Public Opinion Quarterly, Vol. 81, No. 4, Winter 2017, pp. 896–929

WEB SURVEYS BY SMARTPHONE AND TABLETS
EFFECTS ON SURVEY RESPONSES

ROGER TOURANGEAU*
AARON MAITLAND
GONZALO RIVERO
HANYU SUN
DOUGLAS WILLIAMS
TING YAN

Abstract With respondents increasingly completing web surveys on tablet computers and smartphones, several studies have examined the potential effects of the switch from PCs to mobile devices. The studies have looked at a range of outcomes, including completion rates, breakoffs, and item nonresponse. We carried out a field experiment that compared responses obtained by smartphones, tablets, and laptop computers, focusing on the potential effects of the different devices on measurement errors. We examined whether the differences across devices in screen size (and the related need to scroll to see the entire question or the full set of response options) might moderate the effects of response order, affect the strategy respondents used to decide which of two options was preferable, change the effect of question context,

Roger Tourangeau is a vice president of the Statistics Group and codirector of the Survey Methods Unit at Westat, Rockville, MD, USA. Aaron Maitland was a senior survey methodologist at Westat, Rockville, MD, USA. Gonzalo Rivero is a senior data scientist at Westat, Rockville, MD, USA at the time this study was done. Hanyu Sun, Douglas Williams, and Ting Yan are senior survey methodologists at Westat, Rockville, MD, USA. The work reported here was supported by a grant from the National Science Foundation [SES-1261340 to R.T.], with additional support from the Bureau of Justice Statistics. The authors thank the Methodology, Measurement, and Statistics Program at the National Science Foundation, the Bureau of Justice Statistics, and Dr. Cheryl Eavey for their support. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Bureau of Justice Statistics. The authors are grateful to Rose McAloon for her research assistance, to Tamara Bruce and Naoko Okuma for their oversight of the fieldwork, and to Abie Reifer, Bryan Davis, and Andrew Heller for their management of the technical and programming aspects of the study. *Address correspondence to Roger Tourangeau, Westat, 1650 Research Blvd., Rockville, MD 20850, USA; e-mail: RogerTourangeau@westat.com.

doi:10.1093/poq/nfx035 Advance Access publication December 10, 2017

or influence the use of definitions. Our experiments were based on the principle of visual prominence—the idea that respondents are more likely to notice and consider information that is easy to see. The experiments were deliberately designed to maximize the impact of screen size on the results, since the screen size would affect the visual prominence of key information. However, like many of the prior studies examining mobile devices, although response order, context, and evaluation strategy affected the answers respondents gave, few device effects emerged.

Introduction

According to Tourangeau, Conrad, and Couper (2013), web surveys have good measurement properties. For example, web surveys reduce socially desirable responding (because a computer rather than an interviewer administers the questions) and impose less cognitive burden than, say, telephone surveys (because the respondent can easily reread difficult questions). In addition, web surveys can be interactive, offering respondents help when they seem to be struggling with a question or intervening when they seem to be rushing through the survey (Conrad et al. 2009). However, virtually all of the work by Tourangeau, Conrad, and Couper (2013) focuses on web surveys done on laptop or desktop computers, and it is clear that many web surveys are now being completed on smartphones or tablets (Mavletova 2013; Lugtig and Toepoel 2015; Struminskaya, Weyandt, and Bosnjak 2015). Although much of the work on smartphones for web surveys has been done in Europe, where smartphone use varies substantially from to country, many web surveys are done on tablets or smartphones in the United States as well.

The literature to date suggests that smartphones may increase various forms of nonresponse to web surveys as compared to PCs (laptops or desktops). Sample members asked to complete the survey by smartphone are less likely to start the survey, more likely to break off before completing the questionnaire if they do start it, and more likely to skip individual items without answering them than sample members who access the survey on a PC (Couper, Antoun, and Mavletova 2017). In terms of nonresponse, tablets tend to be closer to PCs than to smartphones. Couper, Antoun, and Mavletova (2017) provide a good overview of the issues raised by web surveys done on smartphones and other mobile devices from the total survey error perspective.

In addition to the impact of smartphones on the various forms of nonresponse to web surveys, researchers have raised concerns about the quality of the data collected on smartphones. At least two features of smartphones are the basis for these concerns. First, smartphones have much smaller screens than tablets or PCs and the limited screen "real estate" may affect the answers. More specifically, the need to scroll in order to see all the response options or to see the entire question could affect the processing of the answer categories or the question, leading to heightened response-order effects or more superficial processing of the questions. Couper et al. (2004) showed that web respondents were more likely to select response options that were immediately visible than options that required an action (such as scrolling) to render them visible. Similarly, Conrad et al. (2006) showed that when respondents had to do something to get a definition for a term in a web questionnaire (for example, clicking on the term), they were less likely to access it. The need to scroll on a smartphone may make it less likely that respondents will see and consider all the options or other information needed to give an adequate answer to the question. The small screens on smartphones may also make it more difficult to see and compare answers to related questions on a single screen. Tourangeau, Conrad, and Couper (2013) argue for importance of the *visibility*, or visual prominence, of information. Respondents are more likely to notice and use information (such as a definition) the less they have to do to make the information visible. The need to scroll with a smartphone may render information crucial to formulating an answer less visible to respondents. Still, it is possible the smartphone users expect to have to scroll based on their prior experience with their smartphones.

The other smartphone feature that has been a source of concern has been the touch screen interface. Although tablets use the same user interface, the small screen size of smartphones means that smartphone respondents may have a difficult time registering their answer, leading to item nonresponse or inadvertently choosing the wrong answer.

Several studies have examined how collecting data on smartphones affects the answers respondents give. These studies have generally taken two approaches. Some have collected data only on smartphones but have included experiments within that mode, by manipulating, for example, the number of items per screen or whether the survey followed a paging or scrolling design (Peytchev and Hill 2010; de Bruijne and Wijnant 2014; Mavletova and Couper 2014, 2016). Others have compared smartphone responses with those obtained by PC or tablet, sometimes in randomized experiments (de Bruijne and Wijnant 2013; Mavletova 2013; Wells, Bailey, and Link 2014; Mavletova and Couper 2015; Keusch and Yan 2016) and sometimes in quasi-experimental designs (Lugtig and Toepoel 2015; Struminskaya, Weyandt, and Bosnjak 2015).

Table 1 summarizes the results of the studies that have produced peer-reviewed publications. These studies support several general conclusions. First, respondents on smartphones sometimes skip more items than those completing the same survey on tablets or PCs (Lugtig and Toepoel 2015; Struminskaya, Weyandt, and Bosnjak 2015; Keusch and Yan 2016; but see Buskirk and Andrus 2014; Andreadis 2015). Second, respondents take longer to complete surveys on smartphones than on tablets or PCs (de Bruijne and Wijnant 2013; Buskirk and Andrus 2014; Andreadis 2015; Struminskaya, Weyandt, and Bosnjak 2015; Keusch and Yan 2016). Third, respondents generally give shorter answers to narrative open-ended questions on smartphones

Table 1. Studies Focusing on Measurement Error in Surveys Done on Mobile Devices, by Study Characteristics

Study	Sample	Experiment	Devices	Indicators/Experiments and results
Andreadis (2015)	Visitors to Voting Advice Application *n* = ~80,000	No: Compare visitors to the application by the type of device used	PC Smartphone	• Response to questions: No differences • Item nonresponse: No differences • Mean duration: Smartphone slower
Buskirk and Andrus (2014)	SSI nonprobability sample *n* = 430	Yes: Random assignment to devices, within age, sex, and education subgroups	Computer iPhone	• Mean duration: Computer slower • Item nonresponse: No differences by device • Response-order effects: No differences by device • Slider-bar questions: No differences by device
de Bruijne and Wijnant (2013)	CentER probability panel *n* = 379	Yes: Random assignment to computer (2 versions) or mobile device	Computer Tablet/Smartphone	• Responses to questions: Differences on 4 of 26 items • Answers to debriefing questions: No differences • Mean duration: Mobile devices slower
de Bruijne and Wijnant (2014)	CentER probability panel *n* = 454	Yes: Random assignment to 4 experiments	Smartphone	• Paging versus scrolling: Mean difference on 1 of 6 items; paging slower; no differences in item nonresponse • Horizontal versus vertical response options: No differences on 5 items; more item nonresponse with horizontal scales • Long versus short scales: Longer is slower • Open (numerical) versus closed item: No difference in item nonresponse or completion time

