale points were not experimentally manipulated. Since the factorial structure for the seven QE-items was not known, we conducted a Maximum Likelihood (ML) Exploratory Factor Analysis (EFA) with an oblique (Promax) rotation using SPSS 27 on all data without differentiating

between the three device groups. The EFA revealed two factors (Appendix C, Table AC-1). The first factor contained four items on the evaluation of questions and questionnaire and the second factor contained three items on respondents' motivation. We used the ML-EFA and the oblique rotation, because an MGCFA (that had to be conducted thereafter) also uses a ML estimation and because the factors of QE-items were expected to correlate, which was in fact the case (correlations between the factors was  $r = .60$ ). The two factors explained 57% of the entire variance. We compared the results obtained for the QE-items with the results of both, 11 rating scale points and 11 points Radio Buttons realizations of the vacation scale. Comparable results provide indications with respect to the stability of results for instruments on different topics.

## 4. Results

### 4.1. Measurement Invariance

For the three devices, the four factor MGCFA for items on vacation experiences provided an acceptable model fit according to *RMSEA* and *CFI* (Table 2). The standardized factor loadings for each factor and each device were high in magnitude (range mobile phone:  $\lambda = .67$  (item 10) to  $\lambda = .97$ ; Tablet:  $\lambda = .66$  (item 10) to  $\lambda = .98$ ; PC:  $\lambda = .68$  (item 10) to  $\lambda = .98$ ). The factor correlations were very high and ranged from  $r = .92$  to  $r = .97$ . For the information on estimated parameters in all MGCFA's, including factor loadings, intercepts, residual variances and correlations between the factors see Supplementary Material. This model is also the first, unrestricted baseline configural model for the test of measurement invariance (Table 2). Restricting factor loadings of pertinent items to be equal in each of the device groups did not significantly change the goodness of fit according to the change of chi-square, *RMSEA* and *CFI*. Metric invariance therefore applied across devices. This result also emerged for the scalar invariance, as restricting indicators' intercepts to be equal among devices did not significantly change the goodness of fit of *RMSEA* and *CFI*, whereas there was a significant change of chi-square. If the socio-demographic variables of gender, age, education, possession of devices and device experiences are considered in the model, configural, metric and scalar invariance are supported by the results, with the latter two confirmed by the non-significant change in all goodness-of-fit statistics (Table 3). If scalar invariance is supported, latent means can be compared across groups to evaluate systematic shift of means when using one device instead of another. In both models, with and without socio-

**Table 2.** Test of Measurement Invariance among Experimental Groups for Vacation Indicators.

Model	CMIN (df)	$\Delta$ CMIN ( $\Delta$ df)	RMSEA	$\Delta$ RMSEA	CFI	$\Delta$ CFI
<b>Device</b>						
configural	892.94 (294)***	—	.035	—	.960	—
Metric	961.56 (326)***	58.56 (217)	.034	-.001	.957	.003
Scalar	1053.65 (358)***	72.67** (28.52)	.034	.000	.953	.004
<b>Format</b>						
configural	3150.42 (588)***	—	.072	—	.933	—
Metric	5093.78 (668)***	2315.71 (95.68)***	.089	.017	.884	.049
<b>Format (without continuous VAS)</b>						
configural	2753.70 (490)***	—	.072	—	.934	—
Metric	2890.21 (554)***	81.58 (79.6)	.069	-.003	.932	.002
Scalar	3060.21 (618)***	103.52** (68.30)	.067	-.002	.929	.003
<b>Sale Points</b>						
configural	2542.19 (392)***	—	.066	—	.912	—
Metric	6029.46 (440)***	3487.25*** (59.04)	.100	.034	.772	.140
<b>Scale Points (without continuous VAS)</b>						
configural	2179.65 (292)***	—	.066	—	.912	—
Metric	3452.77 (326)***	1349.63*** (39)	.080	.014	.854	.058
<b>Format by Device (without continuous VAS)</b>						
configural	4162.33 (1470)***	—	.079	—	.928	—
Metric	4490.51 (1694)***	258.76 (259.28)	.075	-.004	.925	.003
Scalar	4895.81 (1918)***	333.44** (222.18)	.072	-.003	.920	.005
<b>Scale Points by Device</b>						
configural	3468.64 (1176)***	—	.068	—	.909	—

(continued)

**Table 2.** Continued

Model	CMIN (df)	$\Delta$ CMIN ( $\Delta$ df)	RMSEA	$\Delta$ RMSEA	CFI	$\Delta$ CFI
Metric	7171.98 (1352)**	3535.38*** (222.96)	.101	.033	.769	.140
<b>Scale Points by Device (without continuous VAS)</b>						
configural	2829.89 (882)**	—	.067	—	.911	—
Metric	4234.47 (1010)**	1465.13*** (145)	.080	.013	.853	.058

Note.  $\Delta$ CMIN: with scaling correction factor.

demographic variables, there were no significant differences of latent variable means across the devices ( $p > .10$ ).

In the model for six formats (disregarding devices), the factor loading of the item 10 decreased to the size of  $\lambda = .07/.08$  (standardized) and was non-significant for Big Buttons, Slider, and combined Slider/VAS. It was significant ( $p < .01$ ), but small of  $\lambda = .11$  for Radio Buttons and of  $\lambda = .16$  for categorical VAS. For the continuous VAS, the item 10 gained a negative loading of  $\lambda = -.35$ . This item negatively characterizes vacation experiences (it was quite boring there), so that either a low or a negative loading seemed to be plausible. Other standardized factor loadings were of comparably high size in all different formats and ranged from  $\lambda = .69$  to  $\lambda = .95$ . The factor inter-correlations decreased remarkably and ranged from  $r = .69$  to  $r = .94$ . With the continuous VAS, the correlations between the factors were even lower and ranged from  $r = .37$  to  $r = .59$ .

With respect to Measurement Invariance among different formats, the configural invariance could be evaluated as given (Table 2), although the *CFI* was slightly below the benchmark. Metric invariance was violated due to a significant change in *RMSEA* and also due to significant and remarkable change in chi-square and *CFI*. (As metric invariance is a prerequisite for scalar invariance, the test results for the latter are not presented if the first is violated.) This result is also supported, if the effect of socio demographic variables is held constant during the analysis (Table 3). However, continuous VAS made the highest contribution to a large chi-square, which suggests that it was less comparable with other formats. There were also estimation problems in all metric models when the continuous VAS was included. If the continuous VAS was excluded from the analysis, the remaining five formats showed configural and metric invariance. Scalar invariance was

**Table 3.** Test of Measurement Invariance among Experimental Groups for Vacation Indicators Including Socio–Demographic Variables Gender, Age, Education, Possession of Device and Device Experience.

