

The Effect of Seriousness and Device Use on Data Quality

Anne-Roos Verbree¹, Vera Toepoel¹,
and Dominique Perada¹

Social Science Computer Review
2020, Vol. 38(6) 720-738

© The Author(s) 2019



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0894439319841027

journals.sagepub.com/home/ssc



Abstract

Nonserious, inattentive, or careless respondents pose a threat to the validity of self-report research. The current study uses data from the Growth from Knowledge Online Panel in which respondents are representative of the Dutch population in education, gender, and age over 15 years ($N = 5,077$). By doing regression analyses, we investigated whether self-reported seriousness and motivation are predictive of data quality, as measured using multiple indicators (i.e., nonsubstantial values, speeding, internal data consistency, nondifferentiation, response effects). Device group and demographic characteristics (i.e., education, gender, age) were also included in these analyses to see whether they predict data quality. Moreover, it was examined whether self-reported seriousness differed by device group and demographic characteristics. The results show that self-reported seriousness and motivation significantly predict multiple data quality indicators. Data quality seems similar for different device users, although smartphone users showed less speeding. Demographic characteristics explain little of the variance in data quality. Of those, education seems to be the most consistent predictor of data quality, where lower educated respondents show lower data quality. Effect sizes for all analyses were in the small to medium range. The present study shows that self-reported seriousness can be used in online attitude survey research to detect careless respondents. Future research should clarify the nature of this relationship, for example, regarding longer surveys and different wordings of seriousness checks.

Keywords

seriousness, motivation, seriousness checks, self-report, data quality, devices

In today's society, taking surveys seems inevitable (e.g., Chen, 2011; Secolsky & Denison, 2012). For instance, students are asked to evaluate academic matters, employees are asked to complete satisfaction surveys, and consumers of different products and customers of various services are asked about their satisfaction through surveys. Information obtained from surveys can be used to predict learning outcomes, monitor trends, investigate preferences, and the list goes on (Chen, 2011). Nowadays, most surveys are administered online, not only on desktop or laptop computers

¹ Utrecht University, Utrecht, the Netherlands

Corresponding Author:

Anne-Roos Verbree, Utrecht University, Utrecht, the Netherlands.

Email: a.r.verbree@uu.nl

but also on mobile devices such as tablets and smartphones (Lugtig, Toepoel, & Amin, 2016; Toepoel & Lugtig, 2014). Advantages of online surveys, such as convenient collection of data from a large number and diversity of respondents (Aust, Diedenhofen, Ullrich, & Musch, 2013; Johnson, 2005; Reips, 2000, 2002, 2009), are accompanied by the threat of obtaining useless data from nonserious, unmotivated, inattentive, or careless respondents (Maniaci & Rogge, 2014; Meade & Craig, 2012; Reips, 2000, 2002). For example, some respondents might click through a survey out of curiosity instead of providing well-thought answers or rush through the questions because of the effortlessness of online surveys (Johnson, 2005; Reips, 2009). The phenomenon of respondents' behavior characterized by generating a merely satisfactory answer instead of having the motivation to provide accurate and optimal responses is called satisficing (Krosnick, 1991; Zhang, 2013). It must be taken into account that survey results are only valid and valuable if respondents answer the questions seriously resulting in good data quality (Chen, 2011; Maniaci & Rogge, 2014).

Data quality refers to the degree in which raw data, provided directly by respondents, accurately reflect respondents' true levels of the constructs which are supposed to be measured (Meade & Craig, 2012). Obtaining survey results from nonserious respondents may result in poor data quality, threatening the validity of research (Oppenheimer, Meyvis, & Davidenko, 2009; Reips, 2009). Inattentive responses can have negative effects on effect sizes and power, depending on the proportion of inattentiveness in the sample (Maniaci & Rogge, 2014). Despite the great prevalence of research using self-reported data in the literature, not much attention has been spent on the degree of (non)serious responding in self-report research. Consequently, it is crucial to investigate whether a self-reported measure of seriousness is an effective approach to detect nonserious respondents. If so, nonserious respondents could be removed from a data set to improve data quality and to obtain more valid knowledge about various aspects of the world. Therefore, the current article aims to gain insight about the quality of data obtained through online surveys, by using data from the nonprobability online access panel of Growth from Knowledge (GfK). The aim of the current study is to empirically examine (1) whether self-reported seriousness is a significant predictor of data quality, (2) whether data quality varies for respondents who use different devices, and (3) whether data quality differs for respondents with different demographic characteristics. In addition, we examine whether seriousness differs for different device and demographic groups.

Background

Seriousness Checks

A simple way to address the problem of nonserious respondents is to ask them about the seriousness of their participation, often referred to as seriousness checks (Aust et al., 2013; Reips, 2000, 2009). In the study of Aust and colleagues (2013), respondents self-reported their seriousness after completing a survey about political attitudes by answering the following question: "It would be very helpful if you could tell us at this point whether you have taken part seriously, so that we can use your answers for our scientific analysis, or whether you were just clicking through to take a look at the survey?" (p. 530). They found that nonserious respondents were able to identify themselves as such since the seriousness check predicted data quality, measured by correlations between particular survey items and agreement with official voting results. Likewise, in the second study of Oppenheimer and colleagues (2009), self-reported motivation in a paper-and-pencil survey was higher for respondents who passed than for respondents who failed an instructional manipulation check, which is an indicator of satisficing as it measures whether respondents read the instructions. However, motivation did not differ between failing and passing respondents in their first study.

In previous studies, seriousness checks have been conducted both before (Reips, 2002, 2008, 2009) and after (Aust et al., 2013) participation in the survey. Reips (2002, 2008, 2009) has suggested to employ seriousness checks in the early stage of the survey because this has been shown to be the best predictor of dropout rates and thus a measure of motivation. In addition, conducting a seriousness check prior to the completion of the survey can be taken as a precaution to reduce dropout rates (Reips, 2002). Aust and colleagues (2013) implemented a seriousness check after completion of the survey, assuming this was more in line with the true nature of participation by reflecting a potential change of mind during participation. By removing nonserious respondents, such seriousness checks have the potential to improve data quality, which can be measured by different indicators as will be described next.

Data Quality

In the case of attitude measurement through surveys, it is impossible to detect whether a respondent is answering a question truthfully. There is no validation data available, such as information in registers. In these cases, different indicators can be used as an indirect measure of data quality.

Nonsubstantial values: Item nonresponse and “not applicable” answers. Nonsubstantial values include item nonresponse and selecting an option like “don’t know,” “no opinion,” or “no answer,” if provided (Heerwegh & Loosveldt, 2002). Item nonresponse is perhaps the most widely used indicator regarding data quality in the existing literature (e.g., de Leeuw, Hox, & Huisman, 2003; Toepoel & Lugtig, 2014; Weber, Denk, Oberecker, Strauss, & Stummer, 2008) and is characterized by blanks or gaps in the data set for some respondents for some specific questions (de Leeuw et al., 2003). Also, a large number of “don’t know,” “no opinion,” or “not applicable” answers for a single respondent can be considered as an indicator of satisficing and nonserious answering and accordingly of low data quality (e.g., Kaminska, McCutcheon, & Billiet, 2010; Krosnick, 1991; Lenzner, 2012).

Speeding. Additionally, speeding can be examined as an indicator of data quality. Respondents with strikingly low response times, frequently called speeders, may save time by glancing over instructions, making impulsive judgements, performing shallow memory searching, or simply by answering randomly (Aust et al., 2013). The assumption behind this indicator is a nonlinear relationship between response time and data quality. This implies strikingly low response times are assumed to result from satisficing, indicating nonserious respondents. Once a certain threshold is identified, response times below the threshold may or may not be considered nonserious (Meade & Craig, 2012).

