Motivated Misreporting in Smartphone Surveys

Daikeler, Jessica; Bach, Ruben L.; Silber, Henning; Eckman, Stephanie

Article

Social Science Computer Review
2022, Vol. 40(1) 95–107

DOI: 10.1177/0894439319900936
journals.sagepub.com/home/ssc

Motivated Misreporting in Smartphone Surveys

Jessica Daikeler[1], Ruben L. Bach[2], Henning Silber[1], and Stephanie Eckman[3]

Abstract

Filter questions are used to administer follow-up questions to eligible respondents while allowing respondents who are not eligible to skip those questions. Filter questions can be asked in either the interleafed or the grouped formats. In the interleafed format, the follow-ups are asked immediately after the filter question; in the grouped format, follow-ups are asked after the filter question block. Underreporting can occur in the interleafed format due to respondents' desire to reduce the burden of the survey. This phenomenon is called motivated misreporting. Because smartphone surveys are more burdensome than web surveys completed on a computer or laptop, due to the smaller screen size, longer page loading times, and more distraction, we expect that motivated misreporting is more pronounced on smartphones. Furthermore, we expect that misreporting occurs not only in the filter questions themselves but also extends to data quality in the follow-up questions. We randomly assigned 3,517 respondents of a German online access panel to either the PC or the smartphone. Our results show that while both PC and smartphone respondents trigger fewer filter questions in the interleafed format than the grouped format, we did not find differences between PC and smartphone respondents regarding the number of triggered filter questions. However, smartphone respondents provide lower data quality in the follow-up questions, especially in the grouped format. We conclude with recommendations for web survey designers who intend to incorporate smartphone respondents in their surveys.

Keywords

motivated underreporting, mobile data quality, filter questions, follow-up questions, misreporting, measurement error

Many surveys use eligibility questions to ask respondents only those questions that apply to them. For example, asking unemployed respondents about working hours or salary is meaningless as these follow-up questions do not apply. Experimental evidence across several modes, topics, and countries

[1] GESIS-Leibniz Institute for the Social Sciences, Mannheim, Germany
[2] University of Mannheim, Germany
[3] RTI International, Washington, DC, USA

Corresponding Author:
Jessica Daikeler, GESIS-Leibniz Institute for the Social Sciences, B2, 1, Mannheim 68159, Germany.
Email: jessica.daikeler@gesis.org

suggests that the order in which filter and follow-up questions are asked affects data quality. When the structure of the questions makes it obvious to respondents that "no" answers will shorten the survey, they tend to engage in motivated misreporting (Eckman & Kreuter, 2018; Kreuter et al., 2011; Tourangeau et al., 2015). In addition, more and more people and survey respondents are using smartphones instead of computers (European Society for Opinion and Marketing Research, 2018). However, previous research has not studied motivated misreporting in web surveys conducted via smartphone, which is more burdensome than responding on a computer or laptop (which we will group together and call PCs), as the font and selection boxes are often smaller and page loading times are usually longer (Couper & Peterson, 2018). Thus, we expect in this study that respondents on smartphones are more likely to engage in motivated misreporting to avoid follow-up questions and to reduce survey length and burden. The negative impact on data quality may also extend to the responses to the follow-up questions.

In this article, we examine motivated misreporting in filter and follow-up questions in an experimental web survey which randomized respondents to smartphone and PCs. We first investigate whether there is a format effect—Can we replicate previous results of motivated misreporting in filter and follow-up questions? Second, is there a device effect—Do respondents answer filter and follow-up questions differently on smartphones and PCs? Third, is there an interaction between the format and device effects—Is the format effect stronger on smartphones than PCs? We begin by reviewing the literatures on motivated misreporting and smartphone surveys to develop our hypotheses. We then describe our data, methods, and data quality indicators and perform the analyses. Finally, we provide field recommendations for the usage of filter questions in PC and smartphone surveys.

Review of Relevant Literature

This section summarizes previous findings from two relevant strands of research, motivated misreporting and data quality, in web surveys conducted via smartphones.

Response Behavior to Filter Questions

Filter questions are found in many surveys. For example, the U.S. Consumer Expenditure Survey uses filter questions to ask about household purchases: Respondents indicating purchases are asked follow-up questions about the price of the items and for whom they were bought (U.S. Bureau of Labor Statistics, 2016). While filters and other forms of eligibility questions arguably improve survey designs and reduce response burden, they can also increase measurement error. Several studies have demonstrated such motivated misreporting by comparing responses to filter questions asked in two formats (e.g., Bach & Eckman, 2018; Bach et al., 2019; Duan et al., 2007; Eckman et al., 2014; Kreuter et al., 2019; Kreuter et al., 2011). The interleafed format asks the follow-up questions (if applicable) immediately after the relevant filter question. The grouped format asks all filter questions first before asking the follow-up questions that apply (for an illustration, see Table 1). In the interleafed format, respondents can learn that triggering a filter results in additional questions, while it is not possible to foresee the follow-up questions in the grouped format. Comparing the two formats has shown that respondents on average trigger fewer filters in the interleafed format than in the grouped format (Bach et al., 2019). This format effect seems to be due to respondents underreporting in the interleafed format to reduce the burden of the survey (Eckman et al., 2014). Based on this research, we expect fewer triggered filter questions in interleafed question format.

