Sources of error in web surveys: the role of respondent, design and device in data quality

Valerija Kolbas

A thesis submitted for the degree of
Doctor of Philosophy in Survey Methodology

Institute for Social and Economic Research
University of Essex

December 2018

Declaration

No part of this thesis has been submitted to this or any other university for another degree.

Chapter 3 is co-authored with Dr Peter Lugtig and Dr Vera Toepoel.

In memory of my father

Acknowledgements

In my experience, PhD was not just a process of obtaining an academic degree award but an opportunity to grow in ways that until then were not imaginable. Anything worth having does not come easily in life, and so my PhD was a challenging and primarily a lonely journey that tested my perseverance, resilience and self-confidence. Paradoxically, this was the time when more than ever I relied on support of people surrounding me. Whether knowingly or not, actively or simply by being there many people have been a great support for me throughout these years and I cannot imagine being where I am now without their presence in my life.

I would like to express my gratitude to Professor Nick Allum and Dr Tarek Al Baghal, my research supervisors, for their guidance, useful critiques of this research work and assistance in keeping my progress on schedule. I would also like to thank Professor Nick Allum for introducing me to the field of survey methodology and Dr Tarek Al Baghal for sparking my interest in the area of questionnaire design and psychology of survey response.

I would like to express my very great appreciation to Andrew Cleary for his assistance with the collection and access to my data. I would like to offer my special thanks to Gerry Nicolaas for mentoring me during my placement in the Ipsos-MORI and giving me a chance to participate in the exciting research projects and the daily operations of market research company.

I would like to thank the BOOST2018 team: Angus Holford, Emilia Del Bono and Adeline Delavande for letting me use their data for my research project.

I would like to thank my co-authors Peter Lugtig and Vera Toepoel for their guidance, enthusiastic encouragement and useful and constructive recommendations that have extended beyond our research project.

Special thanks should be given to my friend and colleague Alexander Wenz for continuous support, encouragement, useful and constructive recommendations and generous feedback that I have been lucky to receive throughout my PhD. I have greatly benefitted from discussions I have had with Roxana Baltaru, Vilma Agalioti-Sgompou, Brendan Read, Jennifer Kelley and Linh Nguyen over these years.

I would like to thank Sarah Swin - without your support I would not have gotten this far. I would also like to thank my mum and my extended family Laima Taylor and Karolis Butkus for creating a supportive environment throughout the years. Finally, I would like to thank Miruna Ivan for her ongoing support, patience and understanding throughout this challenging journey. Thank you for believing in me.

Summary

This thesis investigates how the design of response scales, the type of device used to respond to the survey, and respondent characteristics affect measurement error in web surveys.

Chapter 1 explores measurement error associated with different response formats in the National Student Survey, a cross-sectional web survey in the UK. Respondents who accessed the survey on a mobile device were randomly allocated to radio-button grids and drop-down response scale designs. Respondents using a PC viewed all questions in a radio-button format. Applying the radio-button design to tablets, smartphones and PCs reduces measurement error between devices. Drop-down lists were shown to reduce straightlining in grids on smartphones but increase primacy effects and affect response distributions.

Chapter 2 looks at scale direction effects in an online panel experiment administering survey questions with either ‘forward’ or ‘reverse’ ordered response scales. The ‘forward’ scale design results in higher selection of the high/positive responses whereas effects for the ‘reversed’ scale are less pronounced. Respondent age, education, gender and extraversion trait are associated with scale direction effects, suggesting the role of satisficing and anchoring mechanisms. The ‘forward’ scale design reduces selection of high/positive responses among conscientious respondents.

Chapter 3 explores the use of mobile devices for survey completion in an online cohort study. I find that respondents with a higher need for cognition and higher extraversion are more likely to use a PC for survey completion or switch between smartphones and PCs. More agreeable respondents are likely to use smartphones or tablets in addition to PCs for

survey completion. When controlling for these respondent characteristics, I find that item non-response is higher for mobile than PC respondents.

Collectively, these findings provide practical implications for survey designers with regards to scale design and mobile optimization, but also emphasize the role of respondent characteristics when predicting measurement error.

Contents

List of tables

List of figures

Introduction

In 1998, Couper and Nicholls stated that the expansion of computers in data collection is an inevitable process that will introduce crucial changes to survey practice (Couper & Nicholls, 1998). Over the following twenty years, computer technology became an essential tool in interviewer and self-administered surveys. Simultaneously, the development of internet technologies has opened new possibilities for web surveys. In 1998, the internet coverage in Great Britain was 9% (ONS, 2012). Over the following twenty years, the coverage has increased tenfold: by 2018 it has reached 90% (ONS, 2018). As a response to this rapid development of information technologies and increasing internet coverage, many major surveys have adopted the web survey mode as an additional or sometimes primary method of data collection and new probability-based online panels have been established. Furthermore, the probability and non-probability online panels have been welcomed by many in market research, public opinion research and academic research as a quicker and cheaper method of data collection (Couper, 2017). The capabilities of online web design have the potential to improve survey experience both for survey takers and survey researchers. Survey researchers could benefit from the opportunity to reach populations dispersed across a wide geographical area or hard-to-reach populations. Web survey mode offers a range of innovative design features: the possibility of creating complex survey designs, a relatively quick way of testing new design features as well as the possibility to collect rich paradata. Furthermore, filtering questions, forced responses, skip logics, tailored responses, and randomized questionnaire elements can improve response rates and quality (de Leeuw, 2008; Couper, 2008; Couper & Bosnjak, 2010). Furthermore, the design capabilities may improve data quality that researchers obtain by receiving less non-response or unclear, wrong answers (Biemer et al., 2017; Couper, Kennedy, Conrad, Tourangeau, 2011). The appearance of online surveys offered design solutions that could be more useful in collecting sensitive information and reduce respondent burden by implementing rich multimedia elements (Couper, 2008; de Leeuw & Hox, 2011). The design could be adapted to remove any ambiguities in wording of questions and responses, use visual and audio aid for additional clarifications, use skips where necessary – all this could potentially reduce response burden.

With the ever-changing capabilities of the internet and technology, the development of good online survey design practice is an ongoing process. However, new online survey design requires testing to ensure data quality is comparable to the previously established design solutions as well as to alternative data collection modes. In addition, the emergence of mobile internet has introduced yet another challenge to survey researchers as surveys could require an optimized design to reduce measurement error and yield comparable responses to surveys completed on PCs, smartphones and tablets. Certain web survey design features such as scale design, length and direction are still under scrutiny. Furthermore, there is still undergoing research studying ways the optimised response scale design can reduce cognitive burden and improve measurement error on mobile devices. More research is also needed to establish how the selection of devices for web survey completion contributes to measurement error in longitudinal web surveys.

Survey design and measurement error

Measurement error is a big part of the Total Survey Error (Biemer & Lyberg, 2003; Biemer et al., 2017; Groves, 1991). In web surveys, measurement error can be influenced by the question design, respondent characteristics (education, age, motivation, cognitive characteristics and perceived response burden), device being used to respond or due to the interaction between these sources. Previous studies demonstrated that small changes in question design or ordering of the questions can have a great impact on response quality. The choice of visual design elements, including presentation of questions on a page – number and spacing (Christian, Dillman, Smyth, 2007; Couper, Traugott, Lamias, 2000; Tourangeau, Couper, Conrad, 2004), use of additional audio and visual information, design of open text questions (Christian & Dillman, 2004), and choice of response scale design and direction (Christian & Dillman, 2004; Couper et al., 2000; Couper, Tourangeau, Conrad, Crawford, 2004; Tourangeau et al., 2004; Tourangeau, Couper, Conrad, 2007) can all influence measurement error (Biemer et al., 2017; Smyth, Dillman, Christian, Stern, 2006). A significant part of online survey research has been dedicated to the best online survey design practices (see Couper, 2008; Dillman, Smyth, & Christian, 2008). Nonetheless, certain web survey design issues– for example response scale design in web surveys (orientation, layout, direction, length of the scale, full or partial labelling, and selection tools) - are still under scrutiny. While some design features are included on the list of the best design practices other features such as scale direction are still under investigation (Liu & Keusch, 2017; Yan, Keusch, & He, 2018). Traditionally research design showed no preference for a particular scale direction (Yan & Keusch, 2015) while Tourangeau et al. (2004) suggested that respondents prefer scales starting with a positive option or the most desirable option first. More recent research identified different response strategies that a positive or a negative scale may encourage. Yan and Keusch (2015) suggested that satisficing and anchoring and adjustment heuristics are behind the scale direction effects – given responses change depending on whether respondents view a scale starting with a positive or a negative response first. Later, studies (Liu & Keusch, 2017; Toepoel, Das & Van Soest, 2009; Yan & Keusch, 2015) found empirical support for presence of satisficing and anchoring mechanisms in scale direction effects. Studies that have found both anchoring and satisficing mechanisms taking place in response behaviour further support the notion that both of these processes may occur when respondents are taking a survey. However, in order to identify the role of anchoring or satisficing in response behaviour and improve response quality more research is needed controlling for the explanatory factors of satisficing and anchoring such as respondent cognition and personality measures and demographics. To the best of my knowledge no studies have yet measured the psychographic moderators of satisficing and anchoring - cognition, motivation, demographics, and personality traits - to observe their unique role in scale effects and determine which response mechanism is taking place.

Device use and measurement error

The constant evolution of the internet and IT services has introduced mobile devices such as smartphone and tablet as an additional web survey element. In the early years of mobile surveys there have been some debates as to whether we should restrict survey access from mobile devices. However, by 2017 76% of UK population has owned a smartphone (Ofcom, 2017) and 78% accessed internet away from home or work, i.e. via a mobile device (ONS, 2018). As a result of the spread of mobile internet and device ownership as well as an ever-increasing mobile survey response Toepoel and Lugtig (2015) suggested that modern web surveys are ultimately mixed-device surveys.

Mobile survey participation has introduced a range of new possibilities as well as challenges for survey designers. Before responses collected from PCs and mobile devices could be analysed together it was important to ensure the comparability of the data. The main issues concerning mixed-device surveys are whether different devices lead to different outcomes: whether responses are affected by the usability of the survey design applied to different devices and the differences between mobile and PC respondents. Initial research using probability and non-probability panels found no significant differences in data quality between tablet and PC responses (Guidry, 2012; Lugtig &

Toepoel, 2015; Tourangeau et al., 2017) perhaps due to tablets sharing some features with PCs. On the other hand, non-optimized smartphone surveys have been causing worse data quality (de Bruijne & Wijnant, 2013; Mavletova, 2013; Mavletova & Couper, 2013; Mavletova & Couper, 2015; Wells, Bailey, & Link, 2014). Introduction of the mobile optimised design had a positive effect on non-response and measurement error but did not negate it completely (Mavletova & Couper, 2014; Mavletova & Couper, 2016a; Mavletova & Couper, 2016b). Still smartphone respondents were somewhat more likely to provide shorter answers (Wells et al., 2014) and demonstrated higher break-off and non-response (Couper & Peterson, 2016).

In order to reduce measurement error influenced by the type of the device researchers have identified several design solutions (see Couper, Antoun, Mavletova, 2017 for a full review). It was found that a higher number of items on a page could increase item non-response and grouping questions into grids produced more non-differentiation, higher break-offs, longer completion times as compared to an item-by-item presentation. In terms of the response scale design the format and layout of the scale have received a considerable attention. Due to the smaller screen size of mobile devices and a typically portrait mode of viewing vertically aligned designs such as drop-down lists were proposed as a potentially convenient design solution. However, early findings revealed that drop-down lists may cause a stronger primacy effect and differences in response distribution when compared to radio-button and slider-bar designs (Courtright, Pashupati, Pettit, & Knowles, 2013; Peterson, Mechling, LaFrance, Swinehart, & Ham, 2013; Stapleton, 2013). Furthermore, the type of the scale – slider bar or spin wheel – caused an increased error among mobile users (Peterson et al., 2013). The direction and alignment of scales could impact response distribution or item non-response (de Bruijne & Wijnant, 2014; Keusch, Yan, Han, & He, 2014). Even though smartphone surveys are an important source of data, it is yet to be established what scale design is not only most usable on smartphones but also produces most comparable to PC responses.

Respondent characteristics and measurement error

Respondents can become an additional source of measurement error when certain cognitive and personality characteristics (cognitive ability, need for cognition, motivation, Big Five personality traits) affect response behaviour and move responses away from the true estimate. It is important to understand how measurement error is affected by these sources in order to draw conclusions from survey data (Biemer, Groves, Lyberg, Mathiowetz, Sudman, 1991; Groves, 1991). Respondent characteristics can influence the survey outcomes at several stages of data collection: first, there might be differences between respondents who participate in the initial and follow-up studies and those who do not, second, there might be a difference between those who respond from PC and those who use smartphones exclusively and third, respondent differences might interact with the survey type and design and produce different quality responses.

Research so far has demonstrated that there are personality differences between respondents who participate in web surveys, who claim to agree to take surveys via mobile device and those who in the end respond from a mobile device (Antoun, 2015; Bosnjak et al., 2013; Butt & Phillips, 2008). Bosnjak et al. (2013) demonstrated that personality traits such as openness, conscientiousness and extraversion can predict respondents' participation in online surveys. Butt and Phillips (2008) showed that extraversion and agreeableness can predict mobile phone use, whereas Antoun (2015) showed that Big Five personality factors and need for cognition can predict respondents' willingness and likelihood to take surveys on mobile devices. These findings outline the importance of further research into the significance of respondent characteristics in the process of recruiting participants and collecting data from different devices. Furthermore, research found that socio-demographic composition of mobile respondents group is quite different from the rest of the online sample. Mobile respondents are more likely to be younger (Arn, Klug, & Kołodziejski, 2015; de Bruijne & Wijnant, 2014; Mavletova, 2013; Peterson et al., 2013; Sommer, Diedenhofen, Musch, 2017; Toepoel & Lugtig, 2014; Wells et al.,, 2014), to be female (de Bruijne & Wijnant, 2014; Peterson et al., 2013; Sommer et al., 2017; Wells et al., 2014), to have a higher income (Mavletova & Couper, 2014; Toepoel & Lugtig, 2014), to define themselves as early technology adopters (de Bruijne & Wijnant, 2014), to rely mainly on mobile internet (Mavletova, 2013; Mavletova & Couper, 2016a; Wells et al., 2014) and to have a higher trust in mobile surveys (Antoun & Couper, 2013; Bosnjak et al., 2010).

In order to understand how to improve design of survey elements it is important to study how respondents process the information they view and what factors affect their participation and response. Respondents' age, education and experience with the survey have demonstrated differences in employed response strategy and consequently quality of produced responses. Respondents who were older, less motivated, and had less experience with the survey were more likely to produce lower quality responses (Fang, Wen, Prybutok, 2013; Krosnick, 1999; Malhotra, 2008; Oppenheimer, Meyvis, Davidenko, 2009; Toepoel et al., 2009). Smartphone is considered to be a device less optimal for taking surveys due to the increased response burden, higher likelihood of distractions or multitasking (Couper et al., 2017). Therefore, taking into account known limiting effects of respondent psychographics on response propensity and quality could provide information for designing a survey that reduces cognitive burden on smartphone respondents.

As previously has been mentioned data quality collected from mobile devices and PCs can be affected by the devise usability and respondent characteristics. One of the approaches towards disentangling self-selection and device effects employs the power of longitudinal panels where panel members take part in several waves of a survey using a device of their choice (Lugtig & Toepoel, 2015; Struminskaya, Weyandt, Bosnjak, 2015). Early findings indicated that throughout the waves a group of respondents persistently have been using a smartphone. Longitudinal studies are able to answer whether these respondents are less concerned about providing good responses or they are more proficient in using mobile devices for surveys. Lugtig and Toepoel (2015) found a very low effect of device on measurement error and attributed differences in quality to the self-selection of respondents into devices. In turn, Struminskaya et al. (2015) found highest item non-response rates for smartphones or switching patterns that involve smartphones and related them to the device use rather than respondents. Recent findings suggest that further investigation is necessary to disentangle device from self-selection effects.

Roadmap to this thesis

The underlying theme I wish to investigate in this thesis is how elements of the survey design, in particular the design of response scales, the device used to respond to the survey, and respondent characteristics are associated with measurement error in web surveys. The first chapter focuses on the scale design effect on measurement error in PC and mobile web surveys. The second chapter explores the effect of the direction of the vertical Likert-type scales on response behaviour. The third chapter explores the role of the device on data quality and participation in a longitudinal online survey. The following paragraphs will briefly describe the objectives of each chapter.