Model	CMIN (df)	ΔCMIN (Δdf)	RMSEA	ΔRMSEA	CFI	ΔCFI
<b>Device</b>						
configural	897.02 (294)***	–	.035	–	.977	–
Metric	963.82 (326)***	58.52 (213)	.034	–.001	.976	.001
Scalar	1051.36 (358)***	40.50 (35)	.034	.000	.974	.002
<b>Format</b>						
configural	3134.30 (588)***	–	.072	–	.946	–
Metric	5053.56 (668)***	2289.32 (95)***	.088	.016	.907	.039
<b>Format (without continuous VAS)</b>						
configural	2746.35 (490)***	–	.072	–	.948	–
Metric	2882.19 (554)***	61.47 (73)	.068	–.004	.946	.002
Scalar	3034.40 (618)***	77.50 (67)	.066	–.002	.944	.002
<b>Sale Points</b>						
configural	2471.13 (392)***	–	.065	–	.925	–
Metric	5912.54 (440)***	3245.98*** (63)	.099	.034	.803	.122
<b>Sale Points (without continuous VAS)</b>						
configural	2117.29 (294)***	–	.064	–	.925	–
Metric	3382.84 (326)***	1265.55*** (39)	.079	.015	.875	.050
<b>Format by Device (without continuous VAS)</b>						
configural	4286.66 (1470)***	–	.080	–	.939	–
Metric	4634.99 (1694)***	276.71 (250)	.076	–.004	.936	.003
Scalar	4992.04 (1918)***	282.46 (232)	.073	–.003	.934	.002
<b>Scale Points by Device</b>						
configural	3494.48 (1176)***	–	.068	–	.919	–
Metric	7161.48 (1352)***	3498.51*** (219)	.101	.033	.797	.122
<b>Scale Points by Device (without continuous VAS)</b>						
configural	2845.67 (882)***	–	.067	–	.921	–
Metric	4270.17 (1010)***	1425.50*** (152)	.081	.014	.868	.053

Note. ΔCMIN: with scaling correction factor.

given due to the non-significant change of *CFI* and *RMSEA*, but the change of chi-square was significant again (Table 2). If socio-demographic effects were

held constant, full configural, metric and scalar invariance emerged between the formats if the continuous VAS was excluded from the analysis (Table 3). Therefore, with the exception of the continuous VAS, measurements were not only comparable between devices, but also between different formats. However, the latent means of the factors 1, 2 and 4 were significantly lower in the case of Sliders than for other formats ( $p < .05$  and  $p < .01$ ). Similarly, for the combined Sliders/VAS the mean of the first factor was lower than with other formats ( $p < .05$ ). Comparing latent means on the basis of the MGSEM with socio-demographic variables did not reveal any latent mean differences among different format groups, showing that potential shift of the means is due to different use of these formats by different groups of respondents. This can be concluded because formats did not differ from each other with respect to socio-demographic sample composition. The results show that continuous VAS was not comparable to other formats due to the differences in measurement error as well as Sliders and Combined Slider/VAS may be associated with a systematic bias (latent mean shift).

For the scale points, the item 10 gained negative loadings and a magnitude of approx.  $\lambda = -.40$  in all rating scale groups, while remaining positive loadings ranged between  $\lambda = .54$  and  $\lambda = .92$ . The factor inter-correlations decreased (ranged from  $r = .32$  to  $r = .61$ ) as compared with the MGCFA models for the effect of a device. The configural invariance of scale points (again regardless of the device used) can be questioned due to little *CFI* (Table 2). If configural invariance is nevertheless assumed, restricting factor loadings to be equal significantly decreased goodness of fit referring to a significant increase of chi-square and *RMSEA* and significant decrease of *CFI*. Metric invariance was therefore strongly violated for a different number of scale points. This result did not change if the continuous VAS (100 points) was excluded from the comparison (Table 2) or socio-demographic variables were considered in the analysis (Table 3). Therefore, metric invariance was systematically violated when a different number of scale points was used for data collection.

Looking at the measurement invariance for formats per device (while the continuous VAS was excluded, see Table 2), configural and metric invariance emerged. However, the model fit of the configural model could be evaluated as just acceptable again, as *RMSEA* was on the border of benchmark and *CFI* was slightly below it. Scalar invariance can be assumed according to the change of *RMSEA* and *CFI*; however, chi-square significantly worsened again. If socio-demographic variables were included in the analysis, scalar invariance was not violated anymore and was exhibited for different formats when used with different devices (Table 3). For scale points by

different devices, neither configural, nor metric invariance appeared again, regardless of whether controls were made for socio-demographic variables (Table 2 and Table 3).

## 4.2. Reliability

Table 4 and Figures 1 to 3 summarize the scores of Composite Reliability for the general structure and their pertinent 95% Confidence Intervals (CIL for the lower and CIH for the higher border) obtained for the three devices, the six formats and the four groups with a different number of scale points. In terms of devices, perfect reliability (close to 1) resulted for each, without any notable differences (Table 4, Figure 1). This shows that the given instrument on vacation experiences exhibited a very high level of reliability pointing toward high factor loadings and little error terms.

If differentiations were made between the six formats, the absolute size of reliability decreased slightly (Table 4, Figure 2). Hence, the values of reliabilities remained very high ( $\rho > .95$ ), with the exception of the continuous VAS, where reliability decreased to  $\rho = .92$ . Due to non-overlapping 95% CIs (Table 4 and Figure 2), this size of reliability was significantly lower than

**Table 4.** Composite Reliability for Vacation by Device, Format and Scale Points.

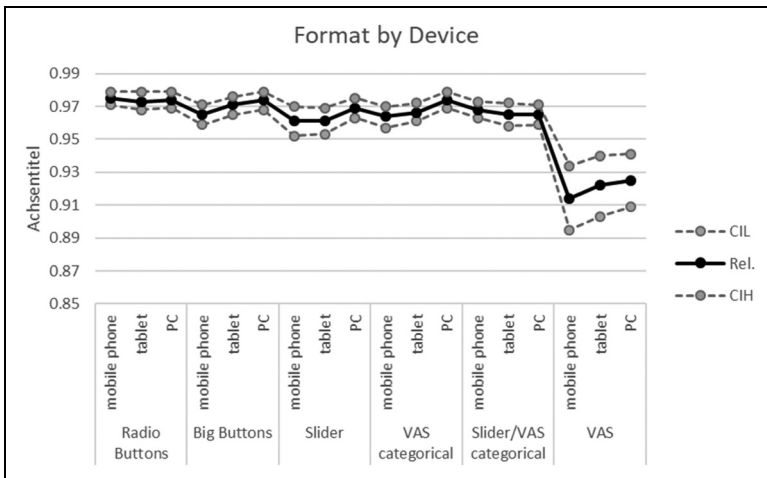
Factors	Value	CIL	CIH
<b>Device</b>			
mobile phone	0.994	0.993	0.995
tablet	0.994	0.994	0.995
PC	0.995	0.994	0.996
<b>Format</b>			
Radio Buttons	0.974	0.971	0.977
Big Buttons	0.970	0.967	0.973
Slider	0.962	0.957	0.966
VAS categorical	0.968	0.965	0.971
Slider/VAS	0.966	0.962	0.969
VAS continuous	0.921	0.911	0.932
<b>Scale Points</b>			
5	0.903	0.895	0.912
7	0.916	0.909	0.923
11	0.921	0.914	0.928
100 (VAS continuous)	0.919	0.908	0.930

Note. CIL and CIH: lower and higher borders of the 95% Confidence Interval.

with other formats. Although the differences in reliability across the remaining formats were not highly pronounced, the 95% CIs were not overlapping (pointing toward significant differences) between Radio Buttons on the one hand and Sliders, categorical VAS and combined Sliders/VAS on the other hand. Therefore, significantly higher reliability coefficients were observed for radio buttons than for most other formats and the continuous VAS exhibited remarkable lower reliability than all other formats.