Hypothesis 1: Respondents in the interleafed format trigger fewer filter questions.

However, researchers who rely on survey data are not only interested in responses to filter questions but also in responses to the follow-up questions. Two studies have examined data quality

Table 1. Filter Questions in Interleafed Versus Grouped Format.

Interleafed Version	Grouped Version
In the past 3 months, have you purchased coffee for consumption at home?	In the past 3 months, have you purchased coffee for consumption at home?
Please briefly describe the most recent coffee you purchased.	In the past 3 months, have you purchased beer or wine for consumption at home?
How satisfied are you with the quality of the coffee?	In the past 3 months, have you purchased tobacco?
For whom was it purchased?	In the past 3 months, have you purchased children's clothing or shoes?
How much did it cost?	In the past 3 months, have you purchased clothing or shoes for yourself?
In the past 3 months, have you purchased beer or wine for consumption at home?	FOR EACH YES
Please briefly describe the most recent shirt you purchased	Please briefly describe the most recent [item] you purchased.
[. . .]	How satisfied are you with the quality of the [item]?
In the past 3 months, have you purchased tobacco?	For whom was it purchased?
[. . .]	How much did it cost?

Table 2. Definition of Data Quality Indicators.

Indicator	Definition, reference and type of question affected	
Heaping	Definiton	Reported value is divisible by 10/binary
	Other papers using	Antoun et al. (2017)
	Follow-ups used	How much did it cost?
Categories not selected	Definiton	Number of categories (not) selected/metric
	Other papers using	Lugtig and Toepoel (2016)
	Follow-ups used	For whom was it purchased?
Middle category selected	Definiton	Middle category "neither nor" was selected/metric
	Other papers using	Krosnick (1991)
	Follow-ups used	How satisfied are you with the quality of the [product]?
Item-nonresponse	Definiton	Item-nonresponse or don't know
	Other papers using	Lugtig and Toepoel (2016) and Antoun et al. (2017)
	Follow-ups used	All

in follow-up questions: Kreuter et al. (2011) in filter question follow-ups and Eckman and Kreuter (2018) in related looping questions. In a telephone survey, Kreuter et al. (2011) found more item-nonresponse to the follow-up questions in the grouped format. That is, respondents in the grouped format trigger more filters questions but then respond to fewer follow-up questions. We expect to replicate this effect in our web survey and extend it to additional data quality indicators common in literature (see Table 2 for an overview).

Hypothesis 2: Respondents in the interleafed format provide better data quality in the follow-up questions than respondents in the grouped format.

Response Behavior in PC and Smartphone Surveys

Just as question format can affect data quality, so can the device used to respond to the survey. Respondents use a variety of device types to participate in web surveys (e.g., desktop PCs, laptops,

tablets, or smartphones). Response behavior is relatively similar when respondents complete the survey on their PCs, laptops, or tablets. Taking a survey on smartphones, however, can lead to differences in response behavior (e.g., Antoun et al., 2017; de Bruijne & Wijnant, 2013; Gummer & Rossmann, 2015; Schlosser & Mays, 2018; Tourangeau et al., 2018).

Some studies find no difference in response behavior between respondents using smartphones and those using other devices. Smartphone respondents are at least as likely to provide conscientious and thoughtful answers and to disclose sensitive information on smartphones as on PCs (Antoun et al., 2017). They provide no substantial data quality differences in terms of item-nonresponse, straightlining, scale reliability, and validity (Tourangeau et al., 2018). Yet, other studies do find evidence of differences in response behavior. Smartphone respondents perceive surveys as shorter (de Bruijne & Wijnant, 2013); however, it takes them longer to answer a questionnaire (Keusch & Yan, 2017). The risk of break off is 2.8 times higher in web surveys completed via smartphone than PC (Mavletova & Couper, 2015). Moreover, smartphone respondents have more trouble executing tasks such as using small sliders and date-picker wheels (Antoun et al., 2017) and tend to provide shorter answers to open-ended questions (Couper et al., 2017). The smaller display size on smartphones, which may prevent respondents from seeing the entire screen at once, may explain some of these findings (Couper & Peterson, 2018; Mavletova, 2013). Loading times may also be longer on smartphones than on other devices (Couper & Peterson, 2018), and respondents may be more distracted (de Bruijne & Wijnant, 2013; Pinter, 2015; Poynter, 2015).

