mda
methods, data, analyses
Journal for Quantitative Methods and Survey Methodology

Volume 9, 2015 | 2

The Collection of Survey Data using Mixed Devices

Vera Toepoel & Peter Lugtig (Editors)

methods, data, analyses is published by GESIS – Leibniz Institute for the Social Sciences.

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects. We especially invite authors to submit articles extending the profession's knowledge on the science of surveys, be it on data collection, measurement, or data analysis and statistics. We also welcome applied papers that deal with the use of quantitative methods in practice, with teaching quantitative methods, or that present the use of a particular state-of-the-art method using an example for illustration.

All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. Mda appears in two regular issues per year (June, December).

Please register for a subscription via http://www.gesis.org/en/publications/journals/mda/subscribe

Print: Verlag Pfälzische Post GmbH, Neustadt, Germany
Printed on chlorine-free paper.

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)

Content

methods, data, analyses | Vol. 9(2), 2015, pp. 155-162 DOI: 10.12758/mda.2015.009

Online Surveys are Mixed-Device Surveys. Issues Associated with the Use of Different (Mobile) Devices in Web Surveys

Vera Toepoel & Peter Lugtig
Utrecht University

1 Issues in Mixed-Device Surveys

Survey research is changing in a more rapid pace than ever before, and the continuous and exponential growth in technological developments is not likely to slow down. Online surveys are now being completed on a range of different devices: PC, laptops, tablets, mobile phones or hybrids between these devices. Each device varies in screen sizes, modes of operationalization and technological possibilities. We define online surveys that are in practice being completed on different devices as mixed-device surveys. This special issue discusses issues in the design and implementation of mixed-device surveys, with the aim to bring survey research to the next level: in our view all web surveys should from now be thought of as mixed-device surveys.

Theory and best practices for mixed-device surveys are still in its infancy. The current state of knowledge about the dynamics of taking surveys on mobile devices is not as advanced as necessary in times of rapid change. While current technology opens great possibilities to collect data via text, apps, and visuals, there is little scientific research published about the actual uses and best practices of these applications to increase data quality. Researchers and survey methodologists in particular need to find ways to keep up with fast changing technologies.

1.1 Mobile Penetration Rates and Mobile Survey Completion

The penetration rate of mobile phones with Internet connection has increased dramatically in the last couple of years. Europe tops the global market on smartphone penetration. In the Netherlands, for example, there has been an increase from around 36% of the population owning a mobile phone with Internet access in 2010 to 72% in 2013 (SN, 2013). In the United States, figures increased from 35% in 2011 to 56% in 2013 (PEW, 2013). Although the majority of the population owns a smartphone, only a small part of the population is actually using it for survey completion. This is probably related to the fact that online surveys are often not yet adapted to be completed on small devices. However, if the questionnaire is dynamically programmed and suitable for completion on small devices, more people are inclined to use a mobile device for survey completion. We found for example that 57% of panel members with a mobile phone used it when being prompted in a dynamically programmed survey (Toepoel & Lugtig, 2014).

1.2 Mixed Device Surveys – a Research Agenda

Representation

The main drawback of online surveys has always been the lack of a sampling frame of email addresses for the general population. Mobile devices, and especially mobile phones, may in the future be used to overcome this problem, because they offer so many channels of communication.

For example, mobile surveys can draw on the advantages of probability-based sampling via Random Digit Dialling (RDD). Second, mobile surveys can easily switch between self-administered and interviewer-administered questions and approach respondents using multiple methods (apps, sms, e-mails and calls). This can be especially useful in the context of a panel study. When respondents are interviewed multiple times, respondents can be approached in multiple ways. On top of this, the mode of survey administration can also be switched within measurements. We know little about what works in practice, and formal studies that document the combined effects on coverage and nonresponse error of different sampling methods for mobile devices are to our knowledge non-existent.

Direct correspondence to
Vera Toepoel, Utrecht University
E-mail: v.toepoel@uu.nl

Measurement

Earlier studies have shown that some survey questions are better asked in a particular survey mode. Data quality is generally higher in self-administration modes (Saris & Gallhofer, 2007; Campanelli et al., 2013), especially when the topic of interest is in some way sensitive (Kreuter et al.). On the other hand, an interviewer may lead to better data when questions are complicated; for example when working out a respondent's life history.

Mobile surveys can draw on technological innovations that come with big data. Sensor data such as GPS, accelerometers, or biomarkers are available on almost all mobile phones and tablets. They offer new and better ways to collect data on specific questions, and can be used to investigate how context affects data quality (see Link et al., 2014). In addition, sensor data can alleviate the burden of survey completion for respondents in time-consuming time budget, health and travel studies.

Although mobile devices offer new possibilities, they are not without their pitfalls. The screen size is smaller than on traditional computers, there is a variability in how questions are displayed (depending on the type of device, personal preferences and browsers) and entering data works differently.

In addition, people use mobile devices differently from traditional computers. People are used to using mobile phones for short messaging, not for taking long surveys. This means that questionnaires should probably be shortened, or split into multiple short questions. In the future, surveys on mobile phones may consist of only a few questions at a time, asked in several bursts.

The fact that available studies often show mixed findings on for example response timings, break-off rates, and survey evaluation in mixed-device studies , can be (partly) an outcome of the rapid changes in technology over time in addition to increased societal learning and growing comfort with devices and their many features (AAPOR Taskforce Report on Mobile Technologies, 2014). The fact that respondents complete online surveys on traditional desktop PCs as well as new mobile devices makes designing surveys a challenge. Issues associated with mixed-mode surveys – for example whether the questionnaire should be optimized for each mode versus a generalized design- can be extended to a mixed-device context.

1.3 Moving from Online Surveys to Mixed-Device Surveys

In order to adapt our surveys to new technologies, we need to redesign our surveys. For example, we have to rethink the use of some question formats. It took time before survey methodologists understood how to redesign paper-and-pencil surveys to online surveys. Now, we have to redesign online surveys to become mixed-device surveys.

Long matrix questions are not suitable for small devices. For example, slider bars, and especially Visual Analogue Scales that work on a point-and-click-principle, save space on the screen (Toepoel, 2016). In addition to question formats, questionnaire length is important to take into account when designing a multi-device survey. Research shows mixed results when it comes to measurement differences between devices (e.g., Bosnjak et al., 2013; de Bruijne & Wijnant, 2013; Buskirk, 2015; Busse & Fuchs, 2012; Lynn & Kaminska, 2013; Peytchev & Hill, 2010; Lugtig & Toepoel, 2015; Vehovar, Berzelak & Lozar-Manfreda, 2010). If questions are dynamically programmed and designed for mixed-devices, measurement differences seem to be minimal.

Mobile phones are rapidly replacing key tasks formerly done on PC and laptops. It seems a matter of time that mobile phones or mobile devices in general are preferred for survey completion over regular desktop PCs. For example, Toepoel (2016) shows that respondents evaluate the completion of surveys on mobile phones better when they have more experience in mobile phone survey completion.

2 Papers in this Special Issue

The papers in this special issue on mixed-device surveys all study the issues mentioned in the previous section, and provide a start for understanding how to design mixed-device surveys. They offer a unique view on questionnaire design in an era where researchers will not know in advance what device a respondent is going to use to complete a survey, let alone how the questionnaire looks on the respondent's device. We can, however, try to predict respondent behavior, in addition to designing our online questionnaires with care.

The first paper in this special issue by Axinn, Gatny, and Wager is titled "maximizing data quality using mode switching in mixed-device survey design". Since the advantages of the web mode for studies with frequent re-interviews can be offset by the serious disadvantage of low response rates and the potential for nonresponse bias, the authors examine the potential for a mixed-device approach with active mode switching to reduce attrition bias. The Relationship Dynamics and Social Life (RDSL) study design allows panel members to switch modes by integrating telephone interviewing into a longitudinal web survey with the objective of collecting weekly reports. The authors found that allowing panel members to switch modes kept more participants in the study compared to a web only approach. In addition, they found that the characteristics of persons who ever switched modes were different from those who did not. Mode options and mode switching can therefore be important for the success of longitudinal web surveys to maximize participation and minimize attrition.

In the second paper, Arn, Klug, and Kolodziejski look at the challenge of optimizing survey layout in online research to enable multi-device use. This paper presents results of the implementation of a new adapted design at the panel of DemoSCOPE that allows the participants to take part in a survey on multiple (especially mobile) devices. To evaluate this adapted design, the authors compare interview data and question timings of panellists who participated before and after the design transition. The key outcomes in this study are the completion rate, item non-response, open questions, straightlining, timing of single question and the length of the total interview are presented. In addition, the authors have presented examples of both old and new designs to the panel community and invited them to assess these examples concerning orientation, colour, design and usability. The authors evaluate the differences in these assessments before and after the design transition for smartphone and desktop users. They end with suggestions for best practices for online studies on different devices.

Andreadis shows in the third contribution to this special issue that computer users and smartphone users give responses of almost the same quality. Combining a design of one question in each page and innovative page navigation methods, we can get high quality data by both computer and smartphone users. The two groups of users are also compared with regard to their precisely measured item response times. The analysis shows that using a smartphone instead of a computer increases the geometric mean of item response times by about 20%. The data analyzed in this paper were collected by a smartphone-friendly web survey. As a result, there are no significant interactions between smartphone use and either the length of the question or the age of the respondent. Thus, the longer response times among smartphone users should be attributed to other causes, such as the likelihood of smartphone users being distracted by their environment.

Buskirk, Saunders, and Michaud note that survey researchers are still trying to understand which online design principles directly translate into presentation on mobile devices and which principles have to be modified to incorporate separate methods for these devices. One such area involves the use of input styles such as sliding scales that lend themselves to more touch centric input devices such as smartphones or tablets. Operationalizing these types of scales begs the question of an optimal starting position and whether these touch centric input styles are equally preferred by respondents using less touch capable devices. While an outside starting position seems optimal for slider questions completed via a desktop computer, this solution may not be optimal for completion via mobile devices. The experiment presented in the paper by Buskirk, Saunders and Michaud moves the mixed device survey literature forward by directly comparing outcomes from respondents who completed a collection of survey scales using their smartphone, tablet or computer. Within each device, respondents were randomly assigned to complete one of 20 possible versions of scale items determined by a combination of three experimental

factors including input style, length and number formatting. Results from this study suggest more weaknesses than strengths for using slider scales to collect survey data using mobile devices and also suggest that preference for these touch centric input styles varies across devices and may not be as high as the preference for the more traditional radio button style.

Struminskaya, Weyandt, and Bosnjak use the data from six online waves of the GESIS Panel, a probability-based mixed-mode panel representative of the German population to study whether the responses provided using tablets or smartphones differ on indicators of measurement and nonresponse errors than responses provided via personal computers or laptops. They extend the scope of past research by exploring whether data quality is a function of device-type or respondent-type characteristics using multilevel intercept-only models. Overall, they find that responding with mobile devices is associated with a higher likelihood of measurement discrepancies compared to PC/Laptop survey completion. For smartphone survey completion, the indicators of measurement and nonresponse error tend to be higher than for tablet completion. However, the effects are relatively small and some indicators (such as straightlining) are not related to a device but are attributable to a respondent.

In all, this special issue on mixed-device surveys in *methods, data, analyses* offers food for thought on how to design surveys in the modern era. The future will tell us whether the design principles discussed in this issue will hold when new devices arise. Until then, we are happy that we live in exciting times for survey methodology.

References

AAPOR Taskforce Report on Mobile Technologies. (2014). Mobile Technologies for Conducting, Augmenting and Potentially Replacing surveys: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. Retrieved from: http://www.aapor.org/Mobile_Technologies_Task_Force_Report.htm

Buskirk, T. D. (2015). The Rise of Mobile Devices: From Smartphones to Smart Surveys. *The Survey Statistician*, 72, 25-35.

Busse, B., & Fuchs M. (2012). Recruiting Respondents for a Mobile Phone Panel. *Methodology,* 1-10.

Bosnjak, M. et al. (2013). Online Survey Participation via Mobile Devices. Retrieved 29th of July 2013 from: http://www.psyconsult.de/bosnjak/pages/publications/conference-contributions.php

Campanelli, P., Nicolaas, G., Jackle, A., Lynn, P., Hope S., Blake M., & Gray M. (2013). A Classification of Question Characteristics Relevant to Measurement Error and Consequently Important for Mixed Mode Questionnaire Design. Paper presented at the Presented 11 October 2011 at the Royal Statistical Society, London, UK. Retrieved

from: http://www.natcenweb.co.uk/genpopweb/documents/other-resources/RSS-Oct-2011-Handout-Recommendations.pdf

De Bruijne, M., & Wijnand, 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.

Groves, R. M., & Lyberg L. (2010). Total Survey Error. Past, Present, and Future. *Public Opinion Quarterly*, 74, 849-879.

Lynn, P., & Kaminska, O. (2013). The Impact of Mobile Phones on Survey Measurement Error. *Public Opinion Quarterly,* 77, 586-605.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly 72*, 847–65

Link, M., Duan, S., Bristol K., & Lai, J. (2014). *The Generational Technology Divide and Implications for Smartphone Data Collection.* Paper presented at the Conference of the American Association of Public Opinion Research, May 15-18, Anaheim.

Lugtig, P., & Toepoel, V. (2015). Mixed Devices in a Probability Based Panel Survey. Effects on Survey Measurement Error. *Social Science Computer Review.* DOI: 10.1177/0894439315574248

Peytchev, A., & Hill, G. (2010). Experiments in Mobile Web Survey Design: Similarities to Other Modes and Unique Considerations. *Social Science Computer Review.* 28, 319-335.

PEW Research Center Report. Smartphone Ownership. (2013). Update June 5 2013 by Aaron Smith. Retrieved from: http://www.pewinternet.org/files/old-media//Files/Reports/2013/PIP_Smartphone_adoption_2013_PDF.pdf

Schouten, B., Cobben, F., van der Laan, J., & Arends, J. (2014). The impact of contact effort and interviewer performance on modespecific nonresponse and measurement bias. CBS Discussion Paper, 2014/05.

Statistics Netherlands. (2013). Retrieved from February 3 2014 from: http://statline.cbs.nl/statweb/

Toepoel, V. (2016). Buttons of balken, klikken of slepen: wat werkt er nu het beste op mobiele telefoons, tablets of PCs? In A. E. Bronner, P. Dekker, E. de Leeuw, L. J. Paas, K. de Ruyter, A. Smidts, & J. E. Wierenga (Eds.), *Ontwikkelingen in het Marktonderzoek*. Jaarboek 2016 Markt Onderzoek Associatie. Haarlem: Spaar en Hout.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: Wiley.

Toepoel, V., & Lugtig, P. (2014). What happens if you offer a mobile option to your web panel? *Social Science Computer Review,* 544-560.

Vehovar, V. et al. (2010). Mobile Phones in an Environment of Competing Survey Modes: Applying Metric for Evaluation of Costs and Errors. *Social Science Computer Review*, 28, 303-318.

methods, data, analyses | Vol. 9(2), 2015, pp. 163-184 DOI: 10.12758/mda.2015.010

Maximizing Data Quality using Mode Switching in Mixed-Device Survey Design: Nonresponse Bias and Models of Demographic Behavior

William G. Axinn, Heather H. Gatny & James Wagner
University of Michigan Survey Research Center

Abstract

Conducting survey interviews on the internet has become an attractive method for lowering data collection costs and increasing the frequency of interviewing, especially in longitudinal studies. However, the advantages of the web mode for studies with frequent re-interviews can be offset by the serious disadvantage of low response rates and the potential for nonresponse bias to mislead investigators. Important life events, such as changes in employment status, relationship changes, or moving can cause attrition from longitudinal studies, producing the possibility of attrition bias. The potential extent of such bias in longitudinal web surveys is not well understood. We use data from the Relationship Dynamics and Social Life (RDSL) study to examine the potential for a mixed-device approach with active mode switching to reduce attrition bias. The RDSL design allows panel members to switch modes by integrating telephone interviewing into a longitudinal web survey with the objective of collecting weekly reports. We found that in this design allowing panel members to switch modes kept more participants in the study compared to a web only approach. The characteristics of persons who ever switched modes are different than those who did not – including not only demographic characteristics, but also baseline characteristics related to pregnancy and time-varying characteristics that were collected after the baseline interview. This was true in multivariate models that control for multiple of these dimensions simultaneously. We conclude that mode options and mode switching is important for the success of longitudinal web surveys to maximize participation and minimize attrition.

Keywords: attrition, longitudinal (panel) study, mode switching, non-response bias, web survey, journal-keeping

1 Introduction

As internet access spreads worldwide, conducting survey interviews via the web has become an attractive method for lowering data collection costs while increasing the frequency of interviewing, especially in longitudinal studies (Couper, 2008). Web surveys are particularly appealing to researchers studying dynamic behaviors that require detailed, timing-specific measures collected over a long period of time (Axinn, Jennings & Couper, 2015; Stone, Shiffman, Atienza, & Nebeling, 2007; Mehl & Conner, 2013). Besides the cost savings, the advantages of web surveys for these studies include portability, flexibility, and confidentiality – web surveys allow respondents to complete surveys at whatever time and location is convenient and private for them. These properties extend to multiple devices including personal computers, laptops, tablets, and smartphones, further providing respondents with more options for convenience with little difference in measurement error between the devices (Lugtig & Toepoel, 2015). However, the advantages of the web mode for studies with frequent re-interviews can be offset by the serious disadvantage of low response rates and the potential for nonresponse bias to mislead investigators. Web surveys are known to have lower response rates compared to almost any other survey mode (Lozar Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008; Shih & Fan, 2008), and in longitudinal designs these lower response rates can produce serious misinformation regarding the true nature of changes over time (Graham & Donaldson, 1993; Lepkowski & Couper, 2002; Kristman, Manno, & Côté, 2005). In this paper we examine the potential for a mixed-device approach which allows panel members to switch modes – integrating telephone interviewing into a longitudinal web survey – to reduce the potential for attrition bias to produce misleading measures of dynamic behaviors.

We use data from the Relationship Dynamics and Social Life (RDSL) study, which was designed to investigate factors shaping the dynamics of sexual behavior, contraceptive use, and unintended pregnancy in a cohort of young adult women. The RDSL studied a random, population-based sample of 1,003 young women ages 18-19, residing in one county in the state of Michigan, USA. The representative sample of young women in the general population was accomplished by selection of individuals from the state driver's license and personal identification card databases. Investigators conducted a 60-minute face-to-face baseline survey to launch the study and then enrolled women in a 2.5-year panel study that required completion of weekly surveys about contraceptive use, relationships, and prospective pregnancy intentions. Web and telephone modes were selected for the weekly surveys

Direct correspondence to
Heather H. Gatny, Survey Research Center, Institute for Social Research, University of Michigan, PO Box 1248, Ann Arbor, MI 48106-1248, USA
E-mail: hgatny@umich.edu

to maximize respondent privacy by eliminating the need for written records that must be kept and could potentially be discovered by a third party. Additionally, telephone surveys generally achieve higher response rates than either mail or web (Lozar, Manfreda, et al., 2008). Ninety-two percent of women in the baseline survey had internet access and were encouraged to complete the follow-ups surveys by web. Women without internet access were asked to complete the surveys by telephone. However, all women were provided the study website URL and telephone number and were allowed to complete each week's survey by either mode. This protocol actively used mode switching to reduce non-response. Those who were late completing their journals were contacted by email first, then by phone, to complete their surveys. The face-to-face baseline interviews were conducted March 2008-July 2009 and the web-based panel study concluded in February 2012. The response rate for the baseline interview was 84% (RR1; AAPOR, 2011); 99% of those who completed the baseline survey agreed to participate in the panel; and 75% continued to participate in the panel for at least 18 months.

The RDSL design provides an unusually strong opportunity to investigate associations between individual characteristics measured in the baseline interview and subsequent participation in the panel study. Several studies have examined the consequences of changing modes on participation in a single wave in panel surveys (Jackle, Lynn, & Burton, 2015; Lynn, 2012; Hoogendoorn, Lamers, Penninx, & Smit, 2013; Wagner, Arrieta, Guyer, & Ofstedal, 2014). This study is unique in that interviewing was conducted weekly and panel members were allowed to switch between modes as necessary. This allows us to examine the impact on estimates of dynamic behaviors of allowing panel members to switch modes across multiple waves. In this paper we examine the extent to which use of a mixed-device approach and active mode-switching alter results relative to the alternative no-switching approach. Using RDSL measures we estimate the extent to which allowing mode-switching improves participation in the longitudinal measurement for select subgroups and characteristics. First, we use baseline measures to compare the cases who would have been represented if no mode-switching was allowed with the cases who remained in the study by allowing them to switch modes. Second, we use the baseline measures to assess associations between various individual characteristics and the number of mode switches each respondent made during the 18-month panel. Third, we investigate the extent to which the addition of the option to switch modes changes estimates of key behaviors in the panel study, including residential moves, changes in intimate partners, sexual experience, contraceptive use, and pregnancy. We also extend this investigation into estimates of consequences of specific intimate partner dynamics across the panel study to produce mode switching in subsequent journals. Finally, we investigate the extent to which key model parameters from previously published substantive results differ when models are estimated on cases that used the *same* mode for all interviews. Alto-

gether the results provide important new evidence of the ability of mixed-device mode switching approaches to compensate for the weaknesses of single mode web-only approaches by reducing attrition.

2 Mixed-Device Mode Switching

Theoretically, allowing mixed-device mode switching in a panel design may have many advantages for maximizing participation across time. Two different processes define the total success maximizing survey participation: establishing contact with the respondent and the respondent's consent to complete the survey. A crucial issue in obtaining respondent consent and cooperation is the incentive to burden ratio associated with completing the survey (Groves & Couper, 1998). Groves, Singer, and Corning (2000) describe this as the "leverage-saliency" theory of nonresponse. Survey respondents place different values on aspects of the survey request. Groves, Singer, and Corning label these "leverage." Leverage can be either positive or negative. Some panel members place a high positive value on an incentive while others may be interested in completing the survey because they find the topic interesting. A long survey might be a negative leverage for some panel members. On the other hand, the survey design makes particular features of the design "salient." For instance, the survey may emphasize the incentive or the interesting questionnaire in their contacts with panel members. Response rates are maximized when the appropriate set of design features are made salient to those for whom these features have larger leverage. For example, the shorter and easier a survey is to complete, the lower the negative leverage. For those panel members for whom this aspect of the survey is an important feature, making this salient may increase their probability of participating. Keeping survey tasks short always reduces the burden and this is especially important for repeated interviewing over time (longitudinal studies) and the more often the interview is repeated the more important this becomes. But different design features are salient for different respondents. One appeal of mixed-device surveys is the opportunity to allow each respondent to use whatever device is easiest for that respondent. With web surveys, computers, tablets, and smartphones could each be used, allowing each respondent to choose the device that is the least burden for that specific respondent. Allowing respondents to change devices across interviews provides the means for respondents to select the easiest device at each interview, enhancing the ease of the experiences. Easier experiences decrease negative leverage that may reduce the probability of completing the survey and thereby increase respondent participation.

Mode switching is a related design feature. Allowing the respondent to switch modes at each interview allows the respondent to select the easiest mode for the specific circumstances of that interview. Easier modes reduce burden and increase

respondent cooperation. So dynamic life circumstances that make one mode easier one week and a different mode easier the next week support a design that allows mode switching to maximize respondent participation and reduce attrition. Residential moves, employment/financial change, or intimate partner changes are all examples of factors likely to make mode switching appealing. In fact, life circumstances that make daily activities more complicated in any way, including pregnancy, childbirth, poverty, traumatic experience, health limitation, or other crisis circumstances all make ease of completing the survey a high priority in maintaining high respondent cooperation. To the extent mode-switching makes completing the survey easier, any of these circumstances may motivate mode switching as a means to increase participation and reduce attrition.

Mode switching may be equally valuable for establishing contact with respondents across multiple interviews in a longitudinal survey. A key source of attrition in longitudinal surveys is failure to re-contact the specific respondent at future interviews (Groves & Couper, 1998; Schoeni, Stafford, McGonagle, & Andreski, 2013; Couper and Ofstedal, 2009; Ribisl et al., 1996). Many factors make failure to re-contact likely, especially residential moves, but also job loss, divorce, intimate partner breakups, and significant income changes (Lepkowski & Couper, 2002; Trappmann, Gramlich, & Mosthaf, 2015). Life changes that make it more difficult to locate respondents or find them available to complete a survey may reduce re-contact. The portability of both web and phone make them desirable modes in these circumstances, but the ability to switch across these modes may enhance the overall ease of responding. Thus longitudinal surveys that provide mode-switching options may be more successful at keeping respondents with complex or changing life circumstances involved in longitudinal surveys.

3 Data, Mode Switching Measures, and Analysis Plan

3.1 Data

The Relationship Dynamics and Social Life (RDSL) study focuses on 18-19 year old women in a single county in the State of Michigan, USA. The specific county was selected both because several key demographic characteristics of that county fell near the median for the State and because the county had a high degree of variability with respect to income and race, providing high diversity in the general population without requiring over-samples of sub-groups (Barber, Kusunoki, & Gatny, 2011). Sixty-minute face-to-face baseline interviews were conducted with each woman at the start of the study to gather information on her family background; education and career plans; attitudes, values, beliefs, and knowledge about

sexual practices; romantic relationships; and sexual experiences. After the baseline interview, the women were each invited to participate in the weekly journal portion of the study. Over 99% of respondents who completed the baseline survey enrolled in the weekly surveys (n=992) (Barber et al., 2011).

Significant effort was taken to keep these young women enrolled in the weekly journal-keeping study. The burden of each weekly interview was kept low by maintaining an average interview length of seven minutes or less. Emails and/or text messages were sent weekly to remind respondents. Monetary incentives of $1 per weekly journal and a bonus of $5 for having completed five weekly journals on time were given, and small gifts—such as pens and lip balm—were also given to encourage retention (Gatny, Couper, Axinn, & Barber, 2009). Respondents who failed to complete the journal on time were contacted by email and phone, and then eventually by letter. After 60 days of not completing a journal, increased incentives were offered for the next journal entry.[1] At the completion of the journal-keeping study, 84% of respondents who were interviewed at baseline had participated in journal-keeping for at least 6 months, 79% for at least 12 months, and 75% for at least 18 months with some journals missing (Barber et al., 2011).

3.2 Measures of Mode Switching

For this study of mode-switching, we confine our analyses to the 947 respondents who completed 2 or more journals. We analyze journals completed within the first 18 months of journal enrollment (n=39,598) to minimize bias from attrition. At baseline 92% (872/947) of respondents selected to complete the journals by web and 8% (75/947) selected the phone instead. Of the 872 respondents who selected the web, 60% (520/872) completed at least one journal by phone. The range was 1-78 journals completed by phone among these respondents who initially selected the web, and the mean was 8 journals completed by phone. Note this count does not include the mode for journal 1 because that journal was completed with the interviewer.

Of the 75 respondents who selected the phone, 39% (29/75) completed at least one journal by web. The range was 1-64 journals completed by web among these respondents who initially selected the phone, and the mean was 23 journals completed by web. Again this count does not include the mode for journal 1 because that journal was completed with the interviewer.

To construct a measure of the count of the number of mode switches which took place we created a variable counting the number of times a respondent completed a journal in a mode different from the mode used at the previous journal.

1 See Barber et al. (2012) for more information on the design and implementation of the RDSL study.

Note this measure does not include journal 1 in the count because that journal was unlike all others – it was conducted during the baseline interview with the interviewer. This measure also does not include journal 2 because it is the first journal that the respondent completed without the help of the interviewer. Also, a large proportion (84% or 132/157) of those who only had one mode switch had the switch at journal 2. In other words, they did journal 2 in a mode different than what they enrolled in at baseline. The measure of the number of mode switches begins counting switches at journal 3 (n=37,659). Starting at journal 3, a switch is a mode different from the mode used at the previous journal.

The range of mode switching was 0-30 switches. More than half of the sample (504 respondents) had zero mode switches. Though this is a large group of stable single mode users, nearly half of the sample (443 respondents) had at least one mode switch. The mean number of switches was 1.93 and the most common number of switches was two. Over 16% of the sample experienced two mode switches – two switches implies starting in one mode, completing a single journal in the alternate mode, and then returning to the initial mode for the remainder of the study. Nearly 25% of the sample experienced three or more mode switches.

The timing of mode switching as respondents complete more journals implies some switching motivated by the respondent's experience with the initial mode. For example, 29% (18/62) of those who only had one mode switch had the switch at journal 3. Journal one was completed with the interviewer, journal 2 was the respondent's first journal alone, and the journal 3 switches took place during the respondent's second interview alone. In other words, they completed that journal in a mode different than what they used at journal 2, the first journal completed without the help of an interviewer present. Some may have simply wanted to try an alternative to see if it was easier, others may have had a negative experience with their first attempt to complete the journal on their own. Respondents experienced their first mode switch across journals 3 through 71, but by journal 8, more than half of respondents (228/443) who ever experienced a switch had experienced their first switch. Over the 18 months analyzed here respondents could have completed as many as 78 journals, but first mode switches appear to take place early in the process.

3.3 Analysis Strategy

Our analysis proceeds in three steps, each time focusing on mode-switching as the key alternative to attrition from the study. In the first step we use data from the baseline interview before the weekly journal keeping is launched to assess the associations between baseline characteristics and mode switching behavior. This analysis has two parts. In part one we use the comparison of those cases who only used a single mode to those cases who remained in the RDSL by switching modes

to perform t-tests of mean differences, allowing us to identify prior characteristics associated with subsequent mode switching. In part two we estimate multivariate models of the likelihood of ever making a mode switch and of the number of mode switches. This part of this step allows us to assess the independence of associations between various prior background characteristics and respondents' mode switching behaviors.

In the second step we use data from the journal itself to assess the causes and consequences of mode switching rather than attrition from the study. Again, this analysis has two parts. First, we investigate the overall relationship between mode switching behavior and other behaviors reported in the journal. Here we compare key behaviors measured in the journal between those cases who only used a single mode and those cases who remained in the study by switching modes. Second, we investigate the association between measures of weekly relationship dynamics and the likelihood the week ends in an interview mode switch. The investigation uses the special relationship dynamics measures from the RDSL study to highlight how those behaviors themselves may be associated with mode switching.

In the third step, we assess the extent to which substantive conclusions from multivariate models can be altered by eliminating the mode switching alternative to attrition from the study. We use a specific model previously published using RDSL data. We estimate this model as published, and then re-estimate the model assuming the cases that used mode switching would have dropped out of the study (attrition). This comparison highlights the potential substantive research consequences of allowing interview mode switching as an alternative to attrition from the study.

4 Results

4.1 Baseline Characteristics and Subsequent Mode Switches

4.1.1 Comparison of respondents who switch mode with those who do not

Our analysis begins with comparisons between those who switched modes during the 18-month panel study and those who did not (Table 1). We present three versions of each statistic – one for the total sample, one for those who never switched modes, and one for those who ever switched modes (Table 1). Those who switched modes are the most likely to be lost to attrition in a single mode study. The p-values associated with each row indicate the statistical significance of the difference in each statistic between the respondent who never switched modes and those who ever switched modes. For example, there is a statistically significant difference in mode switching with African Americans being more likely to switch at least one time (row 1 of Table 1), but there is not a significant difference in high-school grade point average (GPA) between those who switched modes and those who did not

Table 1 Respondent characteristics

	Total Sample (n=947) %	Subsample who used same mode at every journal (n=504) %	Subsample with at least one mode switch (n=443) %	*p*-value
Sociodemographic Characteristics				
African American	.34	.28	.41	***
Enrolled in school full-time	.51	.54	.47	*
High school GPA [a]	3.12[b]	3.15	3.09	
<$1,000 (1st quartile)	.35	.33	.37	
Currently receiving public assistance	.26	.24	.29	+
Income not enough to make ends meet	.18	.17	.20	
Owns a car	.48	.50	.46	
Childhood Family Background Measures				
Two-parent childhood family structure	.52	.58	.46	***
Biological mother <20 years old at 1st birth	.37	.35	.38	
High religiosity	.57	.54	.61	*
Childhood Socioeconomic Status				
Received public assistance	.37	.33	.41	**
At least one parent has at least some college	.66	.68	.64	
Parents were home owners	.71	.75	.65	***
High parent income	.38	.42	.33	**
Experiences Related to Pregnancy				
Living with partner	.17	.17	.17	
Age at first sex ≤ 16 years	.51	.49	.54	
Two or more sexual partners	.60	.58	.61	
Ever had sex without birth control	.48	.46	.50	
1 or more prior pregnancies	.15	.14	.16	

+ $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (two-tailed independent samples t-tests for significant differences between the two subsamples)

[a] mean GPA presented for sample and subsamples; [b] std. dev.=.61

(row 3 of Table 1). Overall, there are many statistically significant differences in key statistics displayed in Table 1.

We group respondent characteristics into four domains – Sociodemographic Characteristics, Childhood Family Background Measures, Childhood Socioeconomic Status, and Experiences Related to Pregnancy (the main substantive topic of RDSL). All of these characteristics were measured during the baseline inter-

view, before journal-keeping began. None of the measures of experiences related to pregnancy are associated with mode switching during the panel study. By contrast, measures in each of the other domains are associated with significant differences in switching behavior, or potential attrition if switching was not allowed.

Among sociodemographic characteristics, both being African American (compared to being white) and receiving public assistance are associated with significantly higher likelihood of mode switching. We argue any life circumstances that create complexity of social experience are likely to be associated with higher likelihood of mode switching – both results are consistent with that argument. Being enrolled in school full-time is associated with significantly lower likelihood of mode switching. This result is consistent with full-time school promoting stability of experience in early adulthood, in contrast to either part-time or no school. Early adult income levels and car ownership are not significantly associated with mode switching.

Within the domain of childhood family background, growing up in a two-parent family is associated with a significantly lower likelihood of mode switching during the panel study. High religiosity in the childhood family of origin is associated with significantly higher likelihood of mode switching. Experiencing a relatively young mother is not associated with subsequent mode switching. Within the domain of childhood socioeconomic status, growing up in a household that received public assistance is associated with a significantly higher likelihood of mode switching. Growing up in a household in which parents owned their own home or had high incomes were both associated with significantly lower likelihood of mode switching during the panel study. Growing up with parents who had at least some college education is not associated with subsequent mode switching. Again, factors associated with higher mode switching would likely produce attrition if the mode alternatives were not provided.

This initial step in our analysis examines only bivariate associations. In the next step we move on to multivariate models of ever making a mode switch and the number of mode switches – this step allows us to assess the independences of these various associations between individual respondent background and mode switching behaviors.

4.1.2 Associations between respondent background and both likelihood of mode switch and numbers of mode switches

Using the same background characteristics as presented in Table 1, we now estimate multivariate models of mode switching behavior (likelihood of attrition under a single mode design). The first column of Table 2 presents results from a logistic regression model using all the characteristics to predict the likelihood the respondent makes any mode switch. Significant associations documented in this column indicate the specific characteristic is associated with making a mode switch

Table 2 Regression coefficients for models of at least one journal mode switch (logistic) and number of journal mode switches (poisson) (N=947)

	(1) at least one journal mode switch	(2) number of journal mode switches
Sociodemographic characteristics		**POISSON**
African American	**.53 ** **	**.39 *** **
	(.18)	**(.06)**
Enrolled in school full-time	**-.29 +**	**-.28 *** **
	(.15)	**(.05)**
High school GPA	-.13	-.06
	(.13)	(.04)
<$1,000 (1st quartile)	.06	**.16 ** **
	(.16)	**(.05)**
Currently receiving public assistance	.08	**.20 *** **
	(.19)	**(.06)**
Income not enough to make ends meet	.03	**.15 ***
	(.19)	**(.06)**
Owns a car	.15	**-.14 ** **
	(.15)	**(.05)**
Childhood family background measures		
Two-parent childhood family structure	**-.31 +**	-.04
	(.16)	(.05)
Biological mother <20 years old at 1st birth	-.03	.05
	(.16)	(.05)
High religiosity	.08	.08
	(.15)	(.05)
Childhood socioeconomic status		
Received public assistance	.09	.09
	(.17)	(.05)
At least one parent has at least some college	.00	**.16 ** **
	(.16)	**(.05)**
Parents were home owners	**-.33 +**	**-.21 *** **
	(.18)	**(.05)**
High parent income	-.06	**-.24 *** **
	(.17)	**(.06)**
Experiences related to pregnancy		
Living with partner	.01	-.05
	(.21)	(.07)
Age at first sex < 16 years	.11	**.13 ***
	(.19)	**(.06)**
Two or more sexual partners	.00	.02
	(.19)	(.06)
Ever had sex without birth control	.04	.06
	(.18)	(.06)
1 or more prior pregnancies	.10	**-.27 *** **
	(.22)	**(.07)**

	(1) at least one journal mode switch	(2) number of journal mode switches
Other		
Time in study	**.13 *** (.02)**	**.13 *** (.01)**
χ^2	120.49	
Pseudo-R^2	.09	
R^2		.15

Standard errors in parentheses.

† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (two-tailed tests)

independent of the other bivariate associations documented in Table 1. Among these characteristics, being African American, enrolled in school full-time, from a two-parent family, or having parents who owned their own home, each has an independent statistically significant association with ever switching interview modes during the panel study (column 1, Table 2). This means that panel studies of this type which do not allow mode switching may underrepresent respondents who are African American, who are not enrolled in school full-time, who do not come from a two-parent family, and who have parents who did not own their own home. Such attrition bias has the potential to undermine substantive results based on studies that do not allow mode switching. Finally note that in these multivariate models we also control for the length of time in the study before the mode switch – remaining in the study longer significantly increases the likelihood of a mode switch. Consistent with predictions, efforts to keep respondents in longitudinal panel studies for longer periods of time will be more successful when mode switching is designed into the data collection.

Next we use the same measures of respondent background to estimate models of the number of times each individual switches interview modes. Here we use Poisson regression (column 2 of Table 2) because the high skew in the count measure fits a Poisson distribution. The distributional assumptions of the Poisson regression are more consistent with this count of number of switches. This is important because the results in column 2 of Table 2 demonstrate that the majority of background characteristics we measure (11 of 19) have statistically significant and independent associations with the number of mode switches a respondent makes during the 18-month panel study. Failure to allow mode switching in such a panel study greatly increases the chance that the resulting measures will be selective on many different dimensions of social life.

Table 3 Respondent behaviors reported in the journal

	Total Sample (n=947) %	Subsample who used same mode at every journal (n=504) %	Subsample with at least one mode switch (n=443) %	*p*-value
Received public assistance	.25	.19	.32	***
Changed residence	.40	.33	.49	***
Sex	.78	.73	.82	**
Sex without contraception	.50	.41	.59	***
Sex with a new partner	.45	.38	.52	***
Sex with someone other than current partner	.18	.13	.24	***
Conflict with a partner	.16	.11	.21	***
Lived with a partner	.41	.35	.48	***
Pregnant	.13	.10	.18	***

+ $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (two-tailed independent samples t-tests for significant differences between the two subsamples)

4.2 Journal Measures and Journal Mode Switching

4.2.1 Comparing journal measures for those who switched modes and those who did not

Next we examine data from the journal itself. We begin by comparing reports of key substantive behaviors measured in RDSL between respondents who never switched modes and respondents who ever switched modes. The behaviors we investigate include if the respondent received public assistance, changed residence, had sex, had sex without contraception, had sex with a new partner, had sex with more than one partner, had conflict with a partner, lived with a partner, or became pregnant. Table 3 summarizes our findings.

The p-value indicated in each row describes the statistical significance of each comparison. All of these comparisons are statistically significant and in every case the sample who experienced a mode switch had a higher value on the measures. This table provides a powerful summary of the importance of mode switching. In every type of behavior representing core domains of this study, mode switching was associated with higher levels. Without allowing mode switching, it appears the RDSL study would have significantly underestimated each and every core behavior the study was designed to measure.

4.2.2 Predicting mode switches from key behaviors

Now we investigate the possibility that the core behaviors themselves motivate a mode switch. Behaviors such as change in intimate partner relationship status are believed to increase attrition from longitudinal studies because they make locating respondents and convincing those respondents to participate more difficult. Here we use the weekly behaviors of participants in the RDSL study to predict the chances they end the week with a mode switch. Because receiving public assistance was only measured in RDSL quarterly and place of residence was only measured in RDSL monthly, we do not investigate these two factors. Instead we focus on the weekly dynamics of relationships, including sex, contraception, conflict, and pregnancy. In each case we estimate both a bivariate association and then we re-estimate that association controlling for the full set of baseline interview characteristics we examined earlier. The results are presented in Table 4.

Each column of Table 4 comes from a separate model estimate. In columns 5, 7, and 9 of Table 4 we see that sex without contraception, sex with a new partner, and sex with a second partner are each significantly associated with a mode switch at the end of the week, independent of key baseline characteristics. These events increase the likelihood of a mode switch; these data provide evidence that some sexual events may lead to mode switching in the short term. Single mode studies would likely lose respondents who had just experienced similar events, biasing reports of such events downward.

4.3 Substantive Model with and without Mode-Switching

In this analysis (Table 5), we investigate the potential impact of not allowing mode switching on a multivariate model developed to investigate the impact of ambivalent fertility desires on pregnancy risk (Miller, Barber, & Gatny, 2012 {Table 3, Column 3}). This model included a number of demographic control variables as well as experiences related to pregnancy from the baseline interview, such as being 16 years of age or less at first sex. The model, as reported in published research, includes all of the available data. In this original, published model, the desire to become pregnant is a significant and positive predictor of the probability of actually becoming pregnant. Further, the desire to avoid pregnancy is a significant, independent, and negative predictor of the probability of becoming pregnant. This result provided empirical evidence of the simultaneous influence of contrasting attitudes toward pregnancy – an important theoretical advance in our understanding of the relationship among attitudes, intentions, and young adult pregnancies.

The substantive conclusions from the original estimated model are substantially changed when data collected after the first mode switch are omitted. Had mode switching not been an option, many in the study would have likely stopped providing measures (attrition). When the data these respondents provided after the

Table 4 Logistic regression coefficients for models using behaviors reported in a journal to predict a mode switch in the same journal (N=37,659)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Behavior reported in the journal															
Sex		.16 + (.09)	.10 (.09)												
Sex without contraception				.50 *** (.11)	.22 * (.11)										
Sex with a new partner						.56 *** (.14)	.44 ** (.14)								
Sex with someone other than current partner								.54 ** (.19)	.31 + (.18)						
Conflict with a partner										.38 + (.22)	.09 (.23)				
Living with a partner												.10 (.12)	-.07 (.15)		
New pregnancy														.64 * (.31)	.19 (.31)
Sociodemographic characteristics															
African American	.48 *** (.13)		.49 *** (.13)		.48 *** (.13)		.48 *** (.13)		.48 *** (.13)		.48 *** (.13)		.47 *** (.13)		.48 *** (.13)
Enrolled in school full-time	-.22 * (.10)		-.21 * (.10)		-.20 * (.10)		-.21 * (.10)		-.21 * (.10)		-.22 * (.10)		-.22 * (.10)		-.22 * (.10)
High school GPA	-.18 + (.09)		-.18 + (.09)		-.17 + (.09)		-.18 + (.09)		-.18 + (.09)		-.18 + (.09)		-.18 + (.09)		-.18 + (.09)
<$1,000 (1st quartile)	.12 (.11)		.12 (.11)		.12 (.11)		.12 (.11)		.12 (.11)		.12 (.11)		.12 (.11)		.12 (.11)
Currently receiving public assistance	.27 + (.15)		.28 + (.15)		.28 + (.15)		.27 + (.15)		.27 + (.15)		.27 + (.15)		.27 + (.15)		.27 + (.15)
Income not enough to make ends meet	.19 (.14)		.19 (.14)		.18 (.14)		.19 (.14)		.19 (.14)		.19 (.14)		.19 (.14)		.19 (.14)
Owns a car	-.10 (.12)		-.11 (.12)		-.10 (.12)		-.10 (.12)		-.10 (.12)		-.10 (.12)		-.10 (.12)		-.10 (.12)
Childhood family background measures															
Two-parent childhood family structure	-.09 (.12)		-.09 (.13)		-.09 (.12)		-.09 (.12)		-.09 (.12)		-.09 (.12)		-.09 (.12)		-.09 (.12)
Biological mother <20 years old at 1st birth	.11 (.12)		.11 (.12)		.10 (.12)		.11 (.12)		.11 (.12)		.11 (.12)		.11 (.12)		.11 (.12)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
High religiosity	.13 (.12)		.13 (.12)		.13 (.12)		.13 (.12)		.13 (.12)		.13 (.12)		.13 (.12)		.13 (.12)
Childhood socioeconomic status															
Received public assistance	.11 (.13)		.11 (.13)		.10 (.13)		.11 (.13)		.11 (.13)		.11 (.13)		.11 (.13)		.11 (.13)
At least one parent has at least some college	.08 (.12)		.08 (.12)		.08 (.12)		.08 (.12)		.08 (.12)		.08 (.12)		.08 (.12)		.08 (.12)
Parents were home owners	-.21 (.13)		-.21 (.13)		-.22 + (.13)		-.21 (.13)		-.21 (.13)		-.21 (.13)		-.21 (.13)		-.21 (.13)
High parent income	-.26 * (.13)		-.26 * (.13)		-.25 * (.13)		-.26 * (.13)		-.26 * (.13)		-.26 * (.13)		-.26 * (.13)		-.26 * (.13)
Experiences related to pregnancy															
Living with partner	-.01 (.17)		-.03 (.17)		-.03 (.17)		.00 (.17)		-.01 (.17)		-.01 (.17)		.03 (.19)		-.01 (.17)
Age at first sex ≤ 16 years	.14 (.16)		.14 (.16)		.14 (.16)		.14 (.16)		.14 (.16)		.14 (.16)		.15 (.16)		.14 (.16)
Two or more sexual partners	.10 (.16)		.08 (.16)		.08 (.16)		.09 (.16)		.10 (.16)		.10 (.16)		.10 (.16)		.10 (.16)
Ever had sex without birth control	.12 (.14)		.11 (.14)		.11 (.14)		.12 (.14)		.12 (.14)		.12 (.14)		.13 (.14)		.12 (.14)
1 or more prior pregnancies	-.21 (.16)		-.22 (.16)		-.21 (.16)		-.21 (.16)		-.21 (.16)		-.21 (.16)		-.21 (.16)		-.21 (.16)
Other															
Time in study	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)	-.04 *** (.01)
χ^2	241.22	34.05	245.45	56.02	255.41	49.03	255.76	42.13	248.71	35.29	243.22	32.66	242.39	36.33	242.55
Log-likelihood	-6933.04	-7279.52	-6931.28	-7260.13	-6928.25	-7276.90	-6928.08	-7281.13	-6931.77	-7282.81	-6932.93	-7283.82	-6932.65	-7283.07	-6932.87

Standard errors in parentheses.

† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (two-tailed tests)

Table 5 Logistic regression estimates of the effects of positive and negative pregnancy desires on the hazard of pregnancy

	Original Model	Subsample without mode switches
Desire to become pregnant	.22 *	.17
	(.10)	(.12)
Desire to avoid pregnancy	-.24 **	-.26 *
	(.09)	(.10)
Sociodemographic characteristics		
African American	.25	.46
	(.26)	(.36)
Enrolled in school full time	-.15	-.01
	(.22)	(.32)
Graduated high school	.36	.46
	(.25)	(.35)
Receiving public assistance	.43 +	.63 *
	(.25)	(.32)
Importance of Religion	.21	.22
	(.13)	(.17)
Biological mother <20 years old at first birth	.19	-.09
	(.22)	(.29)
One biological parent only (ref=2 parents)	.06	-.13
	(.25)	(.33)
Other (ref=2 parents)	.16	.32
	(.36)	(.46)
Mother‘s education <high school graduate	.09	-.37
	(.34)	(.56)
$15,000-44,999 (ref<=14,999)	-.60 *	-.87 *
	(.31)	(.42)
$45,000-74,999 (ref<=14,999)	-.68 +	-.47
	(.38)	(.50)
$75,000 or greater (ref<=14,999)	-.56	-.51
	(.43)	(.53)
Don‘t know/refused (ref<=14,999)	-.34	-.26
	(.30)	(.39)
Age at first sex 16 years or less	.67 *	.36
	(.30)	(.40)
Lifetime number of sexual partners two or more	.70 *	.48
	(.31)	(.40)
Ever had sex without birth control	.23	.38
	(.27)	(.38)
Number of previous pregnancies	.17 +	.24 *
	(.10)	(.12)
Cohabiting	.38	.87 **
	(.24)	(.32)
Age	-.26	-.17
	(.20)	(.26)
Other		
Time-to-pregnancy	.29 ***	.30 *
	(.08)	(.12)
Time-to-pregnancy squared	-.01 **	-.02 *
	(.00)	(.01)

	Original Model	Subsample without mode switches
Number of journals	-.02 *** (.00)	-.02 *** (.00)
χ^2	174.93	133.97
Log likelihood	-700.58	-340.69
Pseudo-R^2	.14	.17
Journal N	34,377	21,573
Respondent N	887	758

Standard errors in parentheses.
† $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ (two-tailed tests)

mode switch are excluded, the originally significant relationships, although similar to the originally estimated effects, change in ways that would alter substantive conclusions. For example, comparing row one across the two models, the size of the association with desire to become pregnant drops by more than 20% and is no longer statistically significantly different from zero association. Had the model been estimated on these truncated data, estimates would not have provided any empirical support for substantive conclusions that contrasting attitudes may simultaneously shape behavioral choices in opposing directions.

Some of these differences are due to sampling error. The number of journals included in the original model was 34,377. After excluding journals that were completed after the first mode switch, there were 21,573 completed journals. The other explanation for the changed estimates is the changing composition of the response. For example, we see a change in the estimate of the coefficient for cohabiting, which is now significant after deleting journals collected after the first mode switch. Further, some of the baseline characteristics related to pregnancy that were only marginally significant in the original model are now significant in the model on the subset of journals collected before the first mode switch. These include receiving public assistance and the number of previous pregnancies.

For this published model, the data collection strategy allowing respondents to switch modes at multiple points in the data collection process prevent attrition among enough respondents to make a difference in substantive conclusions. Although some of these differences are related to a reduction in sample size, which would likely occur under a single mode strategy, others are due to the composition of who responds when mode switching is available to avoid attrition.

5 Discussion

We know from previous research that attrition from panel studies can be caused by important life events, such as changes in employment status, relationships, or moving (Lepkowski & Couper, 2002; Trappmann, et al., 2015). When these events are the topic of the study, this attrition can lead to significant attrition related bias. The potential extent of such bias in studies featuring frequent measurement to document rapidly changing attitudes and behaviors (such as RDSL) is not well understood.

We found that in a panel survey that collects data weekly, allowing panel members to switch modes was an important approach for reducing attrition bias. The characteristics of persons who ever switched modes are different – including not only demographic characteristics, but also baseline characteristics related to pregnancy and time-varying characteristics that were collected after the baseline interview. This was true even for multivariate models that control for many of these dimensions. The fact that the data from the journal predicts whether or not a mode switch was made is a strong indication that estimates that are based on a procedures that do not allow respondents to switch modes would be characterized by attrition bias.

Of course all studies have limitation, including the one we report here. This study focused on women only and focused on women in a narrow age range. Although the results cannot be extrapolated to men or those at older ages, it is quite likely that many of the same issues apply. The longitudinal study described here featured weekly measurement – longitudinal studies with less frequent interviewing may not be able to use mode switching to reduce attrition as effectively. The study reported here also focused on relationships, sex, contraception, and pregnancy – again it is possible that studies of other topics show fewer potential effects of attrition from failure to allow mode switches. Nevertheless, it is quite likely the same issues described here face longitudinal studies of most topics. From the results presented above, we conclude that not allowing users to switch modes in studies with frequent measurement of attitudes or behaviors increases the risk of attrition bias in estimates.

Our research suggests that it may be possible to profile panel members using data from the baseline interview in order to identify cases for whom mode switching may be an effective tool for combating attrition. Lugtig, for example, uses a factor analysis to define profiles of classes of attriters (2014). Armed with early predictions of which cases may fit the profile of "mode-switchers," survey designers may deploy an "adaptive" design (Wagner, 2008; Schouten & Calinescu, 2011) that tailors the survey design to the characteristics of the sampled unit. In this case, the goal of this design would be to prevent attrition bias.

Web surveys are particularly appealing to researchers studying dynamic behaviors that require detailed, timing-specific measures collected over a long

period of time (Axinn et al., 2015; Stone et al., 2007; Mehl & Conner, 2013). Besides the cost savings, the advantages of web surveys for these studies include portability, flexibility, and confidentiality – web surveys allow respondents to complete surveys at whatever time and location is convenient and private for them. These properties extend to multiple devices including personal computers, laptops, tablets, and smartphones, further providing respondents with more options for convenience with little difference in measurement error between the devices (Lugtig & Toepoel, 2015). Even though web surveys are known to have lower response rates compared to almost any other survey mode (Lozar Manfreda et al., 2008; Shih & Fan, 2008), in this paper we demonstrate the potential for a mixed-device approach to compensate for this weakness and strengthen the web survey approach for frequent, repeated measurement. The approach we advocate allows panel members to switch modes – integrating telephone interviewing into a longitudinal web survey – to reduce the potential for attrition bias to produce misleading measures of dynamic behaviors. Overall, the mixed-device approach brings respondents into the study who are significantly different, making conclusions from the mixed-device panel study more robust.

Previously published methodological results from the special RDSL mixed-mode panel are complementary to the results we present here, all indicating this important tool has many advantages. Other investigations of the method not only provide more detailed descriptions of the study (Barber et al. 2011), but also demonstrate that frequent interviewing does not bias measures (Axinn et al. 2015; Barber, Gatny, Kusunoki, & Schulz, Forthcoming), that the web-phone mix has the potential for integrated biomarker collection (Gatny, Couper, & Axinn, 2013), and that the use of electronic debit cards to pay respondent incentives can greatly enhance the feasibility of this approach (Gatny et al. 2009). Overall this body of methodological research demonstrates many advantages of the mixed-mode, mixed device RDSL approach to frequent repeated survey measurement.

Acknowledgments

This research was supported by two grants from the National Institute of Child Health and Human Development (R01 HD050329, R01 HD050329-S1, PI Barber), a grant from the National Institute on Drug Abuse (R21 DA024186, PI Axinn), and a population center grant from the National Institute of Child Health and Human Development to the University of Michigan's Population Studies Center (R24 HD041028). The authors gratefully acknowledge the Survey Research Operations (SRO) unit at the Survey Research Center of the Institute for Social Research for their help with the data collection, particularly Vivienne Outlaw, Sharon Parker, and Meg Stephenson. The authors also gratefully acknowledge the intellectual contributions of the other members of the original RDSL project team, Jennifer Barber,

Yasamin Kusunoki, Mick Couper, and Steven Heeringa, as well as the Advisory Committee for the project: Larry Bumpass, Elizabeth Cooksey, Kathie Harris, and Linda Waite.

References

American Association for Public Opinion Research (AAPOR). (2011). Standard definitions: Final dispositions of case codes and outcome rates for surveys. http://aapor.org/Content/NavigationMenu/AboutAAPOR/StandardsampEthics/StandardDefinitions/StandardDefinitions2011.pdf.

Axinn, W.G., Jennings, E.A., & Couper, M.P. (2015). Response of sensitive behaviors to frequent measurement. *Social Science Research*, 49, 1-15.

Barber, J.S., Gatny, H.H, Kusunoki, Y., & Schulz, P. (Forthcoming). Effects of intensive longitudinal data collection on pregnancy and contraceptive use. *International Journal of Social Research Methodology.*

Barber, J.S., Kusunoki, Y., & Gatny, H.H. (2011). Design and implementation of an online weekly survey to study unintended pregnancies. *Vienna Yearbook of Population Research*, 9, 327-334.

Couper, M.P. (2008). Designing effective web surveys. New York: Cambridge University Press.

Couper, M.P., & Ofstedal, M.B. (2009). Keeping in contact with mobile sample members. In P. Lynn (Ed.), *Methodology of Longitudinal Surveys*. Chichester, UK: John Wiley & Sons, Ltd. doi: 10.1002/9780470743874.ch11

Gatny, H.H., Couper, M.P., & Axinn, W.G. (2013). New strategies for biosample collection in population-based social research. *Social Science Research,* 42, 1402-1409.

Gatny, H.H., Couper, M.P., Axinn, W.G., & Barber, J.S. (2009). Using debit cards for incentive payments: Experiences of a weekly survey study. *Survey Practice*, November 2009.

Graham, J.W. & Donaldson, S.I. (1993). Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology,* 78(1), 119-128.

Groves, R.M., & Couper, M.P. (1998). Nonresponse in household interview surveys. New York: John Wiley & Sons.

Groves, R.M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly,* 64(3), 299-308.

Hoogendoorn, A.W., Lamers, F., Penninx, B.W., & Smit, J.H. (2013). Does a stepped approach using mixed-mode data collection reduce attrition problems in a longitudinal mental health study? *Longitudinal and Life Course Studies,* 4(3), 242-257.

Jackle, A., Lynn, P., & Burton, J. (2015). Going online with a face-to-face household panel: Effects of a mixed mode design on item and unit non-response. *Survey Research Methods,* 9(1), 57-70.

Kristman, V.L., Manno, M., & Côté, P. (2005). Methods to account for attrition in longitudinal data: Do they work? A simulation study. *European Journal of Epidemiology,* 20(8), 657-662.

Lepkowski, J.M., & Couper, M.P. (2002). Nonresponse in the second wave of longitudinal household surveys. *Survey Nonresponse*, 259-272.

Lozar Manfreda, K.L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, 50(1), 79-104.

Lugtig, P. (2014). Panel attrition separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research*, 43(4), 699-723.

Lugtig, P., & Toepoel, V. (2015). The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error. *Social Science Computer Review*, 0894439315574248. http://doi.org/10.1177/0894439315574248

Lynn, P. (2012). Mode-switch protocols: How a seemingly small design difference can affect attrition rates and attrition bias. ISER Working Paper Series, No. 2012-28.

Mehl, M.R., & Conner, T.S. (2013). *Handbook of research methods for studying daily life*. New York: Guilford Press.

Miller, W.B., Barber, J.S., & Gatny, H.H. (2012). The effects of ambivalent fertility desires on pregnancy risk in young women in the USA. *Population Studies*, 67(1), 25-38. doi:10.1080/00324728.2012.738823

Ribisl, K.M., Walton, M.A., Mowbray, C.T., Luke, D.A., Davidson, W.S., & Bootsmiller, B.J. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning,* 19(1), 1-25.

Schoeni, R.F., Stafford, F., McGonagle, K.A., & Andreski, P. (2013). Response rates in national panel surveys. *The ANNALS of the American Academy of Political and Social Science,* 645(1), 60-87.

Schouten, B., Calinescu, M., & Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39(1), 29-58.

Shih, T.H. & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field methods,* 20(3), 249-271.

Stone, A.A., Shiffman, S., Atienza, A.A., & Nebeling, L. (2007). The science of real-time data capture: Self-reports in health research. New York: Oxford University Press.

Trappmann, M., Gramlich, T., & Mosthaf, A. (2015). The effect of events between waves on panel attrition. *Survey Research Methods,* 9(1), 31-43.

Wagner, J. R. (2008). Adaptive survey design to reduce nonresponse bias. Ann Arbor: ProQuest.

Wagner, J., Arrieta, J., Guyer, H., & Ofstedal, M.B. (2014). Does sequence matter in multimode surveys: Results from an experiment. *Field Methods,* 26(2), 141-155.

methods, data, analyses | Vol. 9(2), 2015, pp. 185-212 DOI: 10.12758/mda.2015.011

Evaluation of an Adapted Design in a Multi-device Online Panel: A DemoSCOPE Case Study

Birgit Arn, Stefan Klug & Janusz Kołodziejski
DemoSCOPE

Abstract

In this paper, we look at the challenge of optimizing survey layout in online research to enable multi-device use. Several studies provide useful advice on target-oriented implementation of web design for CAWI surveys. This paper presents results of the implementation of a new adapted design at the panel of DemoSCOPE that allows the participants to take part in a survey on multiple (especially mobile) devices. To evaluate this adapted design, we compare interview data and question timing of panellists who participated in an insurance study before and after the design transition. Central key figures concerning the completion rate, item non-response, open questions, straightlining, timing of single questions and the length of the total interview are presented. In addition, we have presented examples of both old and new design to the community and invited them to assess these examples concerning orientation, color, design and usability. We evaluate the differences in these assessments before and after the design transition for smartphone and desktop users. We end with suggestions for best practice for online studies on different devices.

Keywords: CAWI surveys, design guidelines, simplicity, tile design

1 Introduction

The visual design of CAWI surveys has become a pivotal topic within the area of market research. With the internet as the main form of communication and the extensive dissemination of mobile devices such as smartphones and tablets computers, market researchers need to adapt more than ever (Revilla et al., 2014; Brujine & Wijnant, 2014). The current technological and cultural conditions suggest a trend towards self-administration (Stern et al., 2014). If the use of self-administered surveys increases, so will the importance of a convenient and convincing visual design of those allowing for a multi-device mobile use.

It has long been recognized that because of the absence of an interviewer in self-administered surveys, respondents search for guidance within the questionnaire itself (Schwarz et al., 1991; Schwarz, 1995). Therefore, design elements such as symbols and graphical elements (spacing, font size, location, color and so forth) are crucial in guiding respondents through a questionnaire the way we want them to. During the 1990s the industry's focus was on a question's wording and how that affects the response process (e.g. Tourangeau et al., 2000). However, several studies had already indicated that visual changes of a survey questionnaire produce different outcomes. The importance of design features for the resulting data quality has been documented long since (e.g. Wright & Barnard, 1975; Wright & Barnard, 1978; Rothwell, 1985; Sanchez, 1992; Jenkins & Dillman, 1997).

With the wide distribution of the Internet during the 2000s and the subsequent proliferation of online research, the visual design of self-administered surveys and its consequences on different stages of a survey process has led to further studies on this matter. These studies support the notion that different design elements affect how people answer questions in self-administered surveys. There is much evidence that certain design choices, such as layout of a question (Christian & Dillman, 2004; Christian et al., 2007) or question-order effects (Krosnick & Alwin, 1987; Couper et al., 2001) are as important as the wording of a question. Furthermore, there is a great variety of issue-specific studies on survey design. For example, Dillman et al. (1993) tested the correlation between response rates and questionnaire design and found that shortening the questionnaire and utilizing a user-friendly design improved response rates of the U.S. decennial census (for an overview on response rates and questionnaire design see Vicente & Reis, 2010). A major part of these studies analysed different effects certain design choices had on surveys. For example: the placement, spacing, and sequence of answer options (Tourangeau et al., 2004), the use of images (Couper et al., 2004; Couper et al., 2007; Deutskens et al., 2004; Shropshire et al., 2009) or the question layout (Dillman & Christian, 2002;

Direct correspondence to
Birgit Arn, DemoSCOPE, Klusenstrasse 17, 6043 Adligenswil, Schweiz
E-mail: birgit.arn@demoscope.ch

Christian & Dillman, 2004) There are many more specific topics being looked at, like grid questions and web surveys (e.g. Couper et al., 2013), questionnaire design and nonresponse bias (Vicente & Reis, 2010) or invitation design (e.g. Whitcomb & Porter, 2004; Kaplowitz et al., 2012). Another practiced approach is to evaluate whether design effects differ with respondents' socio-demographic characteristics (e.g., Krosnick & Alwin, 1987; Knäuper et al., 2004; Fuchs, 2005; Stern et al., 2007; Tourangeau et al., 2007). The past years would suggest that the visual design of web-based surveys is as influential to a respondent's answers as any documented interviewer or wording effect (cf. Stern et al., 2014, p. 294).

Thus, the importance of a good web survey design seems to be evident. But, what is a good web survey design? In a fast-paced multi-device environment and changing user habits, surveyors need to be up to speed and recognize the importance of a state-of-the-art survey design. Besides the question of a good design, there are also technological constraints and nuances to take into account. In the next section, we discuss some specific, more technical challenges when it comes to using mobile devices.

In Section 3 we outline our attempt to offer an optimized web design to our online community. DemoSCOPE (www.demoscope.ch) is the third-largest market research company of Switzerland. To fulfil the high standard of the requirements of our clients we have built up a large online panel that we call the DemoSCOPE community. This community consists of about 30,000 active panellists which come from very diverse socio-demographical strata. The panellists are asked about twice a month to take part in an online survey. To keep the community members at it, we want to offer an optimal web design and the possibility to communicate with each other and directly with the community support at DemoSCOPE. To fulfil these aims we formulated the design guidelines which are presented in Section 3. Note that already 41% of our community participate in the surveys using a mobile device (27% smartphone, 14% tablet). Hence, our specific attention is on users of mobile devices.

In Section 4 we propose two ways to evaluate the adapted online design. First, we propose methods to compare the response behaviour of panellists which participated in an insurance study before and after the adaption of the new design based on the design guidelines. As a second idea we invited the community members to take part in a design evaluation, where we showed examples of the old and the adapted design. The task of the participants was to evaluate the shown screen using 4 different criteria: orientation, color, design and usability. In Section 5 and 6 we present the results of this evaluation. Section 7 contains our conclusions.

2 Specific Challenges when Using Mobile Devices

It is expected that in the near future internet traffic among mobile devices will exceed that of desktop computers (Buskirk & Andrus, 2012a). Smartphones represent a convenient tool for survey data collection, as they are a multimode device accessible through voice, text or web, including synchronous multimedia messaging (SMS) and an ever-increasing variety of apps. Not to mention the possibility to take a survey on the spot. However, the very same opportunities smartphones give also imply great variability with their different devices, operating systems and browser capabilities. As a result, the complication level for the implementation of online surveys for mobile versus desktop computers increases (Buskirk and Andrus 2012b). As the spread of smartphone usage is a relatively recent phenomenon, there is still only little literature on mobile surveys using smartphones and other devices (e.g. Raento et al., 2009; Fuchs & Busse, 2009; Buskirk & Andrus, 2012a; Buskirk & Andrus, 2012b; Mavletova, 2013; de Bruijne & Wijnant, 2013; Wells et al., 2014; Buskirk & Andrus, 2014).

With respect to online questionnaires, researchers nowadays must anticipate the diversity among the end user's device. Designing questionnaires for usage across such a variety of devices is not a matter of can-do attitude but rather already a must-do, as the end-user is also the one deciding on which device an online survey will be taken on.

Obviously, the main constraint is the screen size of the respective device used for survey participation. Screen sizes range from 14-40 inches for computers and laptops, 6-13 inches for tablets and 4-6 inches for smartphones, with the boundary values beginning to overlap across these categories. A web survey should therefore keep its functionality and desired look from the smallest smartphones to the widest TV-like PC screens. Web designers solved the multi-screen problem by following the rules of two main schools, namely adaptive web design (AWD) (Gustafson, 2012) as well as responsive web design (RWD) (Marcotte, 2010). With AWD a server sends the same data packages to each device and the browser of the corresponding device decides which of the upfront designed layouts to choose. Unlike the predefined device specific layouts AWD relies on, RWD uses fluid layout grids, flexible images and media queries to treat every viewport (device) the same way and adapts the layout according to the device's features. Without going into details, in both scenarios the layout of a webpage or, like in our case, an online survey is adapted to the screen used. The main difference is in how this adaption takes place – if it's using predefined solutions to exactly corresponding devices (AWD) or if it responds to any device thanks to a more fluid (flexible) way of defining one layout only (RWD). Despite the promise of an easy sounding solution, as the designer of our online survey we face a multitude of challenges when it comes to putting theory into practice. First, do we want to design a web survey device-specifically or do

we want to design a survey that adapts automatically to every viewport (device)? Furthermore, if we opt for the RWD solution, we still need to consider most of the imperatives on web questionnaires in general, irrespective of the nuances mobile research poses.

3 The Design Guidelines

In the beginning of 2014, DemoSCOPE changed its web questionnaire layout. On the one hand, the aim was to provide respondents with an enjoyable, convenient and mobile optimized design; on the other hand, it was as important to ensure functionality, feasibility and good data quality. Before we have a look at the new survey design and its properties, we firstly present the DemoSCOPE design guidelines.

We consciously decided to use an RWD approach where you develop one questionnaire design that then automatically adapts to the different devices and their parameters. In order to provide a mobile-optimized survey design, different constraints regarding the relatively small screen of smartphones had to be considered.

Firstly, to enable a reasonable legibility for smartphone users, we turned away from using a fixed font size. We changed the pixel-based size definition, which means from an absolute and rigid unit of measurement, to "em" – a relative unit equal to the currently specified point size (in any device or browser). The name used to refer to the width of the capital "M" in the typeface and size being used (the same as the point size). This enables to choose a reasonable ratio where the font size adapts to the actual screen size in use.

Secondly, given that only vertical scrolling is acceptable for smartphones, the use of grid questions should be avoided. There is no technical reason for the preference of vertical scrolling over horizontal scrolling, but it has emerged as the preferred usage and almost all apps and mobile friendly web pages are designed for vertical scrolling. Additional to the omission of horizontal scrolling, we decided to use a one-screen-per-page approach, where normally only one question per screen is displayed. This assures that respondents experience a stable and convenient survey flow. Apart from the no-scrolling advantage of a one-screen-per-page-approach, Couper et al. (2001) and Tourangeau et al. (2004) found that the intercorrelations between items presented on the same page are higher than when items are displayed sequentially on one screen per page. These authors also state that, although the effect as such does not seem to be severe, there is evidence that respondents use proximity among the items as a hint to their meaning, which results in a faster advancement within the survey. However, Couper et al. (2001) found that the one-item-per-screen-approach takes respondents more time to complete the survey than a multiple-item-per-screen approach.

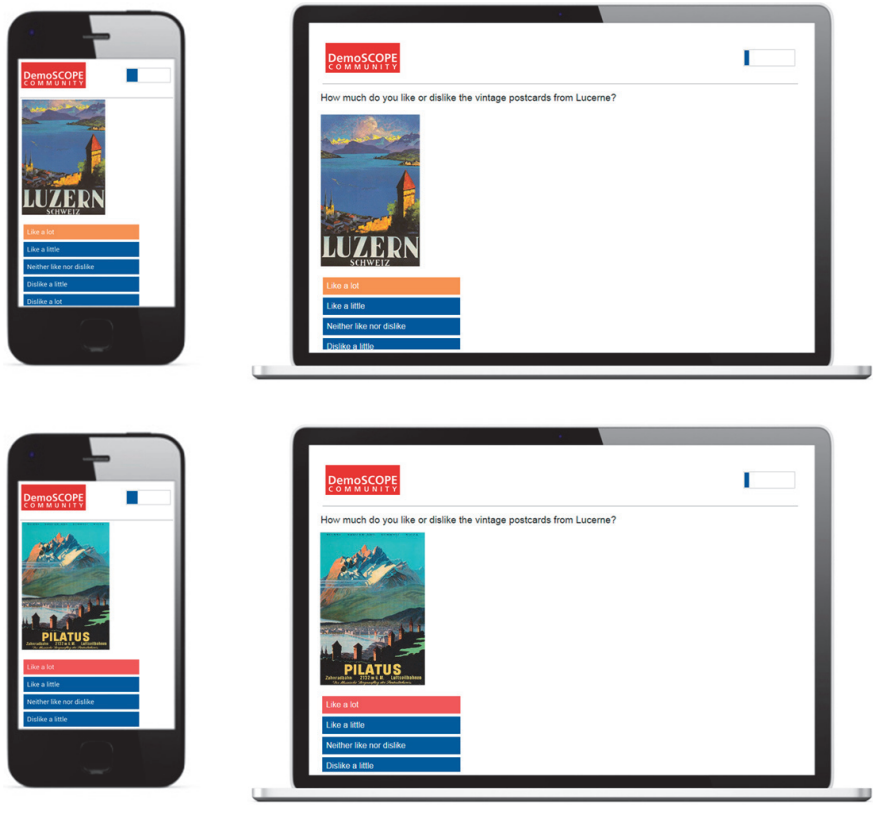


Figure 1 The HSM format for 2 examples: Vintage postcards of Luzern and Pilatus (smartphone and desktop version)

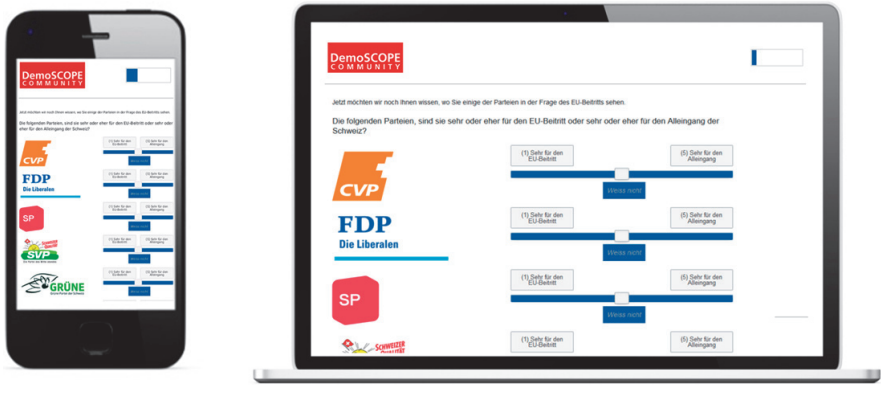


Figure 2 The visual scale sliders for an example with Swiss parties (smartphone and desktop version)

The usage of grid-questions imposes a problem not only to mobile-devices. It is a remainder of research with paper-and-pencil questionnaires, where print-outs were costly. Visually grid-questions make the questionnaire appear shorter, but have the disadvantage of non-careful reading and other negative effects such as straight-lining (Schaeffer & Presser, 2003). Klausch et al. (2012) tested a format where the answer-scale stays on one screen and the question is replaced by horizontally replacing one question with the next one (HSM: horizontal scrolling matrix format, not to be confused with "horizontal scrolling" by the respondent). These authors proof positive effects on data quality when using HSM formatted questions instead of grid-questions. As an example see Figure 1: The response scale stays the same for both examples ("Luzern" and "Pilatus"), but the shown vintage postcard is different. Sometimes a visual comparison between answers given is desired. For such cases, alternatively to the HSM format, we propose visual scale sliders that reduce the scaling-dimension such that it can be displayed on one screen together with the line of statements (see an Example in Figure 2).