Continued

Table 1. *Continued*

Study	Sample	Experiment	Devices	Indicators/Experiments and results
Keusch and Yan (2016)	Amazon MTurk nonprobability sample $n = 1{,}186$	Yes: Random assignment to computer or iPhone among those who started survey on computer	PC iPhone	• Breakoff: Highest among those who started survey on iPhone • Item nonresponse: Higher among those who completed survey on iPhone • Mean duration: Longer for those who completed survey on iPhone • Straightlining: More for those who completed survey on PC • Acquiescence: No differences • Mid-point responding: No differences • Substantive answers: No differences
Lugtig and Toepoel (2015)	6 waves of the LISS probability panel $n = 6{,}226$	No: Compare Rs who switch devices over time	PC Tablet Smartphone	• Item nonresponse: Smartphone highest, PC lowest • Percent answering open-ended item: PC highest, tablet lowest • Open-item mean length: PC longest, tablet shortest • Straightlining: PC highest, tablet lowest • Primacy: Smartphone highest, PC lowest • Mean number of answers on check-all item: No differences

Continued

Table 1. *Continued*

Study	Sample	Experiment	Devices	Indicators/Experiments and results
Mavletova (2013)	Russian online non-probability panel $n = 1{,}013$	Yes: Random assignment to device and to survey length	PC Mobile phone	• Response-order effects: Order effect on 1 out of 6 items, effect larger for PC • Responses to questions: Device differences on 2 of 6 items • Social desirability bias: No differences • Non-substantive answers: No differences • Income reporting: Mobile respondents more likely to skip question • Length of open-ended answer: PC longer
Mavletova and Couper (2013)	Russian online non-probability panel $n = 884$	Yes: 2-wave crossover design	PC Mobile phone	• Income reporting: Higher income reported on PC • Sensitive items: More alcohol consumption reported on PC; no differences on other items
Mavletova and Couper (2014)	Russian online non-probability panel $n = 2{,}110$	Yes: Paging versus scrolling design	PC Tablet Smartphone	• Scrolling versus paging: Mean duration longer with paging; no differences in item nonresponse
Mavletova and Couper (2016)	Russian online non-probability panel $n = 2{,}032$	Yes: Random assignment to 5 vs. 15 vs. 30 items per page	Smartphone	• Breakoff rate: No differences by version • Item missing data: Lowest with 5 items per page, highest with 30 items • Skip errors: Lowest with 5 items • Mean duration: Fastest with 30 items per page

Continued

Table 1. *Continued*

Study	Sample	Experiment	Devices	Indicators/Experiments and results
Peytchev and Hill (2010)	Probability sample of addresses in North Carolina $n = 52$ to 61 (sample sizes vary by experiment)	Yes: 10 experiments in four surveys	Smartphone	• High- versus low-frequency scales: Replicates usual effects of scale labels • Norm of evenhandedness: Replicates effect of question order • Images: No context effects in one experiment; context effect (assimilation) on one of 7 items • 3 versus 7 options visible: No differences • Horizontal versus vertical response options: No differences
Struminskaya, Weyandt, and Bosnjak (2015)	GESIS (probability) panel—6 waves $n = 2{,}913$	No: Compare respondents who switch modes across waves	PC Tablet Smartphone	• Item nonresponse: Smartphone higher than other 2 modes • Straightlining: Smartphone higher than other 2 modes • Percent answering open-ended item: PC higher than other 2 modes • Mean length of open-ended: Smartphone lower than other 2 modes • Mean answers in early part of scale: Smartphone lower than other 2 modes • Mean duration: PC fastest, smartphone slowest

Continued

Table 1. *Continued*

Study	Sample	Experiment	Devices	Indicators/Experiments and results
Toepoel and Lugtig (2014)	Market Response (probability panel) $n = 439$	No: Compare respondents who complete by smartphone or mobile phone	PC Smartphone (optimized)	• Breakoff rate: No differences by device • Mean duration: No differences by device • Item missing data: No differences by device • Mean length of open-ended: No differences by device • Primacy: Nonsignificantly more on smartphone • Mean answers to check-all item: No differences by device
Toninelli and Revilla (2016)	Netquest (non-probability panel) $n = 1,608$)	Yes: Random assignment in each of 2 waves to device type—PC versus 2 versions of smartphone (optimized versus not optimized)—in a crossover design	PC Smartphone (optimized) Smartphone (not optimized)	• Reporting of sensitive information: No overall differences across groups; less underreporting of income within optimized smartphone group • Place where survey completed (home versus elsewhere): No differences across groups • Presence of bystanders: More likely in smartphone groups; also, more smartphone surveys done in presence of strangers

Continued

Table 1. *Continued*

Study	Sample	Experiment	Devices	Indicators/Experiments and results
Wells, Bailey, and Link (2014)	Knowledge Panel (probability) $n = 1{,}981$	Yes: Random assignment to device, and to various experimental conditions	PC or laptop Smartphone	• High- vs. low-frequency scales: No differences by device • Choice of "other" in closed versus half-open (Other, specify) format: No differences by device • Small versus larger text box: Bigger difference on PC • Randomized versus alphabetized response options: No differences by device

NOTE.—Sample sizes given in terms of completed cases.

than on tablets or PCs (de Bruijne and Wijnant 2013; Mavletova 2013; Struminskaya, Weyandt, and Bosnjak 2015), though not always (Toepoel and Lugtig 2014; Lugtig and Toepoel 2015).

Despite their sheer number, these studies have their limitations. For example, most of them were done in Europe and, with the studies done in Russia, some of the respondents used relatively low-tech feature phones to complete the survey rather than smartphones. In the experimental studies, there is often substantial non-compliance, with respondents randomly assigned to one device instead completing the survey on some other device. On the whole, though, the results are reassuring: Responses by smartphones appear to be very similar to responses obtained by PC or tablet, and the differences that are found (such as differences in item nonresponse) are not very large.

Two main hypotheses have guided this past work. First, using smartphones or other mobile devices (such as tablets) to complete a survey may diminish respondents' sense of privacy. This in turn may reduce respondents' willingness to provide sensitive information (Mavletova and Couper 2013; Mavletova 2013). The diminished sense of privacy may in part reflect *where* respondents answer the survey questions. People carry their mobile devices with them and often complete a survey shortly after they receive the invitation; this means they may answer the questions in a less private setting with bystanders nearby (Mavletova and Couper 2013). The presence of bystanders is known to reduce candid reporting (Tourangeau and Yan 2007). Still, neither Mavletova and Couper (2013) nor a recent replication by Toninelli and Revilla (2016) found much support for these hypotheses. Second, as noted earlier, the limited screen size of smartphones and, to a lesser extent, tablet computers may affect how respondents process the questions, the response options, or both.

We carried out a randomized field experiment that compared web surveys done on smartphones, tablets, and laptop computers. To try to reduce any selection effects, we informed participants which device—smartphone, tablet, or laptop—they were assigned to only *after* they had agreed to complete the questionnaire. Field interviewers recruited the respondents, who were selected from an address-based sample in eight areas. The interviewers provided the device so that all the interviews were done on the same smartphone, tablet, or laptop.