For these reasons, the response burden may be higher for respondents who use a smartphone rather than a PC to complete a web survey. Thus, we expect fewer triggered filters and lower data quality in the follow-up questions for smartphone respondents relative to PC respondents.

Hypothesis 3: Smartphone respondents trigger fewer filter questions than PC respondents.

Hypothesis 4: Smartphone respondents provide lower data quality in the follow-up questions than PC respondents.

No previous studies have investigated an interaction between motivated misreporting and device. Response burden is greater on smartphone devices, and the grouped format makes it easy for respondents to reduce burden by avoiding follow-up questions. Thus, we expect to find more motivated misreporting and lower data quality in the follow-ups among smartphone respondents than PC respondents.

Hypothesis 5: Smartphone respondents in the interleafed filter question format trigger fewer filter questions than smartphone respondents in the grouped format and PC respondents in the interleafed format.

Hypothesis 6: Smartphone respondents in the grouped question format provide lower data quality compared to smartphone respondents in the interleafed format and PC respondents in the grouped format.

Data and Methods

To test our six hypotheses, we conducted a web survey where we experimentally varied both filter question format and device. Below, we describe our data and data quality indicators.

Data Collection

We conducted a web survey in July and August 2018. Respondents were recruited from a German nonprobability online access panel. Quotas were given for gender, education, age, and federal state.

Table 3. Sample by Question Format and Device.

Description	Measure	PC	Smartphone	Total
Invitations	Count (%)	17,486 (35.42)	31,885 (65.58)	49,371 (100)
Screen-outs	Count (%)	275 (1.57)	2,563 (8.03)	2,838 (5.14)
Completes	Count (%)	1,902 (54.12)	1,612 (45.88)	3,514 (100)
Break-off rate	in %	10.31	17.51	13.81
Response rate	in %	10.89	5.06	7.11
Respondents only				
Interleafed	Count (%)	929 (26.43)	845 (24.04)	1,774 (50.47)
Grouped	Count (%)	1,051 (29.92)	795 (22.61)	1,740 (49.53)

One of the requirements for participation was that respondents must own and use both a PC and a smartphone. Before receiving the invitation, eligible cases were randomly assigned to use either a desktop computer/laptop (PC) or a smartphone (mobile) to complete the survey (see Online Appendix section 1 for the text of the invitation). Respondents who did not comply with the device assignment were not allowed to complete the survey.

From the initial 49,371 cases, 6,750 opened the invitation link, 195 broke off, and 2,838 were screened out (see Table 3). The most common reason for screening out was noncompliance with the assignment in the smartphone group (2,563). Noncompliance was much more common among those assigned to use a smartphone to complete the survey. We will return to this point later in this section.

The final sample consisted of 3,517 cases: 54% answered with a PC and 46% via a smartphone (Table 3). All cases that completed the filter question section of the survey, which was in the middle of the questionnaire, were counted as completes.

The questionnaire contained one section of 11 filter questions, which asked about purchases of common goods in the past month: coffee, chocolate, beer or wine, tobacco, children's clothing or shoes, clothing or shoes for yourself, medication, flowers, movies, pet supplies, and music. Each filter triggered three follow-up items (price, recipient, and satisfaction with the product). Each respondent could receive up to 33 follow-up questions. These questions were previously used in the Longitudinal Internet Studies for the Social Sciences (LISS) panel (Bach et al., 2019; Kreuter et al., 2019; Table 1 provides examples; see also Online Appendix section 2 for the question wording).

We randomly assigned respondents to either the interleafed (50.5%) or the grouped (49.5%) format (see Table 3). The questionnaire was optimized for smartphones: We used the suggestions (e.g., resolution and text size) of the survey programming tool "Unipark" for smartphones. We then tested these specifications on several devices and optimized them further. Figure 1 shows how the questionnaire displayed on smartphone and PC devices. The median response time to complete the questionnaire was 29 min and 40 s. Respondents in the interleafed format were on average 30 s faster than those in the grouped format (29:55 vs. 29:40 min, $t = 0.18$). PC respondents were on average 7 min faster to complete the questionnaire than smartphone respondents: 26 vs. 33 min ($t = -7.23$), Q50 PC = 26:43 min, Q50 smartphone = 33:24).

The questionnaire contained several additional experiments on consent, attentiveness, data linkage, and survey enjoyment, which were fully crossed with filters and device. Additionally, questions regarding trust, attitudinal questions, sociodemographics, and other variables were included. Respondents had the opportunity to skip questions but not to return to a previous page. The questionnaire had no auto-forwarding.