For the scale points, a slightly lower reliability was given for five points than other number of scale points, while significant differences were observed between five and eleven points (Table 4 and Figure 3). No notable differences were obtained between seven, eleven and 100 points.

For the next results we compare reliability coefficients between different formats implemented for a device. Since 18 groups had to be compared, reliability coefficients and their pertinent 95% CIs are only graphically presented in Figure 4. As Figure 4 shows, there was no device effect on reliability coefficients (meaning no differences between the devices) for Radio Buttons and combined Slider/VAS. Significant differences between the devices were given for other formats. Reliability was slightly lower for mobile phones and tablets than for PCs with Sliders and categorical VAS. For Big Buttons and continuous VAS, reliability was lower with mobile phones than with



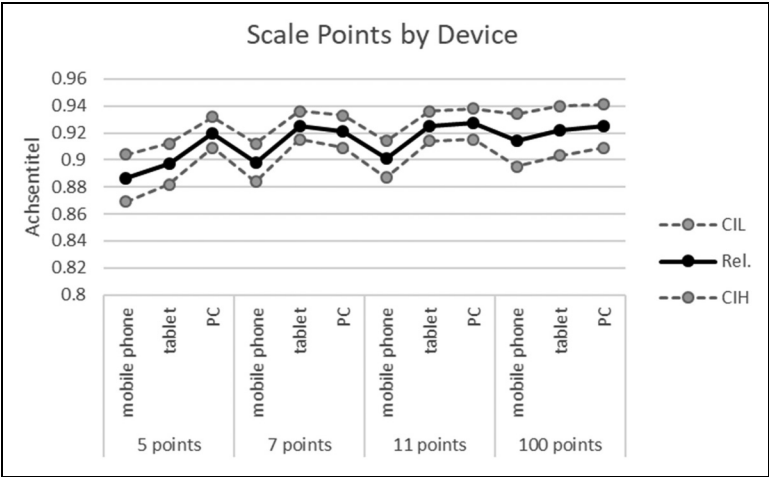
**Figure 4.** Reliability coefficients by format and device.

Note. Rel.: Composite Reliability; CIL and CIH: lower and higher borders of the 95% Confidence Interval.

tablets and PCs. This was a surprising result, since these formats have been thought to be optimizations for mobile phones. For continuous VAS we obtained remarkably lower reliability than for other formats, a result reported on above.

The reliability coefficients were higher for all number of points, if respondents used desktop PCs (Figure 5). With mobile phones, reliabilities were consistently and (with exception of the continuous VAS, meaning 100 points) significantly lower than with other devices. For tablets, reliability was lower than with PCs in the case of five point scales.

In sum, the results show that reliability was not affected by device per se, but differences arise if devices were used with different formats and different number of scale points. Thereby, reliability coefficients varied within the range from  $\rho = .91$  to  $\rho = .99$ , which was a remarkable variability. With the continuous VAS, reliability was notably lower than with other formats. Formats that were expected to improve data quality when using mobile devices, such as Big Buttons, Sliders, or VAS (both, continuous and categorical) performed worse than Radio Buttons in the case of mobile phones. Scale points were relevant in influencing reliability differences for different devices, which demonstrated significantly lower reliabilities for mobile



**Figure 5.** Reliability coefficients by scale points and device.  
Note. Rel.: Composite Reliability; CIL and CIH: lower and higher borders of the 95% Confidence Interval.

phones than other devices. Scales with five categories produced particularly lower reliabilities for both mobile devices.

### 4.3. Cross-Validation

We now look at the results for comparison among devices for the QE-items and compare them with the results for the vacation scale. For the latter, we consider mode effects for the 11 points rating scales and for the 11 point Radio Button format (as QE items were presented using this format and this number of categories). To account for the potential sample differences between devices we consider socio-demographic variables and conduct MGSEM, like the former analyses for the vacation items. The results did not markedly differ between the two types of analysis and we present the results if socio-demographic variables are included (Table 5). For the QE-items, configural invariance can be assumed to be given due to the *CFI*, but *RMSEA* was slightly too high. If metric invariance was nevertheless analyzed, it emerged across devices. Scalar invariance was given as well. Compared to the vacation items (11 points Radio Buttons), configural invariance was somewhat violated due to the results of chi square, *RMSEA* and *CFI*. If configural invariance was nevertheless assumed, metric invariance could be accepted, but scalar invariance was slightly violated due to the significant change of chi-square. These results for the effect of device on Measurement Invariance were strongly comparable between the two different topics pointing to metric and scalar invariance among devices.

The reliability for the general structure of the QE-items was evaluated on the basis of MGCFA without including socio-demographic variables ( $\chi^2_{(3776, 39)} = 441.05$ ,  $p < .001$ ; *RMSEA* = .091; *CFI* = .924). Reliability

**Table 5.** Test of Measurement Invariance among Devices for the Cross-Validation.

model	CMIN (df)	$\Delta$ CMIN ( $\Delta$ df)	RMSEA	$\Delta$ RMSEA	CFI	$\Delta$ CFI
<b>Questionnaire Evaluation (QE-Scale; 11 points, Radio Buttons)</b>						
configural	529.41 (39)***	—	.086	—	.953	—
metric	514.46 (53)***	23.67 (27)	.072	-.014	.956	.003
scalar	557.11 (67)***	10.00 (15)	.066	-.006	.953	.003
<b>Vacation (11 points, Radio Buttons)</b>						
configural	553.99 (294)***	—	.094	—	.889	—
metric	578.40 (326)***	33.40 (47)	.088	-.006	.892	-.003
scalar	630.83 (358)***	50.69 (25)**	.088	.000	.884	.008

Note. Included: Gender, Age, Education, Possession of Devices, Experience in Device Use

coefficients ranged between  $\rho = .70$  (for tablet) and  $\rho = .87$  (for PC, see Table 6). The loadings were  $\lambda \geq .30$  and significant in all groups (see Supplementary Material, Cross-Validation Section, for estimated parameters). Due to the non-overlapping 95% CIs, reliabilities differed significantly between the devices, while the lowest reliability was obtained for tablets, followed by mobile phones, and the highest reliability was observed for PC (Table 6). This result is comparable to the results for 11 points scales of vacation scale, where reliability was slightly, but significantly lower for mobile phones than for PCs (Table 6 and Figure 5). Radio Buttons as a format seemed not to balance the negative mobile device effect for the QE-items evaluated in the 11 point scale, which was the case for the vacation scale in the groups that used Radio Buttons and 11 scale points. In the latter group, reliability differences were not strongly pronounced between different devices, while reliability was high in all three device groups (Table 6). Hence, as comparisons with other formats is not available for the QE-items, we do not know whether the device effect for QE-items would be even stronger with other formats. The difference in the results concerning the potential effects of rating scale format would be due to different reliabilities of the two questionnaires, as the QE-items are associated with a low to medium reliability, but the vacation scale provided medium to high reliability.