Obeying the one-item-per-screen with limited scrolling policy, we introduced an auto-submit function for single-choice items. This enables the respondent to proceed to the next question as soon as he or she selects the answer. However, the use of the auto-submit function carries certain risks, especially when applied on small screens, since some respondents may not notice that they have already progressed to the next question and mix up answers.

Further, we quit using Flash-based elements, complex headers, and website-like tabular depictions.

Altogether, these rules and features form the rules of simplicity which will be the basis for our design guidelines described further down:

- Simple design with as few visual distractions as possible
- One-item-per-screen
- No horizontal scrolling
- No Adobe Flash

The rules of simplicity should enable a quick orientation and easy navigation in an online survey irrespective of the device used.

The following paragraphs conclude the core of what we call the 7 Demo-SCOPE design guidelines:

1. The signature feature of our new survey design is tile-like buttons (tiles), which superseded the allegedly immortal radio-buttons. Over the past decade, tiles have emerged more and more in software of various companies all over the world. Just think of the tiles for apps on iPhones and smartphones based on Android OS. Furthermore, Microsoft has changed its layout to tiles in the latest versions of the Windows software. The tiles we use in our online surveys offer a large area to click on, which is particularly important for small mobile screens. The tile design is

the central, most crucial improvement when comparing the new to the previous survey design. The flat tile design is combined with a modest and steady color concept, which is based on the DemoSCOPE colors red and blue. See Figure 1 for an example; note that the screen is shown for smartphone and desktop users in order to demonstrate the usability and appearance of the tile design on different devices. Also for the following Figures 3-5, the images are shown in both smartphone and desktop modes.

2. Response scales are even, aligned and logical. We follow the considerations of Tourangeau, et al. (2004) that the leftmost or the top item in a scale is seen as the "first", meaning it is expected to represent an endpoint (e.g. "Like a lot"). Further, the listed options are expected to follow some logical order where the final answer option represents the opposite endpoint (e.g. "Dislike a lot"). It was noted by Christian and Dillman (2004) that respondents would answer more quickly and accurately with the scales visually and conceptually kept in logical order.

3. "Don't know" (DK) answer options are visually separated from the substantive answer options, as there is evidence that respondents are misled about the midpoint of a scale when there is no visual distinction. Survey takers tend to be guided by the visual rather than conceptual midpoint of a scale (Tourangeau, Couper, and Conrad, 2004). In our example in Figure 9 this is achieved by a different typography of the "Don't know" text.

4. We are confident that giving the respondent the ability to track his progress within a survey is an absolute must. In that respect online market research is not any different from any web-based endeavour, where it is simply expected to be transparent about any processes people are engaging in while they stay connected. For that reason we use a rather prominent progress bar in the top right of every screen shown. In literature, this issue still causes controversies. Couper et al. (2001) argue that the presence of a progress bar increases the motivation for completing a survey as you get less frustrated by long surveys. However, they also found no significant evidence for this hypothesis. Furthermore, Conrad et al. (2010) find that a progress bar increases the respondents' overall satisfaction with the survey. However, in Villar et al. (2013) a meta-analysis is conducted and the authors find that a permanent progress bar does not actually decrease the drop-off rate. Leaving the discussion aside, we think that it is the researcher's responsibility to offer transparency also on this front. An example of the progress bar is shown in Figure 3.

5. To ensure an engaging and brisk survey experience, we use pictograms for answer options as visual relief from the mere completion of a survey. Figure 4 shows the pictograms that can be used to obtain the most favourite activity for a day in Lucerne.

6. We intentionally deny the use of Flash for any animated or otherwise dynamic questions. The reason for it is that it can be no longer be assumed that Flash is

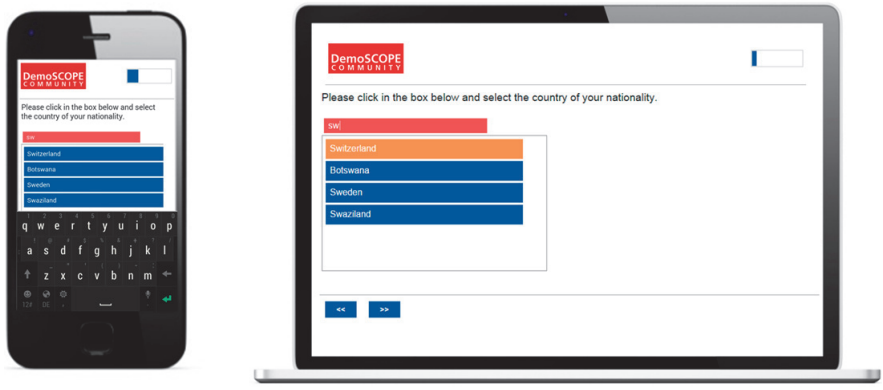


Figure 3 A text search single-choice list with progress bar in page header (smartphone and desktop version)

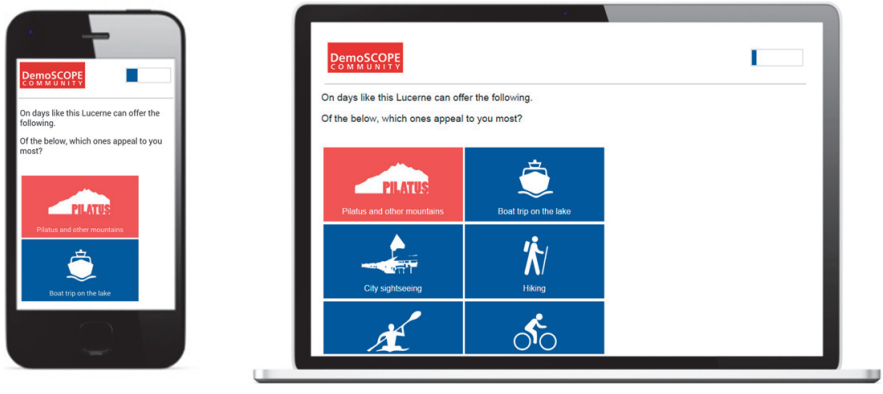


Figure 4 The use of pictograms (smartphone and desktop version)

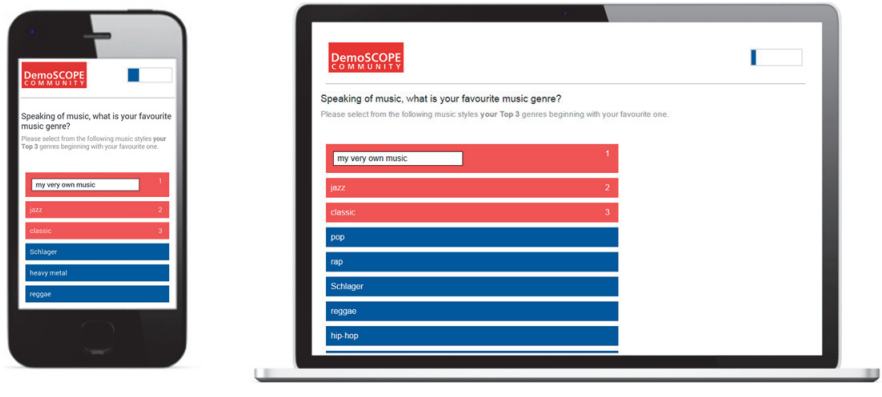


Figure 5 An interactive, yet Flash-free ranking question with built-in text fields (smartphone and desktop version)

Figure 6 Options to connect and share and to contact the support team

installed on people's devices. Hence, we decided to introduce a zero tolerance policy for any Flash-animated elements in our questionnaires. See Figure 5 for a Flash-free ranking question which has animated elements but doesn't require Flash. Hence, we offer interactive questions without the necessity of Flash.

7. Furthermore, we provide our respondents possibilities of linking themselves to different social networks as well as contacting our support staff directly via a prominent support button at any stage of the questionnaire. See Figure 6 for those links.

In summary, these considerations result in the following mainly visual design guidelines:

1. Consistent flat tile design
2. Even, aligned and logical response scales
3. Visually separated "Don't know" and "No answer" options
4. Transparent progress bar
5. Pictograms as answer options or visual relief
6. No use of Flash
7. Direct opportunity at any stage to contact support team

These design guidelines were implemented at DemoSCOPE in spring 2014. Since then, almost all online studies are implemented based on the design guidelines.

4 Methods to Evaluate the Adapted Design

In the following sections we propose two ways to evaluate the design transition. Section 4.1 deals with a comparison of interview data and question timing for the old and the new design and shows differences in respondent behaviour. In Section 4.2 we present the results of a feedback study among community members concerning the old and the new design.

4.1 Analysing Interview Data and Question Timing of Panellists who Participated in an Insurance Study Using the Old and the Adapted Design

Basically, we can use two sources of data. There is the interview data itself, which can give us answers concerning a modified respondent behaviour related to the actual questions in the questionnaire. The second source of conclusions is a question timing file which contains the time needed by the respondent for each screen. Both sources can be used to check if there are any differences in the respondents' behaviour related to the adapted web design.

The first hypothesis concerning the adapted design with its characteristic tiles is that it fits more into the present state-of-the-art environment of software in use for mobile devices. The distraction of the user by an uncommon or complicated design is minimized and it is easier to keep the interest of the respondent in the actual topic of the study high. Thus, we hypothesize that the completion rate for the new design is higher than for the old design. The completion rate is defined as the rate of respondents starting the survey that fully complete all questions. I.e. the completion rate is a quantitative measure for the persistent interest in the study.

A related idea is to measure item non-response for questions which are not obligatory in order to see if the new design stimulates the respondents more to answer also difficult questions properly. Here, we consider especially the interest in pre-formulated multiple and single choice questions with given answer possibilities.

A further topic is open questions. Open questions can be very tiring for the respondent as they have to come up with own proposals or answers. The question is how the respondent can be motivated to give answers to open questions and not to skip them or even leave the study, as the question is conceived as too hard or too long. We propose a tailor-made idea to guide the respondent through an open question by introducing kinds of "motivating" elements.

Another idea is to estimate design effects related to the step from grids to the one-item-per-screen approach: Consider a grid where the single questions or statements are ordered from the top to the bottom and the answering scale is given from the left to the right. In the adapted design we have designed a one-item-per-screen approach where each question is on a single page. Our hypothesis is that the respondents tend to give the same answers when the questions are shown in a grid, as they just go from the top to the bottom clicking on the same radio button. With the one-item-per-screen approach the respondent might be animated to think of a new answer for each statement and less so-called straightlining can be found. In Lugtig and Toepoel (2015) it is discussed that straightlining can be seen as a measure of measurement error and, therefore, it is an issue to think of strategies to reduce this effect.

Conclusions concerning the respondents' behaviour can also be drawn from the question timing file. First, we can look at the total time used for the questionnaire. An idea might be that because of the clearer and easier structure of the questionnaires, the respondent is able to answer the questionnaire in the new design in less time. However, it is worth to examine the issue more detailed: In the questionnaire we have general elements that – as we claim above – clarify the structure of the questionnaire. For example, we use pictograms wherever possible, which might reduce the interview timing. Another issue is the autosubmit-function that is used whenever there is a single choice question. However, the one-item-per-screen approach may induce that the total time for a former grid increases, because several screens are shown.

We try to find empirical evidence for all these hypotheses by analysing key figures (e.g. medians, means, proportions) for the old and the new design. An integral property to guarantee the comparability of an old and an updated version of the questionnaire is that the number and order of questions in the questionnaire have not changed over several or at least two waves of the study. Furthermore, there should be no changes in the sampling process for the potential respondents.

The example chosen here is a multi-client study in the insurance market. This study is conducted quarterly with about 1,250 complete interviews. The topic is the popularity of specific insurances in Switzerland. Furthermore, questions about the use and attitude towards insurances in general are asked.

The available data we have are 9 quarterly waves in total. Four of these waves were presented completely in the old design (2013-1 to 2013-4) and four of these waves were presented completely in the new design (2014-2 to 2015-1). Wave 2014-1 cannot be used for our comparison purposes as some elements of the new design were implemented and some weren't.

As the DemoSCOPE community is a panel, we use respondents that answered the questionnaire using the old and the adapted design. Hence, we can assess key figures for the old and the new design by paring the respective interview and timing files for the same e-mail addresses. We assume that an identical e-mail address means that the questionnaire was filled by the same person.

To obtain an appropriate dataset, we first joined the interview and the timing data for each interview in the waves 2013-1 to 2015-1 (complete and incomplete interviews, excluding 2014-1). Then we merged the datasets for the old and the new design, respectively. However, it is possible that the same person (identified by the e-mail address) answered the questionnaire for the old or the new design more than once. For these cases we reduce the multiple entries to a single entry. This is done by sorting the datasets by a completion indicator and by wave. For multiple entries, we decided to choose the latest, complete interview. After obtaining datasets with single entries we have 7,666 interviews for the old design and 6,370 interviews for

the new design. In the next step, both dataset are joined by the e-mail address. By doing this, we obtained a dataset with 2,032 matching pairs of interviews.

For the analysis of the completion rate we need the complete and incomplete interviews. For the rest of the analyses we need the complete interviews only. Hence, in a second step we chose e-mail addresses with complete interviews in both designs. This results in 1,188 email addresses with paired interview and timing data for the old and the new design.

In Section 5 we will first show a descriptive analysis of several key figures. For statistical analysis of proportions, means and medians we use significance tests for paired samples.

4.2 Analysing Feedback from the DemoSCOPE Community

Another idea was to involve the community and to obtain their opinion about the adapted and the old design. A design test was implemented, where 5 screens from the old and the respective 5 updated screens from the new design were shown in rotated order. For each screen the community members had to assess the following 4 statements on a scale from 1 to 10:

1. The design enables a quick and easy orientation in the questionnaire. (Orientation)
2. I like the color composition of the questionnaire. (Color)
3. I like the design of the questionnaire in general. (Design)
4. The design of the questionnaire is user-friendly. (Usability)

Smartphone screens were shown to the smartphone users. Desktop screens were shown to the laptop and PC users.

In total, 4 * (5+5) assessments had to be made. This results in 20 pairs of scores that can be compared to each other in an analysis. In a further analysis we can sum up the evaluations for the 4 different statements for each design. This results in a total score for the 4 assessed topics for each design. We obtained answers from 112 smartphone and 200 desktop users. Community members from all socio-demographic strata were invited to conduct the study; no filters were set.

Additionally, the community was asked the following question: “Which factors are especially important for you when taking part in an online survey?” The possible choices were:

- Comprehensibility of the questions
- That a quick orientation in the questionnaire is possible
- Appealing visual design
- Interesting topics

- Varied topics
- Feedback on the results of the study, e.g. within a newsletter
- Rewards
- That smartphone or tablet can be used to take part in the study
- That the surveys are short
- That surveys are as much detailed as possible
- General user-friendliness

Each respondent had to select those 3 factors which are most important for them.

5 Results for the Insurance Study

The first proposed key measure is the completion rate. Looking at the completion rates of the 2,032 matching e-mail addresses we find a completion rate of 69.3% for the old design and a completion rate of 78.1% for the new design. A test of proportion for a paired design shows that these two proportions are significantly different on a 95% confidence level (p-value < 0.001). Hence, the completion rate for the new design is significantly different from the completion rate of the new design. We cannot prove that this difference is caused by the new design, but it is a fact that within very short time the completion rate rose by almost 10%.

Another key figure analysing completion behaviour is item non-response. Unfortunately, almost all questions within our insurance study are obligatory (however, most of them offer a "Don't know/No answer" radio button/tile). There are only two questions where we can measure "real" item non-response, i.e. where it is allowed to tick no radio button/tile. The first question analysed is: "Suppose, you want to contract a property insurance (car, furniture). Which insurance would be your first choice?" The second question analysed is: "Suppose, you want to contract a life insurance. Which insurance would be your first choice?". For both items, the percentage of item nonresponse is very low. For the first question, the item nonresponse proportion is 0.9% (old design) and 2.6% (new design), respectively. For the second question, the item nonresponse proportion is 2.0% (old design) and 3.6% (new design), respectively. If we do a statistical test for paired samples (n=1,188 in this and the following paired tests) the proportion of item nonresponse for the old and the new design is found to be significantly different in both cases (p-value < 0.001 and p-value $= 0.026$, respectively). Hence, the item nonresponse is very small, but significantly lower for the old design.

The next topic are open questions. The first "insurance question" after the introductory questions (language, sex, age and post code) is to write down all insurances the respondent remembers spontaneously. In the old design there are nine

boxes offered on the screen where the answers can be written. Our feeling is that many respondents are stressed by the feeling that they have to come up with nine answers that they just decide to quit the study. Looking at the interview data it was found that actually 12.6% of the 2,032 respondents left the study just at this first screen. Hence, an aim of the new design was to reduce the number of incompletes resulting from the layout of this question. The idea was to show only three empty boxes at the beginning and offer an additional empty box when a third, fourth, …, eighth insurance was written down. Using this new strategy, the quitting proportion could be reduced by 1.1% to 11.5%. However, the difference in the quitting proportion is not significantly different on a 95%-level (p-value: 0.227). A second issue is the number of insurances remembered and written down by the respondents. For the old design, the mean number of insurances is 5.51 (median: 5.0) and for the new design the mean number is 5.35 (median: 5.0). A t-test for paired samples shows that the mean number of insurances is significantly different on a 95%-level (p-value: 0.003). Hence, the quitting proportion tends to be lower for the new design. However, the mean number of entries is also lower, as less empty boxes are shown from the beginning.

The next topic considered is the so-called straightlining. The following example is chosen from the insurance study: There are 20 insurances where the respondents have to indicate, if they know the insurance

- well and have personal experience.
- well and have no personal experience.
- know only the name.
- don't know it at all.
- don't know/No answer.

In the old design, the insurances are shown in a grid, see Figure 11 (left); in the new design this is solved by a one-item-per-screen approach with so-called sliding statements (HSM format), see Figure 11 (right). There are several ideas, how straightlining in the grid/sliding statements can be evaluated: If the statements are always shown in the same order, it can be counted how often the same answer is given for two successive statements. However, in the insurance study the insurances were always shown in a different order for the old and the new design. Hence, we measure the variance within the answers for the whole grid/for all sliding statements. The higher the mean/median of the variance, the less straightlining is present. For the old design, the mean of variances is 0.99 (median: 0.99); for the new design, the mean of variances is 0.96 (median: 0.94). A paired t-test for the variances shows that they are significantly different on a 95%-level (p-value: 0.009). Hence, the straightlining cannot be reduced by the new design.

The last issues to be analysed are the timing questions. Let's first look at the introduction of the auto-submit function for single choice questions. The question

is, if the respondent can navigate quicker through the questionnaire, if the auto-submit functionality is used. We check this for a simple single choice question: "How likely is it that you will check alternative offers to your actual property insurance (car, furniture) or will even look for a new offer within the next 12 months". The mean response time for the old design is 22.4 seconds (median: 20.0 seconds) and the mean response time for the adapted design is 24.9 seconds (median: 20.0 seconds). A paired t-test shows no significant difference of the response times on 95%-level (p-value: 0.45). This means that the auto-submit button doesn't decrease the question timing substantially. However, there is still one click less the respondent has to make and this might be more comfortable for the respondent.

Another topic is the timing for questions presenting pictograms. Our examples are the questions about language (German, French) and sex (male, female). The design of the used pictograms is similar to Figure 4. The mean question timing for the old design for the language question is 27.0 seconds (median: 7.0 seconds). The mean and the median are very different in this case. A look at the data vector shows that there are high outliers. The reason might be that the language question is the first question in the questionnaire and people possibly leave it open for a while before they start the survey. Hence, for a comparison in this instance we use the median. For the new design the values are 13.3 (mean) and 5.0 (median). A Wilcoxon-test for medians in paired samples shows that the medians are significantly different (p-value < 0.001). The second pictogram we look at is the one for sex. For the old design, the mean question timing is 10.8 seconds (median: 6.0 seconds); the mean question timing for the new design is 6.0 seconds (median: 5.0 seconds). A t-test for paired samples shows no significant difference between the means (p-value: 0.138). Hence, although the finding is not significant for sex, there is a tendency that the orientation for the respondent is easier when pictograms are used.

The last issue on question timing is the grid for the evaluation of 20 insurances, we already discussed concerning straightlining. We want to know, which influence the one-item-per-screen design has on the question timing. It has to be noted that in this example the auto-submit function is implemented for each screen in the new design as the assessment of the insurances is based on a single choice selection. Hence, although we have a one-item-per-screen approach, only 1 click per page is needed. The mean question timing for the grid in the old design is 90.4 seconds (median: 73.0 seconds); the mean question timing for the new design is 91.7 seconds (median: 81.0 seconds). Again, the mean and the median are rather different which means that there are some high outliers and the question timing might not be normally distributed. Hence, we prefer to test the median rather than the mean. A Wilcoxon test for paired samples shows that the medians are significantly different on a 95% level (p-value < 0.001). Hence, the question timing for the sliding statements (new design) is significantly higher than for the grid (old design).

Table 1 Means and results of paired t-tests for the 4 scores in the old and the new design

Statement	Mean		Difference (New-Old)
	Old Design	New Design	
Orientation	31.3	39.8	8.4*
Color	27.5	36.4	8.8*
Design	28.9	37.6	8.7*
Usability	31.5	39.9	8.4*

Note. * = p< 0.001

As a last issue we look at the interview timing in total. Note however, as discussed above, that the question timing is influenced by the mix of the interview components. While pictograms in the new design need less time, the one-item-per-screen approach requires longer question timing than the old grids. The mean interview timing for the old design is 1528.3 seconds (median: 1092.0 seconds); for the new design it is 1566.0 seconds (median: 1110.0). A Wilcoxon test for paired samples shows that the total median interview length of the old and the new design are not significantly different. So, looking at the questionnaire as a whole there is no reduction in the interview time by the new design.

6 Design Evaluation Questionnaire

In the following figures 7-11 you can see the screens from the old and the new design that had to be evaluated in the design evaluation questionnaire. In a first analysis, we look at the means for each statement and screen. The mean rating for the new design is always higher than the mean rating for the old design. Looking at the differences between the old and the new design (not shown), they are especially high for screen 2 (larger than 2). For the other screens, the differences in the rating are between 1 and 2. Using a t-test for paired values, all mean differences are significant on a 95%-level (p-value < 0.001). In order to aggregate the data a little, the idea was to sum up the ratings for all screens for the 4 topics. The resulting means are shown in Table 1. For the old and the new design the mean of the Color score is lowest. In general, Color and Design are rated lower than Orientation and Usability. The differences between the two designs are between 8 and 9 for all statements (p-value < 0.001). To find out, if the used device plays a role in the rating of the 4 statements, we conducted an ANOVA for repeated measurements with

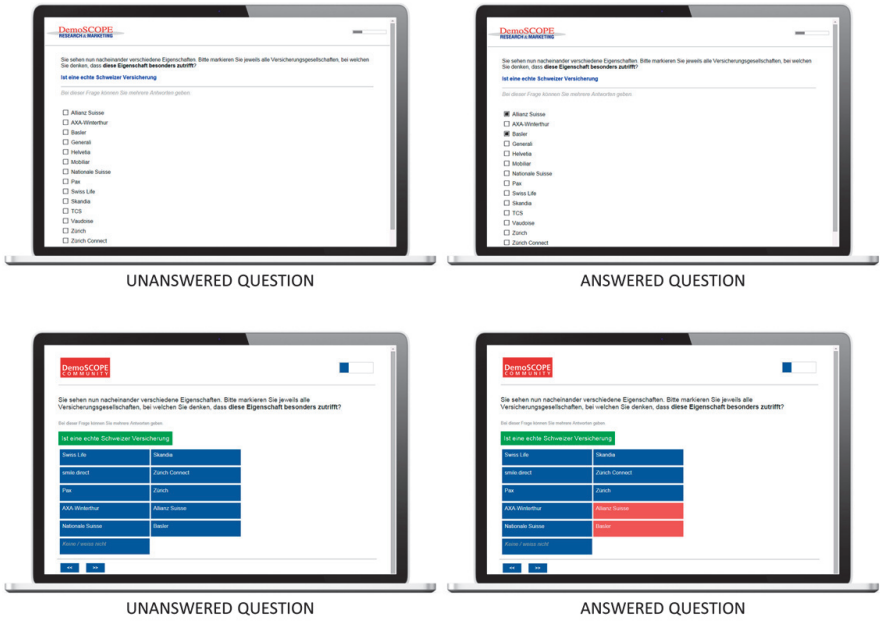


Figure 7 Screen 1 for the design test in old and new design (desktop version)

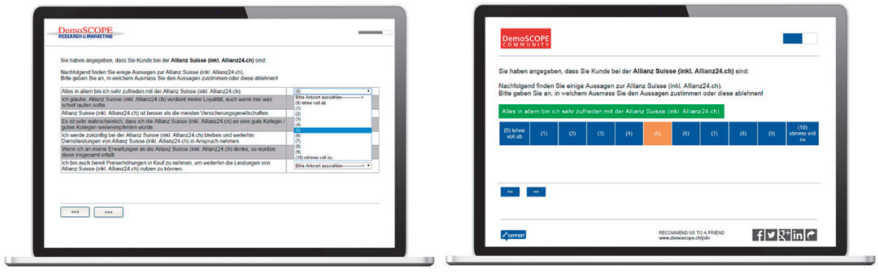


Figure 8 Screen 2 for the design test in old and new design (desktop version)

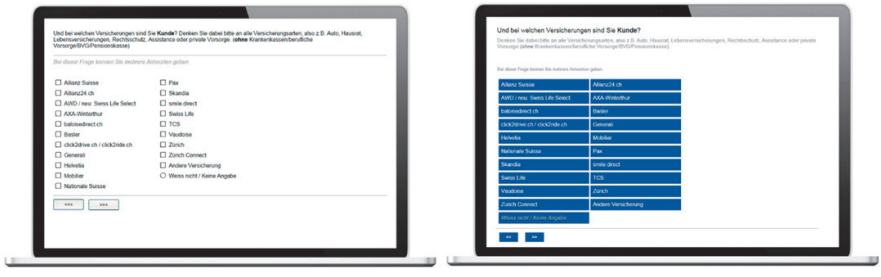


Figure 9 Screen 3 for the design test in old and new design (desktop version)

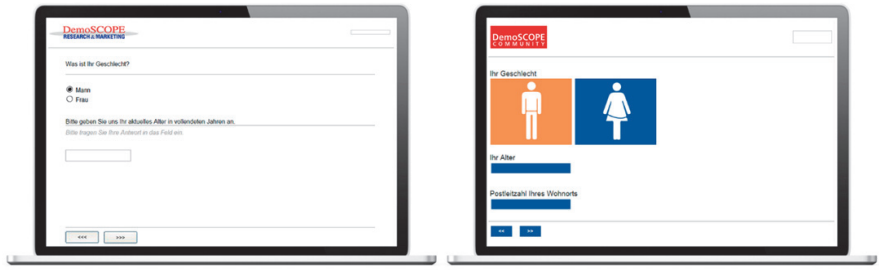


Figure 10 Screen 4 for the design test in old and new design (desktop version)

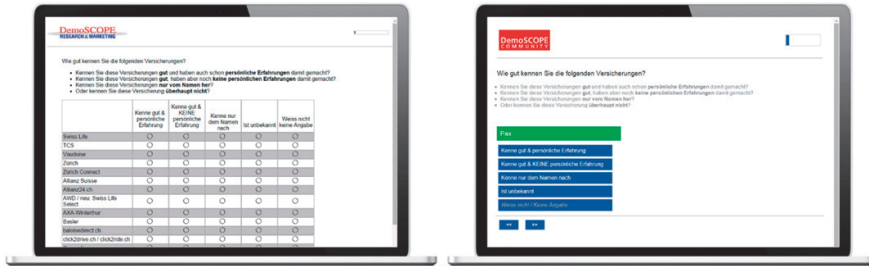


Figure 11 Screen 5 for the design test in old and new design (desktop version)

the used device as an additional factor. For the saturated model for Orientation we obtain a significant design factor ($p < 0.001$). The main effect for device is not significant (p-value 0.06). However, the interaction between design and device is significant (p-value < 0.001). This means that the increase of the rating between the old and the new design is significantly different for smartphone and desktop users. To evaluate this finding a little more detailed, you can see the estimated means for the old and the new design split by device in Figure 12. There is a higher increase in the rating for the smartphone users than for the desktop users. I.e. the benefits of the change from the old to the new design with regard to Orientation are a bit higher for smartphone users. As a result, the Orientation rating for smartphone and desktop users for the new design is almost equal, although the rating of the smartphone users was lower for the old design.

An identical analysis was made for Color. Here the design effect is found significant (p-value < 0.001), but the interaction effect for design * device and the main effect for device are not significant (p-value: 0.44 and 0.28). I.e., when it comes to judging the colors, the new design is rated better than the old one, but this preference is independent of desktop or smartphone usage. The estimated means are also shown in Figure 12.

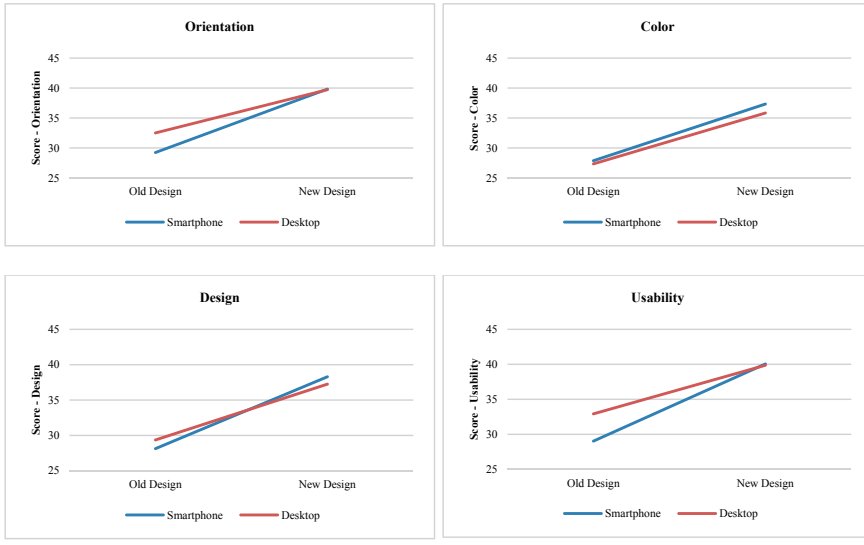


Figure 12 Estimated means for smartphone and desktop users for the old and the new design from the saturated models (two-way ANOVA with repeated measurements) for the 4 statement scores

Regarding the general evaluation of the Design, the new design is rated significantly better than the old design (factor design, p-value < 0.001). The interaction effect is also significant (p-value: 0.05), but the main effect of device is not significant (p-value: 0.92). It can be seen from Figure 12 that the estimated curves for smartphone and desktop users cross each other. Hence, the old Design is liked less by smartphone than desktop users, but the new design is rated better by smartphone users.

For the Usability the design factor is significant (p-value < 0.001). Furthermore, the interaction effect (design * device, p-value < 0.001) and the main effect for device are significant (p-value: 0.03), see Figure 12. This means again, that Usability of the old design is rated worse by smartphone than desktop users, but the new design is rated almost equal by both user groups. There is a very high increase in the Usability rating by the smartphone users. This is the desired effect, because the new design has to be equally well accepted among desktop and smartphone users and it also has to be accepted significantly better than the old design for all user-groups.

In our last analysis we asked the community about their 3 most important components of an online study. In Table 2 you see the proportions of "Yes". It can be seen, that the importance of the different components varies between smartphone and desktop users.

Table 2 Proportions of "Yes" for the importance items for the smartphone and desktop users and results of a significance test

Component		Smartphone	Desktop
1	Comprehensibility of the questions	43.8%	55.5%*
2	Quick orientation in the questionnaire	30.4%	44.5%*
3	Appealing visual design	17.9%	12.5%
4	Interesting topics	41.1%	54.0%*
5	Varied topics	18.8%	21.0%
6	Feedback on the results of the study, e.g. within a newsletter	2.7%	8.0%
7	Rewards	23.2%	15.0%
8	That smartphones and tablets can be used to take part in the study	43.8%*	6.5%
9	That the surveys are short	32.1%	27.5%
10	That surveys are as much detailed as possible	4.5%	10.0%
11	General user-friendliness	36.6%	44.5%

Note. * = $p < 0.001$

However, the Top 1 property is the same for smartphone and desktop users (Comprehensibility of the questions). For smartphone users the further most important components are that smartphone or tablets can be used and interesting topics. For the desktop users the further most important issues are interesting topics, quick orientation in the questionnaire and the general user-friendliness. A significant difference in the absolute proportions can be found for the comprehensibility of the questions, the quick orientation in the questionnaire, interesting topics and that smartphones and tablets can be used to take part in the study, see Table 2.

It would be interesting to see, if the response behaviour for the online survey components can predict, if somebody is a smartphone or desktop user. As an instrument for such an analysis we use a logistic regression model. The response variable is if somebody is a smartphone user or not (0 = desktop, 1 = smartphone). The online survey components act as independent variables. Additionally, we can add sex and age as socio-demographic, explanatory variables. To find the optimal model, we used forward model selection based on the Likelihood Ratio statistic. The resulting model contains 3 significant variables: Two of the survey components and age. You can find a summary of the optimal model in Table 3.

The estimated coefficients for components 3 and 8 are positive, which means that the odds for being a smartphone user increases if one of them is ticked as one of the 3 most important components. The most dominant item is "that smartphones

Table 3 Results for the logistic regression model for end device usage based on the online survey components, sex and age

Component	est. coeff.	std. error	Wald	*P*-value	exp (est. coeff.)
3 Appealing visual design	.849	.363	5.48	.019	2.34
8 That smartphones and tablets can be used to take part in the study	2.343	.359	42.71	<.001	10.41
Age (in years)	-.037	.010	14.53	<.001	.96

and tablets can be used to take part in the study". If this component is ticked among the 3 most important the odds that a person is a smartphone user is increases by a factor of 10.41. If somebody rates an appealing visual design as important, the respective odds increased by a factor of 2.34. Note that the older the person is, the lower is the odds for being a smartphone user. The predictive probability of this model is 76.9%. I.e. 76.9% of the respondents are categorized correctly by the model as smartphone or desktop user.