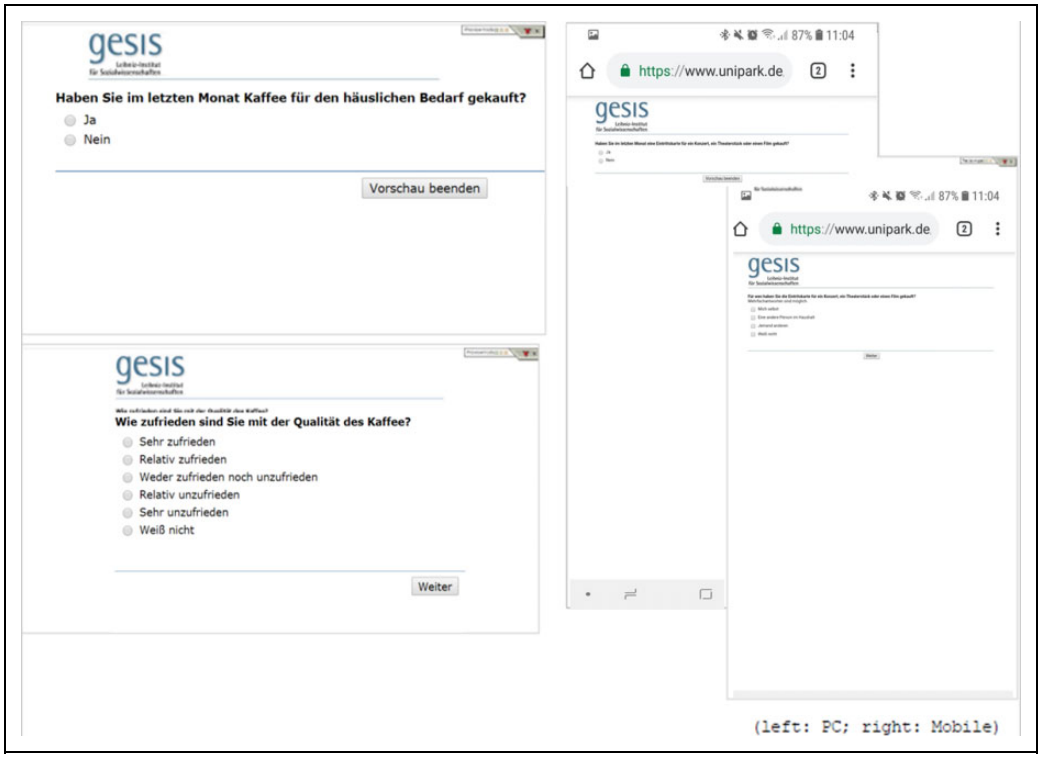

Figure 1. Display of filter and follow-up questions on PC and smartphones.

Check of Randomizations

Random allocation of respondents to device and format was intended to remove all differences between the groups so that any resulting response differences would be due to the experimental manipulation. However, as discussed above and shown in Table 3, panel members assigned to respond via a PC were more likely to respond. This differential nonresponse raises concerns about systematic differences between the PC and smartphone respondents. To check that the randomization worked as intended, we fit two logistic regression models. In the first, the dependent variable was the format (interleafed vs. grouped). In the second, it was the device used (PC vs. smartphone). The independent variables in each model were the sociodemographic information available for all panel members as well as two paradata measures: survey duration and invitation date. We could not use other variables such as attitudes as they were influenced by the various survey methodological experiments. We selected duration to exclude the risk of slower respondents self-selecting into a particular device and thus differing from faster respondents; the same applies to the participation date for late versus early respondents.

The results of these models are shown in Figure 2. In the first two charts, we see that the randomization of the question format worked well: Across all respondent characteristics, we see no significant differences between respondents completing the survey in the two formats. These results reassure us that there were no substantial differences in drop-out between the two formats. As shown in the second chart, however, there are systematic differences in the types of respondents who completed the survey on the two devices: Low income and rural respondents were harder to recruit for the smartphone group. In Germany, as in other countries, these groups have less experience with smartphones (Kongaut & Bohlin, 2016; Puspitasari & Ishii, 2016).

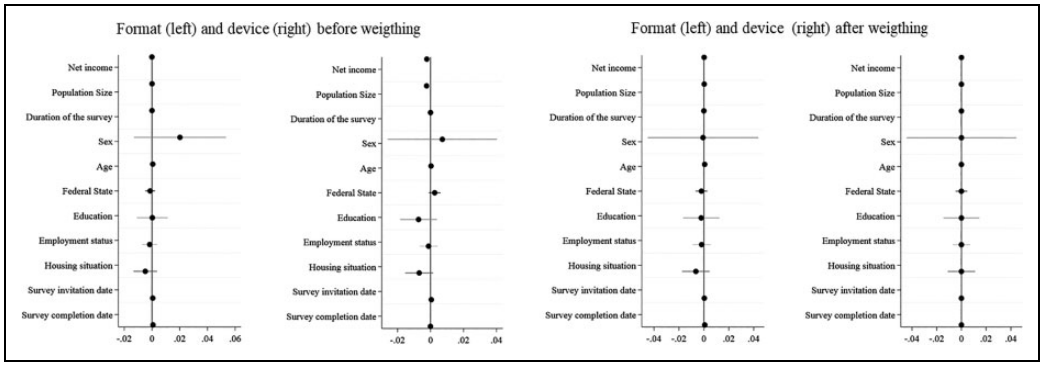

Figure 2. Test of the randomization to format and device.