Our study focused on measurement effects. We attempted to maximize the chances of observing device effects by experimentally varying the format of a number of questions. The experiments randomly varied a) response order and scale length, b) whether two related items were evaluated together on a single screen or separately on successive screens, c) whether items that might be compared were presented in a grid or on separate screens, and d) whether definitions that were likely to affect the answers were presented on-screen (either before or after the response options) or via a hyperlink. All of the experiments are based on well-established phenomena, mostly taken from the survey literature. We chose these five phenomena because we thought they would be particularly prone to the effects of the differences in screen size across the three types of devices and thus particularly likely to show differences by device type.

For example, we thought that data collected by smartphones would show larger response-order effects than those collected on tablets or laptops, especially with longer response scales that would force smartphone respondents to scroll to see all of the response categories. Earlier work has shown that when respondents have to do something to make all the response options visible, larger response-order effects result (Couper et al. 2004). One of the response-order and -length experiments also varied the verbal labeling of the scales (fully labeled versus endpoint-only) to assess whether the full labeling of the options might reduce the size of any response-order effects we found (see Yan and Keusch 2015). Even though the response options were all visible on the smartphone in this experiment, we thought the need to scroll throughout the questionnaire in this smartphone condition might heighten the effect of screen position, making this cue more salient to smartphone respondents.

In the experiment varying whether two items appeared on the same or successive screens, we used items that can be evaluated jointly or separately (Hsee et al. 1999). Hsee and his colleagues found that, when two options are evaluated together (*joint evaluation*), it sometimes produces different preferences from when the same two options are evaluated independently (*separate evaluation*). In joint evaluation, respondents can compare the two options, allowing them to appreciate the value of characteristics that are difficult to evaluate in isolation. For example, in assessing two surgeons, it may be hard to know what to make of the fact that one surgeon has carried out a procedure 20 times unless one also knows that the other surgeon has done the same procedure 200 times. We thought that smartphones would reduce the chance that respondents would evaluate two items jointly, since smartphone respondents would have to scroll to compare the two alternatives even when the items were on a single page. We used items taken from a study by Zikmund-Fisher, Fagerlin, and Ubel (2004), who also found differences between joint and separate evaluation.

Similarly, in the experiment comparing the presentation of items in a grid versus one at a time on separate screens, we thought we would observe a context effect in the grid condition (since respondents could see the items all at once), but that this effect would be diminished for the smartphone respondents (since they would have to scroll to see all the items even when they were in the same grid). Again, this is based on the principle that when respondents have to do something to make information visible, this reduces the likelihood that the information will be used by respondents (Tourangeau, Conrad, and Couper. 2013). Finally, we thought the smartphone might reduce impact of the definitions when they came after the response options, since they would be less visible in the smartphone condition than in the tablet or laptop conditions (Conrad et al. 2006; Peytchev et al. 2010). Thus, the studies share the premise that screen size would affect visibility and that the resulting differences in visibility would be apparent in the results of each experiment.

Method

SAMPLE

Our study used an address-based sample within eight counties in the United States. The eight counties were selected purposively as the primary sampling units (PSUs), based on the availability of experienced field interviewers there. Still, we tried to include a range of different types of areas. Seven of the eight PSUs were located in metropolitan statistical areas and one was rural; one of the metropolitan areas involved a small city. All four Census regions were represented in the set of eight PSUs. Within each PSU, a full probability sample of addresses was selected. Census block groups were used as secondary sampling units (SSUs); six SSUs were selected from each county, with probability proportional to size. After the initial selection, one selected SSU was found to be located entirely on a military base. This SSU was replaced with the next SSU within the county. Once the final 48 SSUs had been selected, we sampled 1,824 addresses, 228 within each sample county. One hundred forty-two of the addresses were vacant, could not be located, or were not dwellings, leaving a final sample of 1,682 occupied addresses.

Once a household was contacted, interviewers completed a short screener with an adult there. In households with a single adult, that adult was selected for the main interview; in households with two adults, one of the two was randomly selected; in households with three or more adults, the last birthday method was used to select one of them for the main interview. Once the sample person was determined, he or she was asked to complete the interview on a device that had been randomly determined ahead of time. Sample persons were offered a $20 incentive to complete the interview.

DATA COLLECTION

Interviews were completed in person by experienced field interviewers from January 21, 2015, through May 4, 2015. A total of 512 interviews were completed across the three device conditions, for an overall response rate (AAPOR RR1) of 30.4 percent. The response rates did not differ significantly across the three device conditions, with the response rate highest for the smartphone condition (smartphone 32.9 percent, tablet 29.6 percent, and laptop 28.8 percent). Despite concerns that cellular coverage might affect access to the web instrument, only two cases could not be completed for this reason (one each in the smartphone and laptop conditions).

Selected respondents were asked to complete the web questionnaire on a device that the interviewer provided while the interviewer waited.[1] During this time the interviewer completed other tasks. The interviewer could not change the assigned device type, and the web instrument was programmed to accept only the device that was assigned. This device type was also verified during data analysis by comparing the device information detected with the experimental assignment. All devices detected matched the one that had been assigned.

The overall topic of the questionnaire was health. Appendix 1 gives the wording of the items analyzed here.

EXPERIMENTAL DESIGN

Data collection device: The main experimental variable was the device type. This affected the amount of information visible to the respondent and the input method the respondent was able to use. The laptop condition featured a mouse and keyboard, whereas the smartphone and tablet used a touch screen interface.

Each interviewer was provided with three devices for accessing the web questionnaire. The devices were a smartphone (iPhone 5s), a tablet (iPad Air), and a laptop (a Fujitsu Lifebook E753). These devices were chosen after an evaluation of available smartphones and tablets that were determined to provide reasonable variation in screen sizes. Figure 1 shows the relative screen sizes of each device. (Paradata indicated that most tablet and smartphone respondents completed the survey in "portrait" orientation, and the screens for those conditions are shown in that orientation.) Because of a constraint built into the software, the browser window for the survey did not use the entire laptop screen (see the dashed area in figure 1). Still, it covered an area more than 50 percent larger than that of the tablet computer screen. The survey was *not* optimized for smartphones.

Each device accessed the web instrument over a cellular network. For the smartphone, this was the telephone's own network. For the tablet and laptop conditions, interviewers used a MiFi device (a small, wireless router that generates a WiFi hotspot) that created a wireless hotspot that the tablet or laptop accessed. Given this method of accessing the Internet, download times should have been similar for all three devices.

Embedded experiments: The web instrument included a number of experiments that we thought might be affected by the type of device on which the survey was

1. As part of another experiment, for half the respondents, the interviewer asked the questions and recorded the answers for the first section of the questionnaire. The experiments described here involved later sections of questionnaire, all of which were self-administered.

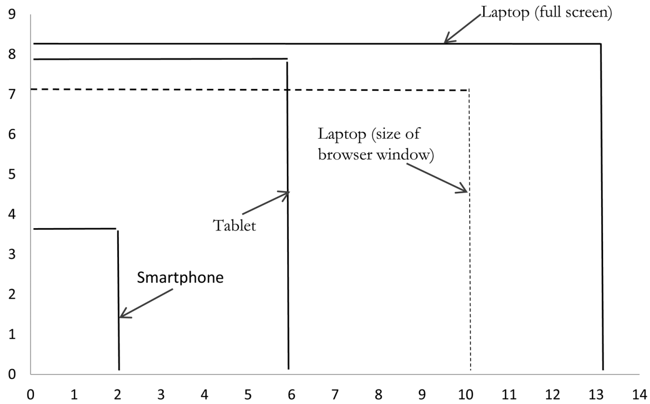

Figure 1. Relative Screen Size of the Three Devices. The dashed outline shows actual web browser size displayed on laptop screen. The labels on the axes show the dimensions in inches.

done. We discuss five of them in this paper. Experiment 1 examined how device type moderated the effects of response order. Respondents were randomly assigned to one of two response orders. This experiment encompassed four items, two items with 13 response options and two with eight. The response options were categorical, and respondents could select only one option from each list. One version of each item presented the options in one order, and the other version presented them in the opposite order. For example, one item asked: "Which of the following health conditions or diseases do you consider to be the biggest health problem in the U.S.?" There were 13 response options to this question. In one version of the item, the list of response options began with "high cholesterol" and ended with "seasonal flu"; in the other version, the list began with "seasonal flu" and ended with "high cholesterol." In the smartphone condition, only the first seven response options were visible. To see the remaining option(s), respondents had to scroll down the page. In the tablet and laptop conditions, all response options were visible. Appendix 1 gives the wording of all four items used in this experiment.