7 Conclusions

Today, a vast majority of Internet users happen to be smartphone users, too. The estimated figure is around 80%, according to GlobalWebIndex. To ignore this fact or to underestimate its importance would be a huge mistake of researchers which try to retrieve information by online surveys. We tried to face this challenge by creating our own design guidelines based on rules of simplicity and wanted to achieve some empirical evidence to evaluate our approach.

To meet this target, we used two approaches: The first approach is to evaluate paired data from members of the DemoSCOPE online panel who have participated in a specific survey before and after a design transition. The second evaluation tool is a study in which panel members were invited to rate examples of the old and the new design.

The analysis of the paired panel data shows that for the new design the completion rate is increased by almost 10%. We see this as a strong hint that a design based on the proposed design guidelines moves a little step towards an optimized online layout. However, we have to consider that this effect could also be caused, e.g., by a novelty effect based on the new setup of the DemoSCOPE online community or a changed general interest in the study topic.

Concerning the topic of question timing, an important finding is that pictograms significantly decrease the question timing. Pictograms reduce time for reading or they increase motivation due to the play-like nature of the pictograms. On the other hand, the one-item-per-screen-approach significantly increases the time needed when compared to the former grid approach. Hence, we can reinforce the findings of Couper et al. (2001) concerning the same issue. Based on our analysis, the introduction of the auto-submit function does not substantially affect the question timing. In total, there is no significant difference in the mean length of interview for both designs for our case study. Thus, the length of the entire interview is influenced by different, often contrary effects of particular interview elements. Therefore, the plain analysis of question timing might not be a useful measure: On the one hand we want participants to carefully read and answer questions, on the other hand we want to support quick and easy navigation through the questionnaire. To get a better understanding of these two conflicting demands, experiments have to be designed where time used for "thinking", and "navigating" is separated from each other.

The question timing for the new design might also be influenced by the introduction of new devices which have a quicker response time. However, the presented insurance study is stripped off from imagery and other media content and as such could not have caused longer loading times even on older devices. Furthermore, throughout the analysed surveying period we have seen no feedback from any respondent to purport this possibility.

Concerning the results from the interview data, we could not show that the new design reduces item nonresponse and straightlining. For an example of an open question in our case study, we showed that tailor-made adaptions can increase the willingness to answer open questions.

When examples of the new and the old design are shown to the DemoSCOPE community as in our rating study, the new design is rated significantly higher when it comes to Orientation, Color, Design and Usability. For Orientation and Design we see that the increase in rating before and after the design transition is significantly different for smartphone and desktop users. Based on this positive feedback from our community we think that we have proposed reasonable guidelines in the direction of an optimized online survey design.

A drawback in interpreting the results of the proposed approaches is that we cannot quantify selection effects which might nuise the result. A selection effect can take place at several stages of the analysis: First, our sample is an online panel which might not represent the true structure of the population. General population's participation in online panels is low and also probability panels are prone to selection bias with potentially large impact on results and decisions: there might be a large group of people who do not like online-research in general or mobile-device adjusted design in particular and are not part of the online panel. This could be

problematic, for example, for the rating study (Sections 4.2 and 6) as people who do not like tile design in general cannot even be invited to participate in the study. On the stage of participation, a self-selection effect might take place: Only people who like the new design participate, those who don't (and liked the old design better) do not participate.

Furthermore, for our paired panel data analysis (Sections 4.1 and 5) we use only the interviews of persons who took part in the insurance study before and after the design transition. This could be a problematic aspect when assessing the high increase in completion rate for the new design. It could be that only the supporters of the new tile design started the survey after the design transition, which results in a higher completion rate as a group of people who refuse the design transition didn't even access the study anymore and are, therefore, not part of our analysis. Furthermore, a lot of results are deduced from interviews of people that completed the insurance study before and after the design transition. This is another stage where selection bias is likely.

However, besides all the possible sources of selection bias we believe that the results of our analyses are valid and can give reasonable hints concerning the setup of an optimized survey layout in online research for multi-device usage. However, real proof for individual aspects of fluid responsive web design has to come from more controlled experiments and true random samples. From within our panel and commercial studies we cannot create controlled experiments with groups that never see the new design or with a random assignment of old and new design to sub-populations, neither can we systematically vary all the design elements mentioned above. But, what can be done is to assess benchmark measures such as the completion rate over time and continuously integrate new and research-based elements into our online design.

In this paper, we offer a handful of ideas how to go about designing online surveys in a new way. We believe it is a constant, ongoing task. This said, some parameters that we consider key in this process, will remain monitored in the day-to-day business. In a playful manner we allow to comment on the topic, user friendliness as well as the length of the survey (as shown in Figure 13). This allows us to ensure that adequate measures are taken in order to maintain a high satisfaction rate amongst our respondents with a positive impact on data quality and response rates.

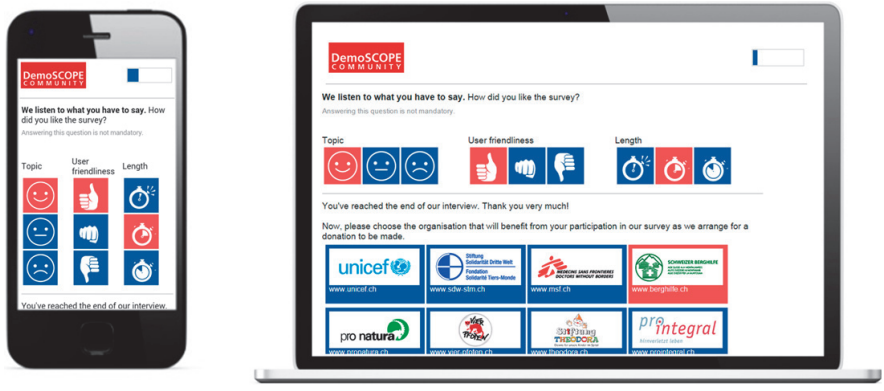


Figure 13 3 Feedback questions asked at the end of each online survey at DemoSCOPE

References

Brujine, M, & Wijnant, M. (2014). Improving Response Rates and Questionnaire Design for Mobile Web Surveys. *Public Opinion Quarterly*, 78 (4), 951-962.

Buskirk, T. D., & Andrus, C. (2012a). Smart Surveys for Smart Phones. Exploring Various Approaches for Conducting Online Mobile Surveys via Smartphones. *Survey Practice*, 5(1). http://www.surveypractice.org/index.php/SurveyPractice

Buskirk, T. D., & Andrus, C. (2012b). Online Surveys Aren't Just for Computers Anymore! Exploring Potential Mode Effects between Smartphone and Computer-Based Online Surveys. Paper presented at the annual meeting of the American Association for Public Opinion Research, Orlando, FL. http://www.amstat.org/sections/SRMS/Proceedings/y2012/Files/400244_500700.pdf

Buskirk, T. D., & Andrus, C. (2014). Making Mobile Browser Surveys Smarter: Results from a Randomized Experiment comparing Online Surveys Completed via Computer or Smartphone. *Field Methods*, 26(4), 322-342.

Christian, L. M., & Dillman, D. A. (2004). The Influence of Graphical and Symbolic Language Manipulations on Response to Self-Administered Questions. *Public Opinion Quarterly*, 68(1), 57-80.

Christian, L. M., & Dillman, D. A. (2004). The Influence of Graphical and Symbolic Language Manipulations on Response to Self-Administered Questions. Public Opinion Quarterly, 68(1), 57-80.

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2007). Helping Respondents Get it Right the First Time: The Influence of Words, Symbols, and Graphics in Web Surveys. Public Opinion Quarterly, 71(1), 113-125.

Conrad, F.G., Couper, M.P., Tourangeau, R, & Peytchev, A. (2010). The Impact Of Progress Indicators on Task Completion. *Interacting with Computers*, 22, 417-427.

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

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture This! Exploring Visual Effects in Web Surveys. *Public Opinion Quarterly*, 68(2), 255-266.

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual Context Effects in Web Surveys. *Public Opinion Quarterly*, 71(4), 623-634.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Zhang, C. (2013). The Design of Grids in Web Surveys. *Social Science Computer Review*, 31(3), 322-345.

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(4), 482–504.

Deutskens, E., De Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response Rate and Response Quality of Internet-Based Surveys. An Experimental Study. *Marketing Letters*, 15(1), 21-36.

Dillman, D. A., Michael D. Sinclair, & Jon R. Clark. (1993). Effects of Questionnaire Length, Respondent-Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys. *Public Opinion Quarterly*, 57(3), 289-304.

Dillman, D. A., & Christian, L. M. (2002). The Influence of Words, Symbols, Numbers, and Graphics on Answers to Self-Administered Questionnaires: Results from 18 Experimental Comparisons. Retrieved February 23, 2015 on http://www.websm.org/db/12/642/rec/.

Fuchs, M. (2005). Children and Adolescents as Respondents. Experiments on Question Order, Response Order, Scale Effects and the Effect of Numeric Values Associated with Response Options. *Journal of Official Statistics*, 21(4), 701-725.

Fuchs, M., & Busse, B. (2009). The Coverage Bias of Mobile Web Surveys Across European Countries. *International Journal of Internet Science*, 4(1), 21-33.

Gustafson, A. (2011). *Adaptive Web Design. Crafting Rich Experiences with Progressive Enhancement*. Chattanooga, Tennessee: Easy Readers.

Jenkins, C. R., & Dillman, D. A. (1997). Towards a Theory of Self-Administered Questionnaire Design. In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., Trewin, D. (Eds), *Survey Measurement and Process Quality* (pp. 165-196). Wiley: New York.

Lugtig, B. & Toepoel, V (2015). The use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review*, 1-17.

Kaplowitz, M. D., Lupi, F., Couper, M. P., & Thorp, L. (2012). The Effect of Invitation Design on Web Survey Response Rates. *Social Science Computer Review*, 30(3), 339-349.

Klausch, T., De Leeuw, E. D., Hox, J., Roberts, A., & De Jongh, A. (2012). Matrix vs. Single Question Formats in Web Surveys: Results From a Large Scale Experiment. Paper presented at the General Online Research (GOR), March 5-7 2012, Mannheim, Germany.

Knäuper, B., Schwarz, N., & Park, D. (2004). Frequency Reports Across Age Groups. *Journal of Official Statistics*, 20(1), 91-96.

Krosnick, J. A., & Alwin, D. F. (1987). An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement. *Public Opinion Quarterly*, 51(2), 201–19.

Marcotte, E. (2010). Responsive Web Design. Retrieved March 2, 2015, from: http://alistapart.com/article/responsive-web-design/.

Mavletova, A. (2013). Data Quality in PC and Mobile Web Surveys. *Social Science Computer Review*, 31(6), 725-743.

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones. An Emerging Tool for Social Scientists. *Sociological Methods & Research*, 37(3), 426-454.

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2014). WP 150 - Who has access to mobile devices in an online commercial panel? From: http://www.uva-aias.net/publications/show/1977

Rothwell, N. D. (1985). Laboratory and Field Response Research Studies for the 1980 Census of Population in the United States. *Journal of Official Statistics*, 1(2), 137-157.

Sanchez, M. E. (1992). Effect of Questionnaire Design on the Quality of Survey Data. *Public Opinion Quarterly*, 56, 206-217.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, 29, 65-88.

Schwarz, N., Fritz, S., & Mai, H. P. (1991). Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis. *Public Opinion Quarterly*, 55(1), 3-23.

Schwarz, N. (1995). What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation. *International Statistical Review*, 63(2), 153-68.

Shropshire, K. O., Hawdon, J. E., & Witte, J. C. (2009). Web Survey Design: Balancing Measurement, Response, and Topical Interest. *Sociological Methods & Research*, 37(3), 344-370.

Stern, M. J., Dillman, D. D., & Smyth, J. D. (2007). Visual Design, Order Effects, and Respondent Characteristics in a Self-Administered Survey. *Survey Research Methods*, 1(3), 121-138.

Stern, M. J., Bilgen, I.,& Dillman, D. D. (2014). The State of Survey Methodology: Challenges, Dilemmas, and New Frontiers in the Era of the Tailored Design. *Field Methods*, 26(3), 284-301.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, Position, and Order: Interpretative Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68(3), 368-393.

Tourangeau, R., Couper, M. P., & Conrad, F. (2007). Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opinion Quarterly*, 71(1), 91-112.

Vicente, P., & Reis, E. (2010). Using Questionnaire Design to Fight Nonresponse Bias in Web Surveys. *Social Science Computer Review*, 28(2), 251-266.

Villar, A., Callegaro, M., & Yang, Y. (2013). Where Am I? A Meta-Analysis of Experiments on the Effects of Progress Indicators for Web Surveys. *Social Science Computer Review*, 31(6), 744-762.

Wells, T., Bailey, J. T., & Link, M. W. (2014). Comparison of Smartphone and Online Computer Survey Administration. *Social Science Computer Review,* 32(2), 238-255.

Whitcomb, M. E., & Porter, S. R. (2004). E-Mail Contacts. A Test of Graphical Designs in Survey Research. *Social Science Computer Review*, 22(3), 370-376.

Wright, P., & Barnard, P. (1975). 'Just fill in this form' – a review for designers. *Applied Ergonomics*, 6(4), 213-220.

Wright, P., & Barnard, P. (1978). Asking multiple questions about several items: the design of matrix structures on application forms. *Applied Ergonomics*, 9(1), 7-14.

methods, data, analyses | Vol. 9(2), 2015, pp. 213-228 DOI: 10.12758/mda.2015.012

Web Surveys Optimized for Smartphones: Are there Differences Between Computer and Smartphone Users?

Ioannis Andreadis
Aristotle University of Thessaloniki

Abstract

This paper shows that computer users and smartphone users taking part in a web survey optimized for smartphones give responses of almost the same quality. Combining a design of one question in each page and innovative page navigation methods, we can get high quality data by both computer and smartphone users. The two groups of users are also compared with regard to their precisely measured item response times. The analysis shows that using a smartphone instead of a computer increases about 20% the geometric mean of item response times. The data analyzed in this paper were collected by a smartphone-friendly web survey. All question texts are short and the response buttons are large and easy to use. As a result, there are no significant interactions between smartphone use and either the length of the question or the age of the respondent. Thus, the longer response times among smartphone users should be explored in other causes, such as the likelihood of smartphone users being distracted by their environment.

Keywords: web surveys, mobile surveys, AJAX navigation, data quality, item response times, smartphones

1 Introduction

The aim of this paper is to study the differences between computer and smartphone users when they complete web surveys optimized for smartphones. The comparison is done on two dimensions. The first dimension refers to the quality of the responses, e.g. the frequencies of no answers or neutral responses. The second dimension refers to the time the respondents spend to answer the questions, i.e. the item response times.

The most recent studies on the effects of mobile use on data quality report limited differences between mobile and computer respondents. Mavletova (2013), analyzing an experiment in Russia, reports that computer respondents type longer responses on open-ended questions. One the other hand, she finds that mobile and computer users have similar levels of socially undesirable and non-substantive responses. In addition, the two groups do not differ significantly in terms of primacy effects. De Bruijne and Wijnant (2013) after running an experiment with participants randomly assigned to three modes (mobile, computer and a hybrid) have not found any significant differences. Toepoel and Lugtig (2014) have not found differences between mobile and desktop users with regard to item nonresponse, the length of answers to a short open-ended question and the number of responses in a check-all-that-apply question. Finally, Wells, Bailey, and Link (2014) have randomly assigned roughly 1,500 online U.S. panelists and smartphone users to either a mobile application or a computer. They have not found any significant response-order effects across modes. However, they report that computer respondents provide significantly longer responses than mobile respondents.

Many web survey researchers have reported that the number of people who use mobile devices to take part in web surveys is increasing rapidly. In addition, the time spent on a web survey is crucial for the quality of the collected data. Longer web surveys suffer from larger break-off rates and greater probability of lower quality responses. Therefore, many recent publications deal with the time spent on responding to web surveys while using mobile devices. Both Mavletova (2013) and De Bruijne and Wijnant (2013) report that mobile device users need more time to complete the questionnaire than computer users. Conversely, Toepoel and Lugtig (2014) find that total response times are almost the same across devices.

Direct correspondence to
Ioannis Andreadis, Laboratory of Applied Political Research, Department of Political Sciences, Aristotle University Thessaloniki, 46 Egnatia St., Thessaloniki, 54625 Hellas (Greece)
E-mail: john@polsci.auth.gr

2 Designing for Smartphone Users

Previous studies on measurement effects have found minimal differences between mobile and computer respondents. The most challenging difference concerns the length of open-ended responses. Nowadays, people become more and more experienced in typing texts using the small keys of their smartphones. Nevertheless, it is still much easier to type using a regular full-size keyboard than using a keypad on a mobile device. As a result, we should continue to expect longer responses on open-ended questions by computer users, especially when the response needs more than 3-4 words.

A good web survey design can remove most of the remaining differences. Survey design always plays an important role. According to Stapleton (2013), horizontal orientation of response choices may increase satisficing by smartphone users, i.e. they are more likely to select one of the first response choices. Vertical scrolling seems to be better than horizontal scrolling. In fact, Mavletova and Couper (2014) argue in favor of using a vertical scrolling design and they report that it leads to significantly faster completion times, and fewer technical problems. As they argue, the smaller number of interactions with the server reduces the risk of dropped connections. On the contrary, Wells, Bailey, and Link (2014) argue in favor of minimal vertical scrolling and support the idea of using one question per page, short questions and short sets of response lists.

A solution that gets the best from both worlds is the use of Asynchronous JavaScript and XML (AJAX) technology. By using AJAX, survey designers can display one question per page while minimizing the risk of failed connections with the server. This is achieved by downloading all pages to the users' browser during the first connection with the server. Then, AJAX takes care of the navigation from page to page. In that way, there is only a second and final connection with the server when the user submits the completed questionnaire. Furthermore, with AJAX technology we avoid any lags between pages[1]. This means that we can have accurate measurement of the time spent between clicks.

3 Data

The findings presented in this paper are based on the analysis of the paradata collected in May 2014 by the Greek Voting Advice Application (VAA) HelpMeVote - VoteMatch Greece (Andreadis, 2013). Voting advice applications are special types of opt-in web surveys that help users find their proximities with the political parties.

1 Couper and Peterson (2015) refer to two kinds of times: between-page (transmission) time and within-page (response) time. Using the AJAX navigation system the former time is zero.

These applications can attract thousands or even millions of users during the pre-electoral period. HelpMeVote is the Greek partner in the multi-national European project VoteMatch (votematch.eu). The target of this project is to run VAAs for the European Parliament elections.

HelpMeVote follows the best practices used in both web and mobile survey design. It runs both on computers and on smartphones. It automatically scales to any screen size and it supports both touch and mouse events. It displays one question per page and supports AJAX navigation. It uses large font size, short texts and the response options are displayed vertically with large buttons.

The questionnaire includes 31 Likert type questions. Each question is displayed on a separate page. Respondents have six answer choices: there are five buttons to express their level of agreement with a statement and a "No answer" button. When a respondent clicks on a button, the timestamp is recorded in a hidden input field and the user is forwarded to the next page. Besides the 31 main questions, HelpMeVote users are asked to fill-in a form. This form includes questions about their gender, age group, education level, and voting behavior. Finally, HelpMeVote captures the user-agent header field, which enables the detection of the users' browser and device type (i.e. smartphone, computer, etc). When the respondent submits the survey, everything is stored to a database. Thus, each database record includes the user responses, the timestamps and the device type.

The HelpMeVote/VoteMatch Greece dataset includes about 80,000 completed questionnaires. The largest part of the dataset consists of computer users (80.7%) and smartphone users (13.5%). The rest of the respondents have used other mobile devices (mostly tablets). The focus of this paper is on the comparison between smartphone and computer users when both groups use a smartphone-friendly web survey. Therefore, users of other mobile devices were not included in the analysis.

4 Methods and Variables

4.1 Quality of Responses

HelpMeVote does not include any open-ended items. Thus, the hypothesis that computer respondents provide longer responses cannot be tested. One the other hand, computer and smartphone users of HelpMeVote can be compared for other data quality patterns.[2] For instance, if smartphone users selected more non-substantive responses (i.e. "Neither agree nor disagree" or "No answer") than computer users, this would suggest that smartphone users provide data of lower quality. Similarly

2 For a list of mode effects related to data quality see Bethlehem and Biffignandi, 2011, p.245

smartphone users can be tested for primacy effects (i.e. selecting the first response choice more often) or any other response-order effects.

When a chi-square test is applied on a large sample, it will almost always give a small p-value. Even when there is no practical difference between expected and observed frequencies, the test will reject the hypothesis of independence. In addition, running a separate test for each of the 31 items included in HelpMeVote would result in multiple comparisons and incorrect rejection of the null hypothesis. Thus, it would be more likely to classify nonsignificant differences as significant.

The aforementioned problems are avoided by creating six new variables. The value of each new variable reflects the number of times the respondent has chosen the corresponding response option ("Frequency of Strongly Disagree" to "Frequency of Strongly Agree" and "Frequency of No Answer"). The range of values of these new variables is from 0 to 31. Each of these variables takes the minimum value (0) when the respondent does not select the corresponding answer in any of the 31 questions. Similarly, it takes the maximum value (31) when the respondent selects the same answer for all questions. With these variables it is easy to analyze mode effects between mobile and computer users. For instance, a comparison of the average values of the variable "Frequency of Strongly Disagree" between mobile and computer users will show if there is a different primacy effect between modes. Similarly, a comparison of the average values of the variables: "Frequency of Neither agree nor disagree" and "Frequency of No Answer" between the two groups will reveal if smartphone users select non-substantive responses more often than computer users.

5 Item Response Times

The analysis of item responses times is much more complicated for two reasons. First, item response times depend on characteristics of both the respondents and the items. As a result, there is a need for a multi-level analysis of the item response times. Second, there is need for data cleaning in order to deal with extremely short or extremely long item response times.

5.1 Multilevel Model

Item response times depend on characteristics of the respondent, e.g. gender, age, education, interest in the theme of the survey and knowledge about the survey topics. Attributes of the items such as the length or the difficulty of the item have also an impact on item response times. Thus, item response times are usually analyzed with a multilevel model. The usual approach is to consider a hierarchical model where the items are nested within the respondents (Van der Linden, 2008), but

there are examples of reversed roles, i.e. the respondents are nested within items (Swanson et al., 2001). Using a non-hierarchical model would underestimate the standard errors of regression coefficients and make nonsignificant coefficients to appear as significant (Hox, 2002; Gelman & Hill, 2006).

For the multilevel analysis, the dataset has to be reshaped in its long format. This way, each of the about 80,000 cases is multiplied by 31 (i.e. the total the number of the items). The outcome of this procedure is a dataset of about 2.5 million cases. Analysis of this huge dataset is difficult even when a strong workstation is used. To overcome this problem, a random sample corresponding to 10% of the complete dataset was selected. The distributions of the main variables are very similar in the sample and in the initial dataset. Replications of the same analysis presented in this paper on other 10% random samples have given very similar findings. The used sample is available by OpenICPSR (Andreadis, 2014b).

5.2 Data Cleaning

Andreadis (2012, 2014a) proposes a method to flag items that were responded in extremely short time. The method is based on the types of reading and the corresponding reading speeds presented by Carver (1992). Scanning is the fastest type of reading. When respondents scan a question, they do not dedicate adequate time to understand the meaning of the text. In addition, according to Bassili and Fletcher (1991), answers to simple attitude questions take between 1.4 and 2 seconds. Adding the minimum needed time to read and comprehend a question and the minimum needed time to answer the question, we get the following formula: threshold = 1.4+[number of characters in the item]/39.375. Users with extremely short times in more than one third of their responses were removed from the dataset. This decision is justified on the hypothesis that these users have responded without paying attention to the questions; these users usually maintain the same attitude throughout the questionnaire.

On the other hand, it is very rare to observe a user spending extremely long to answer most questions. This delay is often a result of an external distraction (e.g. an incoming email, phone call, door knocking, etc). Thus, the recorded time is not the actual time spent on the question. Instead, it is the sum of the time spent on answering the question, plus an unknown amount of time caused by some external distraction. After applying the logarithmic function to the response times to reduce skewness, the extreme values have been identified with the use of boxplot statistics (Tukey, 1977; Hoaglin, Mosteller, & Tukey, 1983; McGill, Tukey & Larsen, 1978). These values were coded as missing (Andreadis, 2015).

5.3 Other Data Preparations

In the following models the logarithm of the response times is used as the dependent variable (i.e. the outcome). Since the main task of this paper is to compare the response times between smartphone and computer users, the binary variable "smartphone" is included into the model as the main treatment variable. This variable gets the value 1 if the respondent is a smartphone user and the value 0 if the respondent is a computer user.

The control variables on the user level are the following: education, gender, age, political interest and a variable (Decided) that gets the value 1 if the respondents had already made their vote choice when they used the VAA and 0 otherwise. Education is used as a categorical variable with five levels: Primary, Lower secondary, Upper secondary, Tertiary and Postgraduate studies. Primary education is used as the reference level and all other levels are compared with it. The expectation is that the higher the education levels are, the less the response time should be. The remaining variables are used as dummy variables. Gender gets the value 0 for female and 1 for male respondents. Some studies have found that female respondents spend more time on web surveys, thus a similar finding is expected from the present analysis. Age gets the value 1 if the respondent is older than 49 years old, and 0 otherwise. According to the literature, older people are expected to spend more time than younger people. For the analysis of item response times, the cases with missing values on the demographic variables have been filtered out. There are two reasons which give support to this decision. First, the percentage of missing values is small. Second, the application of advanced imputation methods, such as imputing the missing value with the predicted value of a regression, would be challenging. Demographic variables (e.g. age) may serve as good predictors of some attitudinal variables (e.g. more conservative views). Trying the opposite, i.e. using attitudinal variables to predict demographic variables would be odd, because attitudinal variables do not have an impact on demographics, such as age or gender.

The variable political interest gets the value 1 if the respondent has indicated an interest in politics. Citizens interested in politics and voters who have already decided about their vote choice should be more familiar with the major issues of the electoral competition. Thus, they are expected to have a clear, pre-formulated opinion about the statements. As a result, they are expected to need less time than people not interested in politics and people who had not decided about their vote choice when they used HelpMeVote.

The control variables on the item level are: the length of the statement (see Andreadis, 2012 and 2014a) and a dummy variable that takes the value of 1 when the statement is about a European Union issue and 0 when the statement is about a national issue. Greek voters are less informed about EU policy issues than they

are on national issues. Thus, they are expected to need more time to express their opinion on EU issues.

6 Findings

6.1 Quality of Responses

Table 1 shows the mean values estimated over the 31 Likert type items of the frequencies of each response option. The p-values in this table show that even with a huge sample of thousands of cases, none of the differences between modes are significant at the 0.01 significance level. At the 0.05 level, one of the differences would be considered as statistically significant, but its magnitude is small and less important for any practical purpose.

The average smartphone respondent and the average computer respondent give similar answers. Both of them select the answer "Strongly Disagree" in 4-5 questions, the answer "Disagree" in 6-7 questions, the answer "Neither … nor" in 4-5 questions, the answer "Agree" in 8-9 questions and the answer "Strongly Agree" in 5-6 questions.

The lack of significant measurement effects is consistent with the findings of previous studies, discussed in the previous sections. According to the t-test output presented in Table 1, there are no significant differences in responses across modes in terms of primacy effects or of response-order effects in general. In addition, there are no significant differences in non-substantive responses across modes. Of course, these findings are not based on an experiment. HelpMeVote users are free to choose the device they use. Thus, it is possible that the effects of self-selection and measurement differences counteract.

Both computer and smartphone users select the response "Agree" more frequently than any other response option. Some scholars may consider this finding as an indicator of acquiescence bias. On the other hand, this could be a result of the specific set of questions, i.e. another set of questions including less popular policy statements could have more "Disagree" and less "Agree" answers. Another observation is that both groups of HelpMeVote users tend to select the less extreme responses ("Disagree" and "Agree") more often than the corresponding extreme responses ("Strongly Disagree" and "Strongly Agree"). For both groups, the frequency of middle category "Neither agree nor disagree" is lower than the expected frequency for a single point when a uniform discrete distribution of 31 5-point items is considered. Finally, the average user of both groups has selected the "No Answer" button in less than one out of 31 items. This is an indicator of lack of satisficing. In general, the data quality is very high in both groups.

Table 1 T-tests for the estimation of mode (computer vs mobile) effects

	Computer		Mobile		t-test	
	Mean	SD	Mean	SD	t	p
Frequency of Strongly Disagree	4.76	4.38	4.47	4.01	2.23	0.03
Frequency of Disagree	6.31	3.59	6.35	3.32	-0.33	0.74
Frequency of Neither … nor	4.79	3.71	4.84	3.35	-0.46	0.65
Frequency of Agree	8.59	4.51	8.80	4.30	-1.47	0.14
Frequency of Strongly agree	5.75	4.29	5.63	3.87	0.93	0.35
Frequency of No Answer	0.80	3.21	0.92	3.62	-1.13	0.26

6.2 Item Response Times

Table 2 shows the multilevel regression model for the logarithm of item response times. The table includes the estimated coefficients and the exponential coefficients along with the outcome of the significance tests. Since the dependent variable is the logarithm of the item response times, the interpretation of the estimated regression coefficients is the following: Suppose that the estimated coefficient for an independent variable X is b. This means that when X increases by one unit the logarithm of the item response time is expected to increase by b units. Consequently, the item response time will be multiplied by e^b. According to Table 2 the constant term of the model is estimated at 2.01. This is the expected mean of the logarithm of the item response times. The exponential value of 2.01 is 7.47. This is the geometric mean of item response times, i.e. the average respondent needs about 7.5 seconds to respond to one item.

The interpretation of the coefficient of the treatment variable shows the impact of using a smartphone on the response times: the coefficient is 0.181 and its exponential value is 1.198. This means that switching from computer to smartphone the geometric mean of response times is expected to increase by 19.8%. An estimate of the treatment effect in seconds is given by the following calculation: 7.47*19.8%=1.5 seconds. This means that the average smartphone user spends about 1.5 seconds more than the average computer user on an item.

The coefficient for the length of the statement is 0.006 and its exponential value is 1.006. This means that, while holding all other predictors constant, for every additional character in the question, the geometric mean of response times increases by 0.6%. According to the model, if a statement refers to a EU policy issue the respondents need more time to give their answer. The corresponding coefficient is 0.104 and its exponential value is 1.11 showing an 11% increase in the geometric mean of response times when switching from a national issue to a EU issue.

Table 2 Multilevel model coefficients and exponential coefficients

Logarithm of times	Coef.	Exp(b)	Std. Err.	z	P>z
Smartphone	0.181	1.198	0.018	9.96	0.000
Item characteristics					
Length	0.006	1.006	0.000	104.98	0.000
EU issue	0.104	1.110	0.002	48.86	0.000
Respondent characteristics					
Male	-0.095	0.909	0.010	-9.94	0.000
Age over 49	0.192	1.212	0.040	4.79	0.000
Education		1.000			
Lower secondary	-0.086	0.918	0.063	-1.35	0.176
Upper Secondary	-0.245	0.783	0.055	-4.49	0.000
Tertiary	-0.335	0.715	0.054	-6.17	0.000
Postgraduate studies	-0.440	0.644	0.055	-8.05	0.000
Political interest	-0.075	0.928	0.010	-7.30	0.000
Decided	-0.055	0.946	0.009	-5.96	0.000
Interactions					
Smartphone#Age	-0.072	0.931	0.041	-1.73	0.084
Smarphone#Length	-0.000	1.000	0.000	-2.36	0.018
Constant	2.010	7.463	0.055	36.44	0.000

The coefficient for male is -0.095 and its exponential value is 0.909. This means that the geometric mean of response times in the group of men is 90.9% of the geometric mean of response times in the group of women. In other words, switching from female to male respondents, the response time is expected to decrease by 9.1%. Following the same model, when we switch from undecided people to people who have already made their choice the geometric mean of response times decreases by 5.4%. Similarly, moving from people who are not interested in politics to people who are interested in politics the geometric mean is expected to decrease by 7.2%. One the other hand, the exponentiated coefficient for older people is 1.21 showing a 21% increase in the geometric mean of response times when switching from younger people to users over 49 years old. Finally, when we switch from primary education to higher education levels, the response time decreases; only the difference between primary and lower secondary education is not statistically significant. The largest difference is observed between the two extreme education levels: the ratio of geometric means between postgraduate studies and primary education is 0.64 indicating that the time spent by the most educated users is 64.4% the time spent by the less educated users, i.e. a decrease of 35.6%.

Finally, the model includes the interaction terms between smartphone use and age, and smartphone use and the length of the question. None of these interaction terms have a significant impact on the item response times at the 0.01 significance level. If the time difference was caused by an unfriendly design, this difference would probably be higher in older people. But the interaction between smartphone use and age is not significant. Therefore, the longer time of smartphone users cannot be attributed to the unfriendliness of the web survey. The lack of a significant interaction between smartphone use and the length of the question does not allow the attribution of the longer times of smartphone users to the smaller display of their devices. This was an expected finding because the survey was carefully designed to fit on the small screens of mobile devices. Thus, all questions are short and the variability of their length is small.

6.3 Validity of the Model and Sensitivity Analysis

According to Table 3, the variance of the random intercept is 0.11 and the estimated error variance is 0.2. The likelihood ratio test shows that the random intercept variance is large. This verifies that the decision to use a multilevel model was correct. Indeed, if a single level model had been used, non significant differences (e.g. the response time difference between primary and lower secondary education levels) would appear as significant.[3]

3 In addition, it was checked whether a different model should be applied instead of the multilevel model. Some of the respondents' characteristics, such as age and education level, known to affect response times, (Yan and Tourangeau, 2008; Couper and Kreuter, 2013), have been reported also to affect mobile web access (Fuchs and Busse, 2009; De Bruijne and Wijnant 2013; Gummer and Roßmann, 2015). This means that, the treatment variable of the model (mobile) is endogenous and it depends on variables that also affect the outcome (e.g. age). This means that an endogenous treatment-effects model should be employed (Greene, 2012; Heckman, 1978; Maddala, 1983; Wooldridge, 2010). In addition, a multilevel model is necessary. Bauer (2003) and Curran (2003) show how to estimate multilevel linear models as structural equation models. It seems that the only feasible way to estimate an endogenous treatment-effects multi-level regression model is to use a generalized structural equation model (Skrondal & Rabe-Hesketh, 2004). Since the treatment is endogenous, it is necessary to test if the correlation between the error terms of the equations of the generalized structural equation model is significant. The value of the correlation coefficient rho is 0.01 and the corresponding test shows that it is not significantly different from 0 (p=0.937). This means that the estimates of a simple multilevel model can be accepted (Andreadis, 2015).

Table 3 Variances of the multilevel model

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
Variance of the random intercept	0.110	0.002	0.106	0.114
Error variance	0.202	0.001	0.201	0.203

7 Discussion

This paper advances mobile survey research in various ways. Firstly, it shows that creating a web survey suitable for both smartphone and computer users involves good design choices. These designs choices (such as the one question per page design) facilitate mobile users to give responses that do not differ from the responses given by computer users. In addition, a good survey design results in high quality data from both groups of users. As the findings presented in this paper show, the data quality of smartphone respondents does not differ significantly from the data quality of the computer respondents. In both groups the level of the data quality is very high and there are no signs of satisficing.

The lack of data quality differences presented in this paper is not based on an experiment. Respondents self-select the device they use to participate in HelpMeVote. Thus, the effects of self-selection may counteract any measurement differences. It is reasonable to believe that respondents have chosen the device that they feel more comfortable to use, and they use it without problems. If the devices were assigned to respondents randomly, the findings could be different. This is a limitation of the presented study. However, the focus of this paper is not on what would happen in a lab after forcing respondents to use a specific device. The focus is rather on what happens in the real world, where respondents are free to choose the device they prefer.

Secondly, this paper offers an innovative method to prepare a dataset of response times for statistical analysis by treating the low and the high extreme values differently. It shows how to flag users who have been answering so fast that they should be removed by the dataset. Moreover, it proposes a way to deal with the extremely large response times by identifying the actual extremes instead of trimming the dataset using arbitrary selected thresholds that lack any theoretical justification.

Finally, this paper offers a precise and thoroughly tested estimate of the impact of using a smartphone on item response times. The comparison was made between computer and smartphone users when they use a smartphone friendly web survey. The analysis has shown that using a smartphone instead of a computer increases the

geometric mean of item response times by 19.8%. This increase was estimated after taking into account item and user characteristics that are known to affect response times, and using the most suitable statistical model.