**Table 6.** Composite Reliability by Topic for the Cross-Validation of the Results.

Factors	Value	CIL	CIH
<b>QE-items</b>			
mobile phone	0.82	0.80	0.84
tablet	0.72	0.71	0.74
PC	0.88	0.86	0.89
<b>Vacation 11 points (also Figure 5)</b>			
mobile phone	0.90	0.89	0.91
tablet	0.93	0.93	0.94
PC	0.93	0.92	0.94
<b>Vacation 11 points, Radio Buttons</b>			
mobile phone	0.93	0.91	0.96
tablet	0.92	0.90	0.95
PC	0.93	0.90	0.96

Note. CIL and CIH: lower and higher borders of the 95% Confidence Interval.

## 5. Discussion and Conclusions

One of the aims of this research was to assess the comparability of data between different devices when using different response formats and a different number of points in rating scales. The paper contributes to the literature with respect to the analysis of functional equivalence (Scheuch 1993) between devices, formats and different numbers of scale points. The analysis presented focuses in particular on under researched aspects in sociological, social and behavioral science methodology. We regard devices as an important, but less researched issue, use measurement invariance analysis that allows for separation of systematic and non-systematic measurement error, and evaluate reliability by means of LVM (Raykov and Marcoulides 2011) that is more appropriate for survey data than other reliability assessment methods.

The results show that devices are comparable (in terms of configural, metric and scalar measurement invariance). This result is similar to the results by Klausch et al. (2013), who found scalar invariance between different self-administered modes of data collection. However, some formats differ from others, either in terms of the amount of non-systematic or systematic measurement error that is a systematic artificial shift of results. In the case of the continuous VAS, metric invariance was violated when compared with other formats, which means that respondents used the continuous VAS in a different way than other formats. One reason would be that a too high level of differentiation of the continuous VAS could not be appropriately used by respondents who would have difficulties precisely scoring a point at a VAS by mouse click or using their finger on a touch screen. Continuous VAS seems to be less easy to use than other formats. Therefore, although continuous VASs make very efficient use of space, and space is very important with small screen sizes such as mobile phones, they are not recommended for mixed-device surveys.

With respect to systematic measurement error, the results obtained with Sliders (also combined with VAS) may differ from those obtained with other formats (Radio Buttons, Big Buttons, Categorical VAS). With the latter, vacation experiences were evaluated as being more entertaining, educational or escapist and in sum more positively. It has been established in cognitive psychology that less familiar stimuli are evaluated more negatively than more frequently experienced stimuli (assuming no negative consequences following exposure), which is also applicable to familiar situations (Zajonc 2001). This can explain the more positive evaluations when using Radio Buttons that are usually implemented in online surveys and represent a

familiar format for respondents. However, this explanation should be subject to further research. Another explanation would be the ease of use of a format, which can foster positive emotions and evoke more positive opinions (Winkielman and Cacioppo 2001).

While data were comparable between devices and formats (except continuous VAS), there was no comparability of data between different numbers of scale points, which was not limited to the comparison of the continuous VAS (100 Points) with other numbers of scale points. Unlike different formats, considering socio-demographic variables as covariates in the analyses did not allow measurement invariance and therefore comparability of data between different number of scale points to be ensured. Therefore, different numbers of scale points can mean different things to respondents, as rating scales differ in the degree of graduation, which in turn impacts the clarity of response options. The lack of comparability between rating scales with a different number of scale points resembles the findings of previous research that has shown that metric measurement invariance was not given between different realizations of rating scales (e.g., Menold and Kemper 2015; Menold and Tausch 2016). Therefore, this finding can also be evaluated as being generalizable to other contents and measurement instruments.

Whereas the results on comparability provide evidence of how devices, formats or different number of scale points can (or cannot) be mixed, they make no claim as to which design is to be preferred in mixed-device studies. Such a decision is supported by the results on measurement quality, i.e. reliability in our case. The results show that devices had no effect on reliability, and reliabilities of the vacation scale were of a high size, without any notable differences among devices. Comparable high measurement quality was also obtained for different formats with one exception. Continuous VAS revealed remarkably lower reliability than other formats. This supports the above explanation on why the data collected with the continuous VAS would not be comparable with other data, namely due to its only apparent precision and potentially difficult use.

With respect to the reliability of formats implemented with a device, it was shown that there was a small device effect for Big Buttons, Sliders and categorical and continuous VAS, where the reliability was lower with mobile phones and sometimes also with tablets than with desktop PC. However, such an effect was absent for Radio Buttons and the combined Slider/VAS, where reliability scores did not differ remarkably across devices. The results with respect to formats are surprising, because Big Buttons, Sliders and VAS were supposed to be of particular advantage for mobile phones (e.g. Antoun et al. 2018). Nevertheless, these formats could be either less familiar to the respondents or less usable than Radio Buttons.

Some effects on reliability were obtained for different devices when using a different number of scale points. Reliability was particularly lower for mobile devices in the case of five points, which can be explained by a lower differentiation and therefore potentially higher item non-response of five points. The result with respect to a higher reliability of seven category rating scales and lower context effects in that case supports the findings of previous studies (De Beuckeleer et al. 2013; Menold and Kemper 2015; Menold and Tausch 2016) and is therefore generalizable over different topics. Since the results did not differ for seven and eleven points, seven categories would be considered a less demanding and reasonable alternative to 11 points (as also de Beuckelaer et al. (2013) have shown).

Altogether, the results show that comparability and measurement quality are less dependent on the device used per se, but much more so on the potential interaction effects of device with response format and with number of scale points. As far as devices are concerned, the results are good news because different devices are used in online surveys and researchers are usually unable to control what device a respondent uses to participate in the survey. Radio Buttons and seven points rating scales allow for low-tech questionnaire design and collection of data that are comparable among devices, free of systematic bias and provide a sufficiently high level of measurement quality (i.e. reliability). Data obtained from continuous VAS are not comparable to the data obtained from other formats and are also less reliable.