Experiment 2 also examined response order and scale length (five versus seven points); however, in contrast to experiment 1, the response options in this experiment were ordinal (that is, scales) rather than categorical. This experiment also compared fully labeled scales with endpoint-labeled scales, resulting in a 2 (scale direction) x 2 (scale length) x 2 (scale labeling) factorial design. The experiment included both behavioral frequency and attitudinal items; respondents either received the behavioral items first or attitudinal items first. The two sets of items came in different sections of the questionnaire. The attitude items in experiment 2 used agree-disagree scales. For the seven-point version, the options were *Strongly agree*, *Agree*,

Agree somewhat, *Neither agree nor disagree*, *Disagree somewhat*, *Disagree*, and *Disagree strongly*. The five-point version dropped *Agree somewhat* and *Disagree somewhat*. The seven-point version of the frequency scale included these response options: *Always*, *Often*, *Frequently*, *Sometimes*, *Occasionally*, *Rarely*, and *Never*. The five-point version omitted *Frequently* and *Rarely*. On all three devices, respondents were able to see all scale points without needing to scroll.

Experiment 3 examined how different screen sizes affected how respondents evaluated two alternatives. This experiment drew on the work of Hsee and his colleagues (1999) on joint and separate evaluation. In joint evaluation mode, the two alternatives are presented simultaneously and one is able to assess the value of attributes that may be hard to evaluate in isolation. For example, in the item below (taken from Zikmund-Fisher, Fagerlin, and Ubel 2004), it is hard to assess the satisfaction rating for the two physicians when they are judged separately, but easier to appreciate this information when the two physicians are evaluated together:

> Imagine you need to select a primary care physician.
> The first option is Dr. Smith, who has a satisfaction rating of 6.8 out of 10 and his office is 15 minutes away.
> The second option is Dr. Jones, who has a satisfaction rating of 8.0 out of 10 and his office is 45 minutes away.

Respondents were asked to rate each doctor on a five-point scale. Zikmund-Fisher et al. (2004) found that participants tended to give Dr. Smith the higher rating when they evaluated the two physicians separately, but to give Dr. Jones the higher rating when they evaluated the two physicians at the same time. Our experiment varied whether the doctors were presented and rated on separate screens (*separate* evaluation) or both doctors were presented on a single screen and both were rated together (*joint* evaluation). In addition, the experiment varied the order in which the two items were presented. Figure 2 displays screen captures (from the laptop and smartphone conditions) for this item.

In the tablet and laptop conditions, respondents were able to see the entire question and response options (as in figure 2). In the smartphone condition, the respondent needed to scroll to see the response options (as in the right half of figure 2b).

Experiment 4 compared the presentation of a battery of similar items one at a time on separate screens versus in a grid. The items asked respondents to rate the severity of 10 illnesses ("a common cold," "epilepsy") using a fully labeled five-point scale ranging from *Not serious at all* to *Extremely serious*. The eighth item on the list was "lung cancer." When the items were presented in a grid, respondents would, we thought, be likely to use "lung cancer" as an extreme

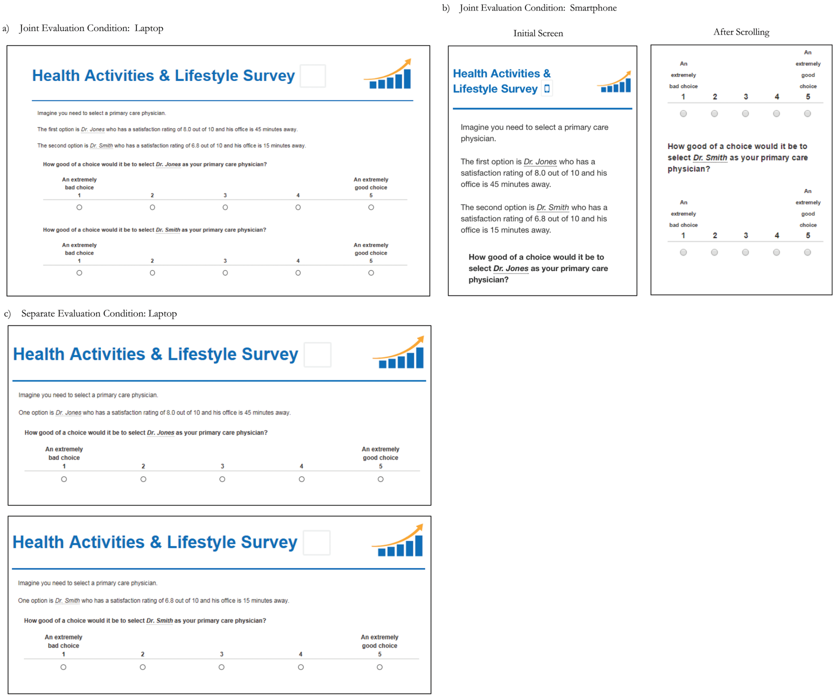

Figure 2. Screen Captures for Joint (panels a and b) and Separate (panel c) Evaluation Conditions. In the separate evaluation condition, the two procedures are presented on different screens. These screen captures are from the laptop (panels a and c) and smartphone (panel b) conditions. Panel b shows the initial screen respondents saw on the smartphone (on the left) and after scrolling to the response options (on the right).

point of comparison and, as a result, give lower severity ratings to the preceding illnesses. When the items were presented one at a time on separate screens, there would not be a similar contrast effect—respondents would already have rated "a common cold" and the other illnesses by the time they saw "lung cancer." We expected the difference between the grid and one-at-a-time presentation to be reduced or eliminated in the smartphone condition since on a smartphone, "lung cancer" was not visible in the grid condition until respondents scrolled down the list. The fourth panel in appendix 1 shows the full list of items in this experiment.

Our fifth experiment compared three methods of presenting definitional information in a series of six items asking about the respondent's diet. Respondents were asked how much they consumed of each of six items on a scale ranging from *Much less than I should* to *Much more than I should*. In one condition, the definitions were presented after the response scale; in a second condition, the respondent had to click on a hyperlink to see the definition; in the final condition, the definition was presented after the question text but before the response options. The definitions were drawn from a study by Conrad et al. (2006) and included relatively surprising information that was likely to affect respondents' answers. For example, the definition of cholesterol read:

> Good cholesterol (HDL) carries bad cholesterol (LDL) away from the arteries. One can increase levels of HDL by consuming vitamin C and niacin, exercising, and not smoking. One can lower LDL by consuming fewer foods such as beef and rich cheeses.

Conrad et al. (2006) found that respondents who read the definitions were more likely to report that they consumed less cholesterol than they should. The other five items in this experiment, also taken from the study by Conrad et al. (2006), followed a similar pattern—if respondents consulted the definitional information, they were likely to change their answers. Our hypothesis was that the definitions would have the greatest impact on the answers when they were displayed just before the response options. Again, we thought that the differences between the different methods of presenting the definitions would be largest in the smartphone condition; in the tablet and laptop conditions, the definitions were visible from the outset even when they were displayed below the response options.