Internal data consistency. Another indicator of data quality is within-person internal data consistency, referring to the consistency of a response string for an individual. The underlying assumption of techniques measuring individual consistency is that attentive respondents provide patterns of responses that are internally consistent (Curran, 2016). Thus, it is expected that respondents provide similar answers to items measuring the same theoretical construct, for example, “I change my mood a lot” (p. 90) and “I have frequent mood swings” (p. 90) from the subscale Neuroticism from the Big Five Personality Inventory (Karim, Zamzuri, & Nor, 2009).

Nondifferentiation and response effects. Although attentive respondents are thought to provide internally consistent data, it is assumed that they do not use the same response option for long periods of time (Curran, 2016). Consequently, responding too consistently to items measuring theoretically distinct constructs indicates nonserious responding and can be used to detect respondents with a

different pattern of nonserious responses than those detected by a lack of internal data consistency (Curran, 2016; Meade & Craig, 2012). Responding too consistently is called nondifferentiation, which is careless respondents' tendency to provide the same or similar rating to many consecutive items (Zhang & Conrad, 2014). This phenomenon is similar to, but not the same as, straightlining, which refers to choosing the same response option for all items in a grid so that the selected answers are in a vertical line (Zhang, 2013). Moreover, nondifferentiation could result from response effects. Those include primacy effects, recency effects, and neutral responding. We take the definition by Lugtig and Toepoel (2015) and define primacy as selecting the first answer option. They treat the selection of the first answer option as an indicator of satisficing and increased measurement error. Recency is the tendency to select the last or later response options more often (Knäuper, 1999). Neutral responding includes that long string responses of nonserious participants tend to center on the midpoint of the response scale (Huang, Liu, & Bowling, 2015). Those response effects, of which a higher degree points to lower data quality, comprise the fifth indicator in our study.

The most effective data screening approach to measure data quality is to utilize several data quality indicators simultaneously (Curran, 2016; Meade & Craig, 2012). To this end, the current study takes multiple indicators of data quality into account: nonsubstantial values, speeding, internal data consistency, nondifferentiation, and response effects. Other indicators of data quality such as the length of open-ended answers, the rounding of numerical responses, and the lack of attention to important exclusions included with a question (Medway & Tourangeau, 2015) are not taken into account because they are not available in the data set.

Devices

As mentioned, respondents use different devices to complete online surveys. In general, previous research showed no device effects on data quality. For example, Sommer, Diedenhofen, and Musch (2017) did not find a difference between desktop and mobile device (i.e., smartphone and tablet) users in data quality, which was measured by the consistency of responses and validation of responses against internal and external data criteria. Schlosser and Mays (2018) also found no noticeable differences between computer and mobile users in terms of break off rate, item non-response, and length of responses to open-ended questions. In contrast, in the study of Struminskaya, Weyandt, and Bosnjak (2015), some differences were found between the completion of online surveys using smartphones or tablets and personal computers. However, these effects were relatively small, and some indicators were not related to a device but attributable to a respondent.

Less research has examined whether self-reported seriousness differs for respondents who complete surveys on different devices. It is well-established that people use mobile devices differently from traditional computers (Toepoel & Lugtig, 2015). Nevertheless, we do not expect that the (different) use of mobile devices results in differences in self-reported seriousness as it has been shown that data quality does not differ among different device users and we expect self-reported seriousness to predict data quality.

Demographic Characteristics

Seriousness may also differ for respondents with different demographic characteristics. For example, younger, less educated, and male respondents are found to have higher levels of inattention (Maniaci & Rogge, 2014). In turn, this may affect data quality. Besides, demographic characteristics may reflect the cognitive effort and capabilities required by the respondents of which a higher level may result in lower data quality (Messer, Edwards, & Dillman, 2012). This may be true for education and age of which a consistent influence on item nonresponse has been found (e.g., de Leeuw et al., 2003; Helasoja, Prättälä, Dregval, Pudule, & Kasmel, 2002; Messer et al., 2012). Age has been

found to predict item nonresponse where higher ages are associated with higher item nonresponse; however, in the study of Struminskaya et al. (2015), older respondents generated lower item nonresponse. Answers of respondents with only a high school degree were associated with higher item nonresponse. For gender, it is less plausible to expect differences in data quality as cognitive capabilities are assumed to be similar for males and females. Indeed, many studies found no gender differences in item nonresponse (e.g., Heerwegh & Loosveldt, 2008; Kwak & Radler, 2002; Messer et al., 2012; Struminskaya, et al., 2015). However, Bech and Kristensen (2009) found that being female predicts item nonresponse which could be due to the age of their participants which ranged from 50 to 75 years. Regarding the selection of a “don’t know” option, Zeglovits and Schwarzer (2016) did not find a significant effect of age, but their results indicate that gender and education influence selecting “don’t know,” where males and higher educated respondents are less likely to select this option than females and lower educated respondents. This effect of education was also found by Young (2012). In addition, she found that respondents older than 48 are more likely to answer “don’t know,” while the effect of gender on selecting this option was found to depend on the topic of the survey questions.

Demographic characteristics can also influence speeding, although less widely studied. It has been found that education and gender do not influence the prevalence of speeding, while speeding is more likely among younger than among older respondents (Zhang, 2013; Zhang & Conrad, 2014)

There is less evidence about the relationship of education, gender, and age with the internal consistency of data. One study found that more educated respondents, females, and older respondents provide more internally consistent data. However, those effects were not significant for all the indicators used and were relatively small (Maniaci & Rogge, 2014). Similarly, in the study of Dunn, Heggstad, Shanock, and Theilgard (2018), being older and being female was correlated with internal consistency, although not significantly.

A relationship between education and response effects is consistently found, where response effects are weaker among highly educated respondents (e.g., Krosnick & Alwin, 1987; Peytchev, 2007; Struminskaya et al., 2015). In general, older respondents are associated with having larger response effects (e.g., Knäuper, 1999; Peytchev, 2007). Regarding gender, the results are mixed with some studies reporting no significant gender effect (e.g., Struminskaya et al., 2015), while, for example, Cole, McCormick, and Gonyea (2012) reported more straightlining among males for most of their item sets. This stands in contrast to Zhang and Conrad (2014) who found that being female predicted straightlining.

Hypotheses

Based on the literature mentioned above, we expect respondents who indicate being more serious to show higher data quality. Second, we do not expect using a different device (i.e., desktop, tablet, or mobile phone) to predict data quality. Accordingly, we do not predict a difference in self-reported seriousness between respondents using those different devices. We expect having completed a lower educational level to be related with lower seriousness and data quality but not for speeding. We do not have a specific expectation regarding the influence of gender on data quality and seriousness, since the literature is quite inconsistent about this relationship. Furthermore, we predict an increase in age to be related to a higher level of seriousness, with higher data quality indicated by a decreased amount of speeding and more internal data consistency, but with lower data quality on the indicators nonsubstantial values, nondifferentiation, and response effects.

Method

Research Design

In the data, a fully crossed $3 \times 5 \times 4$ factorial between-subjects experimental design was used in which respondents were assigned randomly among the conditions of three different factors. The

factors are three different devices (i.e., desktop personal computer, tablet, or mobile phone), five different response formats (i.e., radio buttons, big buttons, slider, visual analogue scale, or a mix of slider and visual analogue scale), and four different scale lengths (i.e., 5-point, 7-point, 11-point, or continuous scale). In this study, we only take device into account and combine the format and scale length conditions. For more information about the effect of these conditions on data quality, see Toepoel and Funke (2018).