Apart from the added value, the limitations of the present study also need to be considered. Our results apply to multi-item scalar questions on opinions only and not to single question measures or measures of behaviors or facts. With respect to opinions, we were able to evaluate only one measurement instrument on vacation experience. The cross-validation using a different topic and comparable multi-factor structure (QE-items) provided comparable results with respect to measurement invariance analyses and showed metric and scalar invariance for devices. With respect to reliabilities, lower reliability for mobile devices could be identified as a stable result. This provides some hints that our results would also be expected to be obtained for scalar questions on other topics. In particular, we assume that results would be generalizable for measurement instruments and questionnaires of a high reliability, as sufficiently high reliability is a prerequisite for the stable results (e.g., Kline 2016). However, because in the cross-validation analysis we could not test the effects for different rating scale formats and different number of scale points, more replicative research when using other concepts and questionnaires is needed to allow for a broader generalization of results. With respect to the scale points, however, it should be mentioned that use of numbers could additionally affect the results. As numbers are used as standard in online

surveys, despite their undesirable properties known from the literature (e.g., Krosnick and Fabrigar 1997; Menold 2020), more research on mixed-device surveys is needed when implementing fully verbalized rating scales. Finally, one critical feature was that an online access panel population was used in the study, which is a selective population due to device ownership, interest and experience in doing online surveys.

Nevertheless, the present study has one particular strength which is unusual in methodological research. We used a randomized three-factor design with a systematic variation of devices, formats and numbers of scale points that allows for casual interpretation of the effects (cf., Hox et al. 2015). We also provided a control for potential sample differences between the analyzed groups with respect to the key socio-demographic variables showing stability of the results provided. The added value of this study is also that it is a first evaluation of measurement invariance between devices in dependence on both, the format of rating scales and the number of scale points. Next, it assesses measurement quality in terms of Composite Reliability, which has rarely been done before.

The study provides clear results and therefore has clear implications for the choices made when designing mixed-device surveys. The devices as such would be only a small source of systematic or non-systematic bias, but less so different formats or particularly different number of scale points. Our results therefore support the position that one should use a unified design for different devices, which was also recommended for mixed modes (DeLeeuw 2018; Dillman, Smyth, and Christian 2014). The mixing of formats and different numbers of scale points should be avoided, in particular. If a seven-point rating scale is used for PCs and five points for smartphones (both with Radio Buttons), for example, the data are unlikely to be comparable and usable for the analyses. Radio Buttons with seven categories as a low-tech solution (as compared to the combination of categorical VAS and Slider) were formats with reasonable measurement quality and little device effects and these realizations therefore appear to be preferable over other choices.

Last but not least, we hope that our study will promote further research on this topic using our method of systematic experimental variation and reliance on the LVM approach.

## 6. Data Availability

Access data and software sources: doi: 10.6084/m9.figshare.17081921

Appendices

Appendix A: Screenshots of Different Experimental Conditions

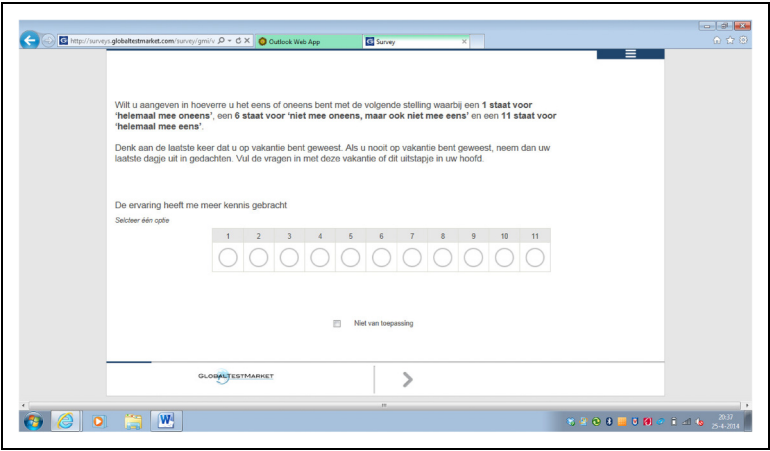


Figure A1. Regular desktop, Radio Buttons, 11-point scale.

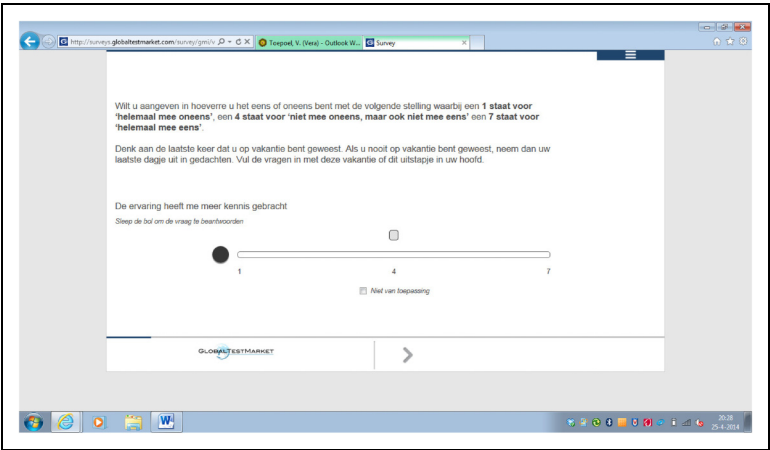


Figure A2. Regular desktop, Slider bar, 7-point scale. This bar works on a drag-and-drop principle. The pointer is initially placed to the left of the bar. Note the exact same length as the 11-point scale with Radio Buttons in Figure 1.

Wilt u aangeven in hoeverre u het eens of oneens bent met de volgende stelling waarbij een 1 staat voor 'helemaal mee oneens', een 4 staat voor 'niet mee oneens, maar ook niet mee eens' en 7 staat voor 'helemaal mee eens'.

Denk aan de laatste keer dat u op vakantie bent geweest. Als u nooit op vakantie bent geweest, neem dan uw laatste dagje uit in gedachten. Vul de vragen in met deze vakantie of dit uitstapje in uw hoofd.

De ervaring heeft me meer kennis gebracht

Klik op de schaal om de vraag te beantwoorden.

1 4 7

☐ Niet van toepassing

GLOUWTESTMARKET

**Figure A3.** Regular desktop, categorical VAS, 7-point scale. Works with a point-and-click principle.

Wilt u aangeven in hoeverre u het eens of oneens bent met de volgende stelling waarbij een 1 staat voor 'helemaal mee oneens', een 4 staat voor 'niet mee oneens, maar ook niet mee eens' en 7 staat voor 'helemaal mee eens'.

Denk aan de laatste keer dat u op vakantie bent geweest. Als u nooit op vakantie bent geweest, neem dan uw laatste dagje uit in gedachten. Vul de vragen in met deze vakantie of dit uitstapje in uw hoofd.

De ervaring heeft me meer kennis gebracht

Klik op de schaal of sleep de bol om de vraag te beantwoorden

1 4 7

☐ Niet van toepassing

GLOUWTESTMARKET

**Figure A4.** Regular desktop, Combination VAS/Slider, 7-point scale. Works on both a point-and-click and drag-and-drop principle.