To address the imbalance between the respondents completing the smartphone and PC versions of the survey, we applied entropy balance weighting. This approach derives weights to balance the observable characteristics of the PC and smartphone respondents (Hainmueller, 2012) and has been used for similar purposes before (Eckman & Haas, 2017). For example, in our sample, more highly educated people have participated via smartphones (see Figure 2); the entropy balancing method creates case-level weights that adjust the mean of the education variable of the PC respondents (control group) to match the smartphone respondents (treatment group). The method solves for the weights that make the means of all the variables shown in Figure 2 match.

After weighting with the entropy balance weights, no significant, observable differences remained between the smartphone and PC survey respondents (see Figure 2, fourth chart). The weights also did not introduce any differences between the interleafed and grouped respondents (see Figure 2, 3rd chart). However, we can only weight for observed characteristics and not for other unobserved characteristics that may differ between the two groups, such as respondents' motivation to participate in the survey. We used these weights in all analyses to remove the small imbalances between the two device conditions and make the two groups of respondents comparable.

Data Quality Indicators

To test for motivated misreporting in the filter questions (Hypotheses 1, 3, and 5), we compare the number of triggered filter questions in the interleafed and grouped formats.

Testing Hypotheses 2, 4, and 6 requires creating indicators of data quality in the follow-ups. The four indicators are summarized in Table 2. The first is heaping and is built from the follow-up question about the product price. When the reported price was divisible by 10, the indicator is 1 ("heaping") and 0 ("no heaping") otherwise. Heaping is an indicator of poor data quality because it takes less cognitive effort to give an approximate price than to remember the exact one, and furthermore, it is easier to enter rounded values without decimals on the keyboard. For the question format effect, we expect respondents in the grouped format to tend more to heaping because they are surprised and might even be annoyed that each affirmative answer to the filter questions has triggered the follow-ups. For smartphone respondents, we expect more heaping as smartphones have a smaller keyboard and entering numbers even on the number keypad is more difficult than on a PC. In 36% (*weighted*) of the triggered price questions, respondents provided heaped responses (see Figure A5 and Table A7 in the Online Appendix).

The second indicator follows studies such as Krosnick (1991) and refers to whether a respondent selected the middle category in a response scale. For the same reasons explained

above, we expect more middle category responses for respondents in the grouped format. Smartphone respondents might be exposed to more distractions and multitasking, which could reduce concentration and increase satisficing. Bypassing the response decision process by selecting the middle category can reduce the burden for respondents. The middle category was selected in only 7.2% (*weighted*) of the items across both devices (see Figure A5 and Table A7 in the Online Appendix).

Item-nonresponse, our third indicator, is a common indicator of poor data quality, used in previous studies such as Lugtig and Toepoel (2016) and Antoun et al. (2017). We again expect more item-nonresponse in the grouped than the interleafed format and for those using smartphones. On average, respondents had 8.0% (*weighted*) missing items in the follow-ups.

The last indicator is the number of categories not selected in multiple-choice items. Lugtig and Toepoel (2016) used the number of categories selected; however, we use the number of items *not* selected so that all our data quality indicators have the same direction: Higher values indicate lower data quality. For this indicator, we used the follow-up question for whom the product was purchased (self, another household member, someone else) which was a check-all-that-apply question. The number of unselected categories can give an indication of data quality for two reasons. On the one hand, selecting the small boxes involves motor and cognitive effort: The respondent should ask herself for each box whether she has purchased the product for this group of people. On the other hand, a single selected category can be an indication of satisficing since the questionnaire accepts an answer as soon as any category has been selected, and thus the processing of the question can be completed quickly. Furthermore, it might be more difficult or burdensome for a smartphone respondent to select more than one category on the small display. Respondents in the grouped format, as explained in the last section, might be more annoyed by the follow-up questions, and thus we expect fewer selected categories for smartphone respondents. Across devices for the triggered filters, 68.5% (*weighted*) of the categories were not selected (see Figure A5 and Table A7 in the Online Appendix).

Analysis Plan

To test our six hypotheses, we use a series of regression models run at the item level. For Hypotheses 1, 3, and 5 concerning the filter questions, we use logistic models where the dependent variable is whether a given filter question was triggered (1) or not (0). In the first model, which test Hypothesis 1, the sole independent variable is the format (interleafed vs. grouped). In the second model, the sole independent variable is the device (smartphone vs. PC). The third model contains the two main effects (format and device) and their interaction.