Explaining why smartphone users need more time than desktop users is not an easy task. There are many potential causes that could explain this difference between the devices, but some of them can be excluded by the design features of HelpMeVote. One of the possible causes is the (usually) slower Internet connections of smartphones. A slower Internet connection would lead to longer transmission times. This could explain the difference between the devices in other studies. However, this factor is not relevant for the data presented in this paper, because there is no lag between pages in HelpMeVote.

Smartphone users may have difficulties responding to a web survey that is not smartphone friendly. For instance, sometimes respondents have to zoom in to read a text written with small fonts. They may have to scroll horizontally in order to read the question. In other cases, respondents have to type their answers. All these actions could delay smartphone users and they could be used to explain longer times in other studies. However, these actions are not required by the smartphone users of HelpMeVote, because it is a smartphone friendly web survey. It uses large fonts and buttons and short question texts. It requires no horizontal scrolling and no typing. As a result, these possible obstacles do not apply to HelpMeVote users. This argument is further supported by the lack of any significant interaction between smartphone use and the age of the respondents. If smartphone users needed more time due to similar difficulties, the situation would probably be worse when older people are involved and the interaction would be significant.

There are two potential causes that could explain the difference between smartphone and computer users of HelpMeVote. It is possible that smartphone users may need more time because they have to scroll vertically before answering the question. Of course, the lack of a significant interaction between smartphone use and the length of the question hinders any blaming on vertical scrolling. But this finding is a result of the small variability of the length of the questions in HelpMeVote. If the variability was larger, the outcome would probably be different. Unfortunately, it is not easy to know if a respondent had to scroll vertically to give a response. This would require the knowledge of the screen resolution of each device and its orientation[4] during all the time the user was completing the survey. For instance, the owner of an Apple iPhone 5 holding the device on its vertical orientation would be able to answer all HelpMeVote questions without any scrolling. On the other hand, if the horizontal orientation of the same device was used, then the respondent would have to scroll vertically most of the times. The recording of the

4 Many smartphones can determine their orientation and automatically rotate the display to present a wide-screen view of the web content. In this case, the vertical space is very limited and the user often has to scroll vertically.

screen resolution and all the changes of the screen orientation is possible, but the additional code complicates and slows down the web survey application. Thus, it is more appropriate for less popular projects and it was not used in HelpMeVote which is used by thousands of voters.

A final reasonable explanation for the longer times among smartphone users is that they are probably completing the survey outdoors and they are more easily distracted than the computer users who complete the survey in a quite room at home or in an office. However, since the dataset used for this paper does not include the parameter of the location where the respondents have completed the questionnaires, this hypothesis cannot be verified.

Many recent publications show a continuous increase of the percentage of respondents who use their mobile devices to respond to a web survey. If web survey designers want to avoid data quality differences between computer and smartphone users, they have to optimize the design of the online questionnaire for smartphones. A good web survey design should definitely eliminate the need for horizontal scrolling. But this feature alone is not enough. Survey designers caring for their smartphone users should also try to minimize the need for vertical scrolling by using short questions and short sets of response lists and by displaying only one question per page. Following these practices, they can expect very similar responses from all their respondents, no matter what device they use.

Acknowledgements

The author would like to acknowledge the contribution of the COST Action IS1004 Webdatanet (Steinmetz et al., 2014) to the development of the presented study. Many of the ideas and methods presented in this paper are the result of the author's participation in Webdatanet and his work on the preparation of the Webadatanet web survey paradata training school.

References

Andreadis, I. (2012). *To Clean or not to Clean? Improving the Quality of VAA Data* Paper presented at XXII World Congress of Political Science (IPSA), Madrid, Spain. Retrieved from: http://www.polres.gr/en/sites/default/files/IPSA12.pdf

Andreadis, I. (2013). Voting advice applications: A successful nexus between informatics and political science. In *Proceedings of the 6th Balkan Conference in Informatics,* 251-258. doi:10.1145/2490257.2490263

Andreadis, I. (2014a). Data quality and data cleaning. In D. Garzia & S. Marschall, (Eds.), *Matching Voters with Parties and Candidates. Voting Advice Applications in Comparative Perspective*, (pp 79-91). Colchester, UK: ECPR Press

Andreadis, I., (2014b). *Paradata from Political Web Surveys*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2014-10-02. doi:10.3886/E17816V3

Andreadis, I. (2015). Comparison of response times between desktop and smartphone users. In D. Toninelli, R. Pinter, & P. Pedraza, (Eds), *Mobile Research Methods*, (pp 63-79). London, UK: Ubiquity Press

Bassili, J. N. & Fletcher, J.F. (1991). Response-time measurement in survey research a method for CATI and a new look at nonattitudes. *Public Opinion Quarterly, 55*(3), 331-346. doi: 10.1086/269265

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135-167. doi: 10.3102/10769986028002135

Bethlehem, J., & Biffignandi, S. (2011). *Handbook of web surveys*. John Wiley & Sons.

Carver, R.P. (1992) Reading rate: Theory, research, and practical implications. *Journal of Reading*, *36*(2), 84-95

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. Journal of the Royal Statistical Society: Series A (Statistics in Society) 176(1), 271–86. doi: 10.1111/j.1467-985X.2012.01041.x

Couper, M.P. & Peterson, G. (2015). Exploring Why Mobile Web Surveys Take Longer. Paper presented at General Online Research, Cologne, Germany.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529-569. doi: 10.1207/s15327906mbr3804_5

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. doi:10.1177/0894439313483976

Fuchs, M., & Busse, B. (2009). The coverage bias of mobile web surveys across european countries. *International Journal of Internet Science 4*(1), 21–33.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel / hierarchical models*. Cambridge University Press.

Greene, W. H. (2012). *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Gummer, T., & Roßmann, J. (2015). Explaining interview duration in web surveys a multilevel approach. Social Science Computer Review, 33(2), 217-234. doi: 10.1177/0894439314533479

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica 46*, 931-959.

Hoaglin, D.C. Mosteller, F. & Tukey, J.W. (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley & Sons.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Psychology Press.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Mavletova, A. (2013). Data quality in pc and mobile web surveys. *Social Science Computer Review, 33*, 725-743. doi: 10.1177/0894439313485201.

Mavletova, A., & Couper, M. P. (2014). Mobile web survey design: Scrolling versus paging, SMS versus E-mail invitations. *Journal of Survey Statistics and Methodology, 2*(4), 498-518. doi: 10.1093/jssam/smu015

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician 32*, 12–16.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.

Stapleton, C. (2013). The smart (phone) way to collect survey data. *Survey Practice, 6*(2).

Steinmetz, S., Slavec, A., Tijdens, K., Reips, U. D., de Pedraza, P., Popescu, A., et al. (2014). WEBDATANET: Innovation and quality in web-based data collection. *International Journal of Internet Science*, *9*(1), 64-71

Swanson, D. B. Case, S. M. Ripkey, D. R. Clauser, B. E., & Holtman M. C. (2001) Relationships among item characteristics, examine characteristics, and response times on USMLE Step 1. *Academic Medicine, 76*(10), S114-S116.

Toepoel, V., & Lugtig, P. (2014). What happens if you offer a mobile option to your web panel? Evidence from a probability-based panel of Internet users. *Social Science Computer Review*, *32*(4), 544-560. doi: 10.1177/0894439313510482

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

van der Linden, W. J. (2008). Using Response Times for Item Selection in Adaptive Testing, *Journal of Educational and Behavioral Statistics, 33*(1), 5–20. doi: 10.3102/1076998607302626

Wells, T., Bailey, J. T., & Link, M. W. (2014). Comparison of smartphone and online computer survey administration. *Social Science Computer Review, 32*(2), 238-255. doi: 10.1177/0894439313505829

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

Yan, T. & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1), 51-68. doi: 10.1002/acp.1331.

methods, data, analyses | Vol. 9(2), 2015, pp. 229-260 DOI: 10.12758/mda.2015.013

Are Sliders Too Slick for Surveys? An Experiment Comparing Slider and Radio Button Scales for Smartphone, Tablet and Computer Based Surveys

Trent D. Buskirk [1], *Ted Saunders* [2] *& Joey Michaud* [2]

1 Marketing Systems Group
2 MaritzCX

Abstract

The continued rise in smartphone penetration globally afford survey researchers with an unprecedented portal into personal survey data collection from respondents who could complete surveys from virtually any place at any time. While the basic research into optimizing the survey experience and data collection on mobile devices has continued to develop, there are still fundamental gaps in our knowledge of how to optimize certain types of questions in the mobile setting. In fact, survey researchers are still trying to understand which online design principles directly translate into presentation on mobile devices and which principles have to be modified to incorporate separate methods for these devices. One such area involves the use of input styles such as sliding scales that lend themselves to more touch centric input devices such as smartphones or tablets. Operationalizing these types of scales begs the question of an optimal starting position and whether these touch centric input styles are equally preferred by respondents using less touch capable devices. While an outside starting position seems optimal for slider questions completed via computer, this solution may not be optimal for completion via mobile devices as these devices are subjected to far more space and layout constraints compared to computers. This experiment moves the mixed device survey literature forward by directly comparing outcomes from respondents who completed a collection of survey scales using their smartphone, tablet or computer. Within each device, respondents were randomly assigned to complete one of 20 possible versions of scale items determined by a combination of three experimental factors including input style, length and number formatting. Results from this study suggest more weaknesses than strengths for using slider scales to collect survey data using mobile devices and also suggest that preference for these touch centric input styles varies across devices and may not be as high as the preference for the more traditional radio button style.

Keywords: Smartphone and Tablet Surveys, Slider Scales, Radio Buttons, Experimental Design, Missing Items

1 Introduction

The continued rise in smartphone penetration globally afford survey researchers with an unprecedented portal into personal survey data collection from respondents who could complete surveys from virtually any place at any time. Indeed, over the past five years, the research in online survey data collection has extended beyond computers to include both smartphones and tablets. Buskirk (2015) describes contemporary trends in survey optimization for these mobile devices but in short, some of the current approaches not only consider how to implement well-established online survey design principles for mobile devices, but also seek to understand which and how any of these principles need to be modified for mobile devices. Mobile devices, in general, represent a type of survey mode in which potential respondents have themselves gained extensive experience using – including checking emails, using apps and browsing the web (Link & Buskirk, 2013). One might conjecture that these respondent experiences might speak to a greater sense and expectation for websites, including survey websites, to be easy to navigate, engaging and interactive.

One type of survey question scale that has been touted in the recent literature as more engaging and more interactive than the more traditional radio button variety is the slider scale. Slider scales, unlike radio buttons, enable both animation and interactivity by requiring the respondent to touch or click a slider handle and slide or drag it along a fixed axis until it reaches the desired answer choice or level. While the usual application of sliders don't go as far as gamification (Keusch & Zhang, 2014), they have been purported to afford respondents a more engaging experience (Cape, 2009; Puleston, 2011). Two related types of scales that have also been explored recently in the survey literature are visual analog scales (VAS) and graphic response scales (GRS). Unlike sliders that usually have a dragging or sliding interactivity, visual analog scales ask the respondent to place a mark for their response along an axis that is anchored by two endpoints while graphic response scales ask respondents to place a mark along an axis that has graded semantic label anchors along the continuum in addition to the two endpoint anchors (Couper et al., 2006). From a required action perspective slider scales require a dragging action

Direct correspondence to
Trent D. Buskirk, Ph.D., Vice President of Statistics and Methodology, Marketing Systems Group
E-mail: TBuskirk@m-s-g.com

Acknowledgements: We would like to thank Michael Sherwood of MaritzCX for his assistance with this project. We would also like to thank Melanie Courtright of Research Now for her assistance in coordinating data collection from the panel. Finally, we would like to thank the very helpful team at Decipher who provided assistance hosting, programming and optimizing the survey.

while both VAS/GRS and radio button scales require a clicking action. From a precision perspective, both VAS/GRS and slider scales may be preferred to radio buttons since theoretically they allow a continuum of answer choices instead of a discrete collection. The category slider represents a more discrete version of slider scales that has gained in popularity as evidenced by ease of availability in widely available pre-package survey software. Much like how graphic response scales add specific descriptors to the underlying range, category sliders add descriptors to break up the underlying continuum of satisfaction, agreement or other construct being represented by the slider. The category sliders represent the "ordinary" response categories that are typically represented by a comparable radio button scale.

The relative merits of VAS/GRS, sliders and radio button scales have been previously explored in the online survey context for computers (Couper et al., 2006) and have recently been explored for both computers and mobile devices (Toepoel & Funke, 2014). More broadly, Sikkel et al. (2014) explored the relative merits of dragging and clicking operations for category sliders, among other scale types, in the context of online surveys completed by PC and find that dragging operations increase user engagement with the survey but only when they are used sparingly. As Derham (2011) pointed out, researchers must make many choices when considering slider scales and these choices can individually and collectively impact data quality. Roster et al. (2015) posited that the considerable variability in the utility of sliders in surveys observed across research studies is in part due to the many aspects of slider construction and presentation that could be considered including among others: scale length, whether the outcome is treated as continuous or discrete, variations of graphics, use of labels and slider starting position. By far the most common starting position that has been tested in the survey literature has been left starting position (see Toepoel & Funke, 2014; Roster et al., 2015; Funke et al., 2011; Sikkel et al., 2014; Buskirk & Andrus, 2014). Petersen et al. (2013) examined sliders with a left start for scale items that had no natural neutral position and a middle start for those with a neutral position but these two starting places were not compared to other possible positions. Slider orientation was examined by Funke et al. (2011) an no discernable differences other than time were noted for vertical versus horizontal versions of the slider and the comparable radio button scale was held at fixed length. Toepoel and Funke (2014) compared sliders and radio buttons based on scales having three different lengths (5, 7 and 11 point items) and found differences between slider and radio buttons for desktop respondents for 5 and 7 point scale items and for mobile respondents for the 11 point scale items. Cape (2008) conducted an experiment comparing four versions of slider scales that varied the formatting of the slider scale but kept the starting position (left most option) and the length of scale (5 point Likert) constant. The results indicated that while different

versions of the slider scale produced different response distributions, the overall mean scores across different versions of the slider scale were similar.

In this study we simultaneously compare three scale aspects for surveys items fielded across smartphones, tablets and computers. An equal number of respondents from each of these device types was recruited and then randomized to complete survey scale items whose format was determined by a combination of three experimental factors including input style, length and number formatting. This experiment moves the mixed device survey literature forward by directly comparing outcomes from respondents who completed a collection of survey scales using their smartphone, tablet or computer. The study also offers one of the more comprehensive comparisons of radio buttons to slider scales in terms of the number of simultaneous attributes of slider scale designs considered within one survey experiment.

2 Recruitment and Experimental Design

Participants for this study were recruited from Research Now's US consumer e-rewards panel which consists of nearly 2.5 million adults making it one of the largest sources of online responses in the U.S.[1] Survey invitations were sent to the panel soliciting participants to complete a short survey using either a smartphone, tablet or computer with the goal of recruiting at least 1,200 respondents from each device type which was tracked using the panelist's device user agent string (Callegaro, 2010). The overall survey consisted of up to 60 possible questions about automobile insurance satisfaction and was designed to be completed in no more than 10 minutes using a web browser. The survey was optimized for mobile devices and according to the taxonomy of Buskirk and Andrus (2012) the mobile versions would be considered active mobile browser surveys. The study fielded in the U.S. between April 4 and 11, 2014 and each respondent received an identical e-incentive that was comparable in value to other panel surveys of similar length.

Because the panel provider's members generally completed surveys online or via tablet computers, we could not randomize device type to each panelist as not all participating panelists had each of these devices. Instead, we allowed device type to be a natural or native blocking variable and made all experimental randomizations within each type of device separately and independently. Specifically, once a panelist clicked on the study link they were taken to an introduction page. At this point we tracked the device type using the device's user agent string (Callegaro, 2010). After clicking start on the introduction page, each panel respondent was then

1 Members of the e-rewards panel are recruited by invitation only from one of many participating partner loyalty programs and respondents who complete surveys while on this panel receive electronic credits that can later be redeemed for various rewards.

randomized to receive scale items for the experiment that were formatted according to one of five possible scale types including: standard radio buttons or sliders with either an outside, left, middle or right starting position as illustrated in Figure 1 A, C-F. Consistent with the recommendations made by Roster et al. (2015) we provided an additional instruction for respondents in any slider scale group to click on the slider handle if their answer was consistent with where the slider began (see Figure 1 C-F). Because this experiment was conducted within the scope of a market research study that required standard radio button scales to produce estimates, the randomization to the scale type used a 4:1 ratio within each type of device with 4 respondents being randomly assigned to standard radio buttons for every 1 randomly assigned to each type of slider scale. In addition to scale type, respondents were equally randomized to one of two scale lengths (5 point vs. 11 point) and equally randomized to one of two scale numbering formats (numbered versus not numbered). All 5-point scales were fully anchored with semantic labels and the numbered versions also included number values below each of the semantic labels (see Figure 1 E, G and A, C, respectively). All 11-point scales were end-anchored with semantic labels and the numbered versions contained number values for each possible choice ranging on the low end of 0 to the high end of 10 (see Figure 1 D and B, respectively). The slider starting position was also relative to the length of scale, so for example, middle start with 5 point scales placed the handle on option 3 and middle start with 11 point scales placed the handle on option 5.

We note that our sample is from an online data source and was not selected by probability sample and was not otherwise intended to represent the broader population of the U.S. But as others have also noted (Buskirk & Andrus, 2014; Couper et al., 2006) our intention here is to compare results across experimental factors (e.g. scale type, scale length and number formatting) as well as the blocking variable of device type. We also note that while some studies have randomized or assigned respondents to device (Peytchev & Hill, 2010; Scagnelli et al., 2012), we allowed respondents to self-select by device. In this way, the experiment is embedded in a setting that is natural to the respondent and likely more consistent with what might be found in practice with respondents completing online surveys using whatever device is available to them.

3 Survey Items and Measures

Twenty three of the 60 possible survey items were considered for this experiment. The remaining questions provided data for two other experiments, both of which have been reported elsewhere (see Buskirk, et al., 2014, Michaud, et al., 2014 and Courtright et al., 2014). The first survey item included in this experiment asked respondents to enter the total number of miles driven within the past year. If the

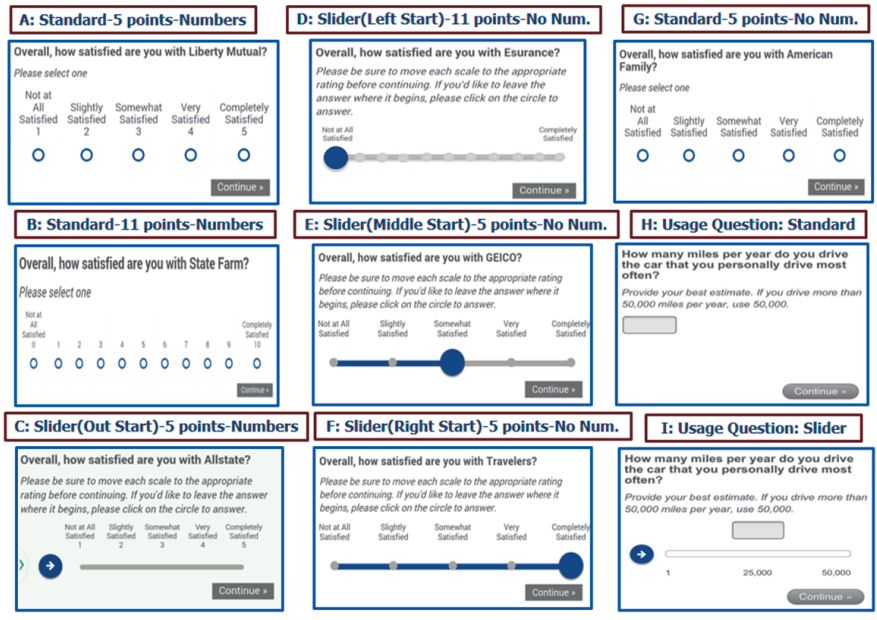


Figure 1 A-G – Visual Examples of the various factor combinations for the 22 core scale items as viewed on a smartphone. H and I refer to text input versus slider input for the single usage item (also viewed on a smartphone)

respondent was assigned to any one of the four slider scale groups, this question was presented as a slider with an outside start; otherwise, it was presented as an open-ended text box as illustrated in Figure 1 H and I, respectively. The remaining 22 questions (henceforth referred to as "core scale items" were presented over 7 separate screens and were organized into three different primary outcome measures including the: Overall Satisfaction Measure (OSM), Brand Performance Measure (BPM) and the Service Preferences Measure (SPM). The OSM was computed as the sum of three scale questions that asked respondents about their overall satisfaction with their Automobile Insurance Providers as well as how likely they were to recommend the provider to friends/colleagues and to renew their policies. The BPM was computed as the sum of ten scale items that asked respondents to rate their primary automobile insurance provider on ease of business transactions, trust, discounts, customer service, convenience, value, and accessibility using a scale that was based on anchors ranging from "Poor" to "Excellent." Finally, the SPM was computed as the sum of 9 scale items that asked respondents to rate the degree of agreement with statements about how they purchase automobile insurance, how they interact with an insurance agency, and the extent to which they want to use

mobile devices for their automobile insurance needs. Each of these scale items was anchored on endpoints that ranged from "Strongly Disagree" to "Strongly Agree." When we discuss the OSM, BPM and SPM measures throughout this paper we will add (5) or (11) to the abbreviation to refer to the number of scale points included in each of the scale items used to compute the measure. For example SPM(5)/SPM(11) refers to the service preference measure computed using scale items with 5 or 11 points, respectively. The actual values assigned to responses for 5 point scale items ranged from 1 to 5 and from 0 to 10 for 11 point scale items.

To examine both preference and consistency of reporting across scale types we also asked every respondent to answer the "overall satisfaction with their insurance provider" item (OSI) a second time at the end of the experiment using a scale presented with the opposite input style.[2] The scale numbering and length were the same across both OSI versions. After the respondent completed the second version of the OSI, they were asked "If you had the choice of how to give us your ratings, which way would you prefer?" with answer choices including "slider", "buttons" (i.e. radio) and "no preference." Using the two OSI items we also computed two versions of concordance. The first measure was simply a binary indicator for an exact match between the two responses (Exact Concordance). The second measure indicated concordance if the two responses differed by no more than 1 category unit up or down (±1Concordance).[3]

4 Analyses and Results

We note that for this study we are interested in comparisons across devices and across the other experimental factors as well as possible interaction effects between these factors for a series of survey related outcomes. At the extreme there could be a total of 60 unique cells, formed by crossing device (3) with scale type (5), scale length (2) and scale numbering (2), that would be compared by a model for any given outcome. Based on this extreme case, we attempted to cap the overall experiment-wise type I error rate to be at worst 30% for a given outcome by setting the *individual* type I error rate to be .005 (0.30/60). Thus for each specific survey outcome, the p-values reported in this paper are not adjusted further for multiple comparisons and we declared statistical significance for any effect or comparison if the unadjusted p-value was less than .005.

2 All respondents initially assigned to the "radio button" scale type were additionally randomly assigned in equal proportion to one of the four slider starting positions for the purposes of the preference and consistency analysis.

3 For example, a respondent who answered 7 for the first OSI and 8 for the second (using an 11 point scale) would not be concordant under exact concordance but would be under ±1Concordance.

4.1 Survey Break-Offs

In general the break-off rates for the experiment were moderately low across the three devices. In total, there were 1,250 computer, 1,340 tablet and 1,449 smartphone respondents who accepted our invitation to participate in the experiment and began the survey.[4] A total of 1,201 computer, 1,199 tablet and 1,198 smartphone respondents completed the experiment for respective break-off rates of 4% for computer, 11% for tablet and 17% for smartphones. While the results are not shown here we did examine break-off rates by the three experimental factors both within device and across devices and found no systematic pattern or practical differences.

4.2 Completion Times

We note that there were technical difficulties with the time tracking algorithm in the first day of fielding rendering the time stamps missing for all survey items for 369 of the 3,598 respondents across the three devices. The distribution of times to complete the single automobile usage item for the 3,229 respondents for which times were available was slightly positively skewed with extreme times observed from 20 PC respondents (2.3%) (exceeding 70 seconds), 30 Tablet respondents (2.6%) (exceeding 62 seconds) and 40 Smartphone respondents (3.4%) (exceeding 68.5 seconds). The longest time observed for this item was from a Tablet respondent who took in excess of 7,115 seconds (or just under 2 hours) to complete this question. Because of the observed skewness, we analyzed the natural log of completion times for the usage item based on a general linear model that includes device and the scale type (e.g. standard open ended text box versus slider-bar) as well as the interaction of these two factors. Based on the model we found that the completion times for the usage item (on the natural log scale) varied significantly by the device ($F(2,3223)=6.55$; p-value=.0015) and type of scale ($F(1, 3223)=27.30$; p-value<.0001). Despite the large outlying observation observed from a Tablet respondent, PC respondents had the largest geometric mean completion time for the usage item which was estimated to be about 9% longer than that from both Smartphone and Tablet respondents (p-values=.0022 and .0009, respectively) as illustrated in Table 1. No significant differences in the geometric means of completion times for the usage item were found between Tablet and Smartphone respondents

4 The total number of survey invitations sent from the sampling provider by device type was not available as device type was determined only upon clicking continue on the initial survey introduction page. A total of 323,259 email invitations were sent to panelists yielding 21,217 opened invitations, which included 441 partial completes/break-offs, 3,598 survey completes and 1,476 panelists who did not persist past the survey intro page. An additional 12,631 panelists opened and responded to the invitation and clicked continue on the survey introduction page but did so using a device for which the quota had already been met and as such were terminated.

Table 1 Descriptive statistics for completion times (rounded and displayed to the nearest second) for the miles driven last year question by device and scale type

Device / Scale Type	n	Mean	Geometric Mean	Median	Std. Deviation	Min.	Max.
PC	884	31	21	20	145	4	3980
Tablet	1161	31	19	18	223	6	7116
Smartphone	1184	27	19	17	81	4	2124
Standard (Radio Buttons)	1613	28	18	17	185	4	7116
Slider	1616	31	21	19	133	4	3980

(p-value>.75). The geometric mean completion time for the usage item for the slider scale group was also estimated to be about 12% longer than that for the standard text box group (p-value<.0001).

The distribution of completion times for the core scale items was also positively skewed for each of the three devices. Some extreme observations[5] were observed from respondents from each of the devices including 20 of the 884 (2.3%) completing via PCs, 29 of the 1,161 (2.5%) completing using Tablets and 49 of the 1,184 (4.1%) completing via Smartphones. Basic summary statistics for the completion times by device type are given in Table 2. We note that the ranges of completion times for PC and Tablet users were generally consistent overall and across scale types, but the range for Smartphone users was quite large in comparison driven by two respondents – one who took more than 68,375 seconds (or just under 19 hours) and the other who took more than 11,873 seconds (or about 3.3 hours) to complete the questions for the experiment on their smartphones. Given the underlying skewness in the distribution, the analysis of differences in completion times across device and the three experimental conditions was conducted using a general linear model applied to the natural log of completion times. We note that the statistical comparisons of completion times for the experiment on the natural log scale were practically identical with and without these two very extreme outliers, so for posterity all analyses included these data points.

Completion times (on the natural log scale) varied significantly by both the device used for completing the survey ($F(2, 3169)=27.27$; p-value<.0001) and by

5 Defined as exceeding 3 times the interquartile range plus the third quartile of completion times, recorded in seconds. Specifically, identified as completion times exceeding 352, 321 and 343 seconds for PC, Tablet and Smartphone respondents, respectively.

Table 2 Time (in seconds) to complete the core scale items (22 questions) by mode of response

Device / Scale Type	n	Mean	Geometric Mean	Median	Std. Deviation	Min.	Max.
PC	884	164	141	137	252	18	6655
Tablet	1161	134	116	110	118	17	1787
Smartphone	1184	242	124	115	2052	20	68376
Standard (Radio Buttons)	1613	202	125	118	1731	30	68376
Slider:Out	386	168	133	126	357	37	6655
Slider:Left	405	190	134	128	617	35	11873
Slider:Mid	396	149	122	114	191	17	2413
Slider:Right	429	139	118	121	132	20	2220

All analyses and statistical hypothesis tests were performed on the natural log scale so we also provide geometric means, since back-transformed means from the natural log scale estimate the geometric means from the raw, untransformed data. All times are rounded and displayed to the nearest second.

the scale type (F(4, 3169)=3.85; p-value=.0040) and these effects were additive in that no interaction between these two factors was detected. None of the other factors nor any second or higher order interactions were significantly related to the natural log of completion times (all remaining p-values >.10). The geometric mean completion time for PC respondents was estimated to be about 19% longer than that of Smartphone respondents (p-value <.0001) and estimated to be about 23% longer than that of Tablet respondents (p-value<.0001). No significant differences were found in completion times for the core scale items between Smartphone and Tablet respondents (p-value>.01). Respondents assigned to the slider left start group had the longest estimated geometric mean completion time (about 135 seconds, on average) and the geometric mean completion time for this group was estimated to be about 11% longer than that for the slider right start group (p-value=.0024). No other significant differences in completion times were found between any of the other scale types.

4.3 Missing Item Rates

While missing values were generally more of the exception than the rule for core scale items, some amount of item missingness was encountered. All in all, roughly 66% of respondents had no missing items for any of the core scale items. Among the third of respondents missing at least one core scale item, the 25th percentile of the number of items missing was 1, the median was 4, the 75th percentile was 9 and the 95th percentile of the number missing was 17. In total, 13 respondents were missing all core scale items. From the negative binomial regression model that explored the number of missing items as a function of device type and the experimental factors and higher order interactions, we determined that the variability in the number of missing items was fundamentally driven by scale type ($\chi^2(4)=2052.12$; p-value<.0001) but the impact of this factor was moderated separately by both scale length ($\chi^2(4)=36.68$; p-value<.0001) and also by device ($\chi^2(8)=43.30$; p-value<.0001).

Essentially, the slider right and middle start groups had significantly higher numbers of missing values, on average, compared to any of the other scale types and the number of missing items is practically (and statistically) consistent across the devices for each of the scale types. The main exception to this trend for device types comes from the slider right start group as shown in Figure 2 B. For this scale type we observed that Smarpthone respondents exhibited significantly higher numbers of missing items, on average, compared to PC respondents (p-value<.0001) but no significant differences were observed between Tablet or PC respondents (p-value>.02) nor between Smartphone and Tablet respondents (p-value>.05). As shown in Figure 2 A, the number of missing items, on average, was fairly consistent across the two scale lengths with the exception being found for the slider middle start group. Here respondents assigned to the 5 point scales had an average number of missing items that was about 75% larger than the 11 point scale group (p-value<.0001) which translated into about 3 additional missing items, on average. The number of missing items for respondents assigned to the 5 point version of the slider right start group exhibited about the same number of missing items than the slider middle start group (p-value>.42), but the number of missing items for the 11 point slider right start group was about twice as large as the 11 point slider middle start group (p-value<.0001). Overall, there was no difference in the number of missing items, on average for either the 5 point or 11 point versions of the slider right start groups (p-value >.30) and this scale type had the largest number of missing items on average (about 8 for the 5 point and 7 for the 11 point versions).

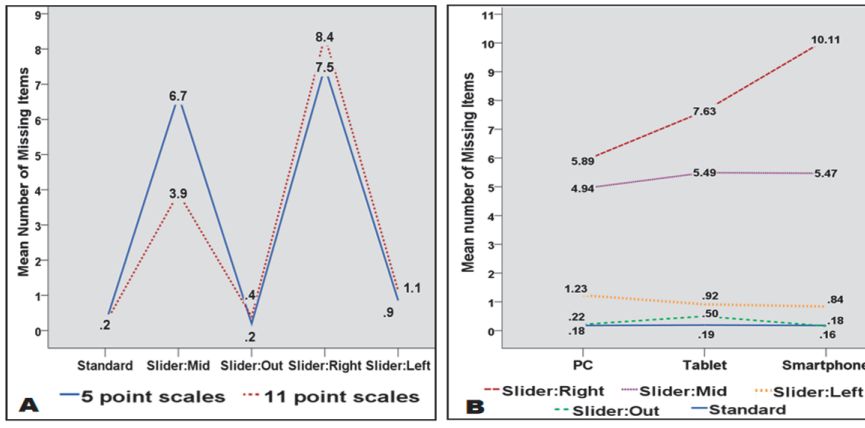


Figure 2 A: Mean number of missing items by scale type and scale length; B: Mean number of missing items by device type and scale type

4.4 Survey Outcomes

4.4.1 Miles driven in the past year

Respondents entered the number of miles they drove within the past year using either a slider scale or an open ended text box. One aspect of sliders that differs from open ended texts, especially for numeric data is that sliders give the respondents a clear sense of the range with labels marking the beginning and ending points of the slider as illustrated in Figure 1 I. Open ended text boxes, on the other hand, can also provide respondents a sense of the range if explicit instructions are included as illustrated in Figure 1 H. Because the slider endpoints are more explicit we expected that more respondents in the slider group would enter values corresponding to the upper or lower endpoints compared to the text group. However, we found no significant differences between the slider and text groups for either the rate of respondents reporting the highest option (i.e. 50,000) or the lowest option (i.e. 0) (both p-values>.47). On the other hand, there were significant differences noted in the proportions of respondents reporting the highest option across devices (p-value<.003) with more Smartphone respondents reporting the maximum allowable amount compared to either PC or Tablet respondents as shown in the rightmost section of Table 3.

To compensate for the positive skewness observed in the miles driven distribution, we analyzed the relationship between the natural logarithm of 1 plus the miles driven and input style, device type and the interaction of these two factors

Table 3 Summary statistics for the number of miles driven in the past year by device type and entry method. Overall statistics for the entire sample are given in the bottom right-hand corner

Input Method	Device type	n	Mean	Std. deviation	Median	Geometric mean	% at 0	% at 50K
Text entry (standard)	PC	602	11327.13	7425.48	10000	8668.09	0.17%	0.50%
	Tablet	595	13442.34	8332.71	12000	10637.59	0%	0.67%
	Smartphone	587	14639.12	9752.80	12000	10192.01	0.67%	2.35%
	Statistics for text entry (across all devices)	1784	13122.36	8652.05	12000	9788.66	0.28%	1.17%
Slider entry	PC	597	11630.87	7368.88	10000	9383.08	0.17%	1.01%
	Tablet	600	13826.69	8708.15	12000	10876.36	0.33%	1.33%
	Smartphone	603	15544.23	9935.90	12500	11373.86	0.50%	2.16%
	Statistics for slider entry (across all devices)	1800	13673.79	8879.16	12000	10512.80	0.33%	1.50%
Summary statistics for devices (across entry types)	PC	1199	11478.37	7395.83	10000	9017.01	0.17%	0.75%
	Tablet	1195	13635.32	8521.88	12000	10756.81	0.17%	1.00%
	Smartphone	1190	15097.76	9852.27	12000	10774.68	0.58%	2.25%
Overall summary statistics for all respondents		3584	13399.30	8769.96	12000	10145.89	0.31%	1.33%

using a general linear model.[6] As was the case for completion times for the experiment, differences in the natural log of miles driven (plus one) varied significantly across device type ($F(2, 3578)=11.38$; p-value<.0001) but not by the style of input ($F(1, 3578)=4.06$; p-value>.04) nor by the interaction of device and input style ($F(2, 3578)=.54$; p-value>0.50). In particular, the geometric mean for the miles driven (plus one) for PC respondents was estimated to be approximately 16% less than that of either Tablet or Smartphone respondents who reported geometric means of roughly 10,757 and 10,775 miles driven within the past year, respectively (p-values<.0001).

4.4.2 High, middle and low option selection patterns for core scale items

Before examining specific substantive outcomes, we first explored general response patterns classified as the selection of "high", "middle" and "low" box options for each of the core scale items. On the five point scale we declared that the respondent selected a: "high option" if their response was either a 4 or 5; a "middle option" if their response was a 3 and a "low option" if their response was either a 1 or 2. For the 11 point scale, "high" options were defined as responses between 7 and 10; "middle" as responses between 4 and 6 and "low" for responses between 0 and 3. We created three separate models to examine the relationship between the selection rates of high, middle and low response options for core scale items and the three experimental factors, device type and all higher order interactions. To adequately compensate for observed over-dispersion for each of these three rates, we used negative binomial regression models with an offset equal to the natural log of the number of core scale items answered.