Results

We first present the results of demographic comparisons across the three device groups and then present the results for each of the experiments in turn. The results are unweighted, but the analyses do take into account the clustering of the data by PSU, using the SAS SURVEY procedures and treating PSU as the clustering variable.

DEMOGRAPHIC DIFFERENCES BY DEVICE

Table 2 shows the composition of the three device groups, by age quartile, education, sex, Hispanic origin, and race (White versus non-White). The smartphone group included a significantly higher proportion of females than the other two groups (Rao-Scott $\chi^2 = 8.58$ with 2 df; $p < .05$). In addition, the three device groups differed significantly by race (Rao-Scott $\chi^2 = 17.4$ with 2 df; $p < .001$), with the laptop group having the highest proportion of Whites and the smartphone group the lowest. Subsequent analyses included race and sex as covariates (because they differed significantly across the device types), as well as age and education (because they were likely to be related to the size of any experimental effects).[2]

2. We also redid all the analyses with whether the respondent had a smartphone as an additional independent variable. We found no main effects for this variable and only scattered (and uninterpretable) interactions with the experimental variables.

Table 2. Demographic Composition of Responding Sample, by Device Condition

	Smartphone	Tablet	Laptop
Age quartile	(*n* = 186)	(*n* = 164)	(*n* = 160)
18–37	26.3	28.7	24.4
38–52	26.3	23.8	23.1
53–67	24.7	25.6	23.8
68 and older	22.6	22.0	28.8
Education	(*n* = 183)	(*n* = 164)	(*n* = 157)
High school graduate or less	31.2	24.4	29.3
Some college/Associate's degree	33.3	29.3	31.2
College graduate or above	35.5	46.3	39.5
Sex	(*n* = 187)	(*n* = 164)	(*n* = 160)
Male	35.8	50.0	50.0
Female	64.2	50.0	50.0
Hispanic	(*n* = 182)	(*n* = 163)	(*n* = 155)
Yes	9.9	12.3	5.8
No	90.1	87.7	94.2
Race	(*n* = 187)	(*n* = 164)	(*n* = 160)
White	78.1	80.1	85.8
All others	21.9	19.9	14.2

NOTE.—Sample sizes vary slightly from item to item due to item nonresponse.

EXPERIMENT 1: RESPONSE ORDER WITH CATEGORICAL RESPONSE OPTIONS

Our first experiment varied the order of the response categories for four items. Table 3 shows the basic results, displaying the percentage of respondents who selected an option that was presented in the first part of the list of response options (when the options were presented in the original order) and in the last portion of the list (when the options were presented in the reverse order). We grouped the first six options and the second seven in the two items with 13 options, and the first four and the last four in the two items with eight options. As expected, we generally found primacy effects. For example, with item B1, 90.6 percent of the smartphone respondents selected one of the six options when they were presented at the beginning of the list (that is, in the original order), but only 74.3 percent selected one of those same six options when they were presented at the end of the list (in the reverse order), a difference of 16.3 percent. Across the four items, the average difference between the options when they were presented at the beginning of the list versus at the end was 13.3 percent in the smartphone condition, 7.4 in the tablet condition, and 8.6 in the laptop condition.

We calculated the proportion of times respondents selected one of the options that appeared in the first portion of the list in the original order. We examined this variable via an analysis of covariance with device type and response

Table 3. Percent of Respondents Selecting Options Appearing First in Original Order, by Item, Response Order, and Device

Item	Smartphone (%)	Tablet (%)	Laptop (%)	Total
B1 (13 options)				
Original order	90.6	90.9	85.7	88.8
Reverse order	74.3	69.3	84.5	75.6
B2 (13 options)				
Original order	55.4	37.9	59.0	51.6
Reverse order	50.7	46.2	50.7	48.9
B3 (8 options)				
Original order	96.1	97.0	87.5	93.3
Reverse order	82.4	90.8	86.7	87.0
B4 (8 options)				
Original order	72.0	65.7	73.2	70.5
Reverse order	53.6	55.7	49.3	53.0

Note.—The percentages encompass the first six options (in the original order) for items B1 and B2 and the first four (in the original order) for items B3 and B4.

order as experimental factors and age, education, sex, and race as covariates. The overall effect of response order was highly significant (F (1,7) = 33.8, $p < .001$). Although the response-order effect observed was larger within the smartphone group as hypothesized, the response order by device type interaction was not significant (F (1,7) < 1). We also ran a similar analysis examining the proportion of respondents selecting one of the first seven options (in the original order) for all four items; these seven options were visible from the outset (in the original order) on all three devices. This analysis produced similar results to those already reported: a significant order effect—F (1,7) = 23.2. $p < .01$—but no interaction with device type. Appendix 2 presents descriptive results analogous to those in table 3.

EXPERIMENT 2: RESPONSE ORDER WITH SCALES

Experiment 2 varied the order of the response categories for 13 items (seven attitudinal and six behavioral). For the attitudinal items, the response options ranged from *Strongly agree* to *Strongly disagree* in one condition (the original order) and from *Strongly disagree* to *Strongly agree* in the other (the reverse order). For the behavioral items, the options went from *Always* to *Never* in the original-order condition and from *Never* to *Always* in the reverse-order condition. This experiment also varied the number (five versus seven) and labeling (fully labeled versus endpoint-only labeled) of the scale points. We found few effects for the scale labeling variable and ignore it in what follows.

The results of experiment 2 (see table 4) were similar to those of experiment 1. We analyzed the attitude items and behavioral items separately, using

Table 4. Mean Percent of Respondents Selecting Options Appearing First in the Original Order, by Item Type, Response Order, Scale Length, and Device

	Smartphone	Tablet	Laptop	Total
Attitude items: 5 options				
Original order	66.7	71.5	64.6	67.2
Reverse order	64.0	76.4	68.0	68.7
Attitude items: 7 options				
Original order	70.4	81.7	80.8	77.7
Reverse order	64.3	68.2	65.7	66.2
Behavior items: 5 options				
Original order	65.9	68.8	62.5	65.3
Reverse order	70.3	71.9	66.1	69.5
Behavior items: 7 options				
Original order	74.5	83.1	72.2	77.0
Reverse order	63.3	63.5	69.6	65.1

NOTE.—The percentages encompass the first two response categories (for the items with five scale points) or the first thee (for the items with seven options). These options make up the *Agree* end of the scale for the attitude items and the *Always* end of the scale for the behavior items. The table entries average across the items.

an analysis of covariance, with scale direction, length, labeling, and device type as experimental factors and age, education, sex, and race as covariates. The dependent variable was the proportion of times the respondent chose a response option in the first half of the scale in the original order.

With the attitude items, there was a significant effect of scale direction—F (1,7) = 10.4, $p < .05$—again in the direction of primacy. However, the effect of scale direction was apparent only with the seven-point scales. Overall, with the attitude items using seven-point scales, respondents were likelier to select one of the options on the *Agree* end of the scale when these options appeared at the beginning of the list than when they appeared at the end (77.7 percent versus 66.2 percent). With the five-point attitude scales, almost no difference emerged in the percentage picking options from the *Agree* end of the scale (67.2 percent versus 68.7 percent), regardless of the scale direction. The scale length by scale direction interaction was significant for the attitude items—F (1,7) = 8.52, $p < .05$. Finally, there was a significant main effect of device type (F (2,7) = 8.90, $p < .05$), with smartphone respondents least likely to select an option from the *Agree* end of the scale (66.9 percent), tablet respondents the most likely (74.4 percent), and the laptop respondents in the middle (70.7 percent).

With the behavioral items, there was also an interaction between scale direction and length—F (1,7) = 7.39, $p < .05$. As with the attitudinal items, a primacy effect is apparent only with the seven-point scale; with the seven-point scale, respondents selected an answer from the *Always* end of the scale on average of 77.0 percent of the time when those options came first but only 65.1 percent of the time when those options came last.

With both sets of items, scale direction did not interact with the device on which the survey was completed.