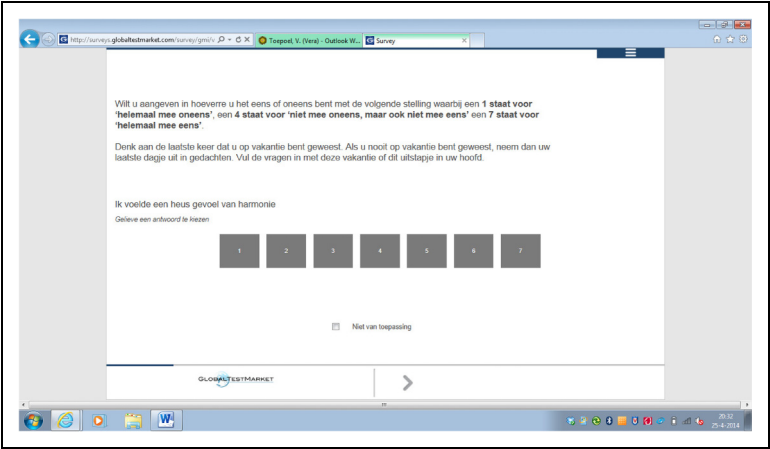


Figure A5. Regular desktop, Big Buttons, 7-point scale

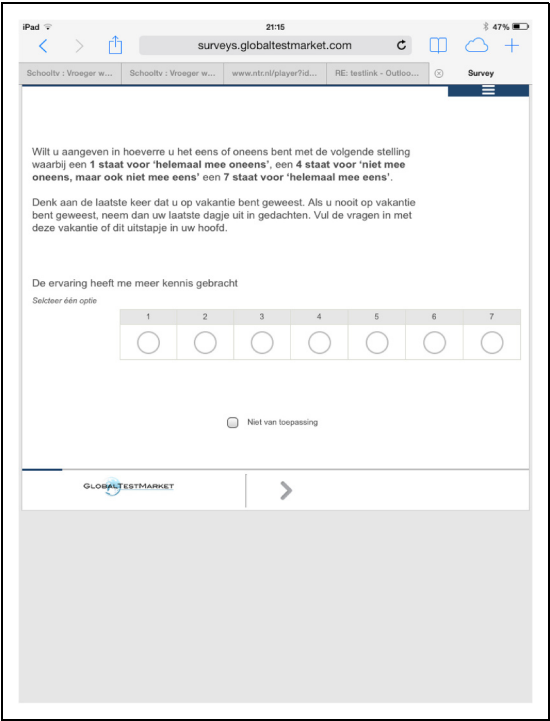


Figure A6. Tablet, Radio Buttons, 7-point scale

The screenshot shows a mobile phone screen with a survey interface. At the top, the status bar displays 'KPN NL 3G', the time '20:49', and a battery level of '68%'. Below the status bar is a dark grey header with the URL 'surveys.globaltestmarket.com' and a refresh icon. A blue navigation bar with a white menu icon is positioned below the header. The main content area is white and contains the following text:

Wilt u aangeven in hoeverre u het eens of oneens bent met de volgende stelling waarbij een **1 staat voor 'helemaal mee oneens'**, een **4 staat voor 'niet mee oneens, maar ook niet mee eens'** een **7 staat voor 'helemaal mee eens'**.

Denk aan de laatste keer dat u op vakantie bent geweest. Als u nooit op vakantie bent geweest, neem dan uw laatste dagje uit in gedachten. Vul de vragen in met deze vakantie of dit uitstapje in uw hoofd.

De ervaring heeft me meer kennis gebracht

*Sleep de bol om de vraag te beantwoorden*

A 7-point slider scale is shown with a horizontal line and tick marks at 1, 4, and 7. A black circular slider is positioned at the value 5, which is also indicated by a small box with the number '5' above it.

☐ Niet van toepassing

At the bottom of the screen is a dark grey navigation bar with five white icons: a left arrow, a right arrow, an upload icon (a square with an upward arrow), a speech bubble icon, and a document icon.

**Figure A7.** Mobile phone, Slider, 7-point scale

## **Appendix B: Vacation Experience Scale**

### *Entertaining Factor*

- A 3. It was nice to watch other people's activities
- A 7. Watching other people was amusing
- A 11. I really enjoyed watching what other people were doing
- A 15. Other people's activities were nice to watch

### *Educational Factor*

- A 1. The experience brought me more knowledge
- A 5. I learned a lot
- A 9. It stimulated my curiosity to learn new things
- A 13. It really was a learning experience

### *Escapist Factor*

- A 4. I felt like a different person there
- A 8. I felt like I was living in a different time or place
- A 12. The experience made me think that I was someone else
- A 16. I really escaped from reality

### *Aesthetic Factor*

- A 2. I felt a real sense of harmony
- A 6. Just being there was really nice
- A 10. It was quite boring there
- A 14. It was very beautiful there

Note. Own translation from Dutch.

## **Appendix C: Items on Questionnaire Evaluation**

### *Factor 1: Evaluation of the Questionnaire*

- X7: On a scale of 0-10, how did you find the design of the questions?
- X8: On a scale of 0-10, how did you feel about the user-friendliness of the questions?
- X5: On a scale of 0-10, were the questions sufficiently clear?

X11: On a scale of 0-10, how difficult was it for you to complete this questionnaire?

*Factor 2: Respondents' Motivation*

X10: On a scale of 0-10, how serious were you about answering this questionnaire?

X1: On a scale of 0-10, how motivated are you to complete this questionnaire?


X9: On a scale of 0-10, how motivated were you with respect to answering this questionnaire?


Note. Own translation from Dutch.

**Table AC\_1.** Patternmatrix of the Exploratory Factor Analysis (ML method) for the Questionnaire Evaluation Items.

Patternmatrix		
Items	Factor	
	1	2
X8: On a scale of 0-10, how did you feel about the user-friendliness of the questions?	0.933	
X7: On a scale of 0-10, how did you find the design of the questions?	0.783	
X5: On a scale of 0-10, were the questions sufficiently clear?	0.530	0.202
X11: On a scale of 0-10, how difficult was it for you to complete this questionnaire?	−0.284	
X9: On a scale of 0-10, how motivated were you with respect to answering this		0.884
X1: On a scale of 0-10, how motivated are you to complete this questionnaire?		0.680
X10: On a scale of 0-10, how serious were you about answering this questionnaire?		0.634

## ORCID iDs

Natalja Menold  <https://orcid.org/0000-0003-1106-474X>

Vera Toepoel  <https://orcid.org/0000-0001-7315-6311>

## Supplemental Material

Supplemental material for this article is available online.