High option selection rates

High option selection rates varied significantly across scale type ($\chi^2(4)=147.72$; p-value<.0001) and scale length ($\chi^2(1)=20.07$; p-value<.0001) and by the interaction of these two effects ($\chi^2(4)=30.23$; p-value<.0001). Neither the main effects of scale numbering nor device type nor any of the other interaction effects were found to be significant (all p-values >.12). In general we found that respondents in the slider middle and right start groups had significantly higher and lower, respectively, high option selection rates compared to other scale type groups as depicted by the solid red lines in Differences between scale type groups for the 11 point scale items

6 We added 1 to all reported miles to avoid irregularities in the natural logarithmic transformation applied to the rather small number of zeroes that were reported for miles driven (Yamamura, 1999).

were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3. Specifically, among respondents assigned to 5-point scales in the slider middle start group selected higher response options at rates that were, on average, nearly 30% more than that those of respondents in the slider left start, slider outside start and standard scale groups (all p-values<.0001). In contrast, respondents assigned to the slider right start group had estimated high option selection rates that were, on average, about 15% less than those of the slider left start, slider outside start and standard scale groups and about 34% less than those of the slider middle start group (all p-values<.0011). The differences observed for the 5 point scale items were generally consistent for the 11 point scale items, as shown in the right panel of Differences between scale type groups for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3, although the magnitude of differences was less and the number of significant differences fewer[7].

Middle option selection rates

The middle option selection rates varied significantly across device type ($\chi^2(2)=13.74$; p-value=.0010), scale type ($\chi^2(4)=483.73$; p-value<.0001), scale length ($\chi^2(1)=8.18$; p-value=.0042) and the interaction of scale type and length ($\chi^2(4)=175.90$; p-value<.0001). Neither the main effect of scale numbering nor any of the other interaction effects from the full model were found to be significant (all p-values>.08). As indicted in Table 4, PC respondents selected middle response options about 14% less often than those for respondents completing by Smartphone, but no other significant differences across devices were noted (p-values>.015). As for scale type differences, generally respondents from the middle slider start group exhibited far lower middle option selection rates compared to any other scale type as depicted by the long-dashed green line in the left and right panels of Differences between scale type groups for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3. More specifically, for the 5 point scale items, respondents in the middle slider start group had middle options selection rates that were, on average, about 85% less than those of respondents from the slider left, outside and right start as well as the standard scale groups (all p-values<.0001). Differences between

7 More specifically, respondents from either the right or outside slider start groups had high option selection rates that were, on average, about 10% less than those of the slider left start and standard scale groups and approximately 20% less than those of the slider middle start group (all p-values<.003) and no other significant differences between scale types were found for the 11 point scales (all p-values>.026).

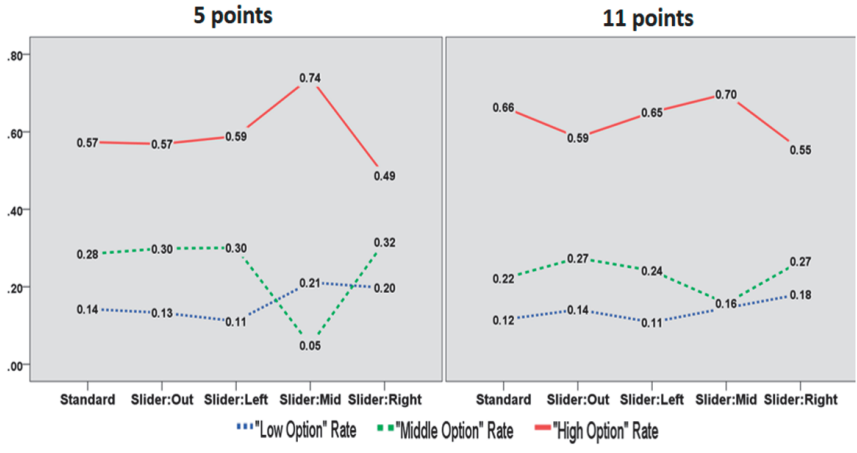


Figure 3 Low, middle and high option selection rates by type and length of scales for core scale items answered

scale type groups for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less[8].

Low Option Selection Rates

The low option selection rates varied significantly across device type ($\chi^2(2)=36.92$; p-value<.0001), scale type ($\chi^2(4)=98.59$; p-value<.0001) and scale length ($\chi^2(1)=16.32$; p-value<.0001) as well as the interaction between scale type and length ($\chi^2(4)=17.98$; p-value=.0012). Neither the main effect of scale numbering nor any of the other interaction effects were found to be significant (all p-values≥.02). As shown in Table 4, PC respondent had low option selection rates that were, on average, about 16% and 30% higher than those of Tablet and Smartphone respondents, respectively (p-values<.0003). With respect to differences in the low option selection rates across scale types, we found that generally respondents in the left slider start group had significantly lower rates while respondents in the middle and right slider start groups had significantly higher rates compared to other scale types as depicted by the blue short-dashed lines in Differences between scale type groups

8 In particular, respondents in the middle start group selected middle response options at rates that were, on average, about 35% less than those for either the slider left start or standard scale groups (p-values<.0001) and about 45% less than those for either the slider outside or right start groups (p-values<.0001). The middle option selection rates for the standard scale group were also about 20% lower, on average, than those of either the slider outside or right start groups (p-values<.0007).

Table 4 Selection of "Low", "Middle" or "High" options across core scale items by device

	Type of device		
	PC (n=1200)	Tablet (n=1192)	Smartphone (n=1193)
Option selection rate	Mean (std. error)	Mean (std. error)	Mean (std. error)
„Low Option"	0.162 (0.004)$^{\dagger,\clubsuit}$	0.135 (0.004)‡	0.126 (0.004)
„Middle Option"	0.224 (0.005)$^{n.s.,\clubsuit}$	0.255 (0.006)$^{n.s.}$	0.259 (0.006)
„High Option"	0.614 (0.006)$^{n.s.,\ n.s.}$	0.610 (0.007)$^{n.s.}$	0.615 (0.007)

† indicates PC user rate is significantly different from Tablet user rate (α=.005)
❖ indicates PC user rate is significantly different from Smartphone user rate (α=.005)
‡ indicates Tablet user rate significantly different from Smartphone user rate (α=.005)
n.s. indicates corresponding comparison is not statistically significant

for the 11 point scale items were generally consistent with those observed for the 5 point scale, although the magnitude of these differences was generally less.

Figure 3. Among respondents randomly assigned to 5 point scales, the slider left start group had low option selection rates that were, on average, at least 40% less than those for the slider bar middle or right start groups (p-values<.0001) and 23% less than those for the standard scale group (p-value=.0001). Respondents in both the slider middle and right start groups had low option selection rates that were, on average, at least 40% higher than those of the slider outside start group (p-values<.0001) and at least 29% higher than those of the standard scale group (p-values <.0008). The pattern of differences across scale types for the 11 point scale items was generally consistent with the findings for the 5 point items, although the overall magnitude of differences was generally lower and the number of significant differences fewer.[9]

9 No significant differences were noted between the slider right, middle and outside start groups (all p-values>.08) nor between the slider left start and standard scale groups (p-value=.230). Respondents assigned to either the slider left start or standard scale groups had low option selection rates that were, on average, at least 20% less than those of either the slider middle or outside start groups (p-values<.0011) and at least 30% less than those of the slider right start group (p-values<.0001).

4.4.3 Satisfaction, brand performance and service preference measures

To explore how the patterns in response option selections might translate into differences in the actual substantive measures of interest, we also examined the relationship between the OSM, BPM and SPM measures and device type, scale type, and scale numbering along with all possible higher order interactions using general linear models computed separately for each measure at each scale length. Normality assumptions were investigated for each of these scales across the experimental conditions and no major issues were detected. The overall reliability for both the five and 11 point scale versions of the OSM and BPM measures, as measured by Chrombach's alpha, exceeded .90 with very little practical variability across the devices. Lower reliability measures were observed for both the SPM(5) and SPM(11) measures (.67 and .72, respectively) but again, very little practical differences in the reliability statistics were observed across the devices.[10]

Due to space considerations we now provide an overall summary of the separate models followed by more specific details for the analyses pertaining to the Brand Performance Measure (BPM). Additional information about any of the models can be obtained upon request from the lead author.

The profile plots for the overall means for the OSM, BPM and SPM outcome measures by scale type and device are displayed separately by scale length in Figure 4. Generally the OSM(5), OSM(11) and BPM(5) measures varied significantly across both scale type and device as main effects. As displayed in Figure 4 A, B and D, PC respondents reported, on average, higher values of these measures compared to Smartphone and Tablet respondents. Moreover, respondents in the slider right start group reported significantly lower measures, on average, than those in the slider middle start group, but both of these groups had significantly lower measures, on average, compared to those for the slider left and outside start and standard scale groups. Similar patterns in differences across scale types were also observed for the BPM(11), SPM(5) and SPM(11) outcome measures, but the but the magnitude and direction of the differences was impacted by th e specific combination of scale type and device (e.g. significant interaction between scale type and device in the models for these outcomes) as depicted in Figure 4 C, E and F. Overall, the findings for both the 5 and 11 point versions of the three scale measures were generally consistent with those reported for the middle and high response selection rate analyses.

10 Lower reliability for the SPM is likely related to the inclusion of at least two items that asked respondents about service preferences that were in direct contrast to one another – namely one item asked whether or not a respondent preferred to work with the insurance agent directly and another question asked whether they would prefer to interact with the insurance company directly without going through an agent.

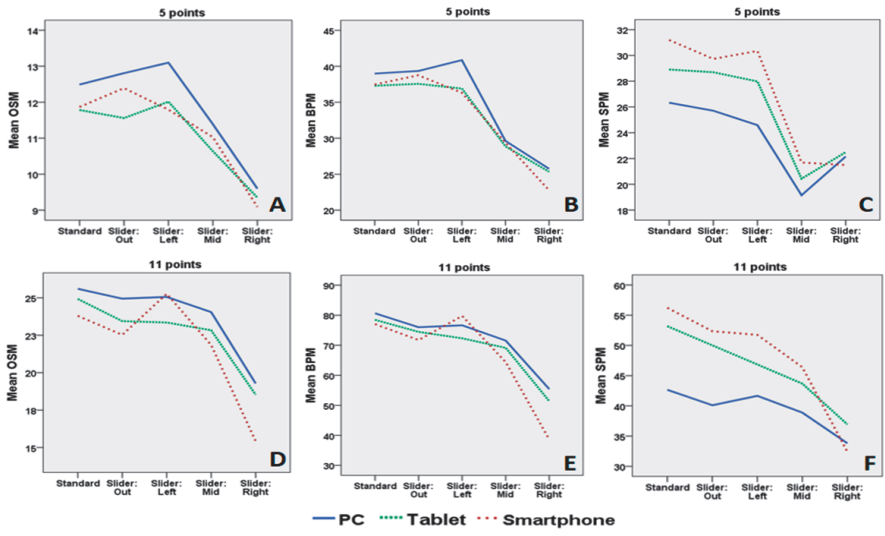


Figure 4 Mean values for the OSM (A (5 point) and D (11 point)), BPM (B (5 point) and E (11 point)) and SPM (C (5 point) and F (11 point)) measures for each scale type and device by scale length

The brand performance measure (BPM)

We found that BPM(5) values varied significantly by scale type (F(4, 1718)=115.17; p-value<.0001) and marginally significantly by device type (F(2, 1718)=5.22; p-value=.0055). None of the other main effects nor any of their interactions were found to be significant (all p-values >.22). As suggested by mean profile plot provided in Figure 4 B, on average PC respondents had BPM(5) values that were estimated to be about 2 scale units higher than those for Smartphone respondents (p-value=.0027) and no significant differences were detected between any other pairs of devices (p-values>.01). Estimated differences in BPM(5) values between scale types were notably larger than those across devices. The average BPM(5) value for the slider right start group was estimated to be roughly 14 units lower than the slider left and outside start and the standard scale groups, about 5 units lower than the slider middle start group (all p-values <.0001). The average BPM(5) value for the middle start group was also estimated to be about 9 points lower than the slider left and outside start groups as well as the standard scale group (all p-values<.0001) and no other significant differences between pairs of scale types were noted.

Differences in scale type for BPM(11), while generally consistent with those found for BPM(5), were moderated by the device used to complete the survey. In particular, we found that BPM(11) values varied significantly by device (F(2,

1726)=8.58; p-value=.0002) and type of scale (F(4,1726)=102.71; p-value<.0001) but also by the interaction of device and scale type (F(8, 1726)=3.26; p-value=.0011). Generally speaking, BPM(11) values were higher for PC respondents followed by Tablet, and then Smartphone respondents on all scale types except the slider left start group which was higher for Smartphone respondents on average, as indicated in Figure 4 E. As for scale types, the slider right start group had significantly lower BPM(11) values, on average, compared to any of the other scale types, but these differences varied in magnitude depending on the type of device. For example, for Smartphone respondents, the slider right start group had an estimated BPM(11) average value that was about 41 units lower than the slider left start group, 38 units lower than the standard scale groups and 33 units lower than the slider outside start groups. The differences in these groups for Tablet users was estimated to be 27, 21 and 23 units, respectively and for PC respondents 25, 21 and 20 units, respectively (all p-values <.0001).

The slider right start group also had significantly lower BPM(11) values, on average, compared to those for the slider middle start groups, but the magnitude of the estimated differences varied from 26 units for Smartphone respondents to 18 units for Tablet respondents to 16 units for PC respondents (all p-values <.0001). Significant differences were also noted for BPM(11) values between the slider middle start and standard group across the three devices and between the slider middle and left start groups for Smartphone respondents. The degree of these differences varied across the devices.[11]

4.4.4 Imputed versions of survey measures using slider starting position

The pattern of differences in the OSM, BPM and SPM measures across both scale type and device is generally consistent with the overall missing item patterns for the core scale items – namely more missing items for the middle and right slider positions with the degree varying by device type. For negatively skewed scale items, it seems reasonable that sliders with a right or middle starting position might have indicated the respondents' desired answer choices more consistently, and as such, respondents might not have realized a need to do anything more to register these choices but to click the "continue" button. To better understand whether some of the differences observed in the three outcome measures could be explained or

11 The slider middle start group also produced significantly lower BPM(11) values, on average, compared to the standard scale group across all three devices with the magnitude of the difference varying from 12 units for Smartphone respondents (p-value<.0001) and 9 units for both Tablet (p-value=.0002) and PC (p-value=.0014) respondents. Finally, the slider middle start group was found to be about 14 points lower, on average, compared to the slider left start group among Smartphone respondents (p-value<.0001).

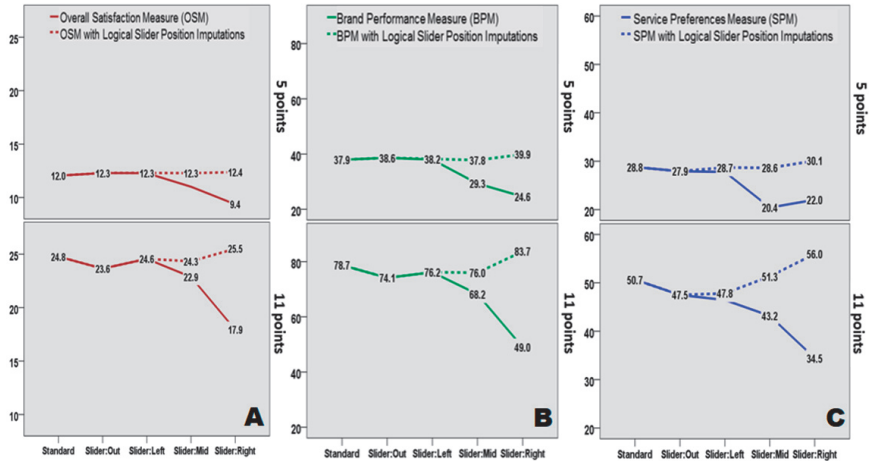


Figure 5 Summary statistics for the three survey outcome measures (A: OSM; B: SPM; and C: BPM) by scale length and scale type (solid lines) and their imputed versions (dashed lines)

adjusted for the impact of item missingness, we imputed the response value that corresponded to the slider's starting position whenever a respondent had a missing item for that scale item. The means for the recomputed "imputed" versions of the OSM, BPM and SPM measures (plotted as dashed lines) are displayed along with those from the original versions (plotted as solid lines) in Figure 5 A, B and C, respectively. For simplicity of display, these plots and analyses aggregated scale measures across device type.

What becomes quickly apparent from Figure 5 for each of the three measures across both the 5 and 11 point scale items is the considerably lower scale values of both the middle and the right slider start groups across for respondents for which scale measures could be computed. These figures represent the key findings of the last section with respect to scale type. What is also apparent is that the imputed versions for each of the three outcome measures generally fall more in line across the scale types. More specifically, for the OSM scale, there were no practical differences across scale types using the imputed versions for both the 5 and 11 point scales as seen in Figure 5 A (top and bottom, respectively). The imputed 5 point version of the BPM still had significant differences between the slider right start group and all other scale types but these differences were practically negligible; moreover no differences were detected between the slider middle start group and any of the other scale types, except for the slider right start group. A similar pattern was found for the 11 point BPM version as well except that the imputed version was significantly higher for the slider right start group compared to all the other scale

types, but the magnitude of the overall differences has been attenuated. Finally for the SPM there are still significant differences between the right slider start scale type and the other scale types for both 5 and 11 point versions but the differences for the 5 point version are now practically negligible. The 11 point versions for the slider middle and slider right groups are still significantly different from the other groups, but the direction has also been reversed and the magnitude has decreased.

4.4.5 Preference for slider scales

To better understand preference and consistency rates (in the next subsection), the scale type factor was separated into two variables – scale input style (e.g. slider or radio button) and slider start position (e.g. outside, left, middle and right). Scale input style specifies the order in which the two versions of the OSI were presented – if scale input style is "slider" then the first OSI (and all other core scale items) was presented on the slider scale using a start position dictated by the slider position variable and the second OSI was presented using radio buttons and vice versa for the "radio buttons" input style. Preference rates for the slider input style were analyzed based on 2,649 (74%) respondents who declared a definitive preference for one of the two input styles using a logistic regression model that included device type, scale length, scale input style and slider position and all higher order interactions among these factors. The scale numbering factor was not included in this analysis to avoid possible sample size issues in the logistic regression model that incorporated the additional scale input style factor and was based only on those respondents who indicated a preference for one of the two input styles.[12]

Preference rates for sliders scales across device and scale input style are given in left side of Table 5 and in total, of the 2,649 respondents included in the analysis, 43% expressed a preference for slider scales. These preference rates varied significantly by both survey scale input style ($\chi^2(1)=319.73$; p-value<.0001) and device type ($\chi^2(2)=202.54$; p-value<.0001) but the differences across device were moderated by both the scale input style ($\chi^2(2)=15.69$; p-value<.0005) and the slider starting position ($\chi^2(3)=20.10$; p-value<.0003). None of the other main effects or higher order interactions were significant (all p-values >.04). The odds for preferring slider scales versus radio buttons across devices showed the same general pattern but were generally larger among respondents who completed the core scale items using slider scales compared to radio button scales. Smartphone and Tablet respondents completing the core scale items using slider scales had significantly

12 We examined slider preferences across the levels of the scale numbering factor as well as separately by device type, slider start position groups and levels of the scale length factor and found no significant differences in the slider preference rates (all p-values>.05). Thus we suspect that pooling across scale numbering would likely have little impact on the substantive findings from the model.

Table 5 Preference rates for the slider input style based on 2,649 respondents who declared a definitive preference for one of the two input styles. Left: Preference for sliders by device type and input style; Right: Slider bar preferences by slider starting position

	Input style used to complete core scale items				Statistics for each device type		Slider input preference by slider start position for the slider version of the overall satisfaction item		
	Radio buttons		Sliders						
Device type	n	Prefer sliders (%)	n	Prefer sliders (%)	n	Prefer sliders (%)	Slider starting position	n	Prefer sliders (%)
PC	416	12.26	383	31.33	799	21.40	Outside	668	35.33
Tablet	455	26.81	443	71.33	898	48.78	Left	636	47.64
Smartphone	459	33.12	493	78.70	952	56.72	Mid	659	47.34
Statistics for each input style	1330	24.44	1319	62.47	2649	43.37	Right	686	43.44

higher odds of preferring a slider scales than PC respondents completing core scale items using sliders (p-values<.0001) with the odds of preferring sliders for Smartphone and Tablet respondents being an estimated 8.8 and 5.9 times the odds for PC respondents, respectively. Among the PC respondents assigned to complete core scale items using slider scales, we note that just less than one-third actually preferred sliders, but nearly three quarters of Smartphone and Tablet respondents assigned to slider scales for the core scale items expressed a preference for sliders over radio buttons (left side of Table *5*). There was also no significant difference in the odds for preferring the slider input style to radio buttons between Smartphone and Tablet respondents completing survey scale items using sliders (p-value>.01). Among those assigned to the radio buttons survey input style significant differences in the odds of preferring slider versus radio buttons were also observed between PC respondents and both Smartphone and Tablet respondents (p-values<.0001) but not between Smartphone and Tablet respondents (p-value>.045). In particular, the odds for preferring slider scales for Smartphone and Tablet respondents were estimated to be 3.5 and 2.6 times that of PC respondents, respectively.

Differences in the odds for preferring slider input to radio button input were also observed between the different starting positions for the slider scales as indicated in the right side of Table 5. In particular the odds for preferring slider input styles among respondents with a left or middle starting slider scale were estimated

to be about 1.7 times those for respondents using a slider scale with an outside start (both p-values=.0001). No significant differences in the odds of preferring slider input styles were observed among respondents completing survey items using slider scales with a left, middle or right starting position (all p-values >.04).

4.4.6 Consistency of responses across slider and radio button scales

From the 3,190 respondents for which concordance measures could be calculated, the exact concordance rate was 68.2% and the ±1concordance rate was 94.2%.[13] Concordance rates using both measures are given in Table 6 by device type and the experimental factors. From the logistic regression model relating exact concordance to device type, slider input style, scale length and slider position and scale numbering we found that these rates varied significantly by device ($\chi^2(2)=20.516$; p-value<.0001) and by scale length ($\chi^2(1)=175.811$; p-value<.0001). The exact concordance rates were not statistically different by scale input style, slider position or scale numbering and none of the higher order interactions between these and other effects were significant (all p-values>.024). The odds for exact concordance for PC respondents were approximately 1.6 times those for Smartphone respondents (p-value<.0001) and about 1.4 times those for Tablet respondents (p-value=.0012). No significant differences were noted for the odds for exact concordance between Smartphone and Tablet respondents (p-value=.2250). The odds for exact concordance for respondents assigned to the 5 point version of the OSI were estimated to be about 3.1 times those for respondents assigned to the 11 point version of the OSI (p-value<.0001) and these differences were consistent across device types.

13 There were 210 respondents who did not answer the first Overall Satisfaction Item (OSI) and another 198 who did not answer the second OSI version. A majority of these missing items come from the slider right starting position group compared to the other starting positions and from respondents completing the survey by smartphone compared to other devices.

Table 6 Observed concordance rates between the overall satisfaction item presented as part of the main survey and again in an alternate format at the end of the survey. The value for the two items matched exactly for the exact concordance rates and matched up to 1 scale unit up or down for the second concordance measure

		Concordance rate between the two versions of the overall satisfaction item	
Group / Experimental factor	n	Exact	±1Concordance
Device type			
PC	1122	73.26%	97.06%
Tablet	1053	66.57%	93.92%
Smartphone	1015	64.24%	91.33%
Slider start position			
Outside	884	67.99%	93.21%
Left	882	67.57%	94.10%
Mid	790	70.51%	94.68%
Right	634	66.40%	95.11%
Scale length			
5 items	1576	80.27%	99.43%
11 items	1614	56.38%	89.10%
Scale numbering			
Numbered	1621	69.96%	95.56%
Not numbered	1569	66.35%	92.80%
Input style for core scale items			
Radio buttons	1610	68.63%	94.53%
Sliders	1580	67.72%	93.86%

5 Discussion

Several studies have found that slider scales, while engaging, can take longer to complete than comparable traditional radio button scales (Sikkel et al., 2014; Roster et al., 2015; Husser & Fernandez, 2013; Funke et al., 2011, among others). However in many of these studies, radio button completion times were compared to sliders with a left starting position. Our results for the completion times for the single continuous item "number of miles driven in the past year" were consistent with these studies in that the slider group had completion times that were longer, on average,

compared to the group which entered their responses directly into an open-ended text box. Our results, for sliders with a left start also echo the findings from prior research in direction but the differences we observed were not statistically significant[14]. However, our findings for the other slider start positions, including most notably sliders with a right or middle starting position were in the opposite direction in that we found completion times for respondents in these two groups to be shorter than those for the standard scales, albeit not statistically significantly different. This opposing result could be directly related to the fact that we observed higher missing items from respondents from both the middle and right starting slider scale groups. In some cases, respondents in the right slider start group who were highly satisfied with their insurance provider might have taken much less time to answer the satisfaction questions simply because their responses corresponded to the slider starting position. As such respondents may not have taken the time to click on each item, but instead hit the next button for the survey to continue, resulting in missing data.

Throughout this paper we have presented empirical evidence showing that the slider starting position can greatly affect the amount of missing items and could impact measurement. As Funke et al. (2011) note "if the handle is placed at the position of a valid answer, intentional response and non-response cannot be distinguished." One starting position that would avoid this issue is outside or off of the slider itself. However, this choice requires more space for the overall slider graphic. While making the slider handle smaller to create more room for the actual slider bar itself might work for mouse interfaces, it might be less optimal for interfaces that rely on finger taps. In our study we also found that respondents completing scale items using an outside starting slider were the *least likely* to prefer slider scales compared to any other starting position[15].

Another option to remedy the missing item issue might require respondents to move the slider away from its starting position and then back to the response category to register the response. Such a requirement would however increase the num-

14 We note had our study used the same Type I error rate for declaring significance as used in both of these studies (α=.05), then we would have also declared differences in completion times to be significantly lower for the left slider start group compared to the radio button group. Moreover, our results were based on the Geometric mean (natural logarithm transformed completion times) rather than the arithmetic mean and our analyses did not eliminate any outliers.

15 Certainly a plausible factor in preference, or lack thereof, for slider scales with an outside start could be related to poor operationalization of this type of slider (slider handle doesn't appear in its entirety on the screen or isn't responsive to respondents actions). However, we believe this factor should contribute as most minimally given that we made every effort possible in the programming phase to ensure that this specific slider scale would be optimized for all three devices including positioning and sizing the slider handle in such a way that it would appear wholly on the screen and not interfere with the legibility of the scale point labels and numbers as displayed in Figure 1: C.

ber of taps required to complete the question from one to two for the slider scales compared to what is required for the radio button scale (Buskirk, 2015b). Such an approach was used by Sellers (2013) who compared slider bars scales with middle, left and right starts to radio buttons. They found that with a forced choice requirement, respondents in the right slider group reported higher right choice options and respondents in the left choice group reported more lower choice options compared to respondents in other groups. Contrary to the method employed by Sellers, we did not force respondents to confirm answer choices for which the slider was neither moved nor clicked and we observed that respondents in the middle and right slider start groups tended to select these answer categories significantly *less often* than any other scale group. Respondents in the right start slider scale group who registered answers for scale items moved the slider away from the starting position but ultimately did not move it back. This pattern was generally consistent across the three devices and both scale lengths; however, the pattern was much stronger with the shorter version of the scale. More specifically, the high option selection rates for those assigned to 5 point scales with middle slider scales were 25% higher than those from any other scale group. Respondents seeing 5 point scale items in the right slider group selected higher categories at rates that were between 8 to 50% *less* than those of any of the other scale groups. We also found that respondents in the middle slider start group also chose lower end options more often than any other scale type except the right slider start group. This finding replicates the pattern observed by Petersen et al. (2013) who reported higher amounts of "2s" and "4s" being selected on a five point slider scale that had a middle start compared to other non-slider presentations. The similarity in the percentage of respondents in the left starting slider and radio button groups choosing higher options for the core scale items echoes what Cape (2009) found in a study comparing left starting sliders with different labelling options to more traditional radio buttons. Specifically, Cape (2009) found that while distributional differences were noted for survey outcomes across different scale types, the "box top" or percentage agreeing with a statement, were nearly identical across the scale types. However, in our study we also saw contrasting results between the radio button group and both the middle slider group where, respondents had significantly higher "box top" rates, and the right slider group, where respondents exhibited significantly lower "box top rates."

In addition to differences in response options and survey outcome measures, we also found differences in preferences for the slider scales. Such differences in preference rates by scale input style might reflect more of a conditioning effect in that respondents may likely prefer what they are comfortable with rather than something new. We expected that some respondents with radio button survey input style would, for example, express higher preferences for radio buttons when faced with a choice between those and a new slider version, and conversely for slider input styles. Indeed others have found somewhat similar results in experiments

that simply asked satisfaction with sliders/radio buttons at the end of the survey experience without requiring respondents to choose between alternate methods of input. For example, While Cape (2008) found that compared to respondents using more traditional Likert scales, respondents who were presented questions using slider scales reported higher levels of satisfaction with it as an instrument to capture their true opinions. In our study we certainly saw evidence of a conditioning effect for preference as well in that those who were presented slider bar questions in the main experiment and then asked to complete an item using radio buttons generally expressed interest in sliders. However, they did not express this interest as consistently as those who completed standard scales in the experiment and then completed one additional slider item did for standard radio buttons (76% of respondents in the radio button version expressed interest for radio buttons compared to 63% of respondents in a slider group expressed interest for sliders. ($\chi^2(1)=53.11$; p-value<.0001). We also found that generally, the preference for sliders increased from PC to Tablet to Smartphone respondents but the degree of differences across devices was still influenced with the input style to which respondents were assigned. More work is needed to better understand whether preferences for sliders might be higher among PC respondents who have touchscreen monitors compared to mouse only input.

In summary, we found consistent patterns in missing item rates and lower, middle and higher response option selections for the respondents in the middle and right slider start groups compared to any of the other slider scale or radio button groups. These trends were generally consistent across devices, and were slightly more pronounced for 5 point compared to 11 point scales. Moreover, these differences were seemingly not impacted by whether scales were additionally numbered or not. The higher missing rates and lower levels of selecting higher categories across the scale items resulted in stark differences in three main survey outcome measures. While the slider start position based imputation resulted in fewer significant differences and practically small differences, it did not fully compensated for the item missingness – especially for the 11 point scale items. For each of the three survey measures, the imputed 11 point version produced overall scale measures for right and middle slider start groups that trended well above the general pattern for the remaining scale types and could give the indication that satisfaction was much higher than reality might suggest. Clearly, without the imputation, the right and middle start slider types generated measures of satisfaction that are likely to be too low. More work is needed to understand if such an approach can be applied uniformly for sliders with missing values or if it should be applied more judiciously. The outcome measures and more specifically, the individual items were generally expected to have a negative skew based on historical trends for similar customer satisfaction/loyalty items. Thus, many of the expected responses were in the upper region of the scales and the direction of item missingness and overall differences

in measures tracked very closely to the expected response pattern. More work is needed to see if comparable results might be obtained for the middle and left slider start groups using scale items with an expected positive skew.

We note that our study has some clear limitations. Our consistent null findings for the scale numbering factor might be related to the fact that the numbering was added to scales that always included semantic labels. The labels, especially for the 5 point scales, might have been sufficient to overshadow any additional impact that numbering could have provided. For the 11 point scales we expected the numbering to have a more pronounced effect since these scales were only labeled at the two anchor points. The difference in scale labeling pattern across the two scale lengths might also confound differences observed for the scale length factor, but we note that the method used to label the 5 point and 11 point scales is generally consistent with typical uses in practice. We also note that while we were able to experimentally randomize respondents to receive different input styles and slider starting positions, scale lengths and scale numbering we had to embed the overall experiment within each of the three devices. Panel expectations and device ownership within the panel sourcing our sample precluded randomizing panelists to device type. Hence, the device used to complete the survey was taken as a natural blocking variable. In light of this, as one might expect, we found some natural differences in the ages of respondents using each type of device with PC respondents being older than tablet respondents and Tablet respondents being older than Smartphone respondents, on average. Differences in other demographic variables that were correlated with age were also found to vary similarly across the three devices and were consistent with other studies that also allowed respondents to self-select their device (see Baker-Prewitt & Miller, 2013 for example). So in sum, when interpreting the device specific comparisons and effects reported in this study one has to consider that they could represent not only device but also the cluster of demographic variables related to the usage of that device.

While sliders may offer more engagement for respondents they come at a cost when thinking about implementing them across many device types with differing space and hardware constraints. And no matter how engaging sliders can be compared to radio buttons, missing items still persist and can certainly be a function of starting position as well as the underlying distribution being estimated. Preference for sliders tends to skew towards those using mobile devices to complete surveys, but this preference doesn't overwhelm previous experience with radio buttons. Even though sliders might be more preferred by smartphone respondents, they also add to the completion times, overall. And given that many studies have consistently shown that surveys tend to take longer on smartphones compared to PCs (Buskirk, 2015b; Wells et al., 2014), it's hard to know whether the positive impact sliders have on engagement would outweigh or be nullified by the negative impact of longer

surveys. More work is needed to understand just how slick a slider needs to be to hit this sweet spot.

References

Baker-Prewitt, J. & Miller, J. (2013). What Happens to Data Quality When Respondents Use a Mobile Device for a Survey Designed for a PC. Paper presented at the 2013 CASRO Online Research Conference, San Francisco, March, 2013. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/0A81BA94-3332-4135-97F6-6BE6F-6CEF475/Paper_-_Jamie_Baker-Prewitt_-_Burke.pdf

Buskirk, T. D. (2015). The Rise of Mobile Devices: From Smartphones to Smart Surveys. *The Survey Statistician, 72*, 25-35. Available at: http://isi-iass.org/home/wp-content/uploads/N72.pdf

Buskirk, T. D. (2015b). Going Mobile with Survey Research: Design, Data Collection, Sampling and Recruitment Considerations for Smartphone and Tablet Based Surveys. Shortcourse presented at the Journal of Official Statistics Anniversary Conference, 2015. Stockholm, Sweden. Available at: http://www.scb.se/Grupp/Produkter_Tjanster/Kurser/_Dokument/JOS-2015/buskirk-FINAL-participant-JOS2015ShortCourseBuskirkJUNE2015.pdf

Buskirk, T. D. & Andrus, C. (2012). Smart surveys for smart phones: Exploring various approaches for conducing online mobile surveys via smartphones. *Survey Practice, 5.* Available at: http://surveypractice.wordpress.com/2012/02/21/smart-surveys-for-smartphones/

Buskirk, T. D. & Andrus, C. (2014). Making mobile browser surveys smarter: Results from a randomized experiment comparing online surveys completed via computer or smartphone. *Field Methods*, 26, 322-342.

Buskirk, T. D., Michaud, J., & Saunders, T. (2014). Swipe, Snap & Chat: Mobile Survey Data Collection Using Touch Question Types and Mobile OS Features. Paper presented at the 39th Annual Conference of the Midwest Association of Public Opinion Research, November 21-22, 2014, Chicago, Il. Available at: http://www.mapor.org/confdocs/absandpaps/2014/1C1_Buskirk_slides.pdf

Callegaro, M. (2010). Do you know which device your respondent has used to take your online survey? *Survey Practice.* Available at: http://surveypractice.wordpress.com/2010/12/08/devicerespondent-has-used/

Cape, P. (2009). Slider Scales in Online Surveys. Paper presented at the 2009 CASRO Panel Conference, Feb. 2-3, 2009 New Orleans. Retrieved on August 31, 2015 from: http://www.surveysampling.com/ssi-media/Corporate/white_papers/SSI-Sliders-White-Pape.image

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, *24*(2), 227-245.

Courtright, M. Saunders, T. & Tice, J. (2014). Innovation in Web Data Collection: How 'Smart' Can I Make My Web Survey? Paper presented at the CASRO Technology and Innovation Event, May, 2014, Chicago. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/97E56036-D4ED-4552-8A5F-E0A75899AEA8/2T1.1_-_T_Saunders_-_Maritz_-_M_Courtright_-_Research_Now_-_J_Tice_-_Decipher.pdf

Derham, P. A. J. (2011). Using preferred, understood or effective scales? How scale presentations effect online survey data collection. *Australasian Journal of Market & Social Research, 19*(2), 13-26.