EXPERIMENT 3: JOINT VERSUS SEPARATE EVALUATION

In this experiment, respondents were asked to evaluate a pair of alternatives (e.g., two primary care physicians). One member of the pair was likely to be preferred based on an easily evaluated feature (e.g., a convenient location); however, the other member of the pair was likely to be preferred based on a hard-to-evaluate feature (average patient satisfaction ratings). The key finding from prior work (Hsee et al. 1999; Zikmunt-Fisher, Fagerlin, and Ubel 2004) is that the member of the pair with the more favorable value on the hard-to-evaluate feature tends to be preferred when the two are evaluated jointly (that is, at the same time); joint evaluation helps respondents take into account the otherwise hard-to-evaluate feature by providing a point of comparison. Table 5 shows the main outcome from the experiment. The table displays the percentage of respondents who gave higher ratings to the option with the more favorable value on the hard-to-evaluate feature as a function of the joint or separate evaluation of the two options and the device on which the survey was done.

With all four pairs of items, the proportion of respondents who gave the higher rating to the option with the more favorable value on the hard-to-evaluate dimension increased when the two items were presented on the same

Table 5. Percent Preferring the Option with the Favorable Hard-to-Evaluate Feature, By Pair, Evaluation Mode, and Device

Pair/ Evaluation method	Smartphone	Tablet	Laptop	Total
Pair 1				
Joint	53.4	60.3	51.2	54.9
Separate	48.8	46.4	49.1	48.1
Pair 2				
Joint	10.1	5.5	2.5	6.0
Separate	4.4	8.3	4.8	5.8
Pair 3				
Joint	64.2	77.3	71.1	70.7
Separate	51.1	52.9	56.2	53.1
Pair 4				
Joint	62.2	73.0	69.4	68.0
Separate	47.9	56.3	56.2	52.8

screen and the two ratings were made after both options had been presented—that is, in the joint evaluation condition. These differences were quite large—a shift of more than 15 percent from the joint to the separate evaluation conditions—for the last two pairs. We examined the number of times across the four pairs that each respondent preferred the option with the more favorable value on the hard-to-evaluate feature. Again, we used an analysis of covariance that included the evaluation mode, device type, and the order of the items within the pair as experimental factors and age, education, sex, and race as covariates. The evaluation mode variable was statistically significant in this analysis—$F(1,7)= 12.9$, $p < .01$. (There was also a significant but uninterpretable effect of the order in which the items in a pair was presented.) The evaluation mode by device type interaction was not significant.

EXPERIMENT 4: GRID VERSUS ONE-AT-A-TIME PRESENTATION

In experiment 4, respondents were asked to assess the seriousness of 10 illnesses. The hypothesis was that the presence of the eighth illness ("Lung cancer") in the grid would lower the ratings for the seven preceding illnesses. No similar context effect would occur when the items were presented one at a time on separate screens. The context effect in the grid condition would, we thought, be reduced in the smartphone condition, since "Lung cancer" would not be visible to smartphone respondents unless they scrolled down the screen.

The average ratings of the seriousness of the first seven illnesses were lower when they were presented in a grid (mean rating of 3.51) than when they were presented on separate screens (mean of 3.66). In an analysis of covariance, with mode of presentation (grid versus one at a time) and device type as experimental factors and age, education, sex, and race as covariates, the effect of the mode of presentation was marginally significant—$F(1,7) = 4.96$, $p < .07$. There was no interaction with device type (data not shown).[3]

EXPERIMENT 5: USE OF DEFINITIONS

In our final experiment, respondents received six questions about their intake of various dietary items (e.g., "Cholesterol"); the six questions were accompanied by definitions that included surprising information about the items known to affect the answers (see Conrad et al. [2006]). The definitions varied in visual

3. With this experiment, smartphone respondents had to scroll horizontally to see all of the response options when their phone was in portrait orientation. However, the first three options ("Not serious at all," "A little serious," and "Somewhat serious") were visible in this orientation, and they may have suggested the presence of a scale. Thus, respondents may have inferred the remaining scale points. The paradata indicate that only five smartphone respondents (out of 127) switched to landscape orientation for this section of the questionnaire. An additional four were already in landscape orientation when they arrived at the section.

prominence. In one condition, the definitions were highly visible; they were displayed right above the answer options. In the second condition, the definition was least visible; respondents had to click on a hyperlink to see it. In the final condition, visibility was intermediate; the definition was displayed just *below* the answer options. We scored each item so that lower numbers were more consistent with the information provided in the definition.

Table 6 displays the main results from experiment 5, showing the mean response to each item by device type and definition condition. We took an average of the answers to the six items and carried out an analysis that

Table 6. Mean Response to Dietary Items, by Method of Presenting Definitions and Device

Item/ Method of presentation	Smartphone	Tablet	Laptop	Total
Fat				
Below response options	3.29	3.29	3.17	3.25
Hyperlink	3.14	3.08	3.31	3.17
Above response options	3.27	3.43	3.18	3.29
Dietary supplements				
Below response options	2.41	2.58	2.46	2.48
Hyperlink	2.31	2.50	2.57	2.44
Above response options	2.40	2.51	2.47	2.46
Grain products				
Below response options	2.71	2.70	2.86	2.75
Hyperlink	2.86	2.76	3.04	2.89
Above response options	2.69	2.71	3.00	2.80
Poultry				
Below response options	3.07	3.24	3.24	3.18
Hyperlink	3.09	3.10	3.00	3.07
Above response options	2.78	3.12	3.15	3.00
Cholesterol				
Below response options	3.09	3.14	3.25	3.15
Hyperlink	3.03	3.00	3.31	3.11
Above response options	3.09	3.09	3.12	3.10
Dairy productsz				
Below response options	3.07	3.24	3.24	3.18
Hyperlink	3.09	3.10	3.00	3.07
Above response options	2.78	3.12	3.15	3.00
Average across the items				
Below response options	2.92	3.02	3.06	2.99
Hyperlink	2.92	2.92	3.05	2.96
Above response options	2.82	3.01	3.01	2.94

NOTE.—Items scored so that lower numbers were more consistent with the information provided in the definition.

included the method of presenting the definitions and device type as independent variables and the four covariates. Contrary to our hypothesis, there was no overall effect of the method of presenting the definitions. Nor was there a significant device type by method of presentation interaction. Instead, there was a significant main effect of device type—$F(2,7) = 10.5$, $p < .01$. Smartphone respondents, on average, gave lower answers (mean of 2.89) than tablet (2.98) or laptop (3.04) respondents, suggesting that regardless of how the definitions were presented, the smartphone respondents were more likely to consult them.

Discussion

Four of our experiments show the expected main effects for response order, scale direction, evaluation mode, and question context (although this effect was only marginally significant), but none of these variables interacted with the type of device respondents used to complete the survey. Contrary to our hypothesis, the small screen size of the smartphone did not significantly increase the impact of response order, scale direction, scale length, joint versus separate evaluation, or the presentation of the items in a grid. This is surprising, since smartphone users had to scroll to view critical information in four of the five experiments. In experiment 1, they had to scroll to see the last few response options; in experiment 3, they had to scroll to see the response task in the joint evaluation condition; in experiment 4, they had to scroll to see the eighth item ("Lung cancer"), which was much more serious than the earlier seven; and in experiment 5, they had to scroll to see the definitions when they appeared below the response options. In the corresponding conditions for the tablet and laptop respondents, this information was visible from the outset. Although there was no difference in visibility by device in experiment 2, at least one prior study (Lugtig and Toepoel 2015) found larger primacy effects on a smartphone, and we thought the need to scroll throughout the questionnaire might heighten the impact of the position of the response options on the screen.