## References

- Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. New York, NY: Wiley.
- Alwin, Duane F. and Jon A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods & Research* 20:139–81. doi: 10.1177/0049124191020001005
- Antoun, C., J. Katz, J. Argueta, and L. Wang. 2018. "Design Heuristics for Effective Smartphone Questionnaires." *Social Science Computer Review* 36(5):557–74. doi: 10.1177/0894439317727072
- Beauducel, Andre and Werner W. Wittmann. 2005. "Simulation Study on Fit Indexes in CFA Based on Data With Slightly Distorted Simple Structure." *Structural Equation Modeling: A Multidisciplinary Journal* 12(1):41–75. doi: 10.1207/s15328007sem1201\_3
- Chen, Fang Fang. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 14(3):464–504. doi: 10.1080/10705510701301834
- Choudhury, S. and D. Bhattacharjee. 2014. "Optimal Number of Scale Points in Likert Type Scales for Quantifying Compulsive Buying Behaviour." *Asian Journal of Management Research* 4(3):432–40.
- Churchill, Gilbert A. and J. Paul Peter. 1984. "Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis." *Journal of Marketing Research* 21(4):360–75. doi: 10.2307/3151463
- Couper, Mick P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly* 75(5):889–908. doi: 10.1093/poq/nfr046
- Couper, Mick P., Roger Tourangeau, Frederick G. Conrad, and Scott D. Crawford. 2004. "What They See is What We Get. Response Options for Web Surveys." *Social Science Computer Review* 22(1):111–27. doi: 10.1177/0894439303256555
- Cronbach, Lee J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16(3):297–334. doi: 10.1007/BF02310555

- De Beuckelaer, A., S. Toonen, and E. Davidov. 2013. "On the Optimal Number of Scale Points in Graded Pair Comparisons." *Quality and Quantity* 47(5):2869–82. doi: 10.1007/s11135-012-9695-2
- De Bruijne, M. and A. Wijnant. 2013. "Can Mobile Web Surveys Be Taken on Computers? A Discussion on a Multi-Device Survey Design." *Survey Practice* 6(4):1–8. doi: 10.29115/SP-2013-0019
- De Bruijne, M. and A. Wijnant. 2014. "Improving Response Rates and Questionnaire Design for Mobile Web Surveys." *Public Opinion Quarterly* 78(4):951–62. doi: 10.1093/poq/nfu046
- De Leeuw, Edith D. 1992. Data Quality in Mail, Telephone and Face to Face Surveys. Amsterdam: Vrije Universiteit (Doctoral dissertation). <https://edithl.home.xs4all.nl/pubs/disseddl.pdf>
- DeLeeuw, E. D. 2018. "Mixed-Mode: Past, Present, and Future." *Survey Research Methods*, 12(2):75–89. doi:10.18148/srm/2018.v12i2.7402
- Dillman, D. A., J. D. Smyth, and L. M. Christian. 2014. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: Wiley.
- Funke, Frederik. 2016. "A Web Experiment Showing Negative Effects of Slider Scales Compared to Visual Analogue Scales and Radio Buttons." *Social Science Computer Review* 34(2):244–54. doi: 10.1177/0894439315575477
- Funke, Frederik and Ulf-Dietrich Reips. 2012. "Why Semantic Differentials in Web-Based Research Should Be Made from Visual Analogue Scales and Not from 5-Point Scales." *Field Methods* 24(3):310–27. doi: 10.1177/1525822X12444061
- Funke, Frederik, Ulf-Dietrich Reips, and Randall K. Thomas. 2011. "Sliders for the Smart: Type of Rating Scale on the Web Interacts With Education Level." *Social Science Computer Review* 29(2):221–31. doi: 10.1177/0894439310376896
- Groves, Robert M., Floyd J. Fowler, P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2nd ed. Oxford: Wiley.
- Gummer, Tobias, Franziska Quoß, and Joss Roßmann. 2019. "Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones?" *Social Science Computer Review* 37(3):371–84.
- Haan, Marieke, Peter Lugtig, and Vera Toepoel. 2019. "Can we Predict Device use? An Investigation into mobile Device use in Surveys." *International Journal of Social Research Methodology* 22(5):517–31. doi: 10.1080/13645579.2019.1593340
- Heerwegh, D. and G. Loosveldt. 2011. "Assesing Mode Effects in a National Crime Victimization Survey Using Structural Equation Models: Social Desirability Bias and Acquiescence." *Journal of Official Statistics* 27(1):49–63.
- Hox, Joop J., Edith D. De Leeuw, and Eva A. O. Zijlmans. 2015. "Measurement Equivalence in Mixed Mode Surveys." *Frontiers in Psychology* 6:87. doi: 10.3389/fpsyg.2015.00087

- Hu, Jingwei. 2020. "Horizontal or Vertical? The Effects of Visual Orientation of Categorical Response Options on Survey Responses in Web Surveys." *Social Science Computer Review* 38(6):779–92. doi:10.1177/0894439319834296
- Hu, Li-tze and Peter M. Bentler. 1999. "Cutoff Criteria for fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus new Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6(1):1–55. doi: 10.1080/10705519909540118
- Jenkins, C. R. and D. A. Dillman. 1995 "Towards a Theory of Self-Administered Questionnaire Design." Pp. 165–96 in *Survey Measurement and Process Quality*, edited by Lars E. Lyberg, Paul P. Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York, NY: Wiley.
- Johnson, Edward Paul. 2015. "Mobile Devices and Modular Survey Design." Survey Sampling International. Accessed May 19, 2017. <http://papor.ipower.com/wp-content/uploads/2016/01/2015-PAPOR-Short-Course-1.pdf>.
- Jöreskog, K. G. 1971. "Simultaneous Factor Analysis in Several Populations." *Psychometrika* 36(4):409–26. doi: 10.1007/BF02291366
- Klausch, Thomas, Joop J. Hox, and Barry Schouten. 2013. "Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions." *Sociological Methods & Research* 42(3):227–63. doi: 10.1177/0049124113500480
- Kline, Rex B. 2016. *Principles and Practice of Structural Equation Modeling*. New York, London: The Guilford Press.
- Krosnick, J. A. and L. R. Fabrigar. 1997 "Designing Rating Scales for Effective Measurement in Surveys." Pp. 141–64 in *Survey Measurement and Process Quality*, edited by Lars E. Lyberg, Paul P. Biemer, Martin Collins, Edith D. de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York, NY: Wiley.
- Lugtig, Peter, Vera Toepoel, and Alerk Amin. 2016. "Mobile-only web Survey Respondents." *Survey Practice* 9(4). doi: 10.29115/SP-2016-0020
- Maitland, Aaron. 2009. "How Many Scale Points Should I Include for Attitudinal Questions?" *Survey Practice* 2(5). doi: 10.29115/SP-2009-0023
- Mayerl, Jochen. 2016 "Environmental Concern in Cross-National Comparison – Methodological Threats and Measurement Equivalence." Pp. 182–204 in *Green European: Environmental Behaviour and Attitudes in Europe in a Historical and Cross-Cultural Comparative Perspective*, edited by Audrone Telesiene and Mattias Groß. London, New York: Routledge.
- Menold, Natalja. 2020. "Rating-Scale Labeling in Online Surveys: An Experimental Comparison of Verbal and Numeric Rating Scales with Respect to Measurement Quality and Respondents' Cognitive Processes." *Sociological Methods & Research* 49(1):79–107. doi: 10.1177/0049124117729694