Respondents

Data come from the GfK Online Panel. This nonprobability online access panel is certified by the International Organization for Standardization. Respondents completing the survey aimed to be representative in education, gender, and age over 15 of the Dutch population. Respondents owning a desktop personal computer, a tablet, and a mobile phone were selected. The questionnaire was extensively tested, so that the layout would function properly on all devices. There were no respondents who used a feature phone (i.e., a phone lacking the advanced functionality of smartphones), therefore the current article shall proceed with the term smartphone henceforth. The response rate was 30%, 34% was nonresponse, and the dropout rate was 4%; 32% of the respondents were not taken into account in the analyses because a device quota was reached. The response rate to invitations on tablets (32.03%) and smartphones (18.67%) was lower than on desktops (63.43%). A reminder was sent for tablets and smartphones. In total, 5,077 respondents completed the survey of whom 1,709 on a desktop, 1,702 on a tablet, and 1,666 on a smartphone. Of the respondents, 48% was male and 30.7% had completed lower education (i.e., prevocational secondary education or less), 41.3% medium education (i.e., senior general secondary education, preuniversity education, or secondary vocational education), and 27.9% higher education (i.e., higher professional education or university education). The mean age was 46.09 years ($SD = 16.38$).

Measurement Instrument

The survey consists of three sections and starts with three questions about attitudes toward surveys; how serious and motivated respondents are regarding completing the survey and how difficult respondents think completing surveys for the panel is in general. Then, 16 questions are asked in the experimental section about respondents' last holiday experience. These items are supposed to measure the four realms of an experience according to Pine and Gilmore's (1998) experience economy. They state that experiences can be sorted into four broad categories: entertaining, educational, escapist, and aesthetic events. See Appendices A and B for the wording of the seriousness questions and experimental questions, respectively. The survey ends with seven evaluation questions; respondents evaluated whether the survey was clear and enjoyable to complete, the design of the survey, the usability of the survey, and indicated for the second time whether respondents were serious and motivated in completing the survey and how difficult it was to complete the survey. For none of the questions, it was required to provide an answer to continue to the next question. Self-reported seriousness asked before the survey was missing for 75 respondents, and seriousness asked after completing the survey was missing for 95 respondents. For self-reported motivation, 51 and 89 respondents were missing, respectively. In total, for 277 respondents, the seriousness factor score was missing (p. 15), which left 4,800 respondents with valid factor scores.

For the 16 experimental questions, three different scale lengths and five different response formats were used. For these questions, we ignored the type of response format, since effects between those formats on data quality are found to be small. This was also true regarding different scale lengths (Toepoel & Funke, 2018). For the remaining questions, a 10-point Likert-type scale was used. Each question had a "not applicable" option presented below the other options.

Furthermore, the GfK Panel added data regarding demographic characteristics and information about the device on which respondents completed the survey.

Procedure

The survey was conducted in April 2014. Completing each survey lasted about 5 minutes. Respondents were randomly asked to complete the survey on a particular device. Respondents were allowed to complete the survey on another device than the one they were assigned to. Among those assigned to a desktop, 24% completed the survey on a tablet or smartphone. About half of the respondents assigned to a tablet or smartphone did not comply and used a different device. However, Toepoel and Funke (2018) showed no selection effect regarding the choice of which device was used. Therefore, we use data based on the device respondents used and ignore to what device they were assigned to. Response times for each question and for the whole survey were recorded. Statistical Package for the Social Sciences, Version 24.0, was used to perform all statistical analyses.

Results

Self-Reported Seriousness and Motivation

To create a score for seriousness, a principal factor analysis was performed on the four seriousness and motivation questions (administered before and after the experimental questions) with oblique rotation (promax). One factor had an eigenvalue over Kaiser's criterion of 1 and explained 66.01% of the variance. Appendix A shows the factor loadings (retrieved from the factor matrix) and the eigenvalue and percentage of the variance explained after extraction for this factor. The items loadings suggest that the extracted factor represents seriousness to complete the survey. Accordingly, Bartlett factor scores were used in all analyses as a measure of seriousness. A repeated measures multivariate analysis of variance comparing seriousness and motivation before and after the experimental questions was significant, Wilks's $L = .01$, $F(2, 4798) = 163,057.21$, $p < .001$, partial $\eta^2 = .99$. There was no significant difference between self-reported seriousness before ($M = 8.69$, $SD = 1.10$) and after ($M = 8.67$, $SD = 1.26$) the survey, $F(1, 4799) = 3.40$, $p = .065$. However, self-reported motivation to complete the survey was significantly higher after ($M = 7.82$, $SD = 1.60$) than before ($M = 7.67$, $SD = 1.27$) the survey, $F(1, 4799) = 68.75$, $p \leq .001$, partial $\eta^2 = .01$, although with a very small effect size. A possible explanation is that respondents became increasingly motivated due to the subject of the survey (i.e., holiday experiences).

In addition, it was analyzed whether the seriousness factor score differed for the device groups and demographic characteristics (i.e., education, gender, age). A one-way analysis of variance showed a significant difference in mean seriousness factor score across device groups, $F(2, 4797) = 8.93$, $p \leq .001$, $\omega^2 = .003$. Tukey's honest significant difference (HSD) test showed that desktop users ($M = 0.08$, $SD = 1.10$) had a significantly higher mean seriousness factor score than smartphone users ($M = -0.09$, $SD = 1.10$), while the difference of both those groups with the tablet group ($M = 0.00$, $SD = 1.08$) was not significant.

Regarding the analysis for the educational groups, Levene's test was found to be significant, $p = .001$, indicating differences in variance across groups. Accordingly, Welch's F was used for the analysis. The mean seriousness factor score differed significantly, $F(2, 2969.97) = 8.56$, $p \leq .001$, estimated $\omega^2 = .003$. The Games-Howell post hoc test, which was used because the assumption of homogeneity of variance was not met, showed that the group which completed higher education ($M = -0.10$, $SD = 1.07$) had a significantly lower mean seriousness score than the groups which completed medium ($M = 0.05$, $SD = 1.06$) and lower education ($M = 0.03$, $SD = 1.16$). These last two groups did not differ significantly from each other.

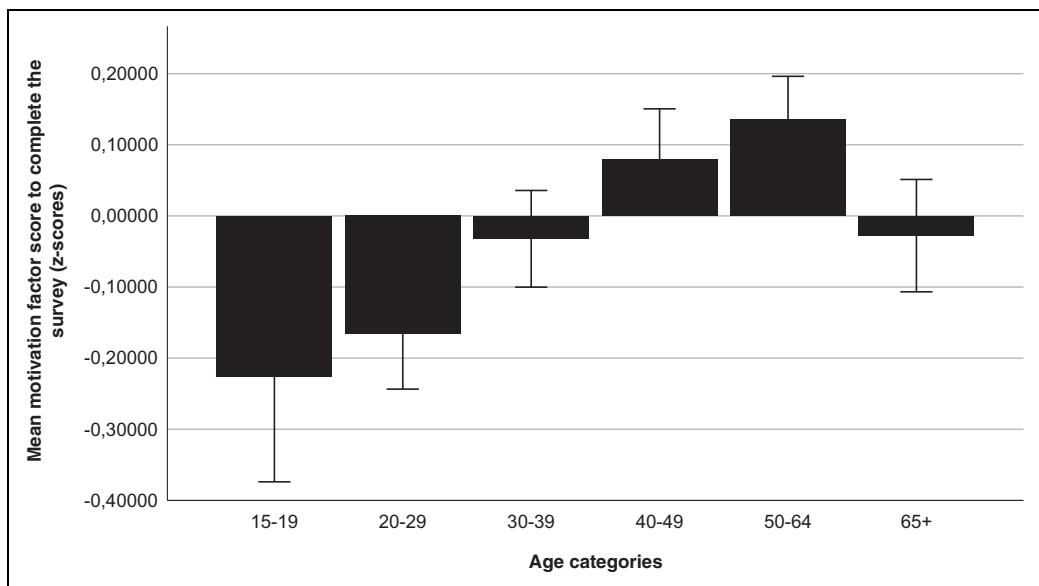


Figure 1. Mean seriousness factor score for the different age categories. Error bars represent 95% confidence intervals.