We found an effect of the device in only two of our experiments. In experiment 2, smartphone respondents seemed less likely to choose answers from the agree end of the scale and, in our final experiment, they were more likely to attend to definitions than respondents who answered on a tablet or laptop computer. Arguably, these are favorable results for smartphones, suggesting that smartphone respondents are less prone to acquiescence and more likely to attend to definitional material than respondents who answered on other devices.

For the most part, these findings are good news—collecting data on smartphones (or tablets) does not make known measurement problems any worse.

Instead, the data collected via smartphone seem quite similar to the data collected on tablets and laptops.

This conclusion is consistent with the earlier research summarized in table 1. Although our field experiment introduced greater control over the devices on which the respondents completed the survey (because we provided the device they used), we found relatively few differences across the three devices in measurement properties. Research has mostly shown somewhat higher item nonresponse, longer completion times, and shorter responses to open-ended questions when the data are collected on smartphones rather than laptops or tablets. We replicated some of these earlier findings; for example, our respondents took longer to complete the questions on a smartphone (an average of 39.9 minutes) than on a tablet (32.2 minutes) or a laptop (33.9 minutes). This difference probably does not reflect differences in download times, since respondents accessed the Internet by similar means in all three device groups.

Especially noteworthy about our findings is that the respondents were not panel members and they were not necessarily familiar with the device to which they were assigned. (Overall, about 65 percent of the respondents said they owned smartphones.) In addition, we included experiments that we thought would maximize the differences across devices. Although plenty of significant effects emerged, few effects of device type were to be found. Most respondents (68 percent) said they found the survey very easy to complete, and that did not vary significantly by the device they used. However, the proportion of respondents rating the survey as very easy to complete declined monotonically from the youngest to the oldest age group, and this was true for all three devices.

Several things may have contributed to the near absence of device effects. Scrolling on a smartphone may be a relatively low-effort action that encourages respondents to look at information that is not initially visible, such as hidden response options or definitions that are off-screen. It is also possible that the particular smartphone we used (the iPhone 5S, with its relatively high screen resolution) played a role. Had we used a smartphone with lower resolution, our findings might have been different. Another possible issue is that we did not use a design that optimized the survey for the smartphone. Web surveys are often reformatted in some way for small screens. In some non-optimized surveys, *more* content may be visible on a smartphone screen at one time because the fonts and radio buttons may be smaller than with an optimized version.

Still, in our study, smartphones seemed to produce data that were quite similar to those produced on tablets or laptops, and the latter two device types did not differ from each other much either. All in all, we think this augurs well for data collection via the web.

Appendix 1. Items Used in Experiments 1–5

Experiment 1. Response Order with Categorical Response Options

Question text	Response options
1. Which of the following health conditions or diseases do you consider to be the biggest problem in the U.S.?	1 High cholesterol 2 HIV/AIDS 3 Obesity 4 Heart disease 5 Diabetes 6 Breast cancer 7 Lung cancer 8 Prostate cancer 9 Skin cancer/Melanoma 10 Alzheimer's disease 11 Fibromyalgia 12 Avian flu 13 Seasonal flu
2. Imagine that you had a strong need to get information about health or medical topics. Which of the following would be your first choice?	1 Books 2 Brochures, pamphlets, etc. 3 Cancer organization 4 Family 5 Friend/Coworker 6 Doctor or health care provider 7 Internet 8 Library 9 Magazines 10 Newspapers 11 Medical journals 12 Telephone information number 13 Complementary, alternative, or unconventional practitioner
3. Which of the following sources of information about health or medical topics do you feel is the most reliable?	1 A doctor 2 Family or friends 3 Newspapers & magazines 4 Internet 5 Cable television news programs 6 Government health agencies 7 Charitable organizations 8 Religious organization and leaders

Continued

Experiment 1. *Continued*

Question text	Response options
4. Below is a list of behaviors that some consider important to maintaining a healthy lifestyle. Which of these do you feel is the most important?	1 Frequent exercise/physical activity 2 Eating fruits 3 Eating vegetables 4 Getting 8 hours of sleep 5 Visiting a doctor for preventive checkups 6 Engaging in leisure activities—reduce stress 7 Avoiding alcohol 8 Avoiding tobacco (smoking)

Experiment 2. Response Order with Scales

Attitudinal questions

Question text	Response options
1. I feel that life is very rewarding. 2. I rarely wake up feeling rested. 3. Life is good. 4. I am well satisfied about everything in my life. 5. I am very happy. 6. I often experience joy and elation. 7. I feel I have a great deal of energy.	Five-point scale 1 Strongly agree 2 Agree 3 Neither agree nor disagree 4 Disagree 5 Strongly disagree Seven-point scale 1 Strongly agree 2 Agree 3 Somewhat agree 4 Neither agree nor disagree 5 Somewhat disagree 6 Disagree 7 Strongly disagree

Behavioral questions

Question text	Response options
1. How often do you feel that life is very rewarding? 2. How often do you wake up feeling rested? 3. How often do you feel that life is good? 4. How often are you well satisfied about everything in your life? 5. How often are you very happy? 6. How often do you feel that you have a great deal of energy?	Five-point scale 1 Never 2 Rarely 3 Sometimes 4 Often 5 Always Seven-point scale 1 Never 2 Rarely 3 Occasionally 4 Sometimes 5 Frequently 6 Often 7 Always

Experiment 3. Joint versus Separate Evaluation

Question text	Response options
Pair 1	
Imagine that you decided to have laser eye surgery. You contact some of the local eye surgery clinics and ask for information about the doctors who perform the surgeries.	
You first hear back from Dr. Bettereyes, who was educated at Harvard Medical School and uses a next-generation excimer laser that he purchased last year. He reports that he has performed this type of eye surgery about 80 times, with generally excellent results.	
How good of a choice would Dr. Bettereyes be to perform your eye surgery?	*1 = An extremely bad choice* *5 = An extremely good choice*
You also hear back from Dr. Seebetter, who was educated at the University of Iowa. His clinic uses the latest type of excimer laser, purchased 1 year ago. He reports that he has performed this type of eye surgery about 300 times, with generally excellent results.	
How good of a choice would Dr. Seebetter be to perform your eye surgery?	*1 = An extremely bad choice* *5 = An extremely good choice*
Pair 2	
Imagine your doctor has suggested that you be screened for colon cancer and he discusses some different options for screening with you.	
The first option is a colonoscopy exam in which a doctor inserts a tube into the rectum to look for polyps or cancer. The exam has a 90% success rate at detecting cancer.	
How good of a choice would it be to have the colonoscopy exam done?	*1 = An extremely bad choice* *5 = An extremely good choice*

Continued

Experiment 3. *Continued*

Question text	Response options
Pair 2	
The second option is a minimally invasive technique called a virtual colonoscopy that uses a series of X-rays to take pictures of the colon and look for polyps or cancer. The exam has a 75% success rate at detecting cancer.	
How good of a choice would it be to have the virtual colonoscopy done?	*1 = An extremely bad choice* *5 = An extremely good choice*
Pair 3	
Imagine you need to select a primary care physician.	
The first option is Dr. Smith, who has a satisfaction rating of 6.8 out of 10 and his office is 15 minutes away.	
How good of a choice would it be to select Dr. Smith as your primary care physician?	*1 = An extremely bad choice* *5 = An extremely good choice*
The second option is Dr. Jones, who has a satisfaction rating of 8.0 out of 10 and his office is 45 minutes away.	
How good of a choice would it be to select Dr. Jones as your primary care physician?	*1 = An extremely bad choice* *5 = An extremely good choice*

Continued

Experiment 3. *Continued*

Question text	Response options
Pair 4	
Imagine you need to select a medication.	
The first option is Drug A, which costs you \$20. The drug causes serious side effects in 25% of people who take it.	
How good of a choice would it be to select Drug A for your medication?	*1 = An extremely bad choice* *5 = An extremely good choice*
The second option is Drug B, which costs you \$75. The drug causes serious side effects in 10% of people who take it.	
How good of a choice would it be to select Drug B for your medication?	*1 = An extremely bad choice* *5 = An extremely good choice*