For Hypotheses 2, 4, and 6 about data quality in the follow-up questions, we use 12 models. The dependent variables are the four indicators of data quality, and the models are logistic or Poisson as necessary. The independent variables are, as above, the two main effects separately and then the model with the interaction. We perform all analyses at the item level and adjust the standard errors for the clustering of the filter questions in respondents. All models are weighted by the entropy balance weights described above. We performed weighted logistic and Poisson regression models in Stata Version 15.1.

Results

Table 4 gives the results from all 15 regression models. The upper part of the table reports results from two separate regression models: one to test the format effect and the other to test the device effect. The lower part of the table reports results from the models that include the main effect and the interaction. We discuss the results in the order of the six hypotheses.

Table 4. Results From All Regression Models.

	Triggered Filter Questions	Heaping	Categories Not Selected	Use of Middle Category	Item-Nonresponse
	Logistic *b* (*SE*)	Logistic *b* (*SE*)	Logistic *b* (*SE*)	Logistic *b* (*SE*)	Logistic *b* (*SE*)
Separate models					
Interleafed	−.080*** (.008)	.025 (.674)	−.260* (.346)	.028 (.110)	−.290*** (.083)
N	38,854	19,121	19,116	19,105	19,131
Smartphone	.015 (.008)	.131* (.060)	−.312 (.349)	.057 (.109)	.236** (.084)
N	38,854	19,121	19,116	19,105	19,131
Interaction models					
Interleafed	−.087*** (.011)	.052 (.086)	−.419* (.483)	.014 (.161)	−.269** (.123)
Smartphone	.017 (.012)	.139* (.079)	.084 (.488)	.118 (.145)	.276** (.115)
Interleafed × Smartphone	.003 (.0149)	.116 (.116)	.354 (.687)	.046 (.215)	.216 (.163)
(Pseudo) R^2	.005	.001	.004	.001	.0003
N	38,854	19,121	19,116	19,105	19,131

Note. Sample sizes differ according to weights used.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

Hypothesis 1: Respondents in the interleafed format trigger fewer filter questions.

Our results replicate the format effect reported in the literature (e.g., Bach et al., 2019; Kreuter et al., 2011): On average, respondents in the grouped format give about one more affirmative answer to the 11 filter questions. The difference between the two formats is statistically significant (see the first model in column 1 of Table 4).

Hypothesis 2: Respondents in the interleafed format provide better data quality in the follow-up questions than respondents in the grouped format.

The results relevant to this hypothesis are in the top row of Table 4 and the second–fifth columns. Each of these columns corresponds to one of the four indicators of data quality in the follow-up questions developed above (see Data and Methods section). The results indicate better data quality in the interleafed format for two of the four data quality indicators. Better data quality, in our case, means that the indicators are significantly lower in the interleafed format (recall that our four data quality variables, defined in Table 2, are each indicators of poor data quality). Respondents in the interleafed format select more items and provide less item-nonresponse (row 1, columns 3 and 5). In the other two models, there is no evidence of an association between format and the follow-up data quality.

Hypothesis 3: Smartphone respondents trigger fewer filter questions than PC respondents.

Contrary to our expectations, we do not find evidence that smartphone respondents trigger fewer filters and so engage in more motivated underreporting. Row 2, column 1 of Table 4 shows that there is no difference in the likelihood of triggering a filter between smartphone and PC respondents (5.1 vs. 5.3, $p = .19$).

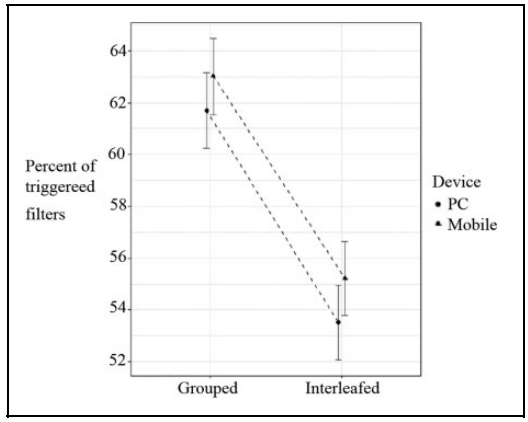

Figure 3. Triggered filter questions by format and device (in %).

Hypothesis 4: Smartphone respondents provide lower data quality in the follow-up questions than PC respondents.

We find evidence of lower data quality among smartphone respondents in two of our four data quality measures: heaping and item-nonresponse (see row 2, columns 2–5 of Table 4). The other two data quality indicators show no evidence for a device effect.

Hypothesis 5: Smartphone respondents in the interleafed filter question format trigger fewer filter questions than smartphone respondents in the grouped format and PC respondents in the interleafed format.