An independent *t* test was done to detect whether the mean seriousness factor score differed between male and female respondents. Male respondents ($M = -0.06$, $SD = 1.09$) reported being less serious and motivated than female respondents ($M = 0.05$, $SD = 1.09$), which was a significant difference, -0.11 , 95% CI $[-.17, -.05]$, $t(4,798) = -3.52$, $p \leq .001$, $d = -.10$.

Significant mean differences in seriousness factor score existed also between age groups, $F(5, 4794) = 10.55$, $p \leq .001$, $\omega^2 = .010$. Tukey's HSD test showed that the means of the 15–19 ($M = -0.23$, $SD = 1.16$) and 20–29 age group ($M = -0.17$, $SD = 1.10$) were significantly lower than the means of the 40–49 ($M = 0.08$, $SD = 1.07$) and 50–64 group ($M = 0.14$, $SD = 1.08$). Moreover, both the 30–39 ($M = -0.03$, $SD = 1.02$) and 65+ group ($M = -0.03$, $SD = 1.17$) had a significantly lower mean seriousness factor score than the 50–64 group (see Figure 1).

Data Quality

Data quality was measured by the following indicators: the amount of nonsubstantial values, speeding, within-person internal data consistency, nondifferentiation, and response effects (primacy, recency, and neutral responding). Regression analyses were used to examine whether seriousness, device group, and demographic characteristics predict data quality. For internal data consistency, a linear regression was performed, while Poisson regressions were performed for the other data quality indicators as these existed of count data. When the assumption of equidispersion was violated, negative binomial regression analyses were used as they are suitable for overdispersed count data.

Nonsubstantial values. Nonsubstantial values were measured by item nonresponse and the number of “not applicable” answers. Item nonresponse, the number of items out of the 16 experimental questions for which a respondent had missing values, was 0.38 on average ($SD = 1.24$). “Not applicable,” the number of items out of 16 on which a respondent chose this answer, was on average 0.54 ($SD = 1.78$). Negative binomial regression analyses were performed to see whether the amount of item nonresponse and number of “not applicable” answers could be predicted by seriousness

Table 1. Regression Analysis Summary for Seriousness, Device Group, and Demographic Characteristics Predicting Item Nonresponse and the Amount of “Not Applicable” Answers.

Predictor Variable	Item Nonresponse			“Not Applicable” Answers		
	B	SE B	Exp(B)	B	SE B	Exp(B)
Constant	−1.53***	.09	0.22 [0.18, 0.26]	−0.77***	.08	0.46 [0.40, 0.54]
Seriousness factor score	−0.14***	.02	0.87 [0.84, 0.92]	−0.31***	.02	0.74 [0.71, 0.77]
Desktop (reference)						
Smartphone	0.44***	.08	1.55 [1.33, 1.81]	−0.09	.06	0.92 [0.81, 1.04]
Tablet	0.80***	.07	2.22 [1.93, 2.56]	−0.27***	.06	0.77 [0.68, 0.87]
Lower education	0.35***	.07	1.42 [1.23, 1.62]	0.42***	.06	1.53 [1.35, 1.71]
Medium education (reference)						
Higher education	−0.03	.07	0.97 [0.84, 1.11]	−0.04	.06	0.96 [0.85, 1.09]
Male (reference)						
Female	−0.03	.06	0.98 [0.87, 1.09]	0.06	.05	1.06 [0.96, 1.17]
15–19	−0.15	.15	0.86 [0.65, 1.15]	−0.35**	.13	0.70 [0.54, 0.91]
20–29	0.28**	.09	1.33 [1.10, 1.60]	0.01	.09	1.01 [0.85, 1.20]
30–39	0.11	.09	1.12 [0.93, 1.34]	0.06	.09	1.06 [0.90, 1.25]
40–49 (reference)						
50–64	−0.24**	.09	0.79 [0.66, 0.94]	−0.18*	.08	0.84 [0.72, 0.98]
65+	−0.23*	.10	0.79 [0.65, 0.97]	0.27**	.08	1.31 [1.12, 1.54]
Overall test statistic	$\chi^2(11, n = 4,800) = 209.92***$			$\chi^2(11, n = 4,797) = 401.98***$		

Note. 95% confidence intervals for Exp(B) are presented in square brackets.

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

factor score, device group, and demographic characteristics. As hypothesized, a lower seriousness factor score predicted higher item nonresponse and more “not applicable” answers. Higher item nonresponse was also predicted by being a smartphone or tablet user and having completed lower education, all with quite large odds ratios. Lower education also predicted more “not applicable” answers. However, in contrast to item nonresponse, being a tablet user had a significant negative relationship with the number of “not applicable” answers. Being a smartphone user had a nonsignificant negative relationship with this indicator. For both indicators, there were no significant relationships with gender. Item nonresponse seemed to decrease with age, although this was not true for the 15–19 group. This decrease with age was not found for “not applicable” answers. The 15–19 and 50–64 group had significantly less “not applicable” answers than the 40–49 group, while respondents of 65+ showed more of these answers. See Table 1 for the results per predictor of both indicators.

Speeding. Response times for completing the whole survey ranged from 1.40 to 7,243.63 minutes, $M = 14.49$ ($SD = 156.09$). The mean response time for the 16 experimental questions was 8.08 minutes ($SD = 150.34$). Speeders were categorized by using a threshold of 300 ms per word, a rough estimate of reading speed (Zhang, 2013). When response times are faster than reading times, it is likely respondents did not give adequate thought to an item. A threshold was calculated by multiplying 300 ms by the number of words in each experimental question. Respondents completing a question faster than this threshold were regarded as speeding for that question. Then, the total number of questions on which respondents sped was calculated and used as indicator of overall speeding. The mean of the questions on which respondents sped was 0.17 ($SD = 0.66$).

A Poisson regression analysis showed slight violations of equidispersion. The log likelihood of the same regression model using a negative binomial distribution indicated an improved fit and was

therefore used to see whether speeding could be predicted by seriousness factor score, device group, and demographic characteristics. A higher seriousness factor score was significantly related to a moderate decrease in the odds of speeding. Being a smartphone user significantly decreased the odds that a respondent sped compared to desktop users with a large effect. The odds ratio indicates that being a smartphone user decreased the risk of speeding by more than 7 times. Educational level and gender were not significantly related to speeding. In contrast, age group was related to speeding in the sense that speeding decreased with age.