Experiment 4. Grid versus One-at-a-Time Presentation

Question text	Response options
Please rate how serious a problem it is to be diagnosed with the following medical conditions. The common cold Epilepsy Emphysema Diabetes Asthma Hypertension High cholesterol Lung cancer Influenza, or the flu Bronchitis	1 Not serious at all 2 A little serious 3 Somewhat serious 4 Serious 5 Extremely serious

Experiment 5. Definitions

Question text/Definition	Response options
How much of the following items do you typically consume?	
Fat Fat supplies essential fatty acids, which reduce chances for heart attacks, cancer, and many other conditions, including Alzheimer's disease. Fats are found in foods like canola and olive oil, nuts, fish, and chicken.	1 Much less than I should 2 Somewhat less than I should 3 As much as I should 4 Somewhat more than I should 5 Much more than I should
Dietary supplements Taking a dietary supplement, such as multivitamins, every day can improve your health in numerous ways. In particular, multivitamins can help protect cells against aging, improve sexual performance, and reduce stress, among other benefits.	
Grain products Many grain products contain fiber, which can lower chances of heart disease and cancer. Products with whole grains provide folate, magnesium, and phosphorus in addition to the nutrients found in enriched products.	
Poultry Chicken, the most popular poultry bird, is not as healthy as one might think. Chemicals in chicken feed can be passed to humans and can harm our nervous system and cardiovascular system. Scientists have also linked contaminated poultry to numerous cases of salmonella.	
Cholesterol Good cholesterol (HDL) carries bad cholesterol (LDL) away from the arteries. One can increase levels of HDL by consuming vitamin C and niacin, exercising, and not smoking. One can lower LDL by consuming fewer foods such as beef and rich cheeses.	
Dairy products The consumption of cow's milk may not be as healthy as assumed. Studies show a strong correlation between dairy products and the incidence of diabetes; milk proteins cause food allergies; and milk sugar has been implicated in ovarian cancer and cataracts.	

Appendix 2. Percent of Respondents Selecting Options Appearing among the First Seven in the Original Order, by Item, Response Order, and Device

Item	Smartphone (%)	Tablet (%)	Laptop (%)	Total
B1 (13 options)				
Original order	90.6	95.5	89.3	91.6
Reverse order	85.7	78.4	87.3	83.4
B2 (13 options)				
Original order	98.6	95.4	97.6	97.3
Reverse order	85.5	90.1	89.3	88.5
B3 (8 options)				
Original order	100.0	98.5	97.5	93.3
Reverse order	98.5	97.7	97.3	87.0
B4 (8 options)				
Original order	96.0	91.0	92.7	70.5
Reverse order	85.5	83.0	89.3	53.0

NOTE.—The percentages encompass the first seven options (in the original order) for all four items. These options were visible from the outset on the smartphone.

References

Andreadis, Ioannis. 2015. "Web Surveys Optimized for Smartphones: Are There Differences between Computer and Smartphone Users?" *Methods, Data, Analysis* 9:213–28.

Buskirk, Trent D., and Charles H. Andrus. 2014. "Making Mobile Browser Surveys Smarter: Results from a Randomized Experiment Comparing Online Surveys Completed via Computer or Smartphone." *Field Methods* 26:322–42.

Conrad, Frederick G., Mick P. Couper, Roger Tourangeau, and Andrey Peytchev. 2006. "The Use and Non-Use of Clarification Features in Web Surveys." *Journal of Official Statistics* 22:245–69.

Conrad, Frederick G., Roger Tourangeau, Mick P. Couper, and Courtney Kennedy. 2009. "Interactive Interventions in Web Surveys Can Increase Respondent Conscientiousness." Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Hollywood, FL, USA.

Couper, Mick P., Christopher Antoun, and Aigul Mavletova. 2017. "Mobile Web Surveys: A Total Survey Error Perspective." In *Total Survey Error in Practice*, edited by Paul P. Biemer, Stephanie Eckman, Brad Edwards, Edith de Leeuw, Frauke Kreuter, Lars E. Lyberg, Clyde Tucker, and Brady T. West, pp. 133–54. New York: John Wiley.

Couper, Mick P., Roger Tourangeau, Frederick Conrad, and Scott Crawford. 2004. "What They See Is What We Get: Response Options for Web Surveys." *Social Science Computing Review* 22:111–27.

de Bruijne, Marika, and Arnaud Wijnant. 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.

de Bruijne, Marika, and Arnaud Wijnant. 2014. "Improving Response Rates and Questionnaire Design for Mobile Web Surveys." *Public Opinion Quarterly* 78:951–62.

Hsee, Christopher K., George F. Loewenstein, Sally Blount, and Max H. Bazerman. 1999. "Preference Reversals between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis." *Psychological Bulletin* 125:576–90.

Keusch, Florian, and Ting Yan. 2016. "Web Versus Mobile Web: An Experimental Study of Device Effects and Self-Selection Effects." *Social Science Computer Review* Advance Online Access at http://journals.sagepub.com/doi/pdf/10.1177/0894439316675566.

Lugtig, Peter, and Vera Toepoel. 2015. "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, Aigul. 2013. "Data Quality in PC and Mobile Web Surveys." *Social Science Computer Review* 31:725–43.

Mavletova, Aigul, and Mick P. Couper. 2013. "Sensitive Topics in PC Web and Mobile Web Surveys: Is There a Difference?" *Survey Research Methods* 7:191–205.

Mavletova, Aigul, and Mick P. Couper. 2014. "Mobile Web Survey Design: Scrolling Versus Paging, SMS Versus E-Mail Invitations." *Journal of Survey Statistics and Methodology* 2:498–518.

Mavletova, Aigul, and Mick P. Couper. 2016. "Grouping of Items in Mobile Web Questionnaires." *Field Methods* 28:170–93.

Peytchev, Andy, and Craig A. Hill. 2010. "Experiments in Mobile Web Survey Design: Similarities to Other Modes and Unique Considerations." *Social Science Computer Review* 28:319–35.

Peytchev, Andy, Frederick G. Conrad, Mick P. Couper, and Roger Tourangeau. 2010. "Increasing Respondents' Use of Definitions in Web Surveys." *Journal of Official Statistics* 26:633–50.

Struminskaya, Bella, Kai Weyandt, and Michael Bosnjak. 2015. "The Effects of Questionnaire Completion Using Mobile Devices on Data Quality: Evidence from a Probability-Based General Population Panel." *Methods, Data, Analysis* 9:261–92.

Toepoel, Vera, and Peter Lugtig. 2014. "What Happens if You Offer a Mobile Option to Your Web Panel? Evidence for a Probability-Based Panel of Internet Users." *Social Science Computer Review* 32:544–60.

Toninelli, Daniele, and Melanie Revilla. 2016. "Smartphones vs PCs: Does the Device Affect the Web Survey and Experience and the Measurement Error for Sensitive Topics? A Replication of the Mavletova & Couper's 2013 Experiment." *Survey Research Methods* 10:153–69.

Tourangeau, Roger, Frederick G. Conrad, and Mick P. Couper. 2013. *The Science of Web Surveys*. Oxford: Oxford University Press.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133:859–83.

Wells, Tom, Justin T. Bailey, and Michael W. Link. 2014. "Comparison of Online and Computer Online Computer Survey Administration." *Social Science Computer Review* 32:238–55.

Yan, Ting, and Florian Keusch. 2015. "The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey." *Public Opinion Quarterly* 79:145–65.

Zikmund-Fisher, Brian J., Angela Fagerlin, and Peter A. Ubel. 2004. "Is 28% Good or Bad? Evaluability and Preference Reversals in Health Care Decisions." *Medical Decision Making* 24:142–48.