Figure 3 illustrates the interaction between format and device effects on filter questions responses. The figure shows the percent of filter questions triggered (y-axis) for the two formats (x-axis) and the two devices. The dashed lines between the two sets of point estimates represent the format effect for each device type: the difference between the grouped and interleafed formats. The two dashed lines have the same slope, indicating that the format effect does not differ across devices, and Hypothesis 5 is not supported by our data. The same result is shown in the coefficient on the interaction term in the third row of the first column of Table 4.

Hypothesis 6: Smartphone respondents in the grouped question format provide lower data quality compared to smartphone respondents in the interleafed format and PC respondents in the grouped format.

The insignificant coefficients on the interaction terms in each of the last four columns in Table 4 lead us to reject this hypothesis.

Discussion

This study randomly assigned web survey respondents to two experimental conditions: filter format (interleafed and grouped) and device (PC and smartphone). With these data, we tested six hypotheses about the performance of filter questions by format and device. We replicated the format effect that is by now well known: Respondents in the grouped format trigger more follow-up questions than those in the interleafed format (Hypothesis 1). However, we did not find a stronger format

effect among smartphone respondents (Hypothesis 3). Nor was there an interaction effect between the format and the device (Hypothesis 5).

Hypotheses 2, 4, and 6 are related to data quality in the follow-up questions rather than responses to the filter questions themselves. Hypothesis 2 was somewhat supported: In two of our four measures of data quality, the grouped format produced lower data quality in the follow-ups than the interleafed format. This result suggests that the grouped format has two somewhat contradictory effects on data quality: It collects more positive responses to the filter questions but lower data quality to the follow-ups. Thus, the net effect of the filter question format on data quality may be more complex than that suggested by previous studies, which focused on the number of triggered filters. For smartphone device respondents, we found lower data quality for two of the four indicators (Hypothesis 4) but no indication of an interaction between format and device (Hypothesis 6).

The study encountered some difficulties in compliance with the device assignment, which we addressed using entropy balance weighting. This approach uses weights that balance the treatment and control groups (here assigned to smartphone and assigned to PC). However, it is possible that there are other (unobservable) differences between the groups that we cannot control for, which could bias our results. Explicitly, there might be differences in the motivation of the respondents, which influence the self-selection effect into the two devices. If the smartphone respondents were more motivated to participate, this could explain the lack of support for some of our hypotheses, all of which relate to respondent motivation. Furthermore, this self-selection effect could be underestimated by the use of an online access panel compared to a probability-based panel because online access panel members may have less survey experience. More evidence is needed on the issue of device effects when answering filter questions and follow-up questions to filter questions. Unfortunately, true random assignment to device is difficult because respondents always have the option not to participate if they do not like the mode and device to which they are assigned.

Despite this shortcoming, the results presented above should concern all researchers using filter questions, especially in web surveys. There is mounting evidence that the format in which filters and follow-ups are asked effects responses in various question types. Researchers should think carefully about whether the responses to the filters or the follow-ups are most important in their research. The grouped format collects higher quality data with respect to the filters themselves (Eckman et al., 2014), but the interleafed format collects higher quality data in the follow-ups. Eckman and Kreuter (2018) argue that the grouped format may be preferable because the missing data in the follow-ups are more visible to analysts and imputation can be used to fill in missing values. However, this study shows that the harm to data quality in the grouped format does not always take the form of missing data. When respondents give a response to a follow-up item, and that response is not correct, analysts are not aware of the error, and it cannot easily be fixed through imputation. Furthermore, this study shows for mixed device studies that smartphone respondents do not provide lower data quality in the filter questions but in the follow-up questions. This effect applies to both question formats. Since this effect occurs particularly with heaping and item-nonresponse, we recommend optimizing the survey design of the follow-up questions for smartphone surveys (e.g., by automatically adjusting the font size or the usage of voice recordings) as well as to implement prompts when it comes to entering heaped numbers. To draw conclusions for the general population, we recommend replicating this study with a representative sample. Furthermore, we recommend controlling the motivation of the respondents on both devices, for example, by asking them directly. This could provide insights into whether more motivated respondents are more likely to participate with a smartphone. Another approach in order to understand how selection bias may have affected the results of this study is to use a laboratory experiment. This would control for the device effect since the threat of self-selection into a specific device is reduced compared to a field experiment.

Data Availability

The code to replicate the results is provided in https://github.com/1234Mannheim/MobileMotivatedUnderreporting. The data used in our study are available on request.

Declaration of Conflicting Interests

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

Funding

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

Software Information

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

Supplemental Material

Supplemental material for this article is available online.

References

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

Bach, R. L., & Eckman, S. (2018). Motivated misreporting in web panels. *Journal of Survey Statistics and Methodology*, *6*, 418–430.