Internal data consistency. To obtain a score for internal consistency, we used the even–odd correlation as recommended by Curran (2016), Huang, Curran, Keeney, Poposki, and DeShon (2012), and Meade and Craig (2012). This measure is based on the assumption that items on the same scale are expected to correlate with each other for each individual (Huang et al., 2012). First, a principal component analysis was conducted on the 16 experimental items with oblique rotation (promax) to detect unidimensional scales within the survey. The values on these items were standardized to eliminate the effects of different scale lengths. One item (i.e., “It was quite boring there”) was intentionally recoded to reflect the direction of the other items. Four factors had eigenvalues over Kaiser’s criterion of 1 and explained together 65.84% of the variance. The scree plot pointed also to extracting four factors. Accordingly, four factors were retained (i.e., A, B, C, and D). Appendix B shows the factor loadings (retrieved from the pattern matrix) after rotation and the eigenvalues and percentages of variance explained after extraction for each factor. The items of each unidimensional scale were divided using an even–odd split based on the appearance of items (i.e., A1, A3 being odd, A2, A4 being even, etc.). Subsequently, for each respondent, the mean of the even items and of the odd items (based on the order of appearance in the survey) for each scale was calculated resulting in an even and an odd subscale score (i.e., the average response on the even questions of Scale A and the average response on the odd questions of Scale A, etc. for the other scales). Then a within-person correlation between those two sets of subscale scores was computed in Excel to obtain the even–odd correlation (Meade & Craig, 2012). Accordingly, the even–odd correlation is a value between -1 and 1 . The average correlation between the even and the odd subsets of unidimensional scales in the data was .53 ($SD = 0.48$).

A linear regression analysis was performed to see whether the even–odd correlation could be predicted by seriousness factor score, device group, and demographic characteristics. As expected, seriousness factor score had a significant positive relationship with the even–odd correlation. Being a smartphone or tablet user and having completed lower education was negatively related to the even–odd correlation, having completed higher education positively, although effects were small. No significant gender and age effects were found, except for the 20–29 group which had a negative relationship with the even–odd correlation compared to the 40–49 group (see Table 2) for the linear model.

Nondifferentiation. Since the items of the survey were not presented in a grid but each on a new page, we included nondifferentiation instead of straightlining (a measure frequently used as an indicator of data quality, e.g., Kaminska et al., 2010; Revilla, Ochoa, & Turbina, 2017; Toepoel & Lugtig, 2015) as fourth data quality indicator. To detect nondifferentiation, the long string index was used, computed as the maximum number of consecutive items to which a respondent answered with the same response option (Johnson, 2005). Accordingly, this indicator has a maximum value of 16. The mean maximum long string was 2.61 ($SD = 1.27$). A Poisson regression analysis was performed to see whether the maximum long string could be predicted by seriousness factor score, device group, and demographic characteristics. Respondents assigned to a continuous scale ($n = 597$) were excluded from this analysis due to the different nature of this answer scale. The maximum long string had a significant negative relationship with seriousness factor score and with being a smartphone or tablet

Table 2. Regression Analysis Summary for Seriousness, Device Group, and Demographic Characteristics Predicting Speeding, Internal Data Consistency, and Nondifferentiation.

Predictor Variable	Speeding			Internal Data Consistency			Nondifferentiation		
	B	SE B	Exp(B)	B	SE B	β	B	SE B	Exp(B)
Constant	−1.49***	.12	0.23 [0.18, 0.28]	0.59*** [0.55, 0.63]	.02		0.96***	.03	2.62 [2.48, 2.76]
Seriousness factor score	−0.16***	.03	0.85 [0.80, 0.91]	0.02* [0.00, 0.03]	.01	.04	−0.07***	.01	0.94 [0.92, 0.95]
Desktop									
(reference)									
Smartphone	−1.95***	.14	0.14 [0.11, 0.19]	−0.05* [−0.08, −0.01]	.02	−.05	−0.10***	.02	0.91 [0.87, 0.95]
Tablet	−0.12	.09	0.89 [0.75, 1.05]	−0.07*** [−0.11, −0.04]	.02	−.07	−0.08***	.02	0.92 [0.88, 0.96]
Lower education	0.04	.10	1.04 [0.89, 1.30]	−0.05** [−0.09, −0.02]	.02	−.05	0.13***	.02	1.13 [1.09, 1.18]
Medium education (reference)									
Higher education	0.07	.10	1.07 [0.89, 1.30]	0.04* [0.01, 0.08]	.02	.04	−0.02	.02	0.98 [0.93, 1.02]
Male (reference)									
Female	−0.13	.08	0.87 [0.75, 1.02]	0.03 [−0.00, 0.05]	.01	.03	0.02	.02	1.02 [0.99, 1.06]
15–19	1.00***	.17	2.71 [1.93, 3.79]	0.03 [−0.04, 0.11]	.04	.02	−0.08	.05	0.92 [0.84, 1.01]
20–29	0.71***	.13	2.04 [1.58, 2.62]	−0.07** [−0.12, −0.02]	.02	−.05	0.01	.03	1.01 [0.95, 1.07]
30–39	0.36**	.13	1.44 [1.11, 1.87]	−0.04 [−0.09, 0.01]	.02	−.03	−0.01	.03	0.99 [0.93, 1.05]
40–49 (reference)									
50–64	−0.24	.13	0.78 [0.61, 1.01]	−0.03 [−0.07, 0.02]	.02	−.02	0.00	.03	1.00 [0.95, 1.06]
65+	−0.62***	.15	0.54 [0.40, 0.73]	−0.04 [−0.08, 0.01]	.02	−.03	0.05	.03	1.05 [0.99, 1.12]
Overall test statistic	$\chi^2(11, n = 4,800) = 434.86^{***}$			$F(11, 4651) = 5.02^{***}, R^2 = .01, n = 4,663$			$\chi^2(11, n = 4,800) = 155.61^{***}$		

Note. 95% confidence intervals for Exp(B) and B are presented in square brackets.

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Table 3. Regression Analysis Summary for Seriousness, Device Group, and Demographic Characteristics Predicting Primacy, Recency, and Neutral Responding.

Predictor Variable	Primacy			Recency			Neutral Responding		
	B	SE B	Exp(B)	B	SE B	Exp(B)	B	SE B	Exp(B)
Constant	.39***	.06	1.48 [1.32, 1.66]	.50***	.06	1.64 [1.46, 1.84]	.87	.06	2.39 [1.14, 2.67]
Seriousness factor score	.05**	.01	1.05 [1.02, 1.09]	.33***	.02	1.38 [1.34, 1.44]	-.11***	.02	0.90 [0.87, 0.93]
Smartphone	-.07	.05	0.93 [0.84, 1.02]	.03	.05	1.03 [0.94, 1.13]	.00	.05	1.01 [0.92, 1.10]
Tablet	-.08	.05	0.92 [0.84, 1.01]	.03	.05	1.03 [0.94, 1.13]	-.01	.05	0.99 [0.90, 1.08]
Desktop (reference)									
Lower education	-.01	.05	0.99 [0.90, 1.09]	.09	.05	1.09 [0.99, 1.19]	.04	.05	1.05 [0.96, 1.14]
Medium education (reference)									
Higher education	.12*	.05	1.12 [1.03, 1.23]	-.09	.05	0.92 [0.83, 1.01]	.02	.05	1.02 [0.94, 1.12]
Male (reference)									
Female	-.03	.04	0.97 [0.90, 1.05]	.19***	.04	1.21 [1.12, 1.30]	.01	.04	1.01 [0.94, 1.09]
15-19	.05	.10	1.06 [0.87, 1.28]	-.08	.10	0.92 [0.76, 1.12]	-.06	.10	0.94 [0.78, 1.14]
20-29	-.10	.07	0.90 [0.79, 1.03]	.07	.07	1.07 [0.94, 1.22]	-.05	.06	0.95 [0.84, 1.08]
30-39	.01	.06	1.01 [0.89, 1.14]	-.11	.07	0.89 [0.79, 1.02]	.04	.06	1.04 [0.92, 1.17]
40-49 (reference)									
50-64	-.02	.06	0.98 [0.87, 1.10]	.17**	.06	1.18 [1.06, 1.33]	.09	.06	1.09 [0.98, 1.22]
65+	-.07	.07	0.93 [0.82, 1.06]	.09	.07	1.09 [0.96, 1.24]	.05	.06	1.06 [0.93, 1.19]
Overall test statistic	$\chi^2(11, n = 4,774) = 26.13^{**}$			$\chi^2(11, n = 4,209) = 412.97^{***}$			$\chi^2(11, n = 4,209) = 48.27^{***}$		