Dobronte, A. (2012, August 21). Likert scales vs. slider Scales in commercial market research. Retrieved June 27, 2015, from https://www.checkmarket.com/2012/08/likert_v_sliderscales/

Funke, F, Reips, U.-D., & Thomas, R. K. (2011). Sliders for the Smart:Type of Rating Scale on the Web Interacts With Educational Level. *Social Science Computer Review*, 29(2), 221-231.

Husser, J. A. & Fernandez, K. E. (2013). To click, type, or drag? Evaluating speed of survey data input methods. *Survey Practice, 6*(2), 1-7.

Keusch, F. & Zhang, C. (2014). A review of Issues in Gamified Survey Design. Paper presented at the 2014 Midwest Association of Public Opinion Research Conference, November 21-22, 2014, Chicago. Available at: http://www.mapor.org/confdocs/absandpaps/2014/4A2_Keusch_slides.pdf

Link, M. W. & Buskirk, T. D. (2012). The role of new technologies in powering, augmenting, or replacing traditional surveys. Short-course presented at the annual meeting of the American Association for Public Opinion Research, Orlando, FL.

Michaud, J., Buskirk, T. D., & Saunders, T. (2014). You CAN Touch This: An Experiment to Compare Computer and Mobile Surveys Using Touch Friendly Question Types." Paper presented at the 69th Annual American Association of Public Opinion Research Concerece, May 15-18, 2014, Anaheim, CA.

Peterson, G., Mechling, J., LaFrance, J., Swinehart, J., & Ham, G. (2013). Solving the unintentional mobile challenge. Paper presented at the CASRO Online Research Conference, March, 2013, San Francisco, CA. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/0A81BA94-3332-4135-97F6-6BE6F6CEF475/Paper_-_Gregg_Peterson_-_Market_Strategies_International.pdf

Puleston, J. (2011, March 14). Sliders: A user guide. Retrieved June 27, 2015, from http://question-science.blogspot.com/2011/02/slider-how-to-use-them.html

Roster, C. A., Lucianetti, L., & Albaum, G. (2015). Exploring Slider vs. Categorical Response Formats in Web-Based Surveys. *Journal of Research Practice, 11*(1), Article D1. Accessed on August 30, 2015 from: http://jrp.icaap.org/index.php/jrp/article/view/509/413.

Sikkel, D., Steenbergen, R., & Gras, S. (2014). Clicking vs. dragging: Different uses of the mouse and their implications for online surveys. *Public Opinion Quarterly, 78*, 177-190.

Sellers, R. (2013). How sliders bias survey data. *Alert!, 53*(3), 56-57.

Toepoel, V. & Funke, F. (2014). Investigating Response Quality in Mobile and Desktop Surveys: A Comparison of Radio Buttons, Visual Analogue Scales and Slider Scales. Paper presented at the 2014 American Association of Public Opinion Research Conference. Anaheim, CA, May, 2014.

Wells, T., Bailey, J., & Link, M. W. (2014). Comparison of Smartphone and Online Computer Survey Administration. *Soc. Sci. Comput. Rev. 32*(2), 238-255. DOI=10.1177/0894439313505829. http://dx.doi.org/10.1177/0894439313505829

methods, data, analyses | Vol. 9(2), 2015, pp. 261-292 DOI: 10.12758/mda.2015.014

The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel

Bella Struminskaya, Kai Weyandt & Michael Bosnjak
GESIS – Leibniz Institute for the Social Sciences

Abstract

The use of mobile devices such as smartphones and tablets for survey completion is growing rapidly, raising concerns regarding data quality in general, and nonresponse and measurement error in particular. We use the data from six online waves of the GESIS Panel, a probability-based mixed-mode panel representative of the German population to study whether the responses provided using tablets or smartphones differ on indicators of measurement and nonresponse errors from responses provided via personal computers or laptops. We follow an approach chosen by Lugtig and Toepoel (2015), using the following indicators of nonresponse error: item nonresponse, providing an answer to an open question; and the following indicators of measurement error: straightlining, number of characters in open questions, choice of left-aligned options in horizontal scales, and survey duration. Moreover, we extend the scope of past research by exploring whether data quality is a function of device-type or respondent-type characteristics using multilevel models. Overall, we find that responding with mobile devices is associated with a higher likelihood of measurement discrepancies compared to PC/laptop survey completion. For smartphone survey completion, the indicators of measurement and nonresponse error tend to be higher than for tablet completion. We find that most indicators of nonresponse and measurement error used in our analysis cannot be attributed to the respondent characteristics but are rather effects of mobile devices.

Keywords: mobile phone surveys, panel survey, mobile devices, nonresponse, measurement

1 Introduction

In web surveys and online panels, it can no longer be expected that respondents participate using desktop computers and laptops only. Survey researchers have reported a growing share of unintended mobile respondents – respondents who use their mobile devices such as smartphones or tablets to access and participate in surveys that were originally designed to be taken on PCs or laptops (de Bruijne & Wijnant, 2014b; Peterson, 2012; Toepoel & Lugtig, 2014; Wells, Bailey, & Link, 2014). In the Dutch online probability-based LISS Panel, the proportion of unintended mobile respondents increased from 3% in 2012 to 11% in 2013, in the CentERpanel, another probability-based general population online panel in the Netherlands, the proportion of unintended mobile respondents increased from 3% in 2012 to 16% in 2013 (de Bruijne & Wijnant, 2014b). In the German mixed-mode GESIS Panel, in 2014 about 17.9% of online respondents completed the questionnaires using mobile devices with 9.2% using smartphones and 8.7% using tablets. In 2015, about 15.6% of online respondents name tablets and 8.1% name smartphones as the preferred mode to answer the questionnaires.[1]

Responding to surveys using various devices, that increasingly become heterogeneous with regard to size and functionality, raises concerns about data quality. Differences between PCs/laptops and mobile devices in screen size and input method as well as the possibility to participate in surveys via mobile devices from a variety of locations and situations where distractions are possible can affect respondents' cognitive processing, increasing the risk of errors (Peytchev & Hill, 2010). Nonresponse error and measurement error are of particular concern.

Respondents using mobile devices for survey completion have demonstrated lower response rates (Buskirk & Andrus, 2014; de Bruijne & Wijnant, 2013), lower completion rates (Mavletova, 2013; Mavletova & Couper, 2013), and higher break-off rates[2] (Callegaro, 2010; Cook, 2014; Mavletova, 2013; McClain, Crawford, & Dungan, 2012; Poggio, Bosnjak, & Weyandt, 2015; Stapleton, 2013). Item-nonresponse has been found to be more pronounced when completing the survey on a mobile device in open-ended questions (Peytchev & Hill, 2010). However, more recent studies did not replicate this result: de Bruijne and Wijnant (2014a) show

1 GESIS (2015): GESIS Panel - Standard Edition. GESIS Datenarchiv, Cologne. ZA5665 Data file version 8.0.0, doi:10.4232/1.12245. Own calculations.

2 We use the term response rate for studies based on a probability samples and completion rate for studies that are not based on probability samples. For studies that focused on break-offs we do not divert from the original terminology used by the authors.

Direct correspondence to
Bella Struminskaya, GESIS – Leibniz Institute for the Social Sciences,
B 2,1, 68159 Mannheim, Germany
E-mail: bella.struminskaya@gesis.org

that respondents using mobile devices are not more likely to provide a half-open "other" answer than to choose a closed "other" option; Wells et al. (2014) find that mobile respondents are not more likely to skip the half-open or open questions. Nevertheless, mobile web respondents have been shown to provide shorter answers to open-ended questions than PC respondents (Mavletova, 2013; Peterson, 2012; Wells et al., 2014).

The second major concern in mobile web surveys is the risk of more pronounced measurement errors. Comparing the responses provided by mobile web respondents to the record data, Antoun (2015) shows that smartphone respondents provide fewer accurate answers when reporting age and date of birth than PC respondents. Cases when validation data is available to the researchers to study measurement errors are an exception rather than a rule. Hence, most researchers use indicators of satisficing behavior that suggests reporting with measurement error. Krosnick (1991) defines satisficing as respondents' failure to consecutively and carefully execute the cognitively demanding stages that precede producing accurate and valid survey responses. These stages include interpreting the meaning of the question, retrieval of relevant information from memory, formation a summary judgement, carefully integrating this information, and clear report of the summary judgement (Tourangeau, 1984; Tourangeau, Rips, & Rasinski, 2000). Satisficing behavior is the result of the interplay of three factors: respondents' ability, motivation and difficulty of the task (Krosnick, 1991, p. 225). Using a mobile device for survey completion can be a difficult task due to technical reasons such as a small screen, a touchscreen, as well as situational characteristics if respondents are outside of home. Providing satisfactory answers instead of accurate answers is indicative of measurement error.

In past studies, the following indicators of satisficing have been used when studying mobile web responses: number of "don't know" answers, non-differentiation (straightlining), primacy effects, rounding, measures of superficial cognitive processing (e.g., answers to cognitive reflection tests), avoiding half-open questions, length of answers to open-ended questions, and answers to sensitive questions (Antoun, 2015; Buskirk & Andrus, 2014; Lugtig & Toepoel, 2015; Mavletova, 2013; Mavletova & Couper, 2013; Wells et al., 2014). Lugtig and Toepoel (2015) find that mobile web respondents report with higher measurement error than PC respondents showing more item missing responses, higher item-nonresponse in open-ended questions, more primacy effects, and fewer response options selected in check-all-that-apply questions. Conversely, in other studies little evidence is found: mobile web respondents are not more likely to demonstrate primacy effects (Buskirk & Andrus, 2014; Mavletova, 2013; Toepoel & Lugtig, 2014; Wells et al., 2014), do not differ from PC respondents in providing socially desirable answers (Antoun, 2015; Mavletova, 2013), do not show increased rounding or superficial cognitive

processing (Antoun, 2015)[3]. Mixed results have been obtained on using the horizontal scales in mobile web surveys. Peytchev and Hill (2010) found that horizontal scrolling generally did not affect responses but a small proportion of respondents failed to scroll and see all possible answer options. De Bruijne and Wijnant (2014a) find that horizontal scale format produces slightly more item missings than the vertical format even when the horizontal scales are fully visible on screen with no need to scroll.

Survey duration, another indicator of satisficing behavior in web surveys with shorter duration being associated with more primacy effects (Malhotra, 2008), has been shown to produce opposite results for mobile web surveys. Using smartphones for survey completion is associated with longer completion times (Antoun, 2015; Cook, 2014; de Bruijne & Wijnant, 2013; Mavletova, 2013; Mavletova & Couper, 2013; Peterson, 2012; Wells et al. 2014). However, the longer duration can be explained by other factors such as connection speed, scrolling, familiarity with the device, or distractions due to respondents' multitasking. Couper and Peterson (2015) show that the connection speed accounts for a small proportion of the difference between PC and smartphone completion. They further argue that multitasking and familiarity with the device are less plausible explanations than the display size and the need for scrolling.

In light of the mixed results about the data quality in mobile web surveys outlined above it is noteworthy that few studies on mobile responding are based on probability-based online panels; and from those that are, several studies are based on the LISS Panel (cf. Antoun, 2015; de Bruijne & Wijnant, 2013; de Bruijne & Wijnant, 2014; Lugtig & Toepoel, 2015), other studies are based on the CentER-panel in the Netherlands (de Bruijne & Wijnant, 2014b) or the Knowledge Panel of GfK Knowledge Networks in the USA (Wells et al., 2014). Mobile web respondents in probability-based panels can differ from mobile respondents in nonprobability panels. Respondents in nonprobability panels can be more technologically sophisticated and able to answer surveys on mobile devices, thereby compensating mea-

3 It can be assumed that finding adverse effects on data quality can be caused by some studies being optimized for survey completion while others are not. Indeed, studies mentioned in this paragraph with the exception of Antoun (2015) were optimized for mobile completion or included experimental conditions that were optimized for mobile devices. However, it does not seem that mixed results presented in this section can be fully explained by mobile optimization as providing shorter answers in open-ended questions, lower completion and response rates are found in both optimized and non-optimized studies. In this review, studies with optimized design (i.e., where special programming for mobile devices was performed), including experimental conditions are: Buskirk & Andrus 2014, de Bruijne & Wijnant 2013, Mavletova 2013, Mavletova & Couper 2013, Peytchev & Hill 2010, Stapleton 2013, Toepoel & Lugtig 2014, and Wells, Bailey & Link 2014. Non-optimized studies are: Antoun 2015, Callegaro 2010, Cook 2014, de Bruijne & Wijnant 2014, 2014a, Lugtig & Toepoel 2015, McClain et al. 2012, Peterson 2012, and Poggio, Bosnjak, & Weyandt 2015.

surement errors with their experience and motivation. For example, in a Russian non-probability panel, Mavletova (2013) finds that more experienced mobile users wrote significantly longer answers to open questions than less experienced mobile users. Furthermore, learning effects can play a role if respondents in nonprobability panels are more experienced than respondents in probability-based panels. It has been shown that professional respondents in nonprobability panels are not more likely to produce data of lower quality (Hillygus, Jackson, and Young, 2014; Matthijsse, de Leeuw, and Hox (2015), but this aspect has not been studied for mobile device vs. PC survey completion.

It is important to investigate the consequences of responding via mobile devices in probability-based general population panels to fully understand whether mobile web response is something survey researchers should be concerned about, given the mixed results provided by the literature reported above. In this article, we concentrate on nonresponse and measurement using several measures of satisficing behavior as indicators of possible measurement errors. We follow an approach chosen by Lugtig and Toepoel (2015) for the LISS Panel data using the data from the GESIS Panel, a probability-based mixed-mode (online and mail) panel of the general population in Germany.

If preferences to answer surveys using a particular device are correlated to the propensity to satisfy, selection and measurement effects will be confounded (Lugtig & Toepoel, 2015). Indeed, past studies have found that respondents answering online surveys via mobile devices differ at least in their demographic characteristics from those who answer online surveys via laptops and PCs (Cook, 2014; de Bruijne & Wijnant, 2013; de Bruijne & Wijnant, 2014b; Toepoel & Lugtig, 2014). Cook (2014), who uses the U.S. data, finds that demographic composition of device groups differ: those who take surveys on tablets are significantly younger, more likely to be female; smartphone respondents are lower educated and have lower income than tablet and PC respondents, both smartphone and tablet use is higher for Hispanics and African-Americans. For the Netherlands, de Bruijne and Wijnant (2013) find small differences in gender between smartphone and PC users with smartphone users more likely to be men; the proportion of those higher educated is significantly higher among smartphone users. Consistent with other studies, mobile web use is highest among young respondents. Toepoel and Lugtig (2014) demonstrate that income, household size, and household composition are predictive of mobile survey completion. Furthermore, de Bruijne and Wijnant (2014b) find that in the LISS Panel sex and age are predictive of unintended access to online surveys via smartphones and tablets. Women and younger respondents are more likely to use mobile devices for survey access. Additionally, living alone is negatively associated with accessing online surveys via tablets while respondents in paid work are more likely to use tablets to access online surveys.

Therefore, it is important to study whether certain respondent behaviors are attributable to a respondent (response style) or are a result of survey completion using mobile devices. This conceptual extension to past approaches involves disentangling device-level and respondent-level determinants of data quality indicators using a multilevel perspective. Overall, our analyses have two goals: (1) to find out to which extent the findings of Lugtig and Toepoel (2015) can be replicated in the GESIS Panel, that is, generalized across different countries and panel configurations, and (2) disentangle the effect of respondent characteristics and device characteristics on measurement-related and nonresponse-related data quality indicators.

2 Data, Measures, and Hypotheses

We use data from six waves of the GESIS Panel – a face-to-face recruited mixed-mode probability-based panel, which is representative of the general population in Germany aged 18 to 70 years at the time of recruitment. About 65 percent of respondents participate online and about 35 percent participate offline via postal mail questionnaires. The recruitment for the GESIS Panel took place in 2013. The first regular wave was fielded in the beginning of 2014. Respondents receive invitations to participate in self-administered surveys every two months. The recruitment rate for the GESIS Panel is 31.6% (AAPOR RR5), the response rate for the profile survey is 79.4%. For 2014 surveys, the completion rates per wave vary between 88.7% and 92.0% for the online questionnaires and between 76.7% and 84.6% for the offline questionnaires. All active panel members receive unconditional incentives of five euros with questionnaire invitations for every wave per post. For our analysis, we use the data for online respondents only. Overall, 3041 online respondents were invited to participate in the first regular GESIS Panel wave in 2014. From those, we exclude 127 persons who did not participate in any of the waves in 2014 as well as one person who switched modes from online to offline. This leaves us with a sample size of 2913 respondents.

The online questionnaires in GESIS Panel are not programmed in a mobile device optimized way, that is, questions are not adjusted for a particular device. For the identification of the device used by a respondent to complete the questionnaire we use the user agent strings (UAS) provided by the panel software. The user agent strings are recoded into the device-variables using a Stata code "parseuas" developed by Rossmann and Gummer (2014). The script distinguishes between mobile phones, tablets and other devices used to complete the questionnaire. The category "other devices" includes desktop computers, laptops and possibly a small proportion of the devices with browser versions that cannot be classified as mobile phones or tablets. Thus, the proportion of PC-completions might be somewhat overestimated in our analyses.

The contents of the questionnaires fielded in the GESIS Panel vary from wave to wave. In order to eliminate the influence of varying questionnaire content on nonresponse and measurement error indicators, our analyses are based on an (mostly) invariant set of questions that are asked in each survey wave. This approach was chosen by Lugtig and Toepoel (2015) for the analyses based on the LISS Panel. The questions that are invariant in every wave are concerned with survey evaluation as they are in the LISS Panel. However, the indicators for the GESIS Panel are slightly different. The evaluation includes various types of questions: a grid question, an open question, and several singe-choice questions. The evaluation part includes overall 14 items about the questionnaire itself, the device used to fill out the questionnaire, whether the respondent completed the questionnaire without a time break, and if not, how long the break lasted, whether the questionnaire was completed at home or outside of the home, whether others were present, and an open field for remarks about the questionnaire.

We use the following indicators of measurement error (ME) and nonresponse error (NR): item-nonresponse (NR), item-nonresponse to an open question (NR), length of answers to an open question (ME), straightlining (ME), choice of left-aligned answer options in horizontal scales (ME), and survey duration (ME). The indicators are operationalized as follows.

Item-nonresponse: We use all of the items for questionnaire evaluation, reported device and conditions under which the respondent filled out the questionnaire to count the number of item missings. We exclude the remark as well as the open question about the duration of the time break if the respondent indicates that he or she did not complete the survey without a break. Thus, the indicator for the number of missing values ranges from 0 to 13. We expect respondents who use smartphones for survey completion to show higher number of item missings. However, we expect no differences in item missings between PC and tablet respondents (Hypothesis H1).

Straightlining: The first question about the questionnaire evaluation is a grid question that contains six items: whether the survey was interesting, diverse, and important for research, long, difficult, or too personal, each measured with a five-point labeled scale. We define straightlining as providing the same answer to all of the items of the grid a respondent answered if the respondent answered at least two items from the evaluation grid. Lugtig and Toepoel (2015) find that straightlining is surprisingly higher for PC respondents. However, the questions they used for analysis were not arranged in a grid. For grid questions, straightlining has been shown to be higher for respondents using mobile phones than for those using tablets and PCs (McClain et al., 2012). Since we use the grid question, we expect to find more straightlining for respondents who answer the questionnaire via smartphones (Hypothesis H2a). For tablets, we expect to find no differences to PCs given the larger screen size (Hypothesis H2b).

Response to an open question: At the end of each questionnaire respondents have the opportunity to provide additional verbal feedback about the questionnaire. We use a binary variable whether a respondent has provided feedback or not. We expect respondents who use smartphones or tablets for survey completion to provide answers to an open question at a lower rate than respondents who complete the survey using PCs (Hypothesis H3).

Length of the answer to an open question: The second indicator that we use related to the open questions is the length of the answer provided by a respondent. In line with the findings from the literature reviewed in the previous section, we expect respondents who fill out their questionnaires via smartphones to provide shorter answers given the small screen size (Hypothesis H4a). We expect to find no differences in answers to open questions or length of these answers provided via PCs and tablets (Hypothesis H4b).

*Choice of left-aligned option*s: The measure of a higher proportion of left-aligned answer options selected is based on the items of the grid evaluation question as well as three single-choice evaluation questions with 5-point horizontal scales. One of these three items is the overall questionnaire evaluation, the other two items vary between the waves: for the first three waves the items ask whether the questions were understandable and whether they made the respondent think about things and in all the following waves the questions asked about how difficult it was to understand the questions and how difficult it was to find an answer. We count the number of times respondent chose the two answer options aligned to the left. Although the questions differ between the waves this should not affect the rate at which respondents using different devices provide options aligned to the left or not. We expect more left-aligned options for responses on smartphones than for PCs and tablets (Hypothesis H5a). No difference between PCs and tablets is expected due to the screen size (Hypothesis H5b).

Duration: The duration is measured in seconds for every wave. We truncated the extreme values of questionnaire duration longer than an hour to an hour. Since our surveys are not optimized for mobile devices, we expect longer completion times both for smartphones and tablets (H6). Note that the indicator for duration does not restrict the questionnaire to the non-changing evaluation part as do the other indicators that we use. For duration, we analyze the time it took respondents to complete the entire questionnaire.

In the first part of our analyses, we follow closely the procedure found in Lugtig and Toepoel (2015). First, we report the overall device use for questionnaire completion in the GESIS Panel in 2014. Second, we look at the indicators of measurement and nonresponse error associated with the usage of a particular device. Third, we concentrate on the longitudinal device use and measurement and nonresponse errors.

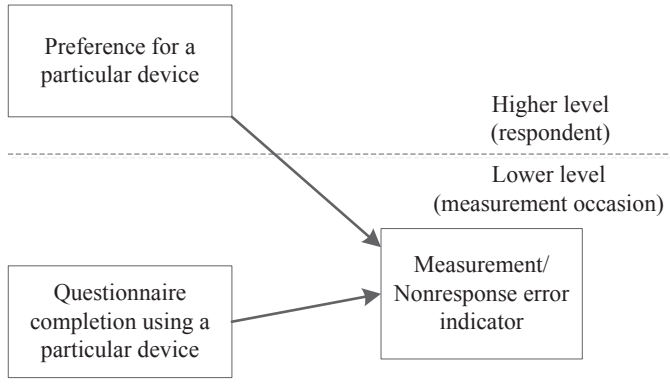


Figure 1 Graphical representation of the two measurement levels

In the second part of our analyses, we attempt to disentangle whether a particular indicator of measurement or nonresponse error is device-related or rather a characteristic of the respondent. For this purpose, for each measurement and nonresponse error indicator we estimate the intercept-only multilevel models, models with indicators of survey completion via tablet or smartphone, and lastly we add respondent characteristics. The intercept-only models do not explain any variance in our dependent variables (i.e., measurement or nonresponse error indicators) but decompose the variance into two independent components for each level (Hox, 2010, p. 15). Our lower level is the measurement occasion (operationalized as each singular survey wave) and respondent is our higher level (see Figure 1).

Measurement occasion is defined as a combination of characteristics of the device that is used to complete the questionnaire and situational characteristics that can be related to the use of this device. The situation characteristics can include distractions, multitasking, changing location, etc. Measurement occasions are nested within respondents. Since different indicators have different scales, we compute logistic models for binary indicators and multilevel regression models for continuous indicators. Adding the device indicators to the models allows us to tease out the device effects from other situational factors that form a measurement occasion. We add respondent characteristics in order to separate the device effects from selection effects. We compare the models based on the intra-class correlation coefficients (ICC), a proportion of the variation at the higher level (respondent) over the total variation (respondent plus measurement occasion).

3 Results

First, we present descriptive results on the device use in the GESIS Panel in 2014. Table 1 shows the absolute counts and proportions of respondents by device as well as transitions from one device to another over the six waves used in our analyses. Most respondents complete the surveys via PCs or laptops, the proportion decreases from 84% in the first wave to 79% in the sixth wave. This indicates an overall increase of mobile device use over time. This result is especially interesting since the online questionnaires in GESIS Panel are not optimized for the completion on mobile devices. The groups who complete the surveys using mobile devices are considerably smaller. The proportion of respondents who complete the surveys via tablets ranges from 7.9 to 10.5%. Smartphone completions have a similar range from 7.6 to 10.5%. The proportions of respondents using tablets for survey completion are about the same as reported by Lugtig and Toepoel (2015) for the LISS Panel, however, the share of respondents who use smartphones to complete panel waves is considerably higher in the GESIS Panel in 2014 than in the LISS Panel in 2013, where it ranged from 1.4 to 3.4%. However, in February 2015 about 6.6% of LISS respondents completed questionnaires via smartphones and about 15.5% of respondents used tablets (Wijnant, 2015). It seems that the differences between the proportions of those completing the surveys via mobile devices in the LISS Panel and in the GESIS Panel can be attributed to the differences in reference periods (i.e., 2013 vs. 2014) and can be explained, for example, by mobile devices becoming more affordable or the public learning to operate such devices.

Transitions from one device to the other are the lowest for PC respondents, ranging from 88.04% (fifth wave to sixth wave) to 90.27% (fourth wave to fifth wave). This result is similar to the results reported by Lugtig and Toepoel (2015) for the LISS Panel in 2013 with less than 5% of respondents switching from PC survey completion to smartphone or tablet.

We calculated the average consistency for each device type. For PC usage, the average device consistency is the highest with 89.09 percentage points. For tablet users, the average device consistency is 67.68 percentage points, ranging from 64.00% to 72.93%. The lowest device consistency is observed for smartphone users: overall, from 58.91 to 61.69% of respondents use smartphone to complete two consecutive waves. The average consistency for smartphone survey completion is 61.46 percentage points. Furthermore, respondents participating via smartphones have higher rates of nonparticipation in the following wave for initial waves. However, these rates become comparable between the devices at later waves (e.g., the fifth and the sixth waves), probably because respondents who participate via smartphones have a higher probability to attrite.[4]

4 In GESIS Panel, after not having participated for three consecutive waves due to either noncontact or nonresponse, participants are excluded from the panel (involuntary attrition). Respondents can also request to be removed from the panel (voluntary attrition).

Table 1 Devices used for questionnaire completion in the six waves of the GESIS Panel (in percent)

		The following wave: Wave x +1					
		PC	Tablet	Smart-phone	Not parti-cipated	N	% of wave respon-dents
First wave	PC	88.86	2.52	3.67	4.95	2342	84.25
2014	Tablet	23.11	64.00	8.89	4.00	225	8.09
(Feb/Mar)	Smartphone	23.94	3.29	61.03	11.74	213	7.66
	Not participated	64.66	4.51	9.77	21.05	—	—
Second wave	PC	89.16	2.73	3.13	4.98	2270	83.00
2014	Tablet	27.31	64.35	2.78	5.56	216	7.89
(Apr/May)	Smartphone	22.49	3.61	65.06	8.84	249	9.10
	Not participated	48.31	4.49	10.67	36.52	—	—
Third wave	PC	89.13	2.92	2.74	5.12	2225	82.38
2014	Tablet	20.18	70.18	4.59	5.05	218	8.07
(Jun/Jul)	Smartphone	23.64	6.98	58.91	10.47	258	9.55
	Not participated	36.79	4.25	6.13	52.83	—	—
Fourth wave	PC	90.27	2.12	3.37	4.24	2168	81.84
2014	Tablet	23.27	66.94	5.71	4.08	245	9.25
(Aug/Sep)	Smartphone	28.39	4.24	60.59	6.78	236	8.91
	Not participated	25.38	3.41	6.82	64.39	—	—
Fifth wave	PC	88.04	3.68	4.00	4.28	2148	81.83
2014	Tablet	14.41	72.93	6.99	5.68	229	8.72
(Oct/Nov)	Smartphone	24.60	7.66	61.69	6.05	248	9.45
	Not participated	22.57	2.43	6.25	68.75	—	—
Sixth wave	PC	—	—	—	—	2050	79.00
2014/2015	Tablet	—	—	—	—	272	10.48
(Dec/Jan)	Smartphone	—	—	—	—	273	10.52
	Not participated	—	—	—	—	—	—
Average device consistency	PC	89.09					
	Tablet		67.68				
	Smartphone			61.46			

N = 2913.

In the second step of our analyses, we report the indicators of measurement and nonresponse error separately for each device type (Table 2). Overall, we observe similar results as Lugtig and Toepoel (2015) that PC respondents report with least measurement and nonresponse error, followed by tablet respondents, and smartphone respondents report with highest measurement and nonresponse error. On

average, responses via smartphones are characterized by higher item-nonresponse and a higher percentage of straightlining in a grid question. Those who respond via smartphones respond to an open question at a lower rate and enter fewer characters when they do answer an open question. Also, smartphone respondents demonstrate longer completion times than PC and tablet respondents.

Our hypothesis concerning item-nonresponse predicted higher levels of item-nonresponse for smartphone respondents and no difference for tablet respondents when compared to PC respondents. We indeed observe higher levels of item-nonresponse for smartphones, which is significantly different from PC and tablet respondents. No statistically significant difference is found for the comparison of item-nonresponse between PCs and tablets.

For straightlining, we also expected to find higher levels for smartphones and no differences between tablets and PCs. Straightlining is highest for smartphone completion and the differences to smartphones and tablets are statistically significant (Table 2), again there are no significant differences between tablets and PCs.

In line with our expectations, both smartphone and tablet respondents provide fewer answers to the open question than PC respondents (about 6% for mobile devices vs. 14% for PCs). There is no difference between providing an answer to the open question when using a smartphone or a tablet for survey completion. The length of the answers to an open question is shortest for smartphones and is followed by tablets, although the difference between tablets and smartphones is not statistically significant. The highest number of characters is provided by respondents who complete the surveys via PC or laptop. This finding can be attributed to the absence of the keyboard to type an answer (although we cannot control whether tablet users have used keyboards, it seems a likely explanation).

Regarding the tendency to choose left-aligned answer options in horizontal scales, smartphone respondents do not show a higher rate than PC or tablet respondents. On the contrary, left-aligned options are chosen more by PC and tablet respondents. Our explanation for this finding is that possibly horizontal scrolling is less of an issue with touch screens of smartphones, and zooming might prompt those who respond via smartphones to choose middle categories at a higher rate. However, this hypothetical explanation deserves further investigation. Concerning survey duration, we find the longest completion times for smartphones, followed by tablets. The differences between each pair of devices in survey duration are statistically significant.

To summarize, we find the highest measurement and nonresponse error indicators levels for smartphones. Although some differences between tablets and PCs are found (e.g., in answering an open question and duration), these differences are rather small and for most of measurement and nonresponse error indicators they are not pronounced. It is noteworthy, that although we find several statistically significant differences between PCs and tablets, and all indicators differ on a statistically

Table 2 Measurement and nonresponse error indicators by device in the six waves of the GESIS Panel in 2014

	PC	Tablet	Smart-phone	Total	ANOVA
Mean count of item nonresponse [b,c]	.189	.177	.472	.213	F(2, 16047)=41.96, p<0.001, η^2= .005
% Straightlining [b,c]	1.47	1.80	3.86	1.71	F(2, 15911)=22.04, p<0.001, η^2= .003
% Answered open question[a,c]	10.10	5.61	5.93	9.33	F(2, 15937)=25.81, p<0.001, η^2=.003
Mean number of characters in open question[a,c]	13.925	6.410	4.910	12.458	F(2, 15937)= 16.53, p<0.001, η^2= .002
Mean number of chosen left-aligned options[b,c]	2.470	2.418	2.248	2.445	F(2, 16085)=23.01, p<0.001, η^2= .003
Mean duration in seconds[a,b,c]	1445.46	1500.27	1862.79	1488.56	F(2, 16085)=190.65, p<0.001, η^2= .023

N pooled = 16085, N persons = 2913. Pairwise contrasts are *t*-tests for continuous variables and tests of proportions for percentages with p<0.01. a – significant difference PC-Tablet; b – significant difference Tablet-Smartphone; c – significant difference Smartphone-PC.

significant level for the comparison smartphones with PCs, the effect sizes for overall comparisons (in Table 2) are relatively small.

Results presented in Table 2 showing that mobile devices are associated with higher measurement and nonresponse errors can be attributed either to the characteristics of the devices or to the characteristics of the respondents. Those respondents who are more likely to use mobile devices for survey completion might be also more likely to cause higher measurement error. In this case, selection effects and measurement effects are intermingled. Following Lugtig und Toepoel (2015), we compare measurement and nonresponse error indicators for respondents who complete the surveys using one device consistently with measurement and nonresponse error indicators of respondents who switch between devices. If the indicators of measurement and nonresponse errors for those who constantly use tablets or constantly use smartphones for survey completion are larger than for those who switch between the mobile devices, it would indicate that measurement and nonresponse errors are more likely device-related than respondent-related. Table 3 presents the indicators of measurement and nonresponse error for groups of respondents who consistently used one device for survey completion, who switched between two

devices, and who used all three device types of devices for survey participation. We restrict the sample to respondents who took part in at least two waves of the panel and thereby had a chance to switch between the devices. From respondents who participated in at least two waves, 67.7% did not switch between the devices and always participated using a PC or a laptop. The proportions of continuous use of a mobile device for survey completion are quite low: 3% of respondents always used tablets and 3.5% always used smartphones for survey completion. About ten percent of respondents used PCs and tablets and about 11.8% used PCs and smartphones. The group of respondents using all three types of devices to complete the surveys was with 2.9% the smallest group.

Table 3 shows that respondents who always use smartphones for survey completion have the highest level of item nonresponse. For those groups that switch between the devices, item nonresponse is highest in groups that involve smartphone completion. Switches between PC and tablet have similar levels of item nonresponse. These findings indicate that item nonresponse is rather device-specific. The indicator for straightlining shows a similar pattern as the indicator for item nonresponse: if switching between devices to complete the surveys involves smartphones or surveys are completed on smartphones exclusively, measurement and nonresponse error indicators are higher than in cases of tablet and PC completion.

Surprisingly, the proportion of respondents who answer the open question is the lowest for those who always complete the surveys using tablets or switch between PCs and tablets. The number of characters entered in an open question is the highest for the groups involving a PC and lowest for groups involving tablets and smartphones. The choice of left-aligned options does not vary much between the groups, and the duration is the highest for groups involving smartphone, except the group in which respondents switch between all three devices to complete the questionnaires.

Overall, from Table 3 we can conclude that as long as survey completion involves smartphones, measurement and nonresponse error indicators are generally higher. However, we cannot draw a conclusion from these results whether reporting with measurement error is due to using a particular device or due to respondent characteristics, since for some indicators (e.g., item nonresponse and straightlining) device properties seem to be one plausible explanation for the decreased data quality and for other characteristics this does not apply.