The first chapter explores the effect of response scale design on measurement error in web surveys taken on smartphones, tablets and PCs. The objective of the first chapter is to compare web surveys with a drop-down or radio-button grid designs presented on smartphones and tablets and identify which design yields data which are most comparable to data collected on a PC. The differences in the layout can be subject to primacy effects and consequently affect differences in response distribution. However, when presented on devices with different screen size and response input methods, some response scale designs might be more user-friendly. For example, question grids are frequently employed in web surveys, however, when used on a smartphone, they often do not fully fit on the screen, resulting in some parts of the question or response options not being visible. Drop-down scales can be better fitted on the smaller screen; however they might encourage respondents to select the earlier options. Using data from the National Student Survey, a web survey on higher education in the UK, I explore how response scale design interacts with the device and affects measurement error. Based on the findings I conclude that smartphone data quality is affected by the instrument design to a greater extent, yet a greater selection of control variables could help to reliably disentangle the effects of device from self—selection on data quality. This chapter sets the stage for chapter two and three which use personality and cognition measurements to expand our understanding of the role of respondent characteristics in online research participation and mobile use in a new and promising context.

The second chapter studies the role of scale direction effects in response selection. Studies exploring the cognitive processes underlying response behaviour suggest that the elements of question design can affect how respondents process answers. The choice of scale direction is believed to stem from tradition rather than research evidence (Rammstedt & Krebs, 2007). There is a nearly equal split between studies employing ‘forward’-ordered scales, with the response scales starting with the highest or the most positive response and ending with the lowest or the most negative response, and studies employing a ‘reversed’-ordered scales, with the response scales starting with the lowest or the most negative response and ending with the highest or the most positive response (Krebs & Hoffmeyer-Zlotnik, 2010). The second chapter will look at scale direction effects that may occur due to the scale direction design in an online survey. Furthermore, I will investigate the cognitive mechanisms that are typically proposed to be responsible for the scale direction effects – satisficing and anchoring. In order to identify the role of anchoring or satisficing in response behaviour I will control for the explanatory factors of satisficing and anchoring such as respondent cognition, personality measures and demographics. The lack of an obvious overlap in explanatory personality variables between satisficing and anchoring frameworks indicates that they both might take place in response behaviour and work as complementary mechanisms. Controlling for the respondent characteristics, I aim to establish to what extent the scale direction effects arise due to satisficing or anchoring-and-adjustment mechanisms. The conclusions from the study will provide guidance on the scale design that produces the least biased results as well as poses less cognitive burden on respondents.

The third chapter looks at the role of respondent and device characteristics on data quality of a longitudinal online survey. Longitudinal online surveys are likely to recruit mobile respondents and switchers – respondents could use different devices to respond to different waves of the survey. Previous research has looked at the differences in demographics between mobile respondents and the effects of device on data quality. So far, no research has looked at the cognitive and personality differences between respondents who use exclusively PC, mobile device or respond from different device types when participating in a longitudinal survey. The third chapter explores the role of the device in the context of a cohort study. We are looking at whether certain respondent characteristics can be linked to a device use pattern and ways in which this can affect data quality. We aim to explore whether there are cognitive and personality differences between respondents who consistently use smartphones for surveys versus those who switch between devices or use exclusively PCs. While controlling for personal differences, we further explore the effect of device use on participation in all waves of the survey, break-off rates and item non-response.

1. The effect of grid and drop-down response formats on measurement error in PC and mobile web surveys

Abstract

The increasing use of mobile devices for web survey completion presents survey designers with new challenges. This study explores measurement error induced by different response formats in a mixed-device online survey. I used data from the National Student Survey (N=269 482), a cross-sectional web survey in the UK. Respondents self-selected into devices. Those who used a smartphone or tablet for survey completion were randomly allocated to either a radio-button grid format or a drop-down format with pre-selected positive, negative, or mid-point response option. Respondents using a PC all received the grid format. Surveys completed on smartphones were found to have longer completion times, higher break-off rates and shorter responses in open-ended questions. Drop-down pre-selected response options were selected more frequently on smartphones. Surveys completed on tablets were less affected by variations in design formats. Employing the same grid design for tablets, PCs and smartphones can reduce measurement error between devices. If drop-down response formats are implemented, then displaying instructions rather than a response in the initial box could reduce measurement error.

1.1. Introduction

Surveys administered via web are growing in popularity, however, a PC is no longer the only option to access web surveys. As the ownership and usage of mobile devices such as tablets and smartphones are ever increasing, limiting survey access to PCs can affect the size and representativeness of the sample (de Bruijne & Wijnant, 2013; Toepoel &

Lugtig, 2014). The trade-off for allowing access from different devices lies in potential differences in measurement error. All web surveys are computerized, self-administered, interactive in nature and rely on the visual channel of communication. Nonetheless, there are important differences between mobile devices and PCs that are assumed to affect measurement error: screen size, methods of navigation and input, speed of internet connection and processing power (Mavletova, 2013).

Over the past years, researchers have paid a lot of attention to the differences in data quality collected from PCs and smartphones. Proposed design solutions aimed at mitigating the effects caused by the design features such as question order (Crawford, 2004; Lugtig & Toepoel, 2015; Peytchev & Hill, 2010; Wells, Bailey, Link, 2014), questionnaire layout (Tourangeau, Couper, Conrad, 2004), scale orientation and response format (de Bruijne & Wijnant, 2014; Healey, 2007; Heerwegh & Loosveldt, 2002). Findings from previous research have not been conclusive and more can be learned about data comparability between surveys designed for PCs and smartphones. Tablets share some features with both PCs and smartphones. Earlier research found no significant differences in data quality between surveys taken on tablets and PCs (Guidry, 2012; Lugtig & Toepoel, 2015) and subsequent research on tablets has been scarce.

In this study, I compare the level of measurement error between PCs and mobile devices and determine the optimal mobile design that yields the most comparable data to PC. In this research, I use data from the National Student Survey 2014 administered in higher education institutions across the UK. Respondents could access the survey on a device of their choice. Those who accessed the survey on a smartphone or a tablet were randomly assigned to survey design conditions featuring different response formats: identical to PC design question grids with a radio-button response format or variations of drop-down formats that displayed either a positive, negative or mid-point pre-selected response option. The data quality is assessed using a set of indirect indicators of measurement error: response time, break-off rate, straightlining, probability of giving an open-ended answer, length of open-ended answers, selection of extreme and middle responses, and selection of non-substantive responses. The results provide insights into appropriate response designs for each of these devices.

1.2. Background

Differences between Internet-enabled devices

The main features that distinguish PCs from smartphones and tablets are the size of the screen and the absence of a mouse and a physical keyboard. Mobile devices (i.e. smartphones and tablets) have smaller screens that can limit the amount of information displayed and affect how the survey is seen and comprehended by respondents (Peytchev & Hill, 2010; Toepoel & Lugtig, 2014). In turn, these features might have different implications on data quality. Couper, Tourangeau, Conrad and Crawford (2004) discovered that web survey responses administered on PCs demonstrated the *visibility* effect – response options that were initially visible were selected with a higher frequency. Peytchev and Hill (2010) compared how variations in questionnaire and scale design affect responses collected from PCs and mobiles. They have concluded that a smaller mobile screen and a keyboard introduce undesirable differences in responses: 23% of mobile respondents reported not noticing that there were further responses or claimed that it was too much effort to scroll to see them. These findings stress the importance of accounting for specific features of each device when designing mixed-device surveys.

There are several ways in which users interact with mixed-device surveys that could account for differences occurring in data quality. The methods of screen navigation and data entry for mobile devices and PCs are quite different. PC users typically rely on a combination of mouse and a keyboard, whereas smartphone and tablet owners select answers by finger-touches and use on-screen keyboards to type responses (Lugtig & Toepoel, 2015). This is particularly evident in surveys with open-ended questions: As using a keyboard to type responses is more convenient than a touchscreen, longer answers would be expected for PC respondents (Buskirk & Andrus, 2014; Mavletova, 2013; Toepoel & Lugtig, 2014). While research has provided evidence that PC users give longer answers to open-ended questions than smartphone users (Lugtig & Toepoel, 2015; Peterson, Mechling, LaFrance, Swinehart, Ham, 2013), considerably fewer studies took tablet respondents into account and the existing evidence does not support the assumption that larger screen size and more convenient entry methods allow for more substantive open-end responses.

Measurement error related to mobile devices in online surveys

In 2017, 76% of the UK population owned a smartphone (Ofcom, 2017). According to the Office for National Statistics 78% of respondents accessed the internet away from home or work, i.e. via a mobile device (ONS, 2018). As the popularity of mobile devices is ever increasing, more researchers have turned their attention towards surveys administered on smartphone and tablet trying to uncover differences between data obtained from such devices and PCs.

There may be different aspects of measurement error associated with smartphone, tablet and PC devices. So far, research has consistently identified that compared to PC, surveys completed on smartphones result in higher break-off rates and longer completion times (Mavletova, 2013; Peterson et al., 2013; Stapleton, 2013). Wells, Bailey and Link (2012) explored whether self-selected tablet, smartphone and PC survey data differ in terms of break-off rates, completion times and item-missing data in a non-probability survey. Their analysis demonstrated that data quality was comparable between tablets and PCs, whereas smartphone surveys yielded higher break-off rates and longer completion times. In a probability web study, de Bruijne and Wijnant (2013) compared response times between tablet, smartphone and PC respondents and found that only smartphone respondents took significantly more time to complete a survey. The reasons behind such results were attributed to survey design and device features: PC is generally more convenient for completing questionnaires, whereas the smaller screen of the mobile device would cause difficulties with navigating the questionnaire, reading the question and, possibly, selecting an answer. Design optimized for smartphones seems to reduce the break-off rates; however, they remain higher than those of PCs. On the other hand, tablet completion times are often comparable to PC.

Furthermore, some differences in responses to open questions provided by smartphone and PC respondents were found due to the data entry format and screen size. Peytchev and Hill (2010) demonstrated that smartphone respondents were more reluctant to provide open-ended responses, whereas other studies performed by Mavletova (2013), and Toepoel and Lugtig (2014) found that even though answers might be shorter, the content of the feedback could still be substantive. Furthermore, Antoun, Couper and Conrad (2017) found that smartphone users provided longer answers to an open-ended question. Possibly, later developments that led to an increase in screen size of mobile devices, the addition of auto-fill and voice input functions as well as respondents becoming more proficient with mobile devices could bridge the gap in response length to open-ended questions between the devices.

Non-optimised surveys are more burdensome for mobile respondents and could lead to a particularly high level of straightlining, however even some optimised designs may fail to reduce straightlining (Peterson et al., 2013). There is less research on data collected from tablets, nonetheless in some instances tablet and smartphone respondents seem to engage in less straightlining than PC respondents (Lugtig & Toepoel, 2015). So far studies have not shown conclusive evidence of whether smartphone, tablet or PC surveys show significantly different levels of straightlining.

Studying data quality and device use in a longitudinal context supported previous findings on device effects. Lugtig and Toepoel (2015) investigated direct and indirect indicators of measurement error (item missings, straightlining, open questions, length of open questions, primacy effect, number of answers checked in a check-all-that-apply question, duration of the survey, evaluation of the questionnaire) in six waves of a survey completed on a PC, smartphone or tablets and observed how switching between the devices affected the data quality. They found that PC surveys demonstrated the least evidence of the measurement error, followed by tablet and smartphone respondents; PC users were more likely to complete optional open questions and give the longest answers. PC respondents also showed less preference for the first options and overall were more positive in their responses. Switching from PC to other devices did not result in a substantive increase of measurement error following the switch. Furthermore, the effect sizes of the observed differences in the measurement error were generally small. The authors have warned that the higher measurement error in tablets and smartphones could be due to the self-selection of the sample into using a particular device rather than a device effect itself.

So far, research suggests that tablets provide better data quality than smartphones and, in some instances, tablet data might be comparable to PC, however more studies involving self-selected tablet respondents would provide better insight into the data quality and demonstrate whether it is plausible to account for tablets when designing and analysing a survey.

Measurement error related to the response format design in mixed device surveys

A considerable amount of research has studied whether surveys designed for PCs are equally suitable to be administered on mobile devices or have tried to ascertain the alternative mobile optimised design. Employing a new survey design for mobile devices could improve response rates and usability of the surveys, however if untested it could become a new source of measurement error (Peterson et al., 2013; Peytchev & Hill, 2010). Data collected from web surveys presented on mobile devices demonstrated that the design of response scales can cause increased measurement error in mobile web as well as PC web surveys. In a randomized mobile experiment Stapleton (2013) found that horizontally presented scales cause mobile respondents to select leftmost options more often. In the study, participants received surveys with a radio-button response format that started either with the extreme negative or extreme positive options. The findings provided evidence of the visibility effect – options initially visible before scrolling were preferred more often. This effect was not found in vertically aligned scales or in drop-down lists. Using an online non-probability panel Peterson et al. (2013) explored whether an optimised mobile survey design can improve user experience while providing results comparable to PC. The experiment used two versions of a PC survey with either radio-buttons or fully labelled vertical grids. There were several versions of mobile surveys: two of them were replicating the PC format, others used drop-down lists, slider bars, numeric responses. The highest level of straightlining was found among PC respondents; mobile format with slider bars showed evidence of midpoint anchoring and drop-down format demonstrated higher preference for first options. Overall, however, analyses revealed no significant differences between PC and mobile versions of survey in data distribution. In the end, researchers concluded that surveys should use the format with a horizontal scale alignment optimized both for PC and mobile versions of surveys.

Radio-button scale design presents all options immediately therefore respondents have no trouble navigating a survey with such format. Radio-button response format takes respondent only one click to choose an answer and does not disrupt the view of other questions presented on a screen. However, in a traditional non-optimised design, radio-buttons cannot be resized, and respondents are presented with a small target area to select a response which may be inconvenient for those with a smaller screen size and are using a touch input (Couper, 2008; Healey, 2007; Heerwegh & Loosveldt, 2002). Heerwegh and Loosveldt (2002) demonstrated that surveys using radio-button format had shorter completion times and less drop-outs, nonetheless researchers argued that the radio-button format is in no way superior to the drop-down list format and a choice between each of them should be made in the context of a particular survey, type of questions and IT proficiency of the target population.

The benefit of drop-down format is that it can save space on the screen as initially it presents only one line that might be left blank, contain instructions or a default option whereas the entire list becomes visible only upon clicking on the box: this format avoids the need for scrolling to the right (Heerwegh & Loosveldt, 2002; Stapleton, 2013). However, drop-down lists were found to cause stronger primacy effects (Stapleton, 2013) and some researchers argued that if drop-down lists cannot be avoided, then placing instructions in the box would avoid privileging the first response option (Dillman et al., 2008). Couper et al. (2004) tested the effect of radio-button versus drop-down lists with five response options initially visible and drop-down list with instructions placed in the initial box. They found that decision making in web surveys is subject to mental shortcuts and found some evidence of order effects – options higher on the list and those initially visible in a drop-down list were more likely to be selected. They concluded that the choice of response format in a survey leads to a different distribution of responses and a radio-button format is not equal to drop-down lists.

Research on the effects of the response format demonstrated that response distributions are equally affected by response formats in mobile as well as PC surveys. However, there is no conclusive evidence whether horizontal or vertical scale alignment leads to a decreased measurement error (de Bruijne & Wijnant, 2014a; Peterson et al., 2013; Peytchev & Hill, 2010; Stapleton, 2013). So far, some ambiguity still remains with regards to the optimal design for smartphone surveys and research has not reached a uniform opinion on whether radio-button or drop-down format yields less bias in data and whether mobile devices are less affected by horizontal or vertical presentation of responses. Even less is known about the optimal survey design for tablets as not many studies involving design manipulations accounted for tablet devices.

Research Objectives

In order to identify a survey design that performs equally well across devices ideally a fully crossed-over experiment where each experimental design is allocated to each device is needed. When working with the large-scale survey used to inform policies I have faced certain restrictions on the experimental design implementation due to the potential loss to the data comparability - it was not possible to implement the experimental designs in the main PC respondent sample. Therefore, the final experimental design presented in this study consists of the default PC design (radio-button grid), default mobile design (positive option pre-selected drop-down list) and three experimental design conditions (radio-button grid, negative option selected drop-down list and mid-point selected drip-down list).

This paper aims to answer how the design of response scales affects measurement error in smartphone and tablet surveys and which response scale design yields data that are most comparable to PCs. When comparing measurement error between devices, I expect longer completion times and higher drop-out rates for mobile devices compared to surveys completed on a PC. I am using two aspects of open question responses in the measurement error framework. First, I record whether respondents answered an optional open question at the end of each questionnaire asking for additional feedback. Along with the fact whether any answer was given, I also code the length of the answer in characters. Shorter or no answers are used as a proxy for a higher measurement error. Based on previous findings, I expect no difference in the likelihood of providing an open text response across devices, yet responses collected from smartphones and tablets to be shorter than those obtained from PC respondents. Respondents who select mobile devices to take surveys should be proficient in providing longer text responses using a virtual keyboard. On the other hand, those who are less skilled in typing text responses on mobile devices will probably not use mobile device to take the survey at all. I expect the scale design to have no direct effect on the length or probability of an open text response. However, if users find a particular experimental design particularly burdensome they might be less willing to provide an optional open text feedback at the end of the survey, or the feedback provided will be significantly shorter.