- Menold, Natalja and Christoph J. Kemper. 2015. "The Impact of Frequency Rating Scale Formats on the Measurement of Latent Variables in web Surveys – an Experimental Investigation Using a Measure of Affectivity as an Example." *Psihologija* 48(4):431–49. doi: 10.2298/PSI1504431M
- Menold, Natalja and Tenko Raykov. 2016. "Can Reliability of Multiple Component Measuring Instruments Depend on Response Option Presentation Mode?" *Educational and Psychological Measurement* 76(3):454–69. doi: 10.1177/0013164415593602
- Menold, Natalja and Anja Tausch. 2016. "Measurement of Latent Variables With Different Rating Scales: Testing Reliability and Measurement Equivalence by Varying the Verbalization and Number of Categories." *Sociological Methods & Research* 45(4):678–99. doi: 10.1177/0049124115583913
- Meredith, William. 1993. "Measurement Invariance, Factor Analysis and Factorial Invariance." *Psychometrika* 58(4):525–43. doi: 10.1007/BF02294825
- Miller, George A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63(2):81–97. doi: 10.1037/h0043158
- Muthén, Bengt O. 2002. "Beyond SEM: General Latent variable Modeling." *Behaviormetrika* 29(1):81–117. doi: 10.2333/bhmk.29.81
- Muthén, Linda K. and Bengt O. Muthén. 2014. *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.
- Pine, B. Joseph and James H. Gilmore. 2011. *The Experience Economy: Work is Theatre & Every Business is a Stage*. Boston: Harvard Business Press.
- Poggio, Teresio, Michael Bosnjak, and Kai Weyandt. 2015. "Survey Participation via Mobile Devices in a Probability-Based Online-Panel: Prevalence, Determinants, and Implications for Nonresponse." *Survey Practice* 8(1). doi:10.29115/SP-2015-0002
- Preston, Carolyn C. and Andrew M. Colman. 2000. "Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences." *Acta Psychologica* 104(1):1–15. doi: 10.1016/S0001-6918(99)00050-5
- Rammstedt, Beatrice, Constanze Beierlein, Elmar Brähler, Michael Eid, Johannes Hartig, Martin Kersting, Stefan Liebig, Josef Lukas, Anne-Kathrin Mayer, Natalja Menold, Jürgen Schupp, and Erich Weichselgartner. 2015. "Quality standards for the development, application, and evaluation of measurement instruments in social science survey research." RatSWD Working Papers 245. Rat für Sozial- und Wirtschaftsdaten (RAT-SWD), Berlin: SCIVERO Verlag.
- Raykov, Tenko. 2012. "Scale Construction and Development Using Structural Equation Modeling." Pp. 472–92 in *Handbook of Structural Equation Modeling*, edited by R. H. Hoyle. New York: The Guilford Press.

- Raykov, Tenko and G. A. Marcoulides. 2011. *Introduction to Psychometric Theory*. New York: Taylor & Francis.
- Revilla, Melanie A. 2013. "Measurement Invariance and Quality of Composite Scores in a Face-to-Face and a web Survey." *Survey Research Methods* 7(1):17–28. doi: 10.18148/srm/2013.v7i1.5098
- Revilla, Melanie, Daniele Toninelli, Carlos Ochoa, and Germán Loewe. 2016. "Do Online Access Panels Need to Adapt Surveys for mobile Devices?" *Internet Research* 26(5):1209–27. doi: 10.1108/IntR-02-2015-0032
- Saris, Willem E. and Irmtraud Gallhofer. 2007. "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions." *Survey Research Methods* 1(1):29–43. doi: 10.18148/srm/2007.v1i1.49
- Schaeffer, Nora Cate and Jennifer Dykema. 2011. "'Questions for Surveys : Current Trends and Future Directions.'" *Public Opinion Quarterly* 75(5):909–61. doi: 10.1093/poq/nfr048
- Scheuch, Erwin K. 1993. "The Cross-Cultural use of Sample Surveys: Problems of Comparability." *Historical Social Research* 18(2):104–38. doi: 10.12759/hsr.18.1993.2.104-138
- Schouten, Barry, Jan van den Brakel, Bart Buelens, Jan van der Laan, and Thomas Klausch. 2013. "Disentangling Mode-Specific Selection and Measurement Bias in Social Surveys." *Social Science Research* 42(6):1555–70. doi: 10.1016/j.ssresearch.2013.07.005
- Toepoel, Vera, Marcel Das, and Arthur van Soest. 2009. "Design of Web Questionnaires: The Effect of Layout in Rating Scales." *Journal of Official Statistics* 25(4):509–28.
- Toepoel, Vera and Don A Dillman. 2011. "Words, Numbers, and Visual Heuristics in Web Surveys: Is There a Hierarchy of Importance?" *Social Science Computer Review* 29(2):193–207. doi: 10.1177/0894439310370070
- Toepoel, Vera and Frederik Funke. 2018. "Sliders, Visual Analogue Scales, or Buttons: Influence of Formats and Scales in Mobile and Desktop Surveys." *Mathematical Population Studies* 25(2):112–22. doi: 10.1080/08898480.2018.1439245
- Toepoel, Vera and Peter Lugtig. 2015. "Online Surveys are Mixed-Device Surveys. Issues Associated with the Use of Different (Mobile) Devices in Web Surveys." *methods, data, analyses* 9(2):155–61. doi: 10.12758/mda.2015.009
- Tourangeau, Roger, Mick P. Couper, and Frederick G. Conrad. 2013. "Up Means Good: The Effect of Screen Position on Evaluative Ratings in Web Surveys." *Public Opinion Quarterly* 77(1):69–88. doi: 10.1093/poq/nfs063
- Winkielman, Piotr and John T. Cacioppo. 2001. "Mind at Ease Puts a Smile on the Face: Psychophysiological Evidence That Processing Facilitation Elicits Positive

- Affect.” *Journal of Personality and Social Psychology* 81(6):989–1000. doi:10.1037/0022-3514.81.6.989
- Zajonc, R. B. 2001. “Mere Exposure: A Gateway to the Subliminal.” *Current Directions in Psychological Science* 10(6):224–28. doi: 10.1111/1467-8721.00154

### Author Biography

**Natalja Menold** completed a Master’s degree in psychology at the University of Tuebingen in 2000. She received her doctorate from the University of Dortmund in 2006. From 2007 to 2019 she was a researcher, senior survey methodologist and from 2011 head of the “questionnaire design and evaluation” team at the GESIS-Leibniz Institute for the Social Sciences in Mannheim, Germany. In 2017, she completed her habilitation at the University of Mannheim, Faculty Social Sciences. She is currently professor of Methods in Empirical Social Research at the Institute of Sociology of the TU Dresden. Her research interests include measurement quality and measurement error.

**Vera Toepoel** is Assistant Professor at the Department of Methodology & Statistics at Utrecht University. She completed her PhD on the Design of Online Questionnaires in 2008 and has been involved in setting up and maintaining several probability and non-probability online panels. Her research interests focus on survey methodology and mobile surveys and passive measurement in particular. Vera is the President of RC33 (Research Committee on Methods and Logistics of the International Sociological Association) and the Vice-Chair of the European Survey Research Association. She is the author of the book “Doing Surveys Online” published by Sage.