Bach, R. L., Eckman, S., & Daikeler, J. (2019). Misreporting among reluctant respondents. *Journal of Survey Statistics and Methodology*, 1–23. http://doi.org/10.1093/jssam/smz013

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys: A total survey error perspective. In P. Biemer, S. Eckman, B. Edwards, E. de Leeuw, F. Kreuter, L. Lyberg, C. Tucker, & B. West (Eds.), *Total survey error in practice* (pp. 133–154). Wiley.

Couper, M. P., & Peterson, G. J. (2018). Why do web surveys take longer on smartphones? *Social Science Computer Review*, *35*, 357–377.

de Bruijne, M., & Wijnant, A. (2013). Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review*, *31*, 482–504.

Duan, N., Alegria, M., Canino, G., McGuire, T., & Takeuchi, D. (2007). Survey conditioning in self-reported mental health service use: Randomized comparison of alternative instrument formats. *Health Research and Educational Trust*, *42*, 890–907.

Eckman, S., & Haas, G. C. (2017). Does granting linkage consent in the beginning of the questionnaire affect data quality? *Journal of Survey Statistics and Methodology*, *5*, 535–551.

Eckman, S., & Kreuter, F. (2018). Misreporting to looping questions in surveys: Recall, motivation and burden. *Survey Research Methods*, *12*, 59–74.

Eckman, S., Kreuter, F., Kirchner, A., Jackle, A., Tourangeau, R., & Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, *78*, 721–733.

European Society for Opinion and Marketing Research. (2018). Global market research report 2018: An ESOMAR industry report. https://www.esomar.org/knowledge-center/library?publication=2898

Gummer, T., & Rossmann, J. (2015). Explaining interview duration in web surveys: A multilevel approach. *Social Science Computer Review*, *33*, 217–234.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, *20*, 25–46.

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.

Kongaut, C., & Bohlin, E. (2016). Investigating mobile broadband adoption and usage: A case of smartphones in Sweden. *Telematics and Informatics*, *33*, 742–752.

Kreuter, F., Eckman, S., & Tourangeau, R. (2019). Salience of survey burden and its effects on response behavior to skip questions. Experimental results from telephone and web surveys. In P. Beatty, D. Collins, L. Kaye, J. Padilla, G. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 213–228). Wiley.

Kreuter, F., McCulloch, S., Presser, S., & Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research*, *40*, 88–104.

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

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.

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*, 725–743.

Mavletova, A., & Couper, M. P. (2015). A meta-analysis of breakoff rates in mobile web surveys. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile research methods: Opportunities and challenges of mobile research methodologies* (pp. 81–98). Ubiquity Press. http://doi.org/10.5334/bar.f

Pinter, R. (2015). Willingness of online access panel members to participate in smartphone application-based research. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile research methods: Opportunities and challenges of mobile research methodologies* (pp. 141–156). Ubiquity Press.

Poynter, R. (2015). The utilization of mobile technology and approaches in commercial market research. In D. Toninelli, R. Pinter, & P. de Pedraza (Eds.), *Mobile research methods: Opportunities and challenges of mobile research methodologies* (pp. 11–20). Ubiquity Press.

Puspitasari, L., & Ishii, K. (2016). Digital divides and mobile Internet in Indonesia: Impact of smartphones. *Telematics and Informatics*, *33*, 472–483.

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.

Tourangeau, R., Kreuter, F., & Eckman, S. (2015). Motivated misreporting: Shaping answers to reduce survey burden. In U. Engel (Ed.), *Survey measurements techniques, data quality and sources of error* (pp. 24–41). Campus.

Tourangeau, R., Sun, H., Yan, T., Maitland, A., Rivero, G., & Williams, D. (2018). Web surveys by smartphones and tablets: Effects on data quality. *Social Science Computer Review*, *36*, 542–556.

U.S. Bureau of Labor Statistics. (2016). *Consumer expenditures and income: Handbook of methods*. https://www.bls.gov/opub/hom/cex/pdf/cex.pdf

Author Biographies

Jessica Daikeler is a postdoctoral researcher in statistics and methodology at GESIS-Leibniz Institute for the Social Sciences. Her research interests are the application of meta-analyses in survey research, the linking of nonresponse and measurement error, and the design of surveys (E-mail: jessica.daikeler@gesis.org).

Ruben L. Bach is a postdoctoral researcher in statistics and methodology at the University of Mannheim, Germany. His current research focuses on the use of new data sources for social research and the methods and tools necessary to analyze them (E-mail: r.bach@uni-mannheim.de).

Henning Silber is a senior researcher and head of the survey operations team at GESIS-Leibniz Institute for the Social Sciences. His research interests include survey methodology, political sociology, and the experimental social sciences (E-mail: henning.silber@gesis.org).

Stephanie Eckman is a fellow with RTI International, Washington, DC. She specializes in understanding data quality and the social construction of data (E-mail: seckman@rti.org).