Note. 95% confidence intervals for Exp(B) are presented in square brackets.

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

user as compared to desktop users. In contrast, having completed lower education predicted a longer maximum long string. No significant gender and age effects were found (see Table 2) for the results per predictor.

Response effects. Finally, we used the number of items for which the first, last, or middle answer option was chosen out of the 16 experimental questions to reflect response effects (i.e., primacy effects, recency effects, and neutral responding, respectively). The mean primacy, recency, and neutral responding were, respectively, 1.50 ($SD = 1.97$), 2.09 ($SD = 2.68$), and 2.54 ($SD = 2.35$). Negative binomial regression analyses were performed to see whether these response effects could be predicted by seriousness factor score, device group, and demographic characteristics. Respondents assigned to a continuous scale ($n = 597$) were excluded from these regression analyses due to the different nature of this answer scale. In contrast to the expectations, seriousness factor score and having completed higher education had a significant positive relationship with choosing the first response option. No significant device, gender, and age effects on primacy were found.

Seriousness factor score also had a positive relationship with choosing the last response option. No device and education effects were found. Being female had a positive relationship with recency. The only significant age effect was a positive relationship of the 50–64 group compared to the 40–49 group with recency.

In contrast to primacy and recency and in line with the expectations, seriousness factor score had a negative relationship with neutral responding. No significant device, education, gender, and age effects were found (see Table 3) for the results per predictor of these analyses.

Discussion and Conclusion

This study has demonstrated that asking survey respondents about the seriousness and motivation of their participation predicts data quality. The extant literature is remarkably silent about the usefulness of seriousness and motivation checks. However, we found a moderate relationship between seriousness and data quality where lower seriousness predicted lower data quality on all data quality indicators included (i.e., nonsubstantial values, speeding, internal data consistency, nondifferentiation, neutral responding), except primacy and recency. For primacy and recency, a positive relationship was found with seriousness and motivation. A possible explanation is that respondents who report higher motivation might have stronger opinions. This could be true, since strong attitudes have more impact on cognition and behavior than weak attitudes (Krosnick & Abelson, 1992). Strikingly, descriptive statistics revealed higher levels of recency than of primacy. This contrasts with the literature which suggests primacy effects occur in online environments, while the results are less clear for recency (Murphy, Hofacker, & Mizerski, 2006). This inconsistency might result from the concurrence (for 15 of the 16 questions) of recency with acquiescence, the tendency to agree with or say yes to items, regardless of their content (Couch & Keniston, 1960). Moreover, the primacy effect may be a function of the nature of the survey topic (Barnette, 2001) and may be less prevalent in the current survey as people generally like holiday experiences, which concurred with recency. Further research should investigate whether and how the survey topic results in differential response effects.

Concerning data quality of respondents using different devices (i.e., desktops, tablets, smartphones), the results were inconsistent. In contrast to previous studies, mobile device users showed lower data quality on the indicators internal data consistency and item non-response, the latter which was considerably increased. However, tablet users showed less “not applicable” answers. The same although nonsignificant effect was observed for smartphone users. It is possible that the Internet connection on mobile devices is slower, resulting in

mobile device users clicking the button for the next page more than one time and consequently skipping an item (Mavletova, 2013) rather than indicating that these users show lower data quality. This is supported by the finding that mobile device users had higher data quality on other indicators. They showed less nondifferentiation than desktop users, as in line with previous studies where items were presented on separate pages (Keusch & Yan, 2017; Lugtig & Toepoel, 2016). And speeding was found to decrease strongly for smartphone but not for tablet users compared to desktop users. This may result from the fact that it takes more time on smartphones to read small print, to zoom, and to select answer options and from a slower Internet connection (Couper & Peterson, 2017; Mavletova, 2013; Wells, Bailey, & Link, 2013). No significant difference in response effects were found, corroborating the existing literature (e.g., Andreadis, 2015; Mavletova, 2013; Toepoel & Lugtig, 2014; Wells, Bailey, & Link, 2014). In sum, device use does not seem to have a large influence on data quality resulting from respondents' answering behavior but rather from using a particular device, visible in a large decrease in speeding on smartphones and a large increase in item nonresponse on mobile devices. This contrasts with that desktop users report being more seriousness and motivated than smartphone users, which could result from surveys on smartphones not being taken as serious as surveys on desktop computers (Weber et al., 2008). However, as our study showed no large negative effect of the use of mobile devices on data quality, this implies that mobile devices can be used as a feasible way of data collection in survey research.

Regarding demographic characteristics, the highest level of education that respondents had completed influenced data quality. As expected, lower educated respondents showed lower data quality compared to medium and higher educated respondents indicated by more nonsubstantial values, less internal data consistency, and more nondifferentiation. No difference in terms of speeding was found, in line with our expectations. Also, there were no differences in response effects. These findings do not correspond with self-reported seriousness. Contrary to our hypothesis, higher educated respondents reported being less serious and motivated than medium and lower educated respondents. Jaccard, McDonald, Wan, Dittus, and Quinlan (2002) found that higher educated respondents show higher self-report accuracy. In line with this and since data quality tend to be higher for respondents who completed higher education, the accuracy of their self-reported seriousness and motivation may be higher. This implies that self-reported seriousness by lower educated respondents could be somewhat higher than in reality. However, this did not erase the predictive value of seriousness, as shown in the current study.

Self-reported seriousness was lower for male than for female respondents, in line with the suggestion that females are more conscientious and willing survey takers (Lambert & Miller, 2015). However, there were no differences in data quality between males and females on all indicators, except recency which is in all probability a result of the survey topic. Consequently, rather than indeed being more serious and motivated, it is likely that female respondents only report being so. This could result from a more pronounced influence of desirability on people's answers among female than among male respondents (Philips & Clancy, 1972).

The influence of age-group on data quality was rather ambiguous. Contrary to our expectations, results for nonsubstantial values were inconsistent and older respondents did not show higher levels of internal data consistency and response effects and lower nondifferentiation, as hypothesized. The most consistent finding was that speeding decreased with age, in line with our expectations. For online survey research, this implies that data quality does not necessarily suffer when older respondents are included or overrepresented. Those effects were not perfectly reflected in self-reported seriousness and motivation, which, as expected, increased with age except for respondents of 65 years and older.