It is possible that a number of factors such as connection speed, multitasking while filling in the survey, or more burdensome survey design may be responsible for a decreased respondent motivation and rates of straightlining and non-substantive responses. I assume smartphone to be a more demanding device for survey completion and predict smartphone survey data to show a higher level of straightlining than PC or tablet. Another measure capturing data quality is the frequency of selecting non-substantive responses.

Higher levels of non-substantive responses were suggested to indicate higher levels of measurement error (Lugtig & Toepoel, 2015). If respondents become more burdened when taking a survey on a particular device, they can opt for selecting '*Not applicable*' instead of putting more cognitive effort and selecting a substantive response. Therefore higher rates of non-substantive responses are expected among smartphone respondents.

Finally, I will look at the difference in the distribution of responses across response formats of mobile devices and compare it to the radio-button grid PC survey design. Radio-button format is generally recommended to be used in surveys administered on PCs whereas drop-down lists are more often recommended for smartphone design due to screen size limitations. The choice of response format, i.e. drop-down versus radio button, demonstrated to have an effect on the response distribution (Peterson et al., 2013). As smartphones and tablets have design characteristics different from PC, it might influence usability. These devices share a touch input method which can affect how convenient each particular design is to navigate and select responses. Smartphones have smaller screens, therefore some formats for this device will require vertical or horizontal scrolling which can potentially affect selection of responses and produce different data. Consequently, I expect that the response scale designs will have different implications on data quality when taken on smartphones and tablets as opposed to PCs.

Smartphone respondents were more likely to select response options that were visible on the screen before scrolling (Peytchev & Hill, 2010). As for tablet respondents Lugtig and Toepoel (2015) found that the primacy effect was slightly higher compared to PC however the difference was not significant. Consequently, the current experiment suggests that in the radio-button grid format the visibility should result in a higher frequency of selecting the first two positive options for smartphone users, whereas no such effect will appear in tablet and PC responses. When looking at the measurement error between different response formats in mobile devices I expect higher levels of selecting first two positive responses in a radio-button format in the smartphone surveys, but no such effect in the tablets.

As smartphone is a less optimal device for taking surveys I expect it to have a higher burden on respondents. As a result, drop-down list designs with pre-selected positive, negative or middle response will result in a higher endorsement of these pre-selected options. As tablets are more convenient devices for survey completion the response distribution of survey data collected from tablets is expected to be less affected by the initially suggested options.

1.3. Data and Methods

Survey Design

This paper uses data from 269 482 students enrolled in 326 Higher Education Institutions in the UK who took part in the online version of the National Student Survey (NSS) fielded between January and April 2014. The survey was fielded by the market research organization Ipsos-MORI on behalf of the Higher Education Funding Council for England. This survey is fielded annually and administered to final year undergraduate students. It serves the purpose of informing prospective student choice, ensuring public accountability and providing feedback for higher education institutions to improve the academic experience of students. The survey uses a mixed-mode approach combining online, post and telephone surveys; invitations are distributed to each student via their university email address supported with the text message prompts (Higher Education Funding Council for England report, 2014). No universal incentive is offered across Higher Education Institutions; however, some Universities may offer an incentive in the form of a prize draw or a reward (Higher Education Funding Council for England report, 2012). This experiment uses an anonymized version of the data set, meaning that response data could not be linked to respondents' sociodemographic variables.

The survey uses 22 core questions split into six blocks that assess attitudes across several domains of student life: Quality of Learning and Teaching, Assessment and Feedback, Academic Support, Organisation and Management, Learning Resources, Personal development and Overall Satisfaction (University of Bristol report, 2016). Questions used paging design with questions assessing one topic presented on a separate page. There are no reverse-scored items, all questions are forced-choice and use a 5-point Likert scale ranging from '*Definitely Agree*' to '*Definitely Disagree*'. '*Not Applicable*' is always offered as a last response option. At the end of the survey, all respondents are able to leave positive and negative feedback of up to 4000 characters in two open text boxes. The open-ended questions are not mandatory to complete.

Experimental Design

Respondents who completed the survey on a PC (desktop or laptop) were always offered a radio-button grid response format (Figure 1.1). The mobile devices were experimentally assigned to the PC grid design or mobile drop-down list design. In this and all previous years of the survey, the drop-down list with the positive initially displayed response ('*Definitely Agree*') was the default response format for surveys completed on mobile devices (Figure 1.2). Additionally, several alternative response designs were introduced to the smartphone and tablet surveys: a radio-button grid design, identical to the PC version (Figure 1.1), or an optimized design presenting a drop-down list with the initial response pre-selected as '*Definitely Disagree*' (Figure 1.3) or '*Neither Agree nor Disagree*' (Figure 1.4).

The smartphone radio-button grid format allowed all questions in a section to be visible straight away; however, the response grid required scrolling to the right, as only the first two most positive answers were visible on the screen right away. The design allowed both portrait and landscape views, therefore if participant decided to rotate the device horizontally to view the survey, the entire grid became visible.

To make a selection in the drop-down design, participants had to click on the initial box to view a vertical list of all options starting with the *'Definitely Agree'* and then select a response. If participants left the pre-selected responses for all questions on the survey page, a message would appear once the 'Next' button is selected asking if the participant is happy with the responses and click the 'Next' button again to confirm their response and proceed to the next page.

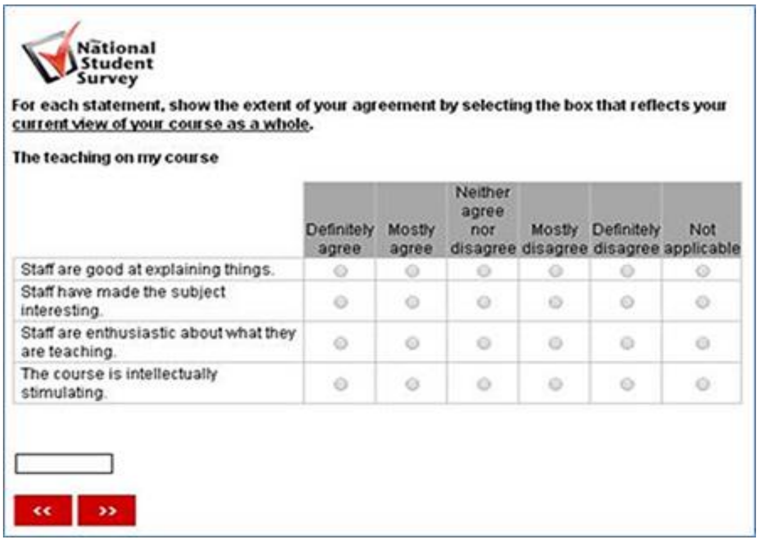

Figure 1.1. Radio-button grid design of the questionnaire presented on a PC and mobile device surveys in NSS, 2014.

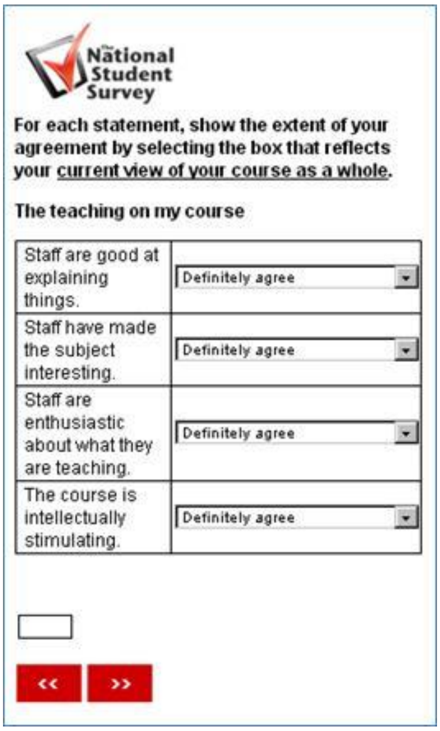

Figure 1.2. Drop-down list design with a pre-selected positive response in a mobile device survey in NSS, 2014.

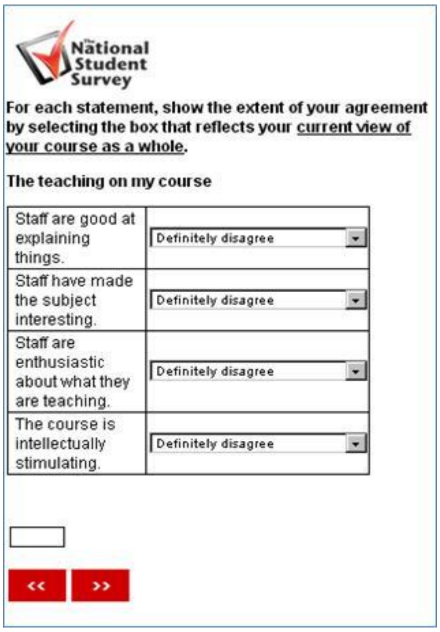

Figure 1.3. Drop-down list survey design with a pre-selected negative response in a mobile device version of NSS, 2014.

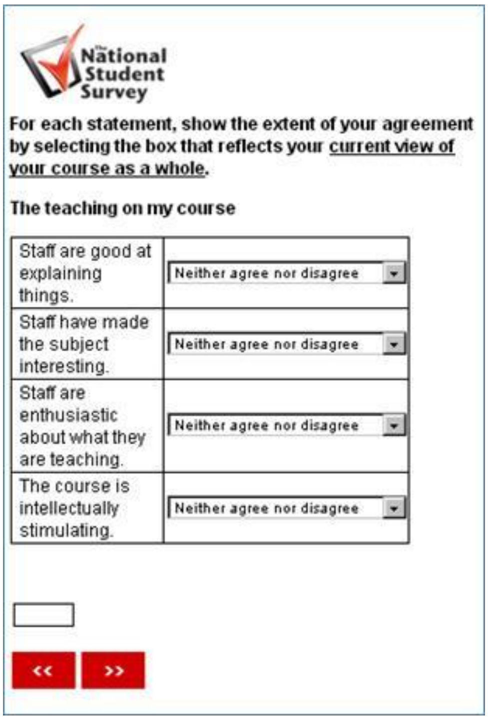

Figure 1.4. Drop-down list survey design with a pre-selected mid-point response in a mobile device version of NSS, 2014.

Figure 1.5 demonstrates the assignment experimental conditions to respondents. Participants could choose any device to access the survey. PC respondents always viewed a default radio-button grid design. Mobile device surveys were presented in a default positive response drop-down list or one of the experimental response design conditions. If respondents opened the survey on a smartphone or a tablet, they were randomly allocated to one of the experimental response designs. All five response format conditions were in the field for a similar duration of time – the mean number of days in the field ranged between 40.4 days and 42.4 days.

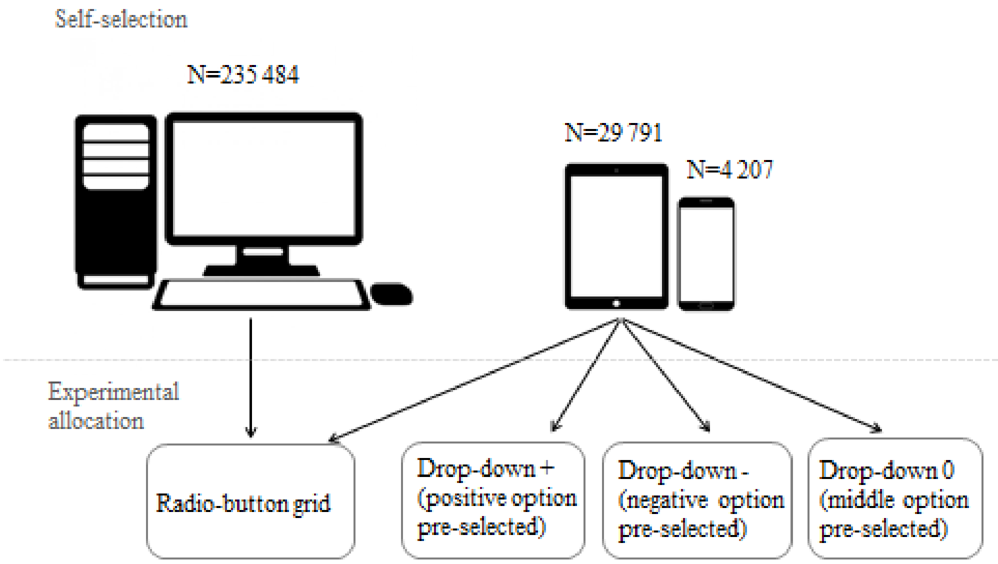

Figure 1.5. Allocation of experimental response scale design conditions to PCs and mobile devices.

1.4. Results

Measurement error indicators for PC and mobile surveys

Table 1.1 presents descriptive statistics for the measurement error indicators for different response scale designs for PC, smartphones and tablets.

Median response time. The longest median response time of 9.9 minutes was observed among smartphone users who viewed drop-down design with the middle response pre-selected. PC users needed the least time – 6.9 minutes to complete a survey with a grid design. Overall, the results suggest that smartphone users needed a slightly longer time to complete the survey across all experimental conditions, as compared to tablet users who submitted responses in a shorter time. The difference in median response times between groups was statistically significant as determined by a one-way ANOVA, F (8, 269 473) = 25.63, $p < 0.001$.

Break-off. Break-off indicated respondents who failed to complete the core part of the survey till the end. The lowest break-off rate was observed for the PC grid design and tablet users with a middle response pre-selected drop-down list – only 5.4% failed to complete the survey. Overall, tablet break-off rates were consistently low across all design conditions and were only slightly higher for the default mobile format – 7.3% break-off in the positive response pre-selected drop-down list. Conversely, the break-off rates of smartphone users were the highest across all conditions. The lowest break-off rate of 14% was in the grid radio-button design whereas the highest break-off rate of 20.1% was observed in a drop-down list with a pre-set middle response condition. A one-way ANOVA showed that the difference in break-off rates between groups was statistically significant, F (8, 269 473) = 887.53, $p < 0.001$.

Response rate and median length of open text response. I recoded the proportion of respondents who responded to the optional questions asking to provide positive and negative additional feedback at the end of the survey. The response rates to the feedback questions were comparable across all device conditions, albeit PC had a slightly higher rate of optional feedback - 66.9% of all respondents provided a response. The response rates on tablets were slightly lower and ranged from 62.2% in the pre-selected middle response drop-down design to 66% in the radio-button design. The response rates on smartphones were only slightly lower and ranged from 55.8% in a middle response drop-down condition to 58.9% feedback rate in both the negative pre-selected option drop-down and radio button grid designs. The difference in proportion of respondents who responded to the optional open text questions between groups was statistically significant as determined by a one-way ANOVA, F (8, 269 473) = 164.92, $p < 0.001$.

Furthermore, I also report the median length of feedback in characters for those who provided a response. Tablet users who received the middle response drop-down design provided a much longer feedback with a median of 470 characters. The other designs received responses that were very similar in length, ranging from 231 characters in a negative response drop-down design for tablet surveys to 298 characters for the radio-button PC survey. An ANOVA showed that the difference in feedback length was statistically significant between device and response design groups, F (8, 176 967) = 80.89, $p < 0.001$.

Straightlining. Straightlining indicated respondents who provided the same response to at least two-thirds of the questions answered in the survey. In the surveys taken on PCs, 39.6% of respondents straightlined. The highest rate of straightlining (43.3%) was among smartphone users who viewed a positive response drop-down format. This response design also caused the highest rate of straightlining among tablet users – 37.9%. A lower straightlining rate among smartphone users was in the drop-down list design with pre-selected negative and middle response options – 32.5% and 32.4% respectively. The lowest straightlining rate was observed among tablet users who received negative option selected drop-down list design – 29.6% of respondents have straightlined in two-thirds of the questions. The difference in straightlining rates between groups was statistically significant as determined by a one-way ANOVA, F (8, 269 473) = 23.46, $p < 0.001$.

Table 1.1. Measurement error indicators for each device and response design in the NSS, 2014 web survey.

	PC	Smartphone				Tablet			
	Radio-button	Radio-button	Drop-down+	Drop-down -	Drop-down 0	Radio-button	Drop-down +	Drop-down -	Drop-down 0
Median Response Time (minutes)	6.9	8.3	9.8	9.6	9.9	8.5	8.4	8.4	9.1
Break-off %	5.4	14.0	18.4	16.3	20.1	5.7	7.3	5.7	5.4
Completed open answers %	66.9	58.9	56.3	58.9	55.8	66	62.7	64.8	62.2
Median open text response length	298	254	243	232	242	285	277	231	470
Straightlining %	39.6	37.2	43.3	32.5	32.4	36.8	37.9	29.6	35.1
Sample size	235 484	772	27 666	695	658	106	3 939	88	74

Extreme and mid-point responding. Non-substantive answers. Figure 1.6 presents the proportion of selected extreme, middle or non-substantive responses within each survey for PCs, smartphones and tablet users. The proportion has been calculated by the number of times respondent selected either an extreme positive, extreme negative or middle response option divided by the total number of questions that the respondent has answered. The highest proportion of extreme positive responses was observed in the smartphone drop-down design with the positive option pre-selected – on average 53% of all responses were *'Definitely Agree'*. Tablet respondents with the grid design experimental condition had the lowest average proportion of positive responses – 31.5%. The difference in proportion of extreme positive responses between groups was statistically significant as determined by a one-way ANOVA, F (8, 269 473) = 50.69, $p <$ 0.001. The highest proportion of extreme negative responses – 9% – was among smartphone users with a negative initial option drop-down response format, whereas the lowest proportion 1.8% of total responses was in the tablet device with negative displayed option drop-down list format. The difference in proportion of extreme negative responses between groups was statistically significant as indicated by a one-way ANOVA analysis, F (8, 269 473) = 683.12, $p < 0.001$.

The highest frequency of middle responding was in the smartphone condition that offered middle response *'Neither Agree nor Disagree'* as an initial option representing 15.8% of all responses in this design group. The lowest proportion – 7% – was among tablet users who got the negative displayed option drop-down list format. The difference in proportion of mid-point responses between groups was statistically significant as determined by a one-way ANOVA, F (8, 269 473) = 72, $p < 0.001$.

The selection of non-substantive options was extremely low among all groups; the highest endorsement of *'Not Applicable'* option was 1.6% in a tablet with a negative option drop-down format, whereas the lowest frequency of 0.3% was in the smartphone middle displayed response drop-down group. A one-way ANOVA showed that the difference in proportion of non-substantive responses between groups was statistically significant, F (8, 269 473) = 18.12, $p < 0.001$. So far, the results for smartphones showed some evidence of a relationship between pre-selected initial response format and a higher frequency of selecting such an option. The responses on tablets were not affected by the pre-displayed option.

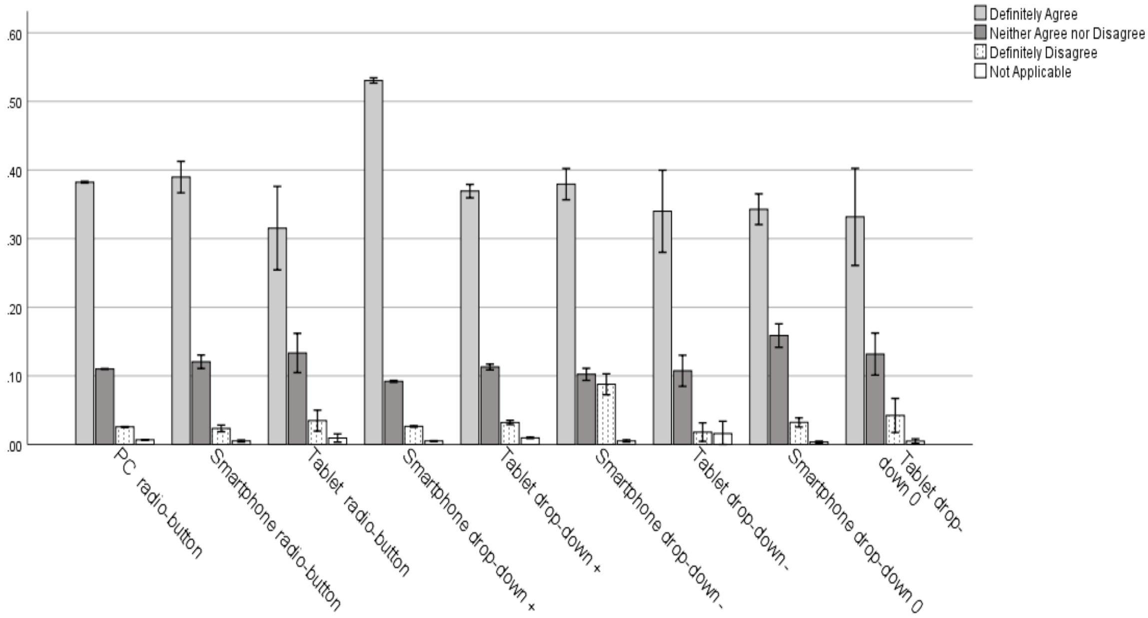

Figure 1.6. Extreme, mid-point and non-substantive response selection frequency grouped by device and response design conditions in the NSS, 2014 web survey.

Multivariate analysis

A series of logistic regressions were fitted to estimate the effect of device and scale design on data quality. The survey controlled for the type of device used to complete the survey and the response format implemented. As the PC survey used only the main radio-button grid design, it was selected as a reference category. Subsequently all the comparisons of data collected from smartphones and tablets with experimentally assigned response designs were made to the PC survey design. Different models were fitted to predict the following data quality indicators: logistic regressions were used to predict whether the survey duration was above the PC median, break-off rate, no open text feedback provided, if the provided feedback length was above the PC median length or below and a binary measure of satisficing. The fractional logistic regression models were fitted to predict the proportion of selected extreme positive, extreme negative, mid-point and non-substantive responses. The fractional logistic regression models were selected as extreme and mid-point responses were recorded as a proportion ranging from 0 to 1. As I am testing several models with each measurement error indicator as an outcome variable I have applied the Bonferroni adjustment to the significance of the regression coefficients. The Bonferroni adjustment allowed controlling for potential Type 1 error occurring due to multiple hypothesis testing.

Table 1.2. Measurement error indicators associated with the device use and response scale design in the NSS, 2014 survey.

	Survey duration above PC median	Survey Break-Off	No open text response	Open text response length above PC median	Straightlining	Extreme positive responding	Extreme negative responding	Mid-point responding	N/A (non-substantive response)
Smartphone									
Radio-button	1.5*** (.1)	2.8*** (.3)	1.4*** (.1)	.7*** (.1)	.9 (.07)	1.03 (.05)	.9 (.1)	1.1 (.05)	.75 (.1)
Drop-down +	2.1*** (.02)	4.00*** (.1)	1.6*** (.02)	.6*** (.01)	1.2*** (.01)	1.8*** (.01)	1.02 (.02)	.8*** (.01)	.8*** (.03)
Drop-down -	2.02*** (.2)	3.4*** (.4)	1.4*** (.1)	.6*** (.1)	.7*** (.1)	.98 (.04)	3.6*** (.4)	.9 (.04)	.8 (.03)
Drop-down 0	2.02*** (.2)	4.4*** (.4)	1.6*** (.1)	.6*** (.1)	.7*** (.1)	.8*** (.04)	1.3 (.1)	1.5*** (.1)	.5*** (.1)
Tablet									
Radio-button	1.8*** (.4)	1.01 (.4)	1.03 (.2)	.9 (.2)	.9 (.2)	.7 (.1)	1.4 (.3)	1.2 (.2)	1.4 (.4)
Drop-down +	1.5*** (.1)	1.4*** (.1)	1.2 (.04)	.8*** (.02)	.9 (.03)	.9 (.02)	1.3*** (.06)	1.02 (.02)	1.4*** (.1)

(continued)

	Survey duration above PC median	Survey Break-Off	No open text response	Open text response length above PC median	Straightlining	Extreme positive responding	Extreme negative responding	Mid-point responding	N/A (non-substantive response)
Drop-down -	1.4 (.3)	1.1 (.5)	1.1*** (.2)	.8 (.2)	.6 (.1)	.8 (.1)	.7 (.3)	.97 (.1)	2.4 (1.3)
Drop-down 0	1.7 (.4)	.9 (.5)	1.2 (.3)	1.4 (.3)	.8 (.2)	.8 (.1)	1.7 (.5)	1.2 (.2)	.7 (.2)

**p*<.05, ***p*<.01, ****p*<.001.

Note. Odds ratios and standard errors (in parentheses) are reported. Reference group is PC with radio-button grid design (N=235 484).

Table 1.2 presents the results of the regression models that predict measurement error indicators using the main PC grid survey design as a reference category. Across all response designs, smartphones showed higher likelihood of longer response time and higher break-off, lower likelihood of giving open text response and lower likelihood of a longer open text response. The scale design had a stronger effect on the response distribution of smartphone rather than tablet responses. There were considerably less differences in measurement error indicators between tablet experimental design conditions and PC.

Smartphones with the radio-button grid design were linked to a longer response time (OR=1.5), higher break-off (OR=2.8), lower likelihood of open text response (OR=1.4) and the provided response was likely to be shorter than PC's (OR=0.7). There were no significant differences in the rates of straightlining or response distribution between smartphones and PCs. Data showed no evidence that the radio-button format on a smartphone caused visibility effects – there was no difference in the selection of positive option between the devices. Tablet radio-button design was linked to longer completion times (OR=1.8) but there were no significant differences across other measurement error indicators. A consistent response scale design among PCs, smartphones and tablets employing radio-button question grids demonstrated no significant differences in response distribution regardless of which device respondents were using.

Larger discrepancies in data quality indicators occurred between the smartphones with drop-down designs and PC radio-button grid responses. The data quality indicators of mobile design with the most positive response presented in the drop-down list showed significant differences between smartphone and PC data. Longer response times (OR=2.1), higher break-off (OR=4.00), no open-text response (OR=1.6), lower likelihood of long open-text response (OR=.6) were linked to smartphone responses. Furthermore, positive option drop-down design was linked to a higher straightlining (OR=1.2) than the PC radio-button design. Smartphone group was linked to higher rates of extreme positive responses (OR=1.8) and lower likelihood of mid-point (OR=.8) and non-substantive responses (OR=.8). The positive response option drop-down list uncovered differences in data quality and response distribution between responses collected from tablet and PC respondents. Longer completion times (OR=1.5), higher survey break-off (OR=1.4), lower likelihood of shorter open-text response (OR=.8) were linked to the table data. Tablet respondents selected more frequently the extreme negative response options (OR=1.3) or non-substantive response options (OR=1.4) than PC radio-button grid design respondents. Overall, data show some evidence that presenting the extreme positive response option leads to a higher endorsement of this option on smartphones but not on tablets.

The data collected from smartphones with the negative response drop-down condition yielded similar differences as the other drop-down formats. Smartphone respondents were likely to take longer to complete the survey (OR=2.02), more likely to break off (OR=3.4), less likely to provide the open text response (OR=1.4) and less likely to provide longer than PC open text response (OR=.6). On the other hand, negative drop-down scale design respondents were less likely to straightline (OR=.7) than PC respondents. The selection of the extreme negative response (‘*Definitely Disagree*’) was significantly higher in the smartphone format (OR=3.6) with the extreme negative initial option displayed than in the PC radio-button grid format. Tablet negative drop-down list viewers were more likely to skip the open text question (OR=1.1), whereas there was no difference in the rest of the data quality indicators. Extreme responding trends show that unlike the tablet users smartphone users were more susceptible to the pre-set initial option.

Next, longer response time (OR=2.02) higher break-off rates (OR=4.4), lower likelihood of providing open text response (OR=1.6), providing shorter open-text response (OR=.6) were associated with the smartphone response design with the initially displayed mid-point option. However, this mobile design condition produced lower straightlining (OR=.7) than PC radio-button grids. Smartphone users who viewed mid-point pre-displayed option scale design were more likely to select the mid-point response (OR=1.5) but less likely to select the most positive response (OR=.8) or the non-substantive option (OR=.5). There was no significant difference in data quality between PC respondents and tablet mid-point response drop-down list design across measurement error indicators.

1.5. Discussion

The present study looked at measurement error occurring in mixed-device surveys and explored the effect of response format design on survey data quality in PC and mobile web surveys. I have evaluated data quality based on indicators of response time, break-off, open text response length, occurrences of extreme and mid-point answering, straightlining and selection of non-substantive options. Based on these observations, I make suggestions for the optimal scale design for smartphone and tablet surveys.

Across all response formats, smartphones were linked to longer response times, higher break-off rates, lower likelihood of providing open text response and shorter open text responses. The results demonstrated significant differences in response distribution across each smartphone drop-down response condition and PC radio-button design: the suggestive effects of the initially displayed option it the drop-down list designs were strongly observed on smartphones but not on tablets. Overall tablet data show fewer discrepancies in response quality between PC and tablet respondents. Tablet response distribution was not affected by the drop-down design to the same extent as smartphone distribution was. Straightlining rates were comparable across all response design conditions administered on tablets.

Radio-button grid design results indicate that when applying the same response design to different devices, users provide largely similar responses. Smartphone users' response times, break-off and open text responses were somewhat worse compared to PC and tablets, however that might be indicative of the general device burden rather than the design effect.

Across all drop-down list versions, I observed that a response initially suggested in a drop-down entry box was selected more frequently on smartphones than on PCs. The higher selection of the extreme positive *'Definitely Agree'* response option was significantly more likely among smartphone users who viewed a positive drop-down list design. Similarly, the negative response drop-down design on a smartphone demonstrated a significantly higher rate of the extreme negative response *'Definitely Disagree'* selected. Furthermore, a higher selection of the *'Neither Agree nor Disagree'* option was strongly linked to the smartphone mid-point response drop-down scale design. Smartphone survey results suggest that exposure to a visible option in a vertical drop-down list does indeed cause higher selection of this option. I can conclude that in smartphone surveys making a certain response option visible (i.e. appear more desirable) affects responses - vertical drop-down lists causes a higher endorsement of the initially displayed option.

The common response burden indicator – straightlining – differed in the smartphone group depending on the response scale design. Compared to the PC responses straightlining rate was higher in the positive option drop-down design condition, but straightlining was lower among smartphone surveys with the negative and mid-point drop-down scale design. High straightlining rates and selection of positive responses indicate that suggesting a positive option first might encourage consistent acquiescent responding.

The distribution of responses and data quality indicators was different between tablet drop-down list experimental conditions and PCs; however, differences were less stark as compared to the smartphone data. The biggest difference in the measurement indicators in terms of break-off, response time, open text feedback length, and extreme positive and non-substantive responses was observed in the positive drop-down list format presented on tablets. Nonetheless, pre-displayed response option was not endorsed with a higher frequency. On the contrary, higher selection of the extreme negative responses and non-substantive responses were linked to tablet respondents who viewed positive-drop-down design. The rest of the drop-down designs revealed fewer discrepancies in response distribution. An increased likelihood of no open text response was observed among mid-point response drop-down format on tablets, whereas drop-down scale with the middle response displayed showed no difference from the PC data. The fact that initial drop-down list options were selected more frequently on smartphones but not on tablets indicates that this effect occurred due to the device burden rather than the design influence.

The study has replicated and advanced measurement error in a mobile surveys framework described by Lugtig and Toepoel (2015). Overall, the results have replicated previous findings on mixed-device data quality as well as contributed to the current field of response scale design for smartphone and tablet surveys. These findings add to the current literature as they demonstrate that smartphone users are more sensitive to the response scale design than PC and tablet respondents. Smartphone data demonstrated that no design offers a solution to the reduced measurement error and there are certain trade-offs to each design.

These findings support the unimode design approach suggesting that the design should be consistent across all data collection modes (Dillman, 2007). In a web survey setting a unimode approach means presenting surveys designed for the PC on mobile devices. The experiment presented in this study demonstrated that consistent response scale design leads to a comparable response distribution and straightlining rates.

Correspondingly with Toepoel and Lugtig (2015) findings, I found lower straightlining among mobile respondents in drop-down negative and mid-point conditions. The drop-down list with the positive option displayed on smartphones caused a higher straightlining rates than PC, whereas radio-button design showed no difference. These findings demonstrate that whilst certain response designs can reduce straightlining on smartphones the PC radio-button grid design might still be encouraging the straightlining behaviour. Current findings support previous recommendations of Dillman et al. (2008) of drop-down list use in a survey design; drop-down design could be applied in smartphone surveys, however, placing instructions in the initial box would be essential to avoid having the pre-selected option endorsed more frequently. Furthermore, I found no evidence of the primacy effect occurring in the vertical scales in a smartphone design – across all experimental conditions only the positive drop-down design showed a higher endorsement of the positive option. When the displayed option was either a negative or the middle option respondents were not affected by the positive direction of the vertical scale. In line with the previous findings, results obtained from tablet experimental conditions indicate that PC design might be the optimal choice for tablet survey design.

The results of the drop-down scale design presented in this chapter offer some insight into the possibility of further exploring the adaptive design approach. The web mode approach that proposes offering a mobile optimized design alongside the main PC design is part of the generalized mode design framework (de Leeuw, 2005; de Leeuw & Hox, 2011). The generalized mode design proposes utilizing the best features of each mode offering adapted designs that would produce the same stimuli across modes. More research is necessary in order to determine the scale design that produces the lowest measurement error both on PC, tablets and smartphones.

The current study is in line with the earlier findings from Wells et al. (2012) and de Bruijne and Wijnant (2013, 2014) who found that compared to smartphones tablet surveys take less time, cause lower break-off rates and demonstrate overall a lower measurement error. In terms of data quality, it is important to notice that smartphone data distribution is highly susceptible to the format used, yet such an effect is not prominent in tablet data. As no comparable effect of response design across smartphones and tablets was observed it is suggested that differences in measurement error are not solely caused by the survey design, but rather they could be due to the respondent self-selection and device usability differences. Current study design did not allow controlling for socio-demographic characteristics however future research could account for the characteristics of smartphone and tablet respondents to uncover usability issues, motivation and cognition differences between these respondent groups.

The robustness of present study has been affected by the lack of experimental conditions tested on the main (PC) device. The misalignment in the ideal and 'realistic' data used for research is common, reflecting the gap between the ideal and practical study settings. Most commonly, surveys are conducted using fixed financial, human and time resources. Furthermore, longitudinal surveys used in comparative analyses and informed policy-making research are particularly sensitive to design manipulations and therefore survey owners are fairly conservative when it comes to tempering with the main respondent population. So often the restrictions posed by the 'realistic' survey field will not allow for a survey design that would be able to test all potential design elements on a wide population sample. As an alternative, researchers may opt to test the experimental conditions in a fully controlled 'lab experiment’. Such a setting will allow to introduce all desirable design elements and test it on an observed group of participants. While acknowledging the obvious benefits of such studies performed in laboratory conditions I would still advocate for the necessity of the 'field' experiments with potential accompanying limitations as field experiment can provide a realistic view of mobile device use in current research field.

The field limitations have posed a certain restriction on the comparison of designs between PC group and mobile devices. Practically, I was able to make comparisons across different mobile designs to the default PC design and conclude which mobile design would produce the most similar responses to the PC. However, I was not able to establish whether alternative PC designs perform better than the default radio-button grid design. The current experimental design restrictions allowed me to test the feasibility of the unimode approach where I could compare the effect of radio-button grid design on responses collected from the PCs and mobile devices. Next, I was able to consider the 'tailored design ' approach where mobile experimental design conditions were compared to the default PC design. In this case the PC design was treated as the benchmark to be compared to. This approach allowed me to establish whether any of the mobile design solutions are able to produce the closest data to the PC. However, I could not test the emerging ‘mobile first’ design approach by first identifying the best performing mobile design and then establishing whether either of the experimental designs perform equally well on a PC. The future research should take the direction of testing the 'mobile first' approach where smartphones are treated as primary devices that collect responses.

One of the limitations of current research is that the results of the analyses are limited as some experimental design groups in tablet and smartphone are relatively small as they are compared to an extremely large PC group. This could have led to some differences between experimental conditions not being detected (King & Zeng, 2001). Next, the survey used for current research did not follow all recommended survey practices; however, it was the comprehensive example of a typical survey used to assess satisfaction in a higher education research. The method used to study response quality and straight-lining in particular would benefit from questions that require reverse coding and instructional manipulation check to detect straightliners. In addition to that, changing the question grids to the ones with a less similar wording of questions across the entire survey and instructions tailored to each particular response format could have improved respondents' ability to concentrate on questions and overall understanding of the survey. The introduction of different ordered response options could provide a solution by counterbalancing the survey, thereby controlling for data consistency across the formats and reliability of responses.

I acknowledge that the use of a non-probability panel as well as allowing self-selection into device for the experiment may have implications for external validity—respondents from the non-probability panel may not be representative of the general population. However, using a survey of respondents who are proficient with the PCs and mobile devices improved the web mode response rates and resulted in a high level of participation from mobile devices. The survey tested experimental designs after respondents selected the response device themselves: the experimental conditions have been administered randomly and the devices were self-selected. Consequently, the data from this study do not allow for the disentangling self-selection and device effects. Allowing respondents to self-select into the device makes it difficult to separate device effects from self-selection effects on response quality. Respondents who choose PC to take a survey are typically different to those who would use a smartphone or a tablet.

Therefore, without the experimental allocation of the device or being able to control for the psychographic parameters of the respondent sample it is difficult to reliably claim from these results whether the differences in responses are attributed to the experimental manipulations on the device or differences between respondents. The findings from the experiments are mostly applicable to surveys with self-selected into device participants. Nonetheless, allowing respondents to choose their own device avoided respondent non-compliance with the assigned device and subsequently increased drop-out rates that have been previously observed in studies that allocated devices to participants. Experimental studies comparing data quality between devices demonstrated that forcing respondents to take a survey from a particular device typically resulted in high non-compliance with the allocated device and higher break-off rates whilst filling in the survey (Mavletova, 2013; Keusch & Yan, 2017). Moreover, a lower rate of mobile response and experimental non-compliance have resulted in discrepancies in sample composition between mobile and PC comparison groups thereby failing to overcome potential generalizability issues. In present study, allowing respondents to select a device of their choice replicated a natural survey setting and reduced the possibility of lower data quality caused by a lower proficiency in using the assigned device.

Certain measurement error indicators (higher break-off, longer response, and lower willingness to provide open text response) were consistent among smartphone respondents across all experimental conditions. This could indicate an increased smartphone burden as compared to PC and even tablets, or, perhaps, respondents who opened a survey on a smartphone were less motivated to provide good responses. Future studies could account for respondent personality and cognitive differences as well as employ a more diverse sample to test whether the smartphone data are largely affected by the limitations of the device or rather are affected by respondent characteristics.

Additionally, studies could capture respondent characteristics before assigning devices to respondents. This way we could get a better understanding of the unique contribution of the device and self-selection effects to the data quality.

2. The effect of scale direction and respondent characteristics on response distributions in web surveys

Abstract

The design of response scales plays an important role in question processing and response formation, potentially affecting differences in response distributions depending on the scale direction. The present study compares the differences in response distribution evoked by two scale directions and the role of satisficing and anchoring mechanisms in the response process. Online panel respondents (N=423) were experimentally assigned to a survey that used response scales ranging from the high/positive to the low/negative option ('forward'-ordered) or vice versa ('reversed'-ordered). Big Five inventory and a Cognitive Reflection Test (CRT) were used as satisficing and anchoring moderators. The 'forward'-ordered scales showed a slightly higher selection of responses from the first half of the scale, whereas the 'reversed'-ordered scale design showed no effect. More extraverted respondents were less likely to select responses from the first half of the scale and more likely to select responses from the low/negative end of the scale. More conscientious respondents selected fewer options from the high/positive end of the scale in a 'forward' scale design. The fact that respondent characteristics affect response selection makes it difficult to control for scale effects, however scale designs could potentially improve response quality. When choosing a scale design, researchers should be aware of the potential effect that each scale direction has on data quality.

2.1. Introduction

Scale design plays an important role in question processing and response formation as respondents often do not have readily available answers and use scales as a cue to understand the question or form a response (Christian, Parsons, & Dillman, 2009; Tourangeau, Couper, & Conrad, 2004; Yan, Keusch, & He, 2018). Questions can be presented with scales starting with the high/positive response labels ('forward'-ordered scale direction) or alternatively use a design where the low/negative labels are placed at the beginning of the scale ('reversed'-ordered scale direction) (Liu & Keusch, 2017; Yan et al., 2018). As new survey questionnaires were developed, the choice of the direction of the scales has been decided by researchers (Krebs & Hoffmeyer-Zlotnik, 2010; Yan & Keusch, 2015) or followed the original design suggested by scale developers (e.g. Robinson, Shaver, & Wisman, 1991; Robinson, Shaver, & Wisman, 1999). After conducting series of web experiments, Tourangeau et al. (2004) concluded that responses are more reliable and given quicker if vertical scales start with a positive option or the most desirable option first as respondents rely on heuristics when processing response scales.

Some studies found that adopting 'forward'-ordered or 'reversed'-ordered scales in web and telephone surveys may have different effects on response distribution (He, Yan, Keusch, Han, 2014; Toepoel, Das, Van Soest, 2009; Yan & Keusch, 2015). The resulting scale direction effects influenced response distributions depending on whether respondents view high/positive or low/negative response labels first. Whereas Liu and Keusch (2017) found stronger effects in a web mode for 'forward'-ordered scales – responses shifted towards a high/positive end of the scales - they observed weaker or no effect for the 'reversed'-ordered scales.

Evidence on scale direction effects is mixed with more instances of ‘forward’ scale direction affecting responses. Research has yet to establish the underlying reasons and circumstances when the choice of scale direction affects responses. Studies that have observed scale effects across all modes proposed anchoring and satisficing as potential mechanisms responsible for such scale direction effects. However, no research so far has directly measured satisficing and anchoring accounting for respondent characteristics typically associated with the likelihood of adopting anchoring and satisficing response strategies.

In this study, I aim to extend earlier research by examining respondent personality characteristics and cognition skills and their role in adopting satisficing or anchoring response behaviour when answering Likert-type questions. The objective of this study is to compare the responses evoked by ‘forward’-ordered and ‘reversed’-ordered scale directions and the role of satisficing and anchoring mechanisms in the scale direction effects. Online survey respondents were experimentally assigned to a survey that presented questions with Likert scales starting with either ‘forward’ (high/positive) or ‘reversed’ (low/negative) options. The survey also included the assessment scales such as Big Five personality traits assessment and a Cognitive Reflection Test (CRT). Respondent characteristics have been observed to have significant effects on the response mechanisms involved in scale processing – satisficing and anchoring-and-adjustment. In order to determine whether scale direction effects are the result of satisficing or anchoring, the measured personality and cognition factors will be used to assess their role in the scale direction effects. First, I aim to answer whether the effect of scale direction on responses differs as measured by the frequency of selecting high/positive, low/negative and mid-point responses. Second, where the scale direction effects occur, I will examine whether respondent characteristics such as personality traits and cognitive reflection are moderating factors in the scale direction effects. The results of the experiment provide guidance on the scale design that produces the least biased results as well as poses less cognitive burden on respondents.

2.2. Background

Scale direction effects

Research suggests that, depending on the choice of the scale direction, the mean values and distribution of responses might change, resulting in scale direction effects (He et al., 2014; Liu & Keusch, 2017). Existing studies provided mixed evidence on the extent of scale direction effects in web surveys. Some studies found no clear effects of scale direction on survey responses or only some of the questions were affected by such design features. Keusch, Yan, Han and He (2014) found no difference in item means between conditions that varied the response scale direction (‘forward’ versus ‘reversed’) or alignment (vertical versus horizontal) on PCs or iPhones. Rammstedt and Krebs (2007) have conducted a within subjects self-administered pen and paper survey experiment and found strong evidence that reversing the order of frequency response scales does not affect response distributions. Christian et al. (2009) have conducted an experiment in a non-probability web survey and further found that presenting the high/positive end of the scale first did not affect responses, but these questions were answered more quickly, suggesting that respondents required less time to perceive and comprehend high/positive scales. Similarly, in a full-factorial randomised experiment using a non-probability web student survey Maloshonok and Terentev (2016) found no difference in response distributions when the survey used either ‘forward’-ordered or ‘reversed’-ordered response scales. Furthermore, Weng and Cheng (2000) have experimentally tested scale effects in a non-probability pen and paper survey and found no difference in data quality between experimental conditions.

Other research has demonstrated that the direction of the scales can affect selection of certain response options. In a web probability household survey experiment Toepoel, Das et al. (2009) found that low/negative direction scales caused a higher selection of negative responses. In a split-ballot pen and paper experiment Krebs and Hoffmeyer-Zlotnik (2010) found that ‘forward’-ordered scales caused a higher selection of high/positive responses whereas a ‘reversed’-ordered scales’ direction did not elicit a similar scale direction effect, yet the options from the low/positive end of the scales were selected less frequently under this condition. Albeit some significant scale effects were detected, the differences in response behaviour did not systematically reflect the direction of the respective scale direction design (Krebs & Hoffmeyer-Zlotnik, 2010). He et al. (2014) provided a summary of empirical studies and found evidence that the answers tend to shift towards the starting point depending on the direction of the scales; however, their own secondary data analysis revealed scale direction effects only in less than half of the items examined (He et al., 2014). In a telephone interview, Yan and Keusch (2015) found that the direction of scales has affected mean values and shifted the overall results towards the starting point of the scales – the scale direction has been observed in an aural mode and could not be generalised to the other modes of data collection without further testing. In a probability survey Yan et al. (2018) have tested several features of the instrument design: location of the question in the questionnaire, direction and length of the response scales. Yan et al. (2018) found stronger scales direction effects in the high/positive ordered (‘forward’) scales design than the low/negative (‘reversed’) scales design: when scales were running from high to low first options were endorsed more frequently than when they were running from low to high options. The scale direction effects differed depending on the administration mode: the effects were more pronounced for items administered via CAPI than in self-administration mode. Furthermore, these findings should be considered with caution as they could be subject to the multiple comparisons error.

The evidence for scale direction effects has been inconclusive; a wide range of probability and non-probability studies using primary and secondary data have revealed mixed evidence of causes of scale direction effects. The mode of administration did not reveal strong effects on the likelihood of observing scale direction effects - scale effects were observed across some self-administered and computer assisted surveys as well as web and pen and paper surveys, whereas an equal part of these surveys failed to find any significant scale direction effects. So far experiments with scales direction design have revealed that a choice of scale direction may have an effect on distributions of responses yet the circumstances and possible causes are not fully explored yet. The observed individual effects of 'forward' and 'reversed' scale design indicates that the processes of reading and evaluating a 'forward' as opposed to a 'reversed' scale may not be the same yet further research is needed into the circumstances that affect scale comprehension.

Satisficing and Anchoring mechanisms

Primacy is the most common scale effect that has been observed in studies examining scale direction effects in visually administered surveys (Yan et al., 2018). Studies that have observed scale direction effects recorded that earlier responses on the scales tend to be selected more frequently under 'forward' than 'reversed' scale directions. Previous studies have used both satisficing and anchoring-and-adjustment frameworks to explain the occurring primacy effects under different scale direction conditions.

The theory of satisficing was first applied to survey methodology by Krosnick (1999). Satisficing can account for primacy effects in visual surveys - respondents save effort by consequently proceeding through response options until they reach the first acceptable one (Smyth, Dillman, & Christian, 2012; Tourangeau, Rips, & Rasinski, 2000). The satisficing framework is widely applied in the area of survey measurement error (Barge & Gehlbach, 2012). As answering surveys poses cognitive demands on respondents, some respondents might satisfice by adopting various strategies minimizing cognitive effort involved rather than providing the true response (optimising). In case of response scale presentation in web surveys, it is likely that the answers of satisficing respondents would be shifted towards the starting point of the scale. When presented with ordered scales respondents would be expected to consider scale points sequentially and those who satisfice would stop at the first plausible option due to memory limitations, decreased motivation or cognitive fatigue (Krosnick, 1999). Such scale effect is considered to be a weak form of satisficing, as respondents engage in all steps of the response process (Callegaro, Manfreda, & Vehovar, 2015; Holbrook, Krosnick, Moore, Tourangeau, 2007). Anchoring is applied as an alternative explanation for the observed scale direction effects. Anchoring is a cognitive heuristic first proposed by Tversky and Kahneman (1974) and later suggested as a mechanism unrelated to cognitive elaboration, memory capacity or satisficing that could account for scale effects observed in surveys (Gehlbach & Barge, 2012; Smyth et al., 2012). According to the anchoring framework, in surveys, responses from the response scales serve as anchors and affect judgements by increasing the availability and construction of features that the scale point (anchor) and a true answer hold in common and reduce the availability of features of the true answer that differ from the suggested response point. The mechanism underlying the anchoring process suggests that respondents use their previous response as an anchor and adjust their subsequent response from that anchor until they reach a plausible estimate (Chapman & Johnson, 1999). Often adjustments to the anchor are insufficient and the final response is shifted towards the anchor. Anchoring-and-adjustment is framed as an involuntary response strategy - a shortcut in one's decision making strategy that cannot be fully negated with awareness or effort.

The role of respondent characteristics in satisficing and anchoring mechanisms

Table 2.1 presents the two proposed mechanisms involved in the response behaviour - satisficing and anchoring - and the role of respondent characteristics in the extent of adopting each response mechanism. The anchoring and satisficing approach in survey responding is distinguished by the contribution of respondent personal characteristics – educational level, age, cognition, motivation and personality factors.

Satisficing is affected by respondent ability, motivation and cognitive burden of the question (Krosnick, 1991). Studies have used respondent characteristics such as lower educational level and cognitive ability (Kaminska, McCutcheon, Billiet, 2010; Krosnick, 1999; Malhotra, 2008; Toepoel, Vis, Das, & van Soest, 2009), lower motivation (Kaminska et al., 2010; Oppenheimer, Meyvis, & Davidenko, 2009) and higher age (Krosnick, 1991; Toepoel, Vis et al., 2009) to explain increased rate of satisficing among respondents. Kaminska et al. (2010) used age, education and cognitive ability to explain satisficing in a face-to-face cross-national survey. The satisficing was captured as a rate of extreme responding, middle responding, non-substantive responding and straightlining. Kaminska et al. (2010) have concluded that satisficing is more dependent on cognitive limitations posed by age and education rather than motivation to provide good answers. However, the face-to-face nature of the survey could have encouraged respondents to put more effort into responding and may produce different results in self-administered surveys. In a probability-based web survey experiment Malhotra (2008) found that low-education respondents who filled out the questionnaire most quickly were the most prone to primacy effects when completing items with unipolar rating scales. These results tap into the satisficing framework as selection of the visible (most easily accessible) response is considered to be a form of satisficing. Toepoel, Vis et al. (2009) have tested the role of respondent characteristics in scale processing in an online probability survey experiment. They found age and education effects in scale processing yet results varied greatly across experimental conditions. The need for cognition and the need to evaluate constructs accounted for variance in survey responding: across most questions respondents with a low need for cognition or the need to evaluate were the most affected by scale design. The researchers suggested that respondent characteristics could indicate tendency to satisfice in a survey yet more controlled experiments are needed. Overall, studies capturing the role of respondent motivation in satisficing returned mixed findings – Kaminska et al. (2010) suggested that lack motivation could be better explained by the limitations in cognitive ability and subsequently higher rates of satisficing. Toepoel, Vis et al. (2009) observed that motivation plays an important role in providing good answers only when memory representation is bad, whereas motivation is not significant when answering questions with more easily retrievable information. Oppenheimer et al. (2009) have explored the usefulness of the instructional-manipulation check questions in dealing with reluctant respondents taking paper survey in a non-probability setting. The results of the study demonstrated that whilst initially motivation could adversely affect satisficing rates when respondents were requested to attend to questions' instructions satisficing was reduced. Overall, studies found a very weak effect of motivation as a standalone predictor of satisficing and any effects were attributed to respondent characteristics such as direct and indirect measures of cognition.

The majority of studies found evidence of anchoring regardless of respondent cognitive ability or motivation which were captured by a range of psychometric scales and self-reported measures – need fo cognition scores, need to evaluate scores, cognitive reflection test scores and SAT scores (Bergman, Ellingsen, Johannesson, & Svensson, 2010; Oechssler, Roider, & Schmitz, 2009; Stanovich & West, 2008). Bergman et al. (2010) have tested the effects of cognition on decision making process in a non-probability experiment of undergraduate students. Bergman et al. (2010) have observed that the anchoring effect decreased with higher cognitive ability test score (CAT test assessed analogies, number series and logical series), but that it was sizeable even in the high cognitive ability group. Furthermore, the cognitive reflection test scores (CRT test assessed level of reflective and deliberate thinking) showed almost no significant role in the anchoring process. These findings were supporting previous research by Stanovich and West (2008) who found no correlation between cognitive ability as measured by the self-reported SAT scores and anchoring. Oechssler et al. (2009) observed anchoring across all experimental groups in a web survey regardless of their CRT scores and even found that the high CRT group were more susceptible to anchoring, although this effect was not significant.

The only set of traits predicting whether respondents rely on the anchoring heuristic so far was found to be respondent personality traits measured by the Big Five personality scale (Eroglu & Croxton, 2010; McElroy & Dowd, 2007). Eroglu and Croxton (2010) explored the role of respondent personality traits in making statistical forecasts whilst observing a sample of 473 employees over a twelve-month period. Eroglu and Croxton (2010) found that individuals who were high in conscientiousness and agreeableness and low on extraversion were more prone to demonstrate anchoring in their estimations. Whilst the effects of personality traits were not directly tested on social survey responses they provide some insight into the role of personality traits in providing factual and opinion-based estimations in a longitudinal context. McElroy and Dowd (2007) examined how the Big Five personality trait of openness to experience as measured by the TIPI (ten-item Big Five personality assessment scale) influenced the effect of previously presented anchors on factual estimations of undergraduate students. The findings indicated that participants high in openness were significantly more influenced by anchoring cues relative to participants low in this trait. The findings were consistent across two different types of anchoring tasks providing convergent evidence for the significant explanatory power of openness trait in anchoring (McElroy & Dowd, 2007). Overall, the research on decision-making in a behavioural economics setting has found that respondent personality traits play a significant role in how they process questions and produce responses, yet, to the best of my knowledge, no published studies in research methodology have used personality traits to account for the anchoring in survey responses.

Apart from the significant role of the Big Five personality traits, studies that explored respondent characteristics' influence in experimental non-probability research found weak to no support for cognitive sophistication as measured by the need for cognition, age or education to have a predictive power of the anchoring-and-adjustment heuristic (Epley & Gilovich, 2006; Eroglu & Croxton, 2010). The lack of an obvious overlap in explanatory personality variables between satisficing and anchoring frameworks indicates that they both might take place in response behaviour and work as complementary rather than competing mechanisms.

Table 2.1. Cognitive and non-cognitive indicators of satisficing and anchoring.

	Respondent characteristics
Satisficing	**Age**. Higher age is positively associated with higher satisficing rates (Kaminska et al., 2010; Krosnick, 1991; Toepoel, Vis et al., 2009). **Education**. Lower education is positively associated with higher satisficing (Kaminska et al., 2009; Toepoel, Vis et al., 2009) and primacy effects (Malhotra, 2008). **Cognitive ability**. Respondents who have lower cognitive performance satisfice more (Krosnick, 1991, 1999; Toepoel, Vis et al., 2009; Tourangeau, Rips, & Raisinski, 2000). **Motivation**. Overall, a very weak effect of motivation as a standalone predictor of satisficing. Cognition and cognitive fatigue measures are better fit for the models (Kaminska et al., 2010; Oppenheimer et al., 2009; Toepoel, Vis et al., 2009). **Big Five personality traits** are not used as a personality predictor to explain satisficing behaviour.
Anchoring	**Big Five personality traits**. Higher openness for experience is associated with a higher anchoring (McElroy & Dowd, 2007). Higher conscientiousness and agreeableness and lower extraversion are associated with higher anchoring (Eroglu & Croxton, 2010). **Cognitive ability**. No significant relationship of cognitive assessment scores with the presence of anchoring in responses (Bergman et al., 2010; Stanovich & West, 2008; Oechssler et al., 2009). **Age**, **education** are not associated with higher or lower appearance of anchoring heuristic (Eroglu & Croxton, 2010).

Applying both satisficing and anchoring-and-adjustment frameworks became a point of interest in more recent survey methods research. Toepoel, Das et al. (2009) found evidence of scale direction effects in a probability web survey experiment and applied satisficing theory to explain the results but also proposed the role of anchoring when the scales starting with the low/negative label induced more low/negative responses. The findings also indicated stronger anchoring in more burdensome questions – questions that included less representative response options produced more biased responses (Toepoel, Das et al., 2009). Yan and Keusch (2015) explored the effect of direction of ranking scales and the cognitive mechanism that would explain such effect in an aural mode survey. The scale direction effect has been observed only for some estimations. Furthermore, moderators typically associated with satisficing such as age, education, motivation and engagement have not interacted with the scale direction to produce different results. Due to this fact, Yan and Keusch (2015) suggested that the differences in response distributions must be due to the anchoring-and-adjustment heuristic rather than satisficing. Albeit, this study has ruled out the role of the satisficing moderators it has not controlled for the anchoring measures to explore the role of anchoring in scale direction effects.

In an eye-tracking experiment Hohne and Lenzner (2015) found support for satisficing as the time spent looking at the first half of vertical response scale was associated with a higher likelihood of selecting a response from that side. In contrast when presented with a horizontal scale, longer times were attributed to the middle option of the scale suggesting that consistent with the anchoring-and-adjustment heuristic, respondents were using the middle point of the scale as a reference point. This study found support for both explanatory mechanisms, yet it did not account for respondent characteristics or whether the writing system participants were using could explain differences in scale processing. Liu and Keusch (2017) explored the scale direction effects and response style in an experiment embedded in a national probability survey. They found that in a 'forward'-ordered scale the acquiescent responding was the highest, however no such effect in the 'reversed'-ordered direction scale. Furthermore, no significant effect of scale direction was detected on the extreme responding. They suggested that such an effect could be due to either satisficing or anchoring yet cautioned that the scale direction effects were not strong. Studies that have found both anchoring and satisficing mechanisms taking place in response behaviour support the notion that both of these processes may occur when responding to a survey. However, in order to identify the role of anchoring or satisficing in response behaviour and improve response quality more research is needed controlling for the explanatory factors of satisficing and anchoring such as respondent cognition and personality measures and demographics.

This study

In this paper, I examine the effect of scale direction on response distribution in a web survey. I explore whether cognition and personality traits and demographics moderate response behaviour under different scale direction conditions. As cognition characteristics are more frequently attributed to satisficing strategies and Big Five personality traits are linked to anchoring-and-adjustment heuristics, the results will demonstrate an extent to which each of these response mechanisms are involved in scale direction effects. This would allow me to evaluate to what extent response selection is affected by either of the response mechanisms and whether a certain type of scale is less burdensome for respondents.

When respondents are not motivated or more burdened by the questionnaire, they might select the first visible answer or the one next to it, hence engaging in satisficing or anchoring-and-adjustment behaviour. Under these circumstances, I expect the ‘forward’ scale to be linked to a higher endorsement of the high/positive end of the scale, whereas the ‘reversed’ scale to be linked to a higher endorsement of the low/negative end of the scale. If satisficing takes place in responding, lower cognition scores and lower indirect cognition indicators (lower education, higher age) will show a significant relationship with stronger primacy and mid-point responses whereas higher cognition scores, younger age and higher education will show a reduced primacy and tendency for mid-responding. A significant interaction between predictors typically attributed to the satisficing framework (measure of cognition, age, education) and scale design would offer further evidence towards the role of satisficing.

In line with previous research on anchoring and satisficing indicators, I expect anchoring to appear irrespective of one’s cognitive ability, therefore the respondent’s cognitive reflection scores should not be predictive of the likelihood to select response options from the first half of the scale; If higher conscientiousness, agreeableness, higher openness or lower extraversion are significantly associated with stronger scale direction effects as measured by primacy and mid-responding I will be able to conclude that the anchoring took place in response behaviour. By testing these hypotheses I will establish whether a certain scale direction is more favourable in a survey design, i.e. produces weaker primacy and mid-point responding both among more and less motivated/proficient respondents.

2.3. Data and Methods

Sample and Study Design

The study employed a between-subjects design with two treatment groups that presented a questionnaire with 'forward'-ordered or 'reversed'-ordered response scales. Respondents from the United States participated in a web survey designed in Qualtrics and fielded in April and May 2016. Respondents were recruited from the Amazon Mechanical Turk (MTurk) respondent pool for a small cash incentive ($0.30 per completed survey). Upon opening the survey link, respondents were randomly assigned to either a 'forward' survey version with Likert response scales running from high to low options (e.g. from *'Always'* to *'Never'* or from *'Extremely'* to *'Not at all'*) or the 'reversed' Likert scales running from low to high response options (e.g. from *'Never'* to *'Always'* or from *'Not at all'* to *'Extremely'*). Only those responding from a PC could access the survey. The total sample consisted of 210 respondents assigned to the 'forward' scale survey design and 213 respondents assigned to the 'reversed' scale survey design.

Questions were taken from the European Social Survey Wave 6 modules; all response scales were adjusted to fit Likert five-point scales and used item-specific responses. Thirty-two questions were used to design two survey versions differing in the direction of the response scales. The core questionnaire consisted of 32 questions on behaviours and attitudes and assessed respondents' environmental position, social life and well-being, health and lifestyle. The full wording of questions is documented in Appendix Table 2.4 and Table 2.5. The first version of the survey questionnaire presented all questions with vertically aligned 'forward' ordered five-point Likert scales. The second version of the survey presented the questionnaire with vertically aligned 'reversed' ordered five-point

Likert response scales. None of the 32 items used reverse-scoring questionnaire items. Only one survey question was presented per page.

Additionally, the survey presented all participants with questions assessing Big Five personality dimensions using a 10-item inventory (Gosling, Rentfrow, & Swann, 2003). The Big Five question grid is presented in the Appendix Table 2.4 and Table 2.5 (Q33). Furthermore, the survey asked respondents to answer three questions to assess their cognitive reflection measured by the Cognitive Reflection Test (CRT) (Frederick, 2005). The full wording of CRT questions is listed in the Appendix Table 2.4 and Table 2.5 (Q34.1-Q34.3). The CRT questions were forced choice and required a numeric text entry. The median response time for the 'forward' scale design was 6.5 minutes and for he 'reversed' ordered design 6.1 minutes.

Participants and Measures

Respondents' mean age was 34 years (SD=12.00), 192 respondents were females (45.3%), 227 males (53.8%) and 4 (0.9%) chose not to disclose their gender. The majority of survey respondents indicated having a college degree (66.9%), 25.1% reported a Master degree or higher and only 8% reported having a High School education or less.

To explore the mechanisms responsible for scale effects, the survey included several socio-demographic characteristics and personality traits that are related to respondents' use of satisficing strategy and anchoring-and-adjustment heuristic.

Cognitive capacity was assessed by the *Cognitive Reflection Test* developed by Frederick (2005). The purpose of the Cognitive Reflection Test (CRT) is to differentiate between more impulsive and more reflective decision-making. The cognitive reflection test shows how reflective respondents are of their own cognition and is associated with the intelligence score and ability to engage in further reflection to answer the task correctly rather than following the intuitive response (Frederick, 2005). The CRT score is also used as a reliable predictor of cognitive ability (Frederick, 2005). The CRT includes three questions designed so that the intuitive spontaneous answer is wrong. It requires respondents to suppress an erroneous (impulsive) answer and think deeper about the solution. Correct answers indicate a higher degree of reflective and deliberate thinking. The questions are designed so that respondents with low need for cognitive reflection are likely to give impulsive answers that seem plausible (24, 10 and 100). Those who are prone to a deeper cognitive reflection will notice that the initial answers may be wrong and upon deeper processing come up with correct answers (47, 5 and 5). Finally, respondents who give completely wrong answers most likely are not motivated to put effort into answering the test questions.

Table 2.2 shows the distribution of responses to all three items of the test, displaying separately correct answers, intuitive but wrong (impulsive) answers and the rest of the answers categorized as the 'Other' group. The median individual-level number of correctly solved CRT items was one. Independent Chi2 tests revealed no association between the scores and experimental conditions.

Table 2.2. Distribution of answers in the cognitive reflection test (N=423).

Question	Correct response (in percent)	Impulsive response (in percent)	Other response (in percent)
Lily pads	39.2	47.3	13.5
Bat and ball	28.1	63.6	8.3
Widgets	40.2	46.6	13.2

Personality can be defined as a set of traits or predispositions that determines an individual's behaviour and is consistent across situations and relatively persistent over time (Levy, Cober, & Norris-Watts, 2004). *Big Five personality structure* has become the most prominent model for operationalising the structure of personality traits. The Big Five model conceptualizes personality as five main traits: conscientiousness, emotional stability, extraversion, agreeableness, and openness to experience. In this study, the Big Five personality traits are assessed using the Ten Item Personality Inventory (TIPI) scale. The TIPI assesses the Big-Five factors using two separate items (e.g. warm, creative, anxious, extraverted, self-disciplined). This measure was utilized because of the accuracy and brevity in assessing individual differences relating to the Five-Factor Model. Despite having somewhat reduced sensitivity in capturing finer psychometric properties due to its brevity, the TIPI has nonetheless shown good test-retest reliability; The TIPI has been found to be a good estimation of the longer version across different methods of data collection as well as across at least some different languages and cultures (Gosling et al., 2003; Rammstedt, Goldberg, & Borg, 2010; Rammstedt & John, 2007). Respondents are presented with statements on their personality asking to rate the extent that they feel each of the traits applies to them on a 7-point Likert-type response scales running from the most positive to the most negative response. Each of the five dimensions (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience) was assessed by two statements, one of which was reverse-coded. For each of the five traits, the individual-level average score, ranging from one to seven, has been recorded. The average score for openness was 3.1 (SD=1.3), for conscientiousness 2.9 (SD=1.4), for extraversion 4.1 (SD=1.5), for agreeableness 2.6 (1.3) and for emotional stability 3.4 (SD=1.4).

Scale direction effects were measured as a selection of the two most positive and two most negative responses. The main survey questionnaire has been using five-point Likert scales. The selection of the first two options in the ‘forward’ ordered scale indicated selection of the higher/positive responses. In the ‘reversed’ scale design the selection of the first two options indicated the selection of the low/negative response options. The selection of the last two options in ‘forward’ scale design was recorded as a selection of the lower/negative options. Whereas in the ‘reversed’ scale design the selection of the last two options indicated selection of the high/positive responses. Mid-point responding captured the frequency of selecting the middle response option in the five-point scale in ‘forward’ and ‘reversed’ scale design conditions.

2.4. Results

Table 2.3 presents the proportion of high/positive, low/negative and mid-point responses in each experimental condition and results from a t-test between these two conditions. The ‘forward’ scale direction demonstrated a slightly higher rate of selection of the high/positive response options compared to the ‘reversed’ scale direction: the average proportion of selection of high/positive responses in the ‘forward’ scale experimental condition was 0.37 as compared to the average proportion of 0.35 in the ‘reversed’ condition, the difference was not statistically significant. Similarly, there was no significant difference in the average proportion of reported low/negative responses: the selection of options from the low/negative end of the scale was similar across both scale directions. The ‘reversed’ ordered scale design reported a slightly higher proportion of mid-point responding: the ‘forward’ scale condition yielded an average proportion of 0.27 mid-point responses whereas the ‘reversed’ scale yielded 0.29 of mid-point responses, but the difference was not significant. The initial results of the t-tests show no difference in response distribution across both scale directions. In addition, there is no indication that either of the scale directions was linked to higher rates of mid-point responding.

Table 2.3. Proportion of high/positive, low/negative and mid-point selected responses in 'forward' and 'reversed' scale conditions.

	'Forward' scale (n=210)	'Reversed' scale (n=213)	t-test statistic
High/positive responses	.37 (.01)	.35 (.01)	$t(421)=-.8$, $p=.4$
Low/negative responses	.36 (.01)	.36 (.01)	$t(421)=-.2$, $p=.8$
Mid-point responses	.27 (.01)	.29 (.01)	$t(421)=1.3$, $p=.2$

Note: Standard errors are reported in parentheses.

Multivariate analyses

The fractional outcome regression models were used to evaluate the proportion of the high/positive, low/negative and mid-point responses. Two nested fractional regression models were fitted to estimate scale direction effects for each of the three response distribution indicators. In each case the first model controlled for respondents' socio-demographic factors including age, gender and education. The second model included the personality-related satisficing and anchoring indicators: test score of the Cognitive Reflection Test and Big Five personality traits. The response distribution differences were further investigated with interaction terms between experimental condition ('forward'-ordered scale direction), demographics, CRT scores and Big Five personality traits.

Table 2.4 presents results of the regression models predicting high/positive, low/negative and mid-point response selection controlling for scale direction and respondent characteristics. Initially, the selection of responses from the high/positive end of the scale was significantly predicted by education ($p<.01$) and gender ($p<.01$). Respondents with high school degree or less were less likely to select high/positive responses (OR=.69). Women also were less likely to select high/positive responses (OR=.8). When satisficing and anchoring predictors were added in the second model the extraversion trait was a significant predictor (p<.001). More extravert respondents are also expected to select less high/positive responses on average for each additional extraversion score (OR=.85).

Next, age, education and gender were used to predict the selection of low/negative responses. Selection of responses from the low/negative end of the scale was higher among older respondents (p<.01), females (p<.01), those with the lowest level of education (p<.01). Older respondents (OR=1.01), those with a high school degree or less (OR=1.37) and females (OR=1.2) were more likely to select responses from the low/negative end of the questionnaire. When personality and cognition measures were included in the model, higher age (OR=1.01; p<.05) and female gender (OR=1.16; p<.05) were still significantly associated with a higher selection of low/negative responses. Respondents higher in extraversion were more likely to select low/negative responses (p<.001): more extravert respondents are expected to select more options from the low/negative end of the scale for each additional score on the extraversion scale (OR=1.18).

The mid-point responding was significantly predicted by age (p<.001). Older respondents were less likely to select middle responses (OR=.99). When personality and cognition predictors were included in the second model age no longer explained mid-point responding. Nonetheless, respondents higher on emotional stability were more likely to select middle responses (OR=1.07; p<.05).

Apart from low/negative responses being significantly more often selected by females, across all models, respondent demographics failed to significantly predict the selection of responses once cognition and personality traits have been included in the model. This indicates the importance of including such predictors as a more direct measure of respondent characteristics.

Albeit there was no overall effect of scale direction or conscientiousness on selection of high/positive responses the selection of the high/positive responses was significantly predicted by the interaction between scale direction and level of conscientiousness (p<.05). The effect of ‘forward’-ordered scale on selection of high/positive responses was the opposite depending on respondents’ level of conscientiousness. The marginal effect of a 1% increase in the conscientiousness showed a 0.016 decrease in high/positive selection for respondents who viewed a ‘forward’ scale order. The ‘reversed’ scale design did not have an effect on response selection depending on the respondents’ conscientiousness score.

Table 2.4. Individual fractional regression models evaluating scale direction and respondent characteristics effects on response distribution.

	High/positive responses		Low/negative responses		Mid-point responses	
	Model 1	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Age	.99 (.003)	.99 (.003)	1.01** (.002)	1.01* (.002)	.99*** (.002)	.99 (.002)
High school degree	.69** (.09)	.83 (.09)	1.37** (.16)	1.36 (.16)	1.02 (.13)	.98 (.13)
Female	.8** (.06)	.83 (.06)	1.2** (.04)	1.16* (.07)	1.04 (.07)	1.05 (.07)
CRT none correct		.95 (.1)		1.01 (.1)		1.04 (.09)
Openness		1.02 (.05)		.95 (.05)		1.02 (.05)
Conscientiousness		1.06 (.05)		.94 (.04)		1.01 (.04)

(continued)

	High/positive responses		Low/negative responses		Mid-point responses	
	Model 1	*Model 2*	*Model 1*	*Model 2*	*Model 1*	*Model 2*
Extraversion		.85***		1.18***		.98
		(.03)		(.04)		(.03)
Agreeableness		.93		1.03		1.04
		(.04)		(.05)		(.03)
Emotional stability		.92		1.03		1.07*
		(.04)		(.24)		(.03)
'Forward' scale direction		.94		1.04		.97
		(.26)		(.24)		(.22)
'Forward' scale x Age		1.00		.99		.99
		(.01)		(.01)		(.01)
'Forward' scale x High school		.74		.9		1.52
		(.21)		(.21)		(.37)
'Forward' Scale x Female		.98		.85		1.25
		(.15)		(.11)		(.16)
'Forward' Scale x CRT none correct		1.14		.98		.9
		(.17)		(.13)		(.12)
'Forward' Scale x Openness		.99		1.08		.93
		(.07)		(.07)		(.05)
'Forward' Scale x Conscientiousness		.88*		1.04		1.1
		(.05)		(.06)		(.06)
'Forward' Scale x Extraversion		1.06		.96		.99
		(.05)		(.04)		(.04)
'Forward' Scale x Agreeableness		1.07		.98		.95
		(.07)		(.06)		(.05)
'Forward' Scale x Emotional Stability		1.01		.96		1.04
		(.06)		(.05)		(.05)
Sample size	423	423	423	423	423	423

*$p<.05$, **$p<.01$, *** $p<.001$.

Note: Odds ratios and standard error terms (in parentheses) are reported.

2.5. Discussion

Several previous studies explored the relationship between scale design and response distribution. The studies attempted to relate the unique effects of scale direction on response distribution from the perspective of the satisficing and anchoring-and-adjustment mechanisms. Present study has used a set of respondent characteristics that are traditionally attributed to either satisficing or anchoring behaviour to explain the scale direction effect occurring in web survey. This study varied the direction of vertical five-point fully labelled Likert rating scales to explore the scale effects on response selection and the role of cognitive and non-cognitive characteristics in response processing.

The results provided no strong evidence of scale direction effects and suggest that certain respondent characteristics are more critical than scale design or cognitive reflection in explaining the occurrence of scale direction effects. Based on these analyses I conclude that respondents' answers are not significantly influenced by the scale direction design they are viewing. Scales starting with a high/positive response options produced a slightly higher number of high/positive responses; however the difference was not significant. When respondent demographical characteristics were controlled for, higher endorsement of the high/positive options was linked to the lower extraversion. Selection of the low/negative responses was linked to an older age, lower educational level and higher extraversion. A higher selection of middle response was predicted by a younger age and higher emotional stability score. These results support previous studies that did not find significant differences in response distribution in non-probability surveys' experimental scale design (Christian et al., 2009; Keusch et al., 2014; Maloshonok & Terentev, 2016; Rammstedt & Krebs, 2007; Weng & Cheng, 2000).

The satisficing and anchoring frameworks suggest that respondent characteristics such as motivation, cognition and educational level play a role in how respondents read a response scale and select a response. An examination of respondent personal characteristics measured in this study suggests that certain characteristics such as age, gender, and Big Five personality traits were significantly associated with a selection of responses from a certain end of the scale.

Only respondents' tendency to select low/negative responses was predicted by demographics - age and gender. However, variables indicative of satisficing (age, education, cognitive reflection) that should predict respondents' likelihood of selecting earlier response options did not interact with the scale direction to predict the selection of responses. This lack of a significant interaction fails to support satisficing as an explanatory mechanism behind scale effects and strengthens the support for the anchoring mechanism taking place.

The variables that were hypothesised to be indicative of the anchoring mechanism – Big Five personality traits – were predicting some of the scale effects indicators. The results lend support towards the anchoring heuristic in the selection of options from the high/positive end of the scale. Respondents' tendency to select high/positive responses was explained by the lower extraversion – a personality trait that is attributed to the use of anchoring. What is more a significant interaction between one of the Big Five traits and a scale direction strengthens the case of anchoring mechanism taking place. Conscientious respondents' responses were affected by the scale direction: more conscientious respondents who were assigned to 'forward'-ordered scales selected less high/positive answers. This result suggests that using 'forward'-ordered scales induced more acquiescent responding among less conscientious respondents. These results provide a response scale suggestion to survey designers; Administering 'reverse-ordered' scales in a survey could reduce the possibility of acquiescent responding among less conscientious respondents and have otherwise no detrimental effects on response quality of more conscientious respondents.

The fact that personality traits and demographic characteristics determine participation and response behaviour makes it difficult for researchers to control for certain response biases. The aim of the present experiment was to investigate whether a certain response scale design leads to a higher selection of certain responses that are not a true reflection of respondents' opinion. In line with previous findings, the evidence of which scale design is best used in the questionnaire is not straightforward – in general, both scales performed comparably. Results of the current study suggest that respondent characteristics interact with the scale design and result in different patterns of high/positive and low/negative responding. Taking current results into consideration, researchers should be aware of the advantages and disadvantages that each scale design will have on data quality. Having no control over respondent characteristics, researchers should look into survey design as a possible way to reduce response burden.

As demonstrated by the analyses, respondent characteristics such as cognitive reflection and Big Five personality traits served as more precise estimation of respondent differences than a traditional set of demographic predictors (age, education, gender). Differences in respondent characteristics can be used to inform survey designs and improve response quality. First, pilot surveys should be tested on a wide range of respondents of different demographics, cognitive ability and personality traits. The process of determining a final survey design should follow a 'principle of the lowest common denominator' – the final survey design should perform well even among respondents with the lowest cognitive ability and those more susceptible to the cognitive fatigue. If the selected design has demonstrated an improved performance among more reluctant respondents, researchers can expect it to perform equally well with the rest of the survey population.

The limitation of administering the questions to capture the Big Five traits as explanatory factors for survey quality lies in the design of the measurement itself. As these scales have been administered as a part of the survey, the responses might be subject to the same response bias as the core questionnaire and fail to objectively evaluate cognitive and psychological traits of each respondent. For example, the Big Five response grids could be affected by participants' motivation (e.g. optimizing and satisficing behaviours) in the same manner as the core questionnaire. Having external, previously collected data on respondent characteristics would assure a more objective evaluation of their characteristics and their role in response selection.

Next, previous studies (for example, He et al., 2014) demonstrated that respondents answering strategy changed depending on the types of questions they were answering, and non-attitudinal items were subject to a stronger scale direction effect than attitudinal items. The current study administered a range of behavioural and attitudinal questions but scale effects' analyses did not explore whether certain types of questions are more likely to be subject to the cognitive shortcuts.

In the present study each question has been presented on a separate page and the survey was relatively short. Furthermore, the questionnaire used item-specific response scales that are less cognitively demanding than agree/disagree response scales (Hohne & Krebs, 2018; Hohne, Revilla, Lenzner, 2018; Hohne, Schlosser, Krebs, 2017). This design feature could have contributed to the lack of the scale effects observed in this study. Combined with incentivised participants the interplay between respondent characteristics and survey design could have resulted in a fairly good quality responses and lower sensitivity to the design manipulations. The lack of scale effects can be applied to the vertical 5-point Likers scales that use item-specific responses. Future studies could test scale effects using longer or horizontally aligned scales or items presented on mobile devices.

The use of an online non-probability panel in current study may have implications for external validity—the sample had a higher education level and IT technology proficiency than the general population and may not be representative of the entire population. In future, the effects of scale direction should be tested on a general population to verify the generalisability of the scale effects. Furthermore, a longitudinal probability survey could increase the power of the findings by providing personality and cognition measures observed at several time points therefore controlling for possibly occurring time-related changes in personality components of individuals. Additionally, future studies could introduce a between-subject experimental design to test the validity of the scale direction effects to ensure that the selection of responses from the high/positive or low/negative end of the scales was due to the scale design and personal differences rather than other factors. The findings from the current study suggest that if questionnaire developers can motivate respondents to provide good responses, provide clear and unambiguous questions and response categories then the choice of the scale direction design may not be a crucial factor in designing a scale.

2.6. Appendix

Table 2.5. Questionnaire for the ‘Behaviour health and wellbeing’ survey. ‘Forward’ scale direction design.

	Question wording
Q1	How environmentally friendly are you? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q2	How concerned are you about knowing what impact you have on the environment? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q3	How worried are you that you have done enough to reduce your impact on the environment? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q4	Do you agree that it takes too much time and effort to do things that are environmentally friendly? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*

	Question wording
Q5	Do you believe that scientists will find a solution to global warming without people having to make big changes to their lifestyle? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q6	Do you ever decide not to buy something because you feel it has too much packaging? • *Always* • *Usually* • *About half the time* • *Seldom* • *Never*
Q7	How often do you take your own shopping bag when shopping? • *Always* • *Usually* • *About half the time* • *Seldom* • *Never*
Q8	How often are you normally happy? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*

	Question wording
Q9	How much of the time during the past week did you enjoy life? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*
Q10	During the past week how much of the time have you felt happy? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*
Q11	How frequently do you go to the movies, concert or theater? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Never*
Q12	How often would you say you take part in social activities with friends, relatives or work colleagues? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*

	Question wording
Q13	Compared to other people of your age, how often would you say you take part in social activities? • *Much more than most* • *More than most* • *About the same* • *Less than most* • *Much less than most*
Q14	How important is it for you to have people with whom you can discuss intimate and personal matters? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q15	How much of the time during the past week did you feel depressed? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*
Q16	How much of the time during the past week did you feel sad? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*

	Question wording
Q17	How much of the time during the past week did you feel lonely? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*
Q18	How much of the time during the past week could you not get going? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*
Q19	How much of the time during the past week did you feel that everything you did was an effort? • *Always* • *Often* • *Sometimes* • *Rarely* • *Never*
Q20	How happy are you with your health in general? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q21	How often do you eat fruit? Please do not include juice. • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*

	Question wording
Q22	How often do you eat vegetables or salad? Please do not include potatoes. • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*
Q23	How often do you smoke? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*
Q24	How regularly do you have a drink containing alcohol? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*
Q25	How frequently do you play sport or go to the gym? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*
Q26	How often are you doing any sports or other physical activity for at least 30 minutes? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*

	Question wording
Q27	Do you stop and talk with people in your neighborhood? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*
Q28	Do you ever seek advice from your neighbor? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*
Q29	To what extent would you say you belong to your neighborhood? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q30	How important are friendships and associations with other people in your neighborhood to you? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*
Q31	How similar are your values to your neighbors' values? • *Extremely* • *Quite* • *Somewhat* • *A little* • *Not at all*

	Question wording
Q32	How similar is your background to your neighbors' background? • *Very often* • *Often* • *Occasionally* • *Rarely* • *Very rarely or never*
Q33	Here are a number of personality traits that may or may not apply to you. Please indicate to which extent you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other. I see myself as… *Agree strongly; Agree moderately; Agree at little; Neither agree nor disagree; Disagree a little; Disagree moderately; Disagree strongly* Extraverted, enthusiastic Critical, quarrelsome Dependable, self-disciplined Anxious, easily upset Open to new experiences, complex Reserved, quiet Sympathetic, warm Disorganised, careless Calm, emotionally stable Conventional, uncreative
Q34	The next three questions will ask you to perform some simple calculations. Please, enter your answer in the text box.
Q34.1	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? *[Textbox]*
Q34.2	A bat and a ball together cost 110 cents. The bat costs 100 cents more than the ball. How much does the ball cost? *[Textbox]*

	Question wording
Q34.3	If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? *[Textbox]*
Q35	What is your age? *[Textbox]*
Q36	What is your gender? *Male; Female; Other*
Q37	What is the highest level of education you have completed? *High School/GED; Some College; 2-year College Degree; 4-year College Degree; Master's Degree; Doctoral Degree; Professional Degree (JD, MD)*

Table 2.6. Questionnaire for the 'Behavior health and wellbeing' survey. 'Reversed' scale direction design.

	Question wording
Q1	How environmentally friendly are you? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*
Q2	How concerned are you about knowing what impact you have on the environment? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*

	Question wording
Q3	How worried are you that you have done enough to reduce your impact on the environment? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*
Q4	Do you agree that it takes too much time and effort to do things that are environmentally friendly? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*
Q5	Do you believe that scientists will find a solution to global warming without people having to make big changes to their lifestyle? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*
Q6	Do you ever decide not to buy something because you feel it has too much packaging? • *Never* • *Seldom* • *About half the time* • *Usually* • *Always*

	Question wording
Q7	How often do you take your own shopping bag when shopping? • *Never* • *Seldom* • *About half the time* • *Usually* • *Always*
Q8	How often are you normally happy? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q9	How much of the time during the past week did you enjoy life? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q10	During the past week how much of the time have you felt happy? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q11	How frequently do you go to the movies, concert or theater? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*

	Question wording
Q12	How often would you say you take part in social activities with friends, relatives or work colleagues? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q13	Compared to other people of your age, how often would you say you take part in social activities? • *Much less than most* • *Less than most* • *About the same* • *More than most* • *Much more than most*
Q14	How important is it for you to have people with whom you can discuss intimate and personal matters? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*
Q15	How much of the time during the past week did you feel depressed? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*

	Question wording
Q16	How much of the time during the past week did you feel sad? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q17	How much of the time during the past week did you feel lonely? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q18	How much of the time during the past week could you not get going? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q19	How much of the time during the past week did you feel that everything you did was an effort? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q20	How happy are you with your health in general? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*

	Question wording
Q21	How often do you eat fruit? Please do not include juice. • *Very rarely or never* • *Rarely* • *Occasionally* • *Often* • *Very often*
Q22	How often do you eat vegetables or salad? Please do not include potatoes. • *Very rarely or never* • *Rarely* • *Occasionally* • *Often* • *Very often*
Q23	How often do you smoke? • *Very rarely or never* • *Rarely* • *Occasionally* • *Often* • *Very often*
Q24	How regularly do you have a drink containing alcohol? • *Very rarely or never* • *Rarely* • *Occasionally* • *Often* • *Very often*
Q25	How frequently do you play sport or go to the gym? • *Very rarely or never* • *Rarely* • *Occasionally* • *Often* • *Very often*

	Question wording
Q26	How often are you doing any sports or other physical activity for at least 30 minutes? • *Very rarely or never* • *Rarely* • *Occasionally* • *Often* • *Very often*
Q27	Do you stop and talk with people in your neighborhood? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q28	Do you ever seek advice from your neighbor? • *Never* • *Rarely* • *Sometimes* • *Often* • *Always*
Q29	To what extent would you say you belong to your neighborhood? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*

	Question wording
Q30	How important are friendships and associations with other people in your neighborhood to you? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*
Q31	How similar are your values to your neighbors' values? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*
Q32	How similar is your background to your neighbors' background? • *Not at all* • *A little* • *Somewhat* • *Quite* • *Extremely*

	Question wording
Q33	Here are a number of personality traits that may or may not apply to you. Please indicate to which extent you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other. I see myself as… *Agree strongly; Agree moderately; Agree at little; Neither agree nor disagree; Disagree a little; Disagree moderately; Disagree strongly* Extraverted, enthusiastic Critical, quarrelsome Dependable, self-disciplined Anxious, easily upset Open to new experiences, complex Reserved, quiet Sympathetic, warm Disorganised, careless Calm, emotionally stable Conventional, uncreative
Q34	The next three questions will ask you to perform some simple calculations. Please, enter your answer in the text box.
Q34.1	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? *[Textbox]*
Q34.2	A bat and a ball together cost 110 cents. The bat costs 100 cents more than the ball. How much does the ball cost? *[Textbox]*
Q34.3	If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? *[Textbox]*
Q35	What is your age? *[Textbox]*

	Question wording
Q36	What is your gender? *Male; Female; Other*
Q37	What is the highest level of education you have completed? *High School/GED; Some College; 2-year College Degree; 4-year College Degree; Master’s Degree; Doctoral Degree; Professional Degree (JD, MD)*

3. The effect of cognitive and non-cognitive characteristics on device preference and response quality in web surveys

Abstract

Respondents in web surveys can participate with several devices: PC, smartphone or tablet. Previous studies have found that in general population studies, age is an important predictor of device use. Yet, beyond age, there are other factors related to the psychosocial background of respondents that may explain why respondents choose a particular device to complete a survey. This study focuses on 1318 respondents who participated in four waves of a student cohort study at the University of Essex, UK. We studied the role of need for cognition, probabilistic numeracy, Big Five personality and behavioural traits as predictors of device preference over the course of the study. The surveys were long, and easiest to complete on a PC. We found that respondents with a higher need for cognition were more likely to respond from a PC. Extravert respondents were more likely to respond from PCs, whereas more agreeable respondents used smartphones or tablets in addition to PCs. Course attendance and higher marks were associated with less switching to smartphones and sticking to PC completion. When controlling for personal differences, item non-response was highest for those who responded from smartphones than only PCs. Additionally, PC respondent group was more likely to complete all four waves of the survey. These findings contribute towards a debate on the usefulness of personality variables in explaining web respondents' behaviour, device preference and implications of multiple device use on data quality. 'Good' survey respondents adapt their choice for a survey device based on the survey experience, while 'Bad' respondents stick with less optimal devices, yet devices may pose additional limitations on data quality.

3.1. Introduction

Research in the past years has shown a consistent increase in the proportion of mobile devices being used alongside desktop and laptop computers for web survey completion (Duggan & Smith, 2013; Hern, 2015; Smith, 2015). According to data from the Pew Research Center, mobile coverage in the general population in the United States has increased between 2010 and 2015, with significantly more respondents accessing the internet via a mobile device than a PC (PEW report, 2015). Similar trends have been observed in the United Kingdom. For example, in 2015 UK smartphone ownership was reported to be 68% whereas more recent reports from 2016 and 2017 found that 76% of UK adults own a smartphone (Eurostat report, 2016; Ofcom, 2017). Similarly, the ownership of tablet computers has been steadily increasing - 53% of adults owned one by 2018 (PEW report, 2018). The proportion of mobile device ownership is particularly high among certain demographic groups: 91% of university graduates and 94% of 18-29–year-olds reported owning a smartphone (PEW report, 2018). Given the discrepancy in rates between mobile device ownership and mobile response rates in surveys (Lugtig & Toepoel, 2015), it becomes evident that not everybody who owns a smartphone would use it for survey completion.

The aim of the present study is to establish to what extent individual differences between respondents can explain device preference in a longitudinal online study conducted among students at the University of Essex. First, we review the literature on the differences in sociodemographic factors between PC and mobile device respondents of an online survey sample. Next, we briefly discuss what the implications are of the fact that respondents self-select into different types of devices for completing web surveys. We will show in our empirical study how several psychosocial variables are related to device preferences over time, and how data quality differs depending on the type of respondent and the device used. We will conclude with a discussion of our findings and implications for longitudinal survey design.

3.2. Background

According to the 2015 Pew Research Center report and 2017 Ofcom report, mobile ownership is ever increasing with the highest rate remaining among richer economies while the rate of mobile penetration in emerging countries is catching up (PEW research, 2016; Ofcom, 2017). This, in turn, causes an increasing rate of smartphone users accessing online surveys from their mobile devices. Corresponding with the increase of smartphone use in the general population, the share of respondents accessing the LISS panel from mobile devices increased from 3% to 11% between 2012 and 2014 (de Bruijne & Wijnant, 2014). Lugtig and Toepoel (2015) reported that 15% of LISS respondents were using tablets and 5% were using smartphones. In comparison, in the German GESIS Panel, 16% of panel respondents indicated tablet as their preferred device and 8% preferred smartphone to complete surveys (Struminskaya, Weyandt, Bosnjak, 2015). Similarly, an increasing share of mobile respondents was observed in European and American online panels (Couper & Peterson, 2016; Lugtig & Toepoel, Amin, 2016; Mavletova, 2013; Struminskaya et al., 2015). Several longitudinal panels demonstrated that despite the survey length and level of optimisation, a certain percentage of respondents consistently take surveys on smartphones and tablets or switch between devices. It is no longer a question whether mobile respondents should be allowed to take a web survey, but rather how to make sure that web surveys are usable for respondents on all kinds of devices.

Another strong argument supporting mobile optimization of web surveys is that mobile respondents are often different from PC respondents in terms of their demographics. One of the most prominent features of the group of mobile respondents is that smartphone and tablet respondents are more likely to be younger (Arn, Klug, & Kołodziejski, 2015; de Bruijne & Wijnant, 2014; Mavletova, 2013; Peterson, Mechling, LaFrance, Swinehart, Ham, 2013; Sommer, Diedenhofen, Musch, 2017; Toepoel & Lugtig, 2014; Wells, Bailey, & Link, 2014), female (de Bruijne & Wijnant, 2014; Peterson et al., 2013; Sommer et al., 2017; Wells et al., 2014), have a higher income (Mavletova & Couper, 2014; Toepoel & Lugtig, 2014), define themselves as early technology adopters (de Bruijne & Wijnant, 2014), and rely mainly on mobile internet (Mavletova, 2013; Mavletova & Couper, 2016a; Wells et al., 2014). Based on 2015 Pew research data on smartphone ownership, Keusch and Yan (2017) concluded that mobile respondents might represent a hard-to-reach part of the population that otherwise cannot be captured in PC web surveys. Lugtig, Toepoel, and Amin (2016) found that mobile-only respondents in the US survey have lower education levels, are more likely to be married and to be non-white. They concluded that mobile respondents represent a group that overlaps with parts of the general population that are otherwise hard to recruit.

Potentially, mobile enabled surveys can help reduce coverage bias by accessing groups that mostly rely on mobile internet (Mavletova & Couper, 2015). Yet, mobile respondents consistently have higher rates of unit non-response (Buskirk & Andrus, 2014; de Bruijne and Wijnant, 2013; Mavletova & Couper, 2014; Wells et al., 2014) in probability and non-probability surveys. Furthermore, break-off rates from mobile respondents are higher compared to desktop users particularly in non-optimized surveys (Bosnjak et al., 2013), but remain high even in surveys with an optimized design (Buskirk & Andrus, 2014; Mavletova 2013; Mavletova & Couper, 2013; Peterson et al., 2013; Wells et al., 2014).

In order to reduce the potential negative effects of mobile completion on data quality, researchers' attention has shifted towards survey design optimisation to reduce non-response and measurement error (Buskirk & Andrus, 2012). Indeed, respondents sometimes evaluate mobile-optimized designs more positively, completion rates are higher (Arn et al., 2015), break-off rates are lower (Mavletova & Couper, 2015; Peterson et al., 2013) and so are item non-response rates (Guidry, 2012), although findings remain mixed across studies (Andreadis, 2015; Antoun, Couper & Conrad, 2017; de Bruijne & Wijnant, 2014; Keusch and Yan, 2017; Mavletova & Couper, 2013, 2014, 2016a, 2016b; Wells et al., 2014). So far, evidence of differences in non-response and measurement error between mobile and PC responses has not reached a uniform conclusion.

Responses collected from mobile devices and PCs may be different due to the limitations posed by each device on response behaviour or the differences between respondents due to the self-selection. Some studies have experimentally assigned devices to participants in order to control for the self-selection and further investigate the effects of the device on responses (de Bruijne & Wijnant, 2013; Keusch & Yan, 2017; Mavletova, 2013; Peytchev & Hill, 2010; Wells et al., 2014). De Bruijne and Wijnant (2013) did not observe strong device effects obtained from mobile household panel respondents who were experimentally assigned to PC, mobile optimised and PC mobile optimised design. Similarly, Wells et al. (2014) have assigned PC and mobile device conditions to the respondents who originally participated in a probability panel from a mobile device and did not find strong device effects in terms of responses given. Keusch and Yan (2017) experimentally assigned respondents who initially responded from a mobile device to a device switching condition in a non-probability panel to capture the effect of the device and respondent characteristics. Keusch and Yan (2017) found significant differences between PC and all iPhone respondents in terms of gender, education and ethnic group further confirming that smartphone users represent a demographically different group. The authors observed more break-offs and item-missing both in assigned and unintentional iPhone condition compared to PC which remained even after controlling for demographics. Differences between PC and smartphones as well as minimal differences between iPhone users and those who switched from PC to iPhone suggest that data quality is predominantly influenced by the design-induced limitations. All the studies above have acknowledged differences between PC and mobile respondents and based their experiments on mobile users' sample only assigning them to mobile and PC conditions. While this study design allows to explore device effects it cannot be generalised to the users who respond from the device they choose and the one most likely they are proficient in. Next, Mavletova (2013) has assigned online PC users to take a survey on a PC or a mobile device. The final sample revealed that mobile respondents were younger, with a higher mobile proficiency and ownership. Even though the devices were experimentally allocated a lower rate of mobile response and experimental non-compliance have resulted in discrepancies between mobile and PC comparison groups. Similarly, the fact that Peytchev and Hill (2010) have assigned respondents to a mobile web survey could be the underlying reason for the observed device effects – respondents who were less proficient to use mobile device for responding produced lower quality data. In the probability panel study, Antoun et al. (2017) experimentally assigned respondents to PC and smartphone and reversed the pattern in the next wave thus controlling for respondents' self-selection bias. While this study did not control for participants' demographics, the sample was comparable to the general LISS panel population. Antoun et al. (2017) have concluded that data from experimentally assigned conditions were of a comparable quality. The studies that request users to participate from a certain device do suffer from the fact that respondents do not stick to the device they were assigned to, they are more likely to break-off or not participate at all. This often leads to the differences in sample composition between PC and mobile comparison groups negating the benefit of an experimental treatment. Moreover, forcing respondents into a device is not a natural situation as in self-administered surveys respondents would normally self-select their device therefore some observed device effects could be attributed to respondents using a device they were less accustomed to.

Another approach towards disentangling self-selection and measurement effects used the power of longitudinal panels where panel members took part in several waves of a survey using a device of their choice (Lugtig & Toepoel, 2016; Struminskaya et al., 2015). The studies observed that device consistency across waves was similar for both panels, smartphone use was associated with a lower likelihood of participating in more waves (Lugtig & Toepoel, 2015; Lugtig et al., 2016; Struminskaya et al., 2015). Lugtig and Toepoel (2015) found the highest item-missing rates for smartphone data but overall a very low effect of device on measurement error and attributed differences to the respondent. In turn, Struminskaya et al. (2015) similarly found highest item non-response rates for smartphones or switching patterns that involve smartphones and related them to the situation rather than respondents' behaviour. Respondent characteristics could explain some of the phenomena: gender had no effect, yet older or more educated respondents showed lower item-non-response.

So far studies on device preference, response propensity and data quality used respondents' demographics to explain participation in the waves and device preference. However, the range of involved personal characteristics might be larger, involving more general psychological parameters such as personality traits and cognitive ability which can then affect respondent behaviour. Several studies used personality traits and cognition to explain respondents' preference for mobile devices and participation in online panels. Butt and Phillips (2008) found that higher extraversion and lower emotional stability and lower agreeableness and conscientiousness were linked to a higher mobile use as measured by making and receiving calls and messages. The study was based on a small non-probability sample and self-reported mobile use data and could have been subject to the multiple testing error, however it provided an early insight into the possible personality differences among heavier mobile users. Bosnjak et al. (2013) has compared general panel respondents to the ones who were willing to participate in the web mode of the study. Bosnjak et al. (2013) found that those who expressed willingness to participate were less conscientious, whereas a final sample of web respondents had lower conscientiousness and higher openness compared to the original panel members. While demographics and personality factors were significant predictors of web survey participation the effect sizes for all personality predictors were rather small. Antoun (2015) has further expanded on this approach and explored the role of Big Five personality factors and need for cognition in respondents' willingness and likelihood to take surveys on mobile devices. Antoun (2015) found that willingness to participate and actual participation were lower among more extravert and more conscientious respondents. Higher education and need for cognition could predict willingness to participate but were not significant among those who actually took the mobile survey.

While this area is yet to be further researched, it seems plausible to assume that the interplay between respondents' demographics and personal traits may have an effect on device preference, response behaviour and quality of responses given. Compared to demographics, cognition and personality traits could be a more direct measurement of respondent characteristics and even serve as proxy variables for a broader range of substantive variables in sociology, psychology, and decision-making that could affect survey response. It is typically still the case that completing web surveys is harder on mobile devices. Respondents who care about giving good answers, or who value a good experience may therefore opt to complete web surveys on the PC. Such a hypothesis would also explain why survey completion rates on smartphones are still so much lower than general Internet use - many respondents may feel that completing a survey on a smartphone is hard or unpleasant. Yet, there is also a group who persistently uses a smartphone. Are these respondents less concerned about providing good data, or are they perhaps better in dealing with a survey task that is harder? Most of recent findings agree that further investigation is necessary to disentangle device from self-selection effects – studying respondents' personality and need for cognition could be one way to gain deeper insight into survey response behaviour. This study seeks to contribute to our knowledge by studying a group of respondents that is homogenous in age in order to avoid strong age effects. We document how device use changes between the waves, and whether the choice of device is affected by psychosocial characteristics.

Our study

Our study uses data from four waves of a longitudinal online survey conducted at University of Essex interviewing a cohort of undergraduate students who started their course in 2015. The online survey was designed with Qualtrics and