Following the analysis of Lugtig and Toepoel (2015), we concentrate on cases where respondents participated in two consecutive waves and code the device transitions as well as changes in error indicators for each transition for each respondent. Then we standardize the distributions of changes in wave-to-wave error indicators, because the indicators have different scales. If the device is the cause of higher nonresponse and measurement error, then for transitions involving device switches the standardized changes in measurement and nonresponse error indicators would not

Table 3 Measurement and nonresponse error indicators across groups of device use patterns

	No device switches			Switch between two devices			Switch between three devices	
	Always PC	Always Tablet	Always Smartphone	PC & Tablet	PC & Smartphone	Tablet & Smartphone	PC, Tablet & Smartphone	Total
Mean count of item nonresponse	.192 (.013)	.149 (.051)	.560 (.142)	.151 (.023)	.448 (.065)	.387 (.206)	.264 (.090)	.234 (.014)
Mean % straightlining	1.49 (.001)	1.76 (.010)	3.62 (.010)	1.31 (.005)	2.97 (.006)	4.02 (.030)	2.47 (.009)	1.78 (.001)
Mean % Answered open question	10.55 (.005)	4.55 (.012)	8.10 (.018)	8.49 (.011)	6.03 (.007)	4.41 (.019)	5.99 (.015)	9.34 (.003)
Mean number of characters in open question	14.498 (1.061)	4.722 (1.694)	5.258 (1.477)	10.689 (1.647)	7.918 (1.572)	3.275 (1.472)	5.029 (1.577)	12.323 (.767)
% Choice of left-aligned options	.275 (.002)	.260 (.010)	.260 (.010)	.287 (.005)	.264 (.005)	.260 (.019)	.255 (.010)	.273 (.002)
Mean duration	1825 (28.9)	1960 (189.8)	3069 (227.6)	1718 (64.4)	2045 (79.1)	2096 (266.6)	1848 (90.3)	1891 (25.3)
Sample size	1918	85	99	282	333	34	81	2832

N = 2832 since 81 observations who participated in only one wave were dropped, standard errors in parentheses.

be different from zero for the groups with transitions to the same device (PC-PC, tablet-tablet, and smartphone-smartphone), while we would expect to find significant differences for groups which involve device changes, especially smartphones. The results are presented in Table 4. Overall, there are 12,598 transitions with nonmissing indicators of measurement error. In line with our expectations, the transitions involving the same device (i.e., PC-PC, tablet-tablet, smartphone-smartphone) are not associated with significant changes in nonresponse and measurement error indicators. Moreover, the magnitude of the changes in standardized nonresponse and measurement error indicators for transitions without the device switches is

small. For the groups involving device switches the most pronounced differences are found in duration: the differences are significant for all transitions with devices switches and for groups involving smartphones the magnitude of the change is considerably larger than for transitions between tablets and PCs. Significant effects are also found for the switches PC→tablet and PC→smartphone in providing answers to the open question and for groups PC→smartphone and smartphone→PC for the choice of left-aligned answer options. The magnitude of these changes, however, is rather small. The manner in which changes in standardized indicators are calculated makes them correspond to standardized mean difference effect sizes (Lipsey & Wilson, 2001, p. 198), so we use the benchmarks provided by Cohen (1992) to interpret the values from Table 4. Overall, we see moderate effects for duration in groups involving smartphones and small effects for duration, tendency to answer the open question and to choose left-aligned options in some groups. Our results are in line with Lugtig and Toepoel (2015), who find that transitions between tablets and PCs show small changes while transitions between smartphones and PCs show the largest changes in measurement indicators, although not significant possibly due to small group sizes. Significant changes were found by Lugtig and Toepoel (2015) for straightlining for transition tablet-tablet and the number of choices made in check-all-that-apply questions (for groups PC-PC, tablet-PC, and smartphone-PC) as well as questionnaire evaluation (for tablet-PC and smartphone-PC).[5]

The analysis presented in Table 4 is based on transitions between the waves, and it controls for respondent characteristics insofar that they stay the same over time while respondents switch between devices. We extend this analysis with multilevel modeling, in which we explicitly control for device effects and respondent characteristics. Since different indicators of nonresponse and measurement error are studied, ideally the models need to include the predictors of reporting with higher levels of item nonresponse, straightlining, or taking longer to complete the surveys, etc. This would make difficult comparing the models with each other. Thus, we use respondent characteristics that were shown to relate to the propensity of responding using a particular device. Since our goal here is not to explain which respondents produce higher nonresponse or higher measurement error but rather to tease out the device effects, this approach seems feasible.

In Table 5, the results of the stepwise procedure of calculating the multilevel models are presented. For this analysis we only include respondents who completed the survey without a break or completed after a break and have no missing values on the dependent variables to be able to compare the models with each other. First, intercept-only multilevel models are presented. The intra-class correlation coefficients (ICCs), the proportion of variance located at the level of the respondent to the total variance (i.e., respondent plus measurement occasion) for the empty models

5 The groups tablet-smartphone and smartphone-tablet were excluded by Lugtig and Toepoel (2015) due to small group sizes.

Table 4 Change in standardized indicators of nonresponse and measurement error associated with different device switches

Group/ Indicator	Item non-response	Straight-lining	Answered open question	Number of characters	Choice of left-aligned options	Duration	N
PC-PC	.000	.003	-.001	-.001	-.003	.001	9824
Tablet-Tablet	.023	-.047	.009	.005	.030	-.056	759
Smartphone-Smartphone	-.016	-.031	.032	.014	.031	.009	704
PC-Tablet	-.078	-.013	-.113*	-.044	-.105	.113*	308
Tablet-PC	.038	.041	.053	.044	.031	-.211**	238
Smartphone-Tablet	-.109	-.121	-.036	-.036	-.049	-.615***	59
Tablet-Smartphone	.114	.086	.085	.108	-.023	.658***	63
PC-Smartphone	.030	.093	-.105*	-.059	-.119*	.689***	358
Smartphone-PC	-.010	-.036	.110	.064	.190**	-.737***	285

N (person-waves) = 12598, N respondents = 2770 (only observations for respondents who took part in two consecutive waves are included); the values are predicted marginal means, significance tests against zero. *p<.05, **p<.01,*** p<.001

show that for some indicators the measurement occasion which includes but is not limited to a device, is more influential and for other indicators the differences are between-person differences.

For item nonresponse model, the intra-class correlation (ICC) of 0.172 means that item-nonresponse is a characteristic of the situation rather than a tendency of a respondent to skip questions. The differences in straightlining (ICC = .754) are rather individual-level differences than the characteristic of the survey situation: some respondents tend to straightline and some do not irrespective of the survey situation. We cannot definitely say that the differences in providing answers to the open question also are individual-level differences rather than the characteristic of the survey situation judging by the intra-class correlation of 0.551. Providing an answer to an open question seems to depend both on respondent preference and on the survey situation. The larger amount of variance for the choice of left-aligned options is located on the level of the measurement occasion, suggesting that choosing left-aligned options at horizontal scales is not a respondent-specific characteristic. Survey duration is as well situation specific, which is consistent with the results presented in Tables 2 and 3.

Table 5 Multilevel models for indicators of measurement and nonresponse error

	Item nonresponse	Straightlining	Answered open question	Choice of left-aligned options	Duration
Null models					
Constant	.124***	-7.390***	-3.510***	2.469***	24.736***
	(.006)	(.608)	(.084)	(.084)	(.164)
Variance at higher level	.052	10.782	4.038	.406	54.356
	(.003)	(2.774)	(.302)	(.016)	(2.028)
Variance at lower level	.251	3.290†	3.290†	.999	111.978
	(.003)	(—)	(—)	(.012)	(1.398)
ICC	.172	.766	.551	.289	.327
Models with device dummies (reference: PC completion)					
Constant	.127***	-6.570***	-3.369***	2.491***	23.804***
	(.006)	(.444)	(.085)	(.015)	(.171)
Tablet completion	-.033	.031	-.763***	-.060	1.552***
	(.018)	(.339)	(.177)	(.039)	(.421)
Smartphone completion	.008	1.255***	-.711***	-.187***	8.716***
	(.018)	(.238)	(.177)	(.038)	(.410)
Variance at higher level	.052	6.220	3.947	.405	53.534
	(.003)	(1.741)	(.297)	(.016)	(1.993)
Variance at lower level	.251	3.260††	3.254††	0.998	108.611
	(.003)	(—)	(—)	(.012)	(1.357)
ICC	.171	.656	.548	.289	.330
Models with device dummies (reference: PC completion) and respondent characteristics					
Constant	.279***	-4.414***	-3.303***	2.141***	29.369***
	(.033)	(.497)	(.314)	(.081)	(.887)
Tablet completion	-.025	.058	-.694***	-.060	1.862***
	(.018)	(.323)	(.177)	(.039)	(.417)
Smartphone completion	.028	.911***	-.410*	-.159***	9.498***
	(.018)	(.239)	(.179)	(.039)	(.414)
Gender (male)	-.015	.319	.051	-.049	.084
	(.012)	(.205)	(.112)	(.029)	(.319)
Age (centered)	.003***	-.038***	.032***	.004***	.133***
	(.001)	(.008)	(.004)	(.001)	(.012)
Education middle	-.070***	-.549	-.063	.182***	-1.028*
	(.019)	(.281)	(.182)	(.047)	(.515)
Education high	-.088***	-1.399***	.330	.214***	-1.556**
	(.018)	(.284)	(.172)	(.045)	(.492)
German	-.080**	-.856*	-.179	.184**	-4.230***
	(.029)	(.397)	(.268)	(.070)	(.766)

	Item nonresponse	Straightlining	Answered open question	Choice of left-aligned options	Duration
Living alone	.002	-.110	.308*	.045	.324
	(.017)	(.295)	(.152)	(.041)	(.449)
In paid work	.002	-.213	-.282*	.016	-.762*
	(.014)	(.231)	(.127)	(.034)	(.369)
Online survey experience	-.022	-.726*	.199	.010	-.229
	(.015)	(.295)	(.138)	(.037)	(.400)
Variance at higher level	.049	4.732	3.707	.396	48.926
	(.003)	(.800)	(.287)	(.016)	(1.865)
Variance at lower level	.252	3.260 ††	3.254††	.998	108.544
	(.003)	(—)	(—)	(.012)	(1.355)
ICC	.163	.620	.533	.284	.311

N (person-waves) = 15623, N (respondents) = 2793; coefficients are betas, ICC short for intra-class correlation, the ICC values higher than 0.5 mean that more variance is located at the higher level; standard errors in parentheses; *p<.05, **p<.01,*** p<.001, † Note that for the logistic models the variance at the lower level is fixed at $\pi^2/3$ (Hox 2010: 128), which equals approximately 3.290. ††rescaled variance to compare logistic models with each other, coefficients are also rescaled – all using meresc Stata command, ICC for the models calculated with rescaled variances. Duration was rescaled to minutes to avoid estimation problems. We excluded the number of characters in open question since it is conditional on providing an answer to an open question and due to estimation problems.

In the second step of our analysis we add the device dummies for tablet and smartphone completion (reference: PC completion) to the models to tease out device effects from other factors forming measurement occasion.[6]

Adding device indicators does not considerably lower the intra-class correlation coefficients, however, significant device effects are found. Completing online surveys using tablets is associated with fewer answers provided in open-ended questions and longer duration, which are statistically significant. Smartphone completion shows significant effects for all indicators with the exception of item nonresponse. Completing surveys with smartphones is associated with higher straightlining, providing fewer answers to the open-ended question, providing fewer

6 Device dummies indicate whether a respondent completed a questionnaire via smartphone, tablet, or PC for each of the measurement occasions. We constrain the effects of the devices to be equal at each measurement occasion since we expect that the content of the survey evaluation items does not influence the indicators of nonresponse and measurement error that we use. One exception is the duration that is a measure for the whole questionnaire. Since the inclusion of measurement occasion dummies did not substantially change the effects of the devices or respondent characteristics, but led to difficulties in the rescaling process for the logistic models, the measurement occasion dummies are not included in the final analysis.

characters in the open-ended question, increased choice of the left-aligned options, and longer duration.

In the final step, we control for respondent characteristics. Adding the respondent characteristics reduces the intra-class correlation coefficients substantially, especially for the indicators of straightlining and providing answers to the open question. We do not observe any implausible results with respect to respondent characteristics. For example, gender shows no significant effects, and it is plausible to not expect differences in data quality indicators we use based on gender. The effects of age are also not contra-intuitive: older respondents tend to provide data of better quality, generating lower item nonresponse, showing lower rates of straightlining, providing answers to an open question and more characters in the responses to open-ended questions. Taking longer to answer online surveys is also plausible. The higher levels of education (lower education being a reference category) are associated with lower likelihood of item nonresponse, straightlining, providing longer answers to the open-ended question and shorter duration, all of which could be the result of higher cognitive abilities. One rather puzzling indicator is the choice of left-aligned options with higher educated and older respondents showing increased choice of left-aligned options.

However, our focus in this analysis is less on the respondent characteristics but rather on their influence on effects of devices used for survey completion. The device effects found in models with device dummies are significant in the models with respondent characteristics. Using tablets for survey completion is associated with lower likelihood to provide answers to an open questions and longer duration. Those who complete surveys on smartphones show more straightlining, are less likely to answer an open question, provide shorter answers to an open question, and show longer duration when controlling for respondent characteristics. Overall, the results of multilevel models signify that completion of the survey on a mobile device has adverse consequences for data quality, especially when smartphones are used. Some indicators of nonresponse and measurement error are more affected than others: for example, the effects are largest for duration, but item nonresponse does not show significant results. Furthermore, the effects of completion of the online surveys using a mobile device cannot be fully explained by the choice of this device by the respondents.

4 Conclusions and Discussion

In this article, we study whether survey completion of online surveys using smartphones and tablets leads to higher measurement and nonresponse errors than when surveys are completed using personal computers or laptops. The analyses replicate and extend the approach chosen by Lugtig and Toepoel (2015), who show that

smartphone survey completion leads to a higher measurement error. In the GESIS Panel, a probability-based mixed-mode panel, the data source for our analyses, more respondents use smartphones for survey completion than in the LISS Panel, a probability-based online panel the data from which is used by Lugtig and Toepoel (2015).

We find that PCs prevail for completion of the online surveys, however, the average consistency for smartphones is about 60%, indicating that if the respondent completes one survey on a smartphone, on average, in 60% of the cases she will complete the next survey on a smartphone as well. For tablets, the average consistency is about 70%. Moreover, there is a slight increase in the proportion of respondents who use mobile devices for survey completion in the course of the six waves. Given that the GESIS Panel questionnaires are not optimized for mobile survey completion, studying the influence of mobile device use for survey completion on data quality is especially important.

We find that most of the indicators of measurement and nonresponse error are higher for mobile devices than for PCs. Online survey completion using smartphones shows higher item nonresponse, higher levels of straightlining in a grid question, lower rate of responding to an open question, and for those who do answer an open question providing shorter answers, as well as longer completion times compared to PC-completion. The differences found between smartphones and PCs are larger than the differences found between tablets and PCs, which is consistent with the results of previous research indicating that PCs and tablets lead to comparable results regarding data quality. For groups of respondents who switch between devices, the highest levels of measurement and nonresponse errors are found in groups, which involve smartphones. Nonetheless, the magnitude of the differences in measurement and nonresponse error indicators for various devices is rather small with the exception of survey duration with both tablet and smartphone respondents taking considerably longer to complete the surveys.

For the LISS Panel, Lugtig and Toepoel (2015) find that measurement errors do not increase when respondents switch from one device to the other. They conclude based on this finding that reporting with measurement error is a respondent-related characteristic. Our analysis of wave-to-wave device transitions shows significant effects in providing fewer answers to an open question for switches from PCs to smartphones or tablets, which is probably due to the absence of the keyboard, increased choice of left-aligned answer options in horizontal scales when switching from smartphone to PC, decreased choice of left-aligned answer options for the switch PC-smartphone, and longer duration for switches from PC to either mobile device. Changes in standardized nonresponse and measurement error indicators such as item nonresponse, straightlining, number of characters in open question are not significant. However, based on the multilevel analysis – an extension to the study we aimed to replicate – only item nonresponse is not predicted by tablet or

smartphone completion. Other indicators of nonresponse and measurement error that we use are affected by the device on which the survey is completed and cannot be attributed to the respondent since we control for respondent characteristics. The results of the multilevel models with device indicators differ somewhat from the replication of Lugtig and Toepoel (2015) analysis. This may be due to the fact that for wave-to-wave transitions only those transitions between two consecutive waves are considered, whereas for multilevel models the basis for analysis are all observations for respondents who took part in at least two waves, meaning respondents could potentially switch devices but did not have to participate consecutively, While the main focus of wave-to-wave analysis is the replication of the strategy chosen by Lugtig and Toepoel (2015), we did not want to exclude respondents who did not switch consecutively between the waves thereby losing the information in multilevel models.

Other reasons why our results only partially align with the results of Lugtig and Toepoel (2015) can be multiple. First, although we also use the evaluation part of the questionnaire so that the content stays the same across the waves, the content of the questions varies between the LISS Panel and the GESIS Panel. Thus, using exact same indicators of nonresponse and measurement error based on the same questionnaire content would be desirable. Another reason for the differences we find might be that the LISS Panel exists longer than the GESIS Panel, so that panel attrition or panel conditioning might be causes of the differences. If respondents who prefer completing surveys on mobile devices attrite at a higher rate and/or respondents who are longer with the panel learn to use mobile devices to report with fewer errors, fewer negative effects on the data quality will be found in the LISS Panel than in the GESIS Panel. This point warrants further investigation. Ideally, two panels existing for the same amount of time should be compared, but this is difficult to realize in practice.

Furthermore, our study is not free from limitations. First, our study does not assign the respondents randomly to a device, which limits our possibilities in studying nonresponse to item nonresponse only. Second, we do not have validation data and can only assess measurement errors using indirect indicators (of satisficing) such as straightlining, choosing left-aligned answer options in horizontal scales, survey duration. Nonetheless, our study provides a robustness check for the results obtained in a probability-based online panel in the Netherlands and extends the replication by including the respondent characteristics. Ideally, to separate selection effects one would use an experimental design. However, in the context of large-scale population surveys it is practically not feasible and studies that assign respondents to the device are confronted with the issues of respondent noncompliance (de Bruijne & Wijnant, 2013; Mavletova, 2013; Wells, Bailey, and Link, 2014) when some respondents complete the survey on their preferred device rather than the device to which they were assigned. One solution to this problem would be to match

respondents on a set of observable characteristics while the devices used for survey completion differ. We could not use this design as the groups completing the surveys on smartphones and tablets are still rather small, but given their rapid growth future studies should explore this option.

What are the practical implications of our analyses based on the results we obtained using the GESIS Panel data? The answer to the question whether survey completion using mobile devices is a problem that survey researchers should be concerned about is yes. Completing surveys with mobile devices, especially smartphones, is problematic. However, our analyses also indicate that for the most part the magnitude of these problems is not large: we find small to moderate effects. Although we cannot provide a definite answer to the question of how should survey designers deal with unintended mobile respondents since our findings are based on observational data, for the moment for GESIS Panel we do not see the need to address the issue of unintended mobile respondents based on the indicators that we use in this article. One notable exception is survey duration: for surveys in time-sensitive situations researchers need to investigate design options such as mobile optimization together with its consequences for data quality.

References

Antoun, C. (2015). *Effects of Mobile versus PC Web on Survey Response Quality: a Crossover Experiment in a Probability Web Panel*. Paper presented at the Annual Conference of the Americal Association for Public Opinion Research, Hollywood, FL.

Buskirk, T. D., & Andrus, C. H. (2014). Making Mobile Browser Surveys Smarter: Results from a Randomized Experiment Comparing Online Surveys Completed via Computer or Smartphone. *Field Methods*. doi: DOI: 10.1177/1525822X14526146

Callegaro, M. (2010). Do You Know Which Device Your Respondent Has Used to Take Your Online Survey? *Survey Practice, 3*(6), 12.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 111, 155-159.

Cook, W. A. (2014). Is Mobile a Reliable Platform For Survey Taking? Defining Quality in Online Surveys From Mobile Respondents. *Journal of Advertising Research*. doi: DOI: 10.2501/JAR-54-2-141-148

Couper, M. P., & Peterson, G. (2015). *Exploring Why Mobile Web Surveys Take Longer.* Paper presented at the General Online Research Conference 19.03.2015, Cologne.

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*(4), 482-504.

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

de Bruijne, M., & Wijnant, A. (2014b). Mobile Response in Web Panels. *Social Science Computer Review, 32*(6), 728-742. doi: 10.1177/0894439314525918

Hilligus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in nonprobability online panels. In M. Callegaro, P. J. Lavrakas, J. A. Krosnick, R. Baker, J. D. Bethlehem, & A. S. Göritz (Eds.), Online panel research: A data quality perspective (pp. 219-237). New York: Wiley.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York: Routledge.

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

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, California: Sage.

Lugtig, P., & Toepoel, V. (2015). The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review, Online first, February 26, 2015*. doi: 10.1177/0894439315574248

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly, 72*, 914-934.

Matthijsse, S. M., de Leeuw, E. D., & Hox, J. J. (2015). Internet panels, professional respondents, and data quality. *Methodology, 11*(3), 81-88. doi: 10.1027/1614-2241/a000094

Mavletova, A. (2013). Data Quality in PC and Mobile Web Surveys. *Social Science Computer Review, 31*, 725-743. DOI: 10.1177/0894439313485201

Mavletova, A., & Couper, M. P. (2013). Sensitive Topics in PC Web and Mobile Web Surveys: Is There a Difference? *Survey Research Methods, 7*(3), 191-205.

McClain, V. C., Crawford, S. D., & Dungan, J. P. (2012). *Use of mobile devices to access computer-optimized web instruments: Implications for respondent behavior and data quality.* Paper presented at the Annual Conference of the American Association for Public Opinion Research, Orlando, FL.

Peterson, G. (2012). *Unintended mobile respondents*. Paper presented at the CASRO Technology Conference 31.05.2012, New York City.

Peytchev, A., & Hill, C. (2010). Experiments in Mobile Web Survey Design: Similarities to Other Modes and Unique Considerations. *Social Science Computer Review, 28*, 319-335.

Poggio, T., Bosnjak, M., & Weyandt, K. (2015). Survey participation via mobile devices in a probability-based online panel: Prevalence, determinants, and implications for nonresponse *Survey Practice, 8*(2).

Rossmann, J., & Gummer, T. (2014). *Stata-ado package "PARSEUAS: Stata module to extract detailed information from user agent strings"*. Retrieved from: http://fmwww.bc.edu/repec/bocode/p/parseuas.ado

Stapleton, C. E. (2013). The Smartphone Way to Collect Survey Data. *Survey Practice, 6*(2).

Toepoel, V., & Lugtig, P. (2014). What happens if you offer a mobile option to your web panel? Evidence from a probability-based panel of Internet users. *Social Science Computer Review, 32*, 544-560.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. Jabine, M. Straf, J. Taur & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: building a bridge between disciplines* (pp. 73-100). Washington DC: National Academy Press.

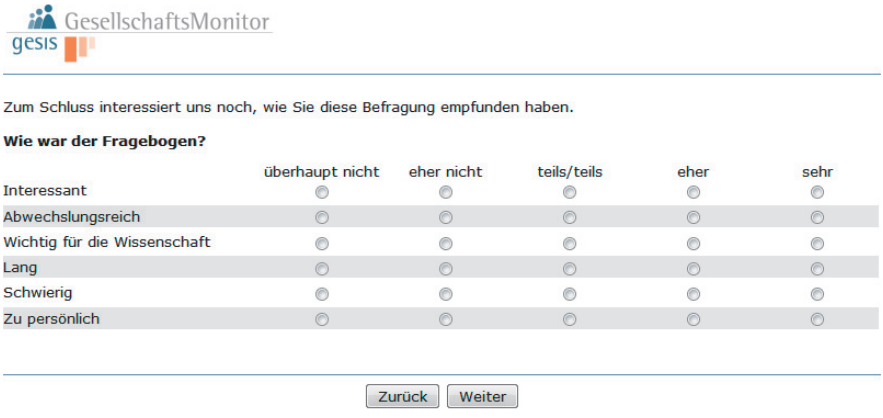
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.

Wells, T., Bailey, J. T., & Link, M. W. (2014). Comparison of Smartphone and Online Computer Survey Administration. *Social Science Computer Review, 32*(2), 238-255. doi: 10.1177/0894439313505829

Wijnant, A. (2015). Mobile devices in a web panel: what are the results of adjusting questionnaires for smartphones and tablets. Paper presented at the 6th Conference of the European Survey Research Association 14.07.2015, Reykjavik, Iceland.

Appendix

Screenshots of the questions used for analysis with translations.

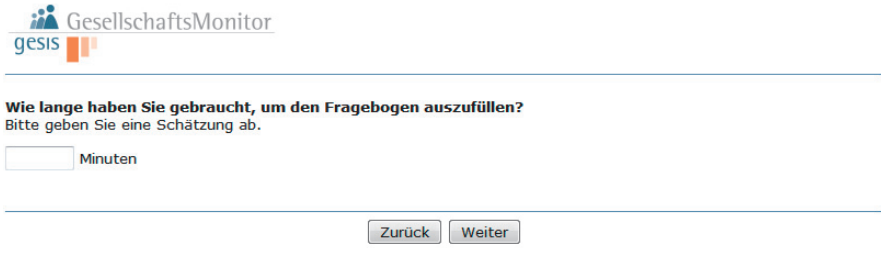


Question text: Finally, we are interested how do you feel about the questionnaire. How was the questionnaire? *Items*: interesting, diverse, important for science, long, difficult, too personal. Scale: not at all, rather not, partly, rather yes, very. (*)

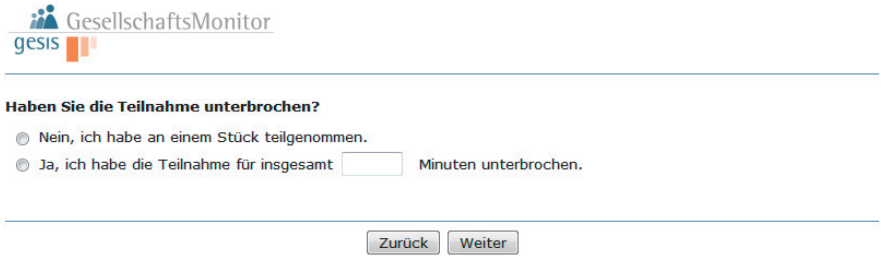
Question text: Did the survey encourage you to think about things? *Scale*: not at all, rather not, partly, rather yes, very. (*)

Question text: Were the questions sufficiently clear? *Scale*: not at all, rather not, partly, rather yes, very.

Question text: Overall, how did you like the survey? *Scale*: not at all, not so good, moderately, good, very good.



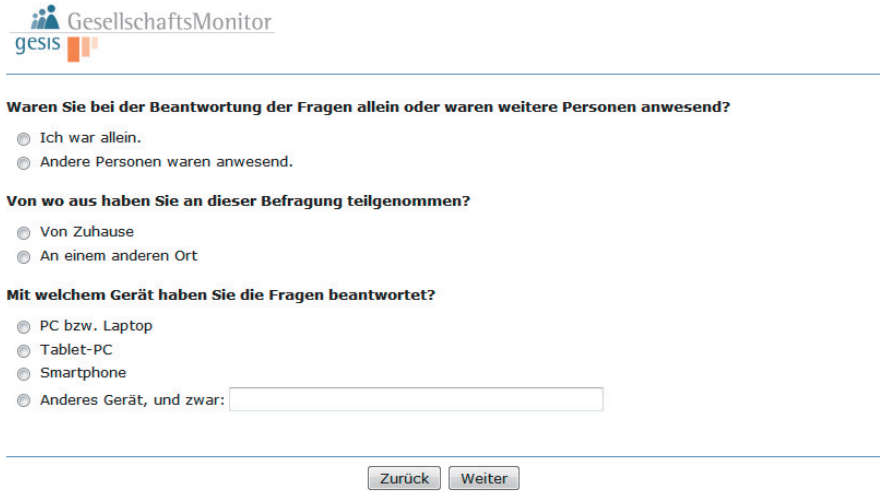
Question text: How long did it take you to complete the questionnaire? Please provide an estimation. __ minutes.



Question text: Did you interrupt your participation?

No, I completed the survey at once.

Yes, I took a break for … minutes.



Question text: Were you alone or were other persons present while you were answering the questions?

I was alone

Other persons were present

From what location did you participate in this survey?

From home

From another place

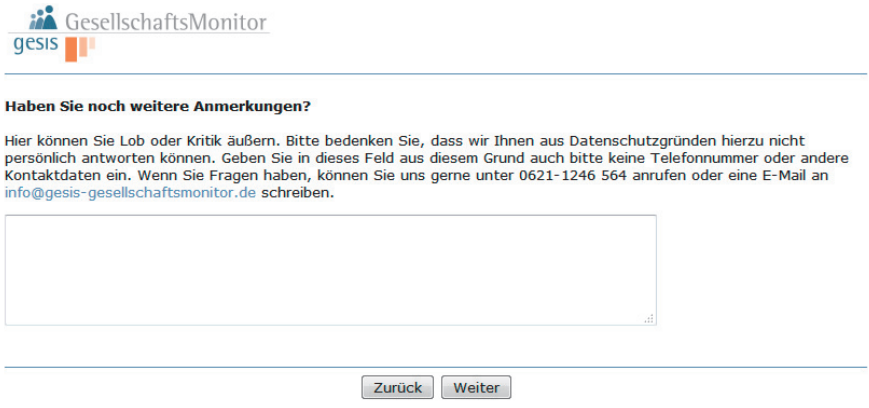
What type of device did you use to answer the questions?

PC or Laptop

Tablet-PC

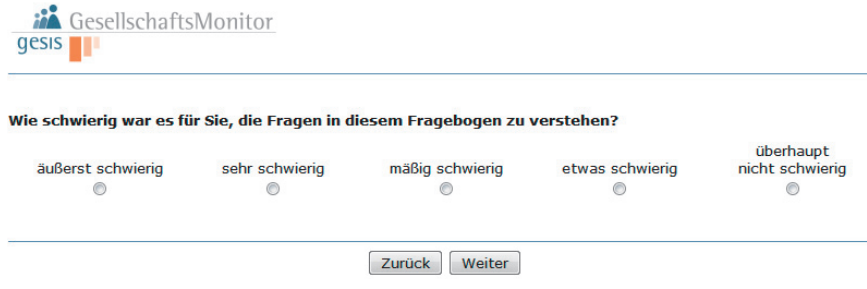
Smartphone

Other device, namely:

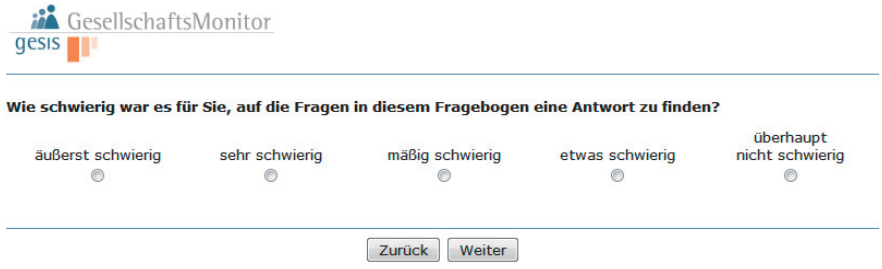


Translation: Do you have any further remarks?
Here you can express praise or critique. Please be aware, that we are not able to react to your comments due to data protection regulations. For these reasons, please do not write your telephone number or other contact information . If you have questions, you can call us on 0621-1246 564 or write us an email to info@gesellschaftsmonitor.de.

(*) Two items that were used for waves 3 to 6 instead of the two items that directly follow the evaluation matrix (marked with an asterisk):



Question text: How difficult was it for you to interpret the meanings of the questions in this questionnaire? *Scale*: Extremely difficult, very difficult, moderately difficult, slightly difficult, not difficult at all



Question text: How difficult was it for you to generate your answers to the questions in this questionnaire? *Scale*: Extremely difficult, very difficult, moderately difficult, slightly difficult, not difficult at all.

Equations for multilevel models:

Empty models for dependent variables "straightlining" and "answered open question":

$$\text{logit}(\pi_{ij}) = \gamma_{00} + u_{0j}$$

Empty models for dependent variables "item nonresponse", "choice of left-aligned options", "duration":

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Models with device dummies for dependent variables "straightlining" and "answered open question":

$$\text{logit}(\pi_{ij}) = \gamma_{00} + \gamma_1 tablet_{ij} + \gamma_2 smartphone_{ij} + u_{0j}$$

Models with device dummies for dependent variables "item nonresponse", "choice of left-aligned options", "duration":

$$Y_{ij} = \gamma_{00} + \gamma_1 tablet_{ij} + \gamma_2 smartphone_{ij} + u_{0j} + e_{ij}$$

Models with device dummies and respondent characteristics for dependent variables"straightlining" and "answered open question":

$$\begin{aligned} logit(\pi_{ij}) = {} & \gamma_{00} + \gamma_1 tablet_{ij} + \gamma_2 smartphone_{ij} + \gamma_3 gender_{ij} + \gamma_4 age_{ij} + \gamma_5 mid\,education_{ij} \\ & + \gamma_6 high\,education_{ij} + \gamma_7 german_{ij} + \gamma_8 living\,alone_{ij} + \gamma_9 in\,paid\,work_{ij} \\ & + \gamma_{10} online\,survey\,experience_{ij} + u_{0j} \end{aligned}$$

Models with device dummies and respondent characteristics for dependent variables "item nonresponse", "choice of left-aligned options", "duration":

$$Y_{ij} = \gamma_{00} + \gamma_1 tablet_{ij} + \gamma_2 smartphone_{ij} + \gamma_3 gender_{ij} + \gamma_4 age_{ij} + \gamma_5 mid\,education_{ij} + \gamma_6 high\,education_{ij} + \gamma_7 german_{ij} + \gamma_8 living\,alone_{ij} + \gamma_9 in\,paid\,work_{ij} + \gamma_{10} online\,survey\,experience_{ij} + u_{0j} + e_{ij}$$

where i is the lowest level (measurement occasion) and j is the highest level (respondent)

Authors Volume 9, 2015

- Ioannis Andreadis, Thessaloniki
- Birgit Arn, Adligenswil
- William G. Axinn, Ann Arbor
- Johann Bacher, Linz
- Michael Bosnjak, Mannheim
- Trent D. Buskirk, St. Louis
- Heather H. Gatny, Ann Arbor
- Aitana Gräbs Santiago, Mannheim
- Lars Kaczmirek, Mannheim
- Stefan Klug, Adligenswil
- Janusz Kołodziejski, Adligenswil
- Edith de Leeuw, Utrecht
- Lars Leszczensky, Mannheim
- Stefan Liebig, Bielefeld
- Peter Lugtig, Utrecht
- Joey Michaud, St. Louis
- Simone M. Schneider, Dublin
- Carsten Sauer, Bielefeld
- Ted Saunders, St. Louis
- Bella Struminskaya, Mannheim
- Vera Toepoel, Utrecht
- Peter Valet, Bielefeld
- James Wagner, Ann Arbor
- Daniela Wetzelhütter, Linz
- Kai Weyandt, Mannheim

Reviewers Volume 9, 2015

We would like to thank the following colleagues for their careful review of the manuscripts published in *mda*, Volume 9, 2015:

- Duane Alwin, Pennsylvania
- Bob Belli, Lincoln
- Michael Braun, Mannheim
- Marika de Bruijne, Tilburg
- Sebastian Bukow, Düsseldorf
- Britta Busse, Bremen
- Fanny Cobben, Den Haag
- Daniel Danner, Mannheim
- Siegfried Gabler, Mannheim
- Peter Grand, Wien
- Joop Hox, Utrecht
- Beat Hulliger, Olten
- Annette Jackle, Colchester
- Rüdiger Jacob, Trier
- Olena Kaminska, Cochester
- Florian Keusch, Mannheim
- Thomas Klausch, Utrecht
- Thomas Krause, Stuttgart
- Ivar Krumpal, Leipzig
- Aigul Mavletova, Moscow
- Morgan Millar, Salt Lake City
- Klaus Pforr, Mannheim
- Manuela Pötschke, Kassel
- Alice Ramos, Lisboa
- Karl-Heinz Reuband, Düsseldorf
- Joseph Sakshaug, Manchester
- Rolf Steyer, Jena
- Tim Spier, Siegen
- Pawel Sztabinski, Warsaw
- Daniele Toninelli, Bergamo
- Fons van de Vijfer, Tilburg
- Tom Wells, San Francisco
- Alexander Wenz, Colchester
- Arnaud Wijnant, Tilburg
- Michael Weinhardt, Bielefeld

Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript by e-mail to mda(at)GESIS(dot)org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should…
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
- Please also send us an abstract of your paper (approx. 300 words), a brief biographical note (no longer than 250 words), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - Tiff
  - Jpeg (uncompressed, high quality)
  - pdf
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.

Please follow the APA guidelines when preparing in-text references and the list of references.

Entire Book:

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: John Wiley & Sons.

Journal Article (with DOI):

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. *Consulting Psychology Journal: Practice and Research*, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

Journal Article (without DOI):

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Chapter in an Edited Book:

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research*. Second Edition (pp. 593-630). Bingley: Emerald.

Internet Source (without DOI):

Lewis, O., & Redish, L. (2011). *Native American tribes of Wisconsin*. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).

gesis
Leibniz Institute for the Social Sciences

ISSN 1864-6956 (Print)
ISSN 2190-4936 (Online)