A possible reason for the discrepancy between differences in seriousness among device and demographic groups and the relationship of seriousness with data quality is that part of the non-serious respondents may answer seriousness checks in a satisficing manner, limiting the predictive value of these checks and making it difficult to establish cutoff values for seriousness checks that identify careless respondents. Nevertheless, we found a relationship of seriousness checks with data quality, which indicates the importance of adopting such checks in surveys. They can be used at the very least as quality check for the obtained data and accordingly data quality may be improved by removing nonserious participants. This could be done regardless whether those checks are included before or after a survey, as this study suggests there are no large differences of incorporating seriousness checks before or after surveys, which future research needs to verify. The finding that seriousness checks predict data quality corroborates the finding of Aust and colleagues (2013) that respondents who report being serious answer questions in a more consistent and valid manner. In contrast to the current study, their seriousness check included a reference to the importance of serious answers for the validity of research. Potentially, the effect of seriousness and motivation checks may be larger using a type of wording referring to the importance of serious answers by minimizing the number of respondents satisficing the seriousness checks, which should be investigated by future research.

Future research should also examine the relationship of self-reported seriousness and motivation with data quality in longer surveys; as in the present study, the administered survey was fairly short. This is important, since motivation may decrease over the length of the survey (Galesic & Bosnjak, 2009). In turn, this can influence data quality. For example, more item nonresponse, shorter open-ended answers, shorter response times, and lower variability of answers to question in grids have been found to increase over the length of the survey (Cole et al., 2012; Galesic & Bosnjak, 2009). Another limitation is that it could not be investigated whether certain respondents are more inclined to use certain devices, since data resulted from an experiment where respondents were assigned to a device. Further research should address this issue. Also, we ignored different scale lengths to which respondents were randomly assigned. Although Toepoel and Funke (2018) did not show large effects of scale length on data quality, scale length could have influenced the results. Nondifferentiation and response effects in particular could be subject to this.

This study contributed to the literature by showing that a self-report measure of seriousness and motivation in online attitude surveys predicts data quality. However, as shown in an analysis by Aust and colleagues (2013), few studies include seriousness checks in surveys. The findings in the current study indicate the importance of incorporating such checks in order to monitor data quality before analysis as well as to help identify and remove careless respondents and, in turn, to improve data quality.

Appendix A

Table A1. Summary of the Exploratory Factor Analysis Results of the Evaluation Questions.

Item	Factor Loadings
On a scale from 0–10, how serious are you regarding filling in this survey?	.67
On a scale from 0–10, how motivated are you to fill in this survey?	.74
On a scale from 0–10, how serious were you regarding answering this survey?	.77
On a scale from 0–10, how motivated were you regarding answering this survey?	.78
Eigenvalues after extraction	2.19
Percentage of variance	54.86

Note. *n* = 4,800. Factor loadings over .40 appear in bold. Items are translated from Dutch.

Appendix B

Table B1. Summary of the Exploratory Factor Analysis Results of the 16 Experimental Questions.

Item	Rotated Factor Loadings			
	Entertaining Events	Educational Events	Escapist Events	Aesthetic Events
I really enjoyed watching what other people were doing	.91	.00	.01	-.09
Other people's activities were nice to watch	.88	.00	.01	-.03
Watching other people was amusing	.74	-.07	.01	.02
It was nice to watch other people's activities	.60	.09	-.08	.15
I learned a lot	-.04	.90	-.04	.03
It really was a learning experience	.03	.80	.14	-.14
The experience brought me more knowledge	-.02	.73	-.14	.14
It stimulated my curiosity to learn new things	.04	.68	.09	.06
The experience made me think that I was someone else	.01	.07	.83	-.20
I really escaped from reality	-.02	-.10	.76	.10
I felt like a different person there	-.02	-.05	.70	.08
I felt like I was living in a different time or place	.00	.10	.61	.09
Just being there was really nice	.05	-.12	.06	.76
It was very beautiful there	-.03	.11	.03	.57
It was quite boring there	-.05	.09	-.09	.50
I felt a real sense of harmony	.07	.05	.14	.44
Eigenvalues	5.21	1.46	1.18	0.99
Percentage of variance	32.56	9.11	7.38	6.21

Note. $n = 4,390$. Factor loadings over .40 appear in bold. Items are translated from Dutch.

Authors' Note

Data can be obtained from the author upon request. The author can be contacted by Email: a.r.verbree@uu.nl. Statistical Package for the Social Sciences (SPSS), Version 24.0, modules such as IBM SPSS Base, IBM SPSS Advanced Statistics, IBM SPSS Bootstrapping, and Microsoft Excel were used to conduct the analyses.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Andreadis, I. (2015). Web surveys optimized for smartphones: Are there differences between computer and smartphone users? *Methods, Data, Analyses*, 9, 213–228. doi:10.12758/mda.2015.012
- Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavioural Research*, 45, 527–535. doi:10.3758/s13428-012-0265-2
- Barnette, J. J. (2001, October). *Practical measurement issues associated with data from Likert scales*. Paper presented at the American Public Health Association, Atlanta, GA.
- Bech, M., & Kristensen, M. B. (2009). Differential response rates in postal and web-based surveys in older respondents. *Survey Research Methods*, 3, 1–6. Retrieved from <https://ojs.ub.uni-konstanz.de/srm/>

- Chen, P. S. D. (2011). Finding quality responses: The problem of low-quality survey response and its impact on accountability measures. *Research in Higher Education*, 52, 659–674. doi:10.1007/s11162-011-9217-4
- Cole, S., McCormick, A. C., & Gonyea, R. M. (2012, April). *Respondent use of straight-lining as response strategy in education survey research: Prevalence and implications*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.
- Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, 60, 151–174. doi:10.1037/h0040372
- Couper, M. P., & Peterson, G. J. (2017). Why do web surveys take longer on smartphones? *Social Science Computer Review*, 35, 357–377. doi:10.1177/0894439316629932
- Curran, P. G. (2016). Methods for detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. doi:10.1016/j.jesp.2015.07.0060
- de Leeuw, E., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, 19, 153–176. Retrieved from <https://www.degruyter.com/view/j/jos>
- Dunn, A. M., Heggstad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationship with individual differences. *Journal of Business and Psychology*, 33, 105–121. doi:10.1007/s10869-016-9479-0
- Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73, 349–360. doi:10.1093/poq/nfp031
- Heerwegh, D., & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in web surveys. *Social Science Computer Review*, 20, 471–484. doi:10.1177/089443902237323
- Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, 72, 836–846. doi:10.1093/poq/nfn045
- Helasoja, V., Prättälä, R., Dregval, L., Pudule, I., & Kasmel, A. (2002). Late response and item nonresponse in the Finbalt Health Monitor Survey. *The European Journal of Public Health*, 12, 117–123. doi:10.1093/eurpub/12.2.117
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27, 99–114. doi:10.1007/s10869-011-9231-8
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100, 828–845. doi:10.1037/a0038510
- Jaccard, J., McDonald, R., Wan, C. K., Dittus, P. J., & Quinlan, S. (2002). The accuracy of self reports of condom use and sexual behaviour. *Journal of Applied Social Psychology*, 32, 1863–1905. doi:10.1111/j.1559-1816.2002.tb00263.x
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39, 103–129. doi:10.1016/j.jrp.2004.09.009
- Kaminska, O., McCutcheon, A. L., & Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74, 956–984. doi:10.1093/poq/nfq062
- Karim, S. A., Zamzuri, N. H. A., & Nor, Y. M. (2009). Exploring the relationship between Internet ethics in university students and the big five model of personality. *Computers & Education*, 53, 86–93. doi:10.1016/j.compedu.2009.01.001
- Keusch, F., & Yan, T. (2017). Web versus mobile web: An experimental study of device effects and self-selection effects. *Social Science Computer Review*, 35, 751–769. doi:10.1177/0894439316675566
- Knäuper, B. (1999). The impact of age and education on response order effects in attitude measurement. *The Public Opinion Quarterly*, 63, 347–370. doi:10.1086/297724
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213–236. doi:10.1002/acp.2350050305
- Krosnick, J. A., & Abelson, R. P. (1992). The case for measuring attitude strength in surveys. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases for surveys* (pp. 177–203). New York, NY: Russell Sage Foundation.

- Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *The Public Opinion Quarterly*, 51, 201–219. doi:10.1086/269029
- Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern respondent profile, and data quality. *Journal of Official Statistics*, 18, 257–273. Retrieved from <https://www.degruyter.com/view/j/jos>
- Lambert, A. D., & Miller, A. L. (2015). Living with smartphones: Does completion device affect survey responses? *Research in Higher Education*, 56, 166–177. doi:10.1007/s11162-014-9354-7
- Lenzner, T. (2012). Effects of survey question comprehensibility on response quality. *Field Methods*, 24, 409–428. doi:10.1177/1525822X12448166
- Lugtig, P., & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review*, 34, 78–94. doi:10.1177/0894439315574248
- Lugtig, P., Toepoel, V., & Amin, A. (2016). Mobile-only web survey respondents. *Survey Practice*, 9, 1–8. Retrieved from <http://www.surveypractice.org/>
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61–83. doi:10.1016/j.jrp.2013.09.008
- Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, 31, 725–743. doi:10.1177/0894439313485201
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437–455. doi:10.1037/a0028085
- Medway, R. L., & Tourangeau, R. (2015). Response quality in telephone surveys: Do prepaid cash incentives make a difference? *Public Opinion Quarterly*, 79, 524–543. doi:10.1093/poq/nfv011
- Messer, B. L., Edwards, M. L., & Dillman, D. A. (2012). Determinants of item nonresponse to web and mail respondents in three address-based mixed-mode surveys of the general public. *Survey Practice*, 5, 1–8. Retrieved from <http://www.surveypractice.org/>
- Murphy, J., Hofacker, C., & Mizerski, R. (2006). Primacy and recency effects on clicking behavior. *Journal of Computer-Mediated Communication*, 11, 522–535. doi:10.1111/j.1083-6101.2006.00025.x
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45, 867–872. doi:10.1016/j.jesp.2009.03.009
- Peytchev, A. A. (2007). *Participation decisions and measurement error in web surveys* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3253382).
- Philips, D. L., & Clancy, K. J. (1972). Some effects of “social desirability” in survey studies. *American Journal of Sociology*, 77, 921–940. doi:10.1086/225231
- Pine, B. J., & Gilmore, J. H. (1998). Welcome to the experience economy. *Harvard Business Review*, 76, 97–105. Retrieved from <https://hbr.org/>
- Reips, U. D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–117). San Diego, CA: Academic Press.
- Reips, U. D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, 49, 243–256. doi:10.1026//1618-3169.49.4.243
- Reips, U. D. (2008). How internet-mediated research changes science. In A. Barak (Ed.), *Psychological aspects of cyberspace: Theory, research, applications* (pp. 268–294). Cambridge, England: Cambridge University Press.
- Reips, U. D. (2009). Internet experiments: Methods, guidelines, metadata. *Human Vision and Electronic Imaging XIV*, 7240. doi:10.1117/12.823416
- Revilla, M., Ochoa, C., & Turbina, A. (2017). Making use of Internet interactivity to propose a dynamic presentation of web questionnaires. *Quality & Quantity*, 51, 1321–1336. doi:10.1007/s11135-016-0333-2
- Schlosser, S., & Mays, A. (2018). Mobile and dirty: Does using mobile devices affect the data quality and the response process of online surveys? *Social Science Computer Review*, 36, 212–230. doi:10.1177/0894439317698437

- Secolsky, S., & Denison, D. B. (2012). *Handbook on measurement, assessment and evaluation in higher education*. New York, NY: Taylor & Francis.
- Sommer, J., Diedenhofen, B., & Musch, J. (2017). Not to be considered harmful: Mobile-device users do not spoil data quality in web surveys. *Social Science Computer Review*, 35, 378–387. doi:10.1177/0894439316633452
- Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The effect of questionnaire completion using mobile devices on data quality. Evidence from a probability-based general population panel. *Methods, data, analyses*, 9, 261–292. doi:10.12758/mda.2015.014
- Toepoel, V., & Funke, F. (2018). Sliders, visual analogue scales, or buttons: Influence of formats and scales in mobile and desktop surveys. *Mathematical Population Studies*, 25, 112–122. doi:10.1080/08898480.2018.1439245
- Toepoel, V., & Lugtig, P. (2014). What happens if you offer a mobile option to your web panel? Evidence from a probability-based panel of internet users. *Social Science Computer Review*, 32, 544–560. doi:10.1177/0894439313510482
- Toepoel, V., & Lugtig, P. (2015). Online surveys are mixed-device surveys. Issues associated with the use of different (mobile) devices in web surveys. *Methods, Data, Analyses*, 9, 155–162. doi:10.12758/mda.2015.009
- Weber, M., Denk, M., Oberecker, K., Strauss, C., & Stummer, C. (2008). Panel surveys go mobile. *International Journal of Mobile Communications*, 6, 88–107. doi:10.1504/ijmc.2008.016006
- Wells, T., Bailey, J. T., & Link, M. W. (2013). Filling the void: Gaining a better understanding of tablet-based surveys. *Survey Practice*, 6, 1–9. doi:10.29115/SP-2013-0002
- Wells, T., Bailey, J. T., & Link, M. W. (2014). Comparison of smartphone and online computer survey administration. *Social Science Computer Review*, 32, 238–255. doi:10.1177/0894439313505829
- Young, R. (2012). *Don't know responses in survey research* (Doctoral dissertation). Retrieved from <https://etda.libraries.psu.edu/catalog/13934>
- Zeglovits, E., & Schwarzer, S. (2016). Presentation matters: How mode effects in item non-response depend on the presentation of response options. *International Journal of Social Research Methodology*, 19, 191–203. doi:10.1080/13645579.2014.978560
- Zhang, C. (2013). *Satisficing in web surveys: Implications for data quality and strategies for reduction* (Doctoral dissertation). Retrieved from <http://deepblue.lib.umich.edu/handle/2027.42/97990>
- Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8, 127–135. Retrieved from <https://ojs.ub.uni-konstanz.de/srm/>

Author Biographies

Anne-Roos Verbree finished a bachelor's degree in Educational Sciences and is currently enrolled in the research master Educational Sciences: Learning in Interaction. Email: a.r.verbree@uu.nl

Vera Toepoel is an assistant professor at the Department of Methods & Statistics at Utrecht University, the Netherlands. She did her PhD on online questionnaire design and wrote the book *Doing Surveys Online* (Sage). She published many papers on survey research in *Public Opinion Quarterly*, *Sociological Methods and Research*, *Survey Research Methods*, and so on. She is the president of the Research Committee of Methodology and Logic of the International Sociological Association. Email: v.toepoel@uu.nl

Dominique Perada is currently in her last year of study in the bachelor's program educational sciences at Utrecht University. In June 2019, she will graduate with a Bachelor of Science in Educational Sciences, with a particular focus on instructional design. Email: d.a.perada@students.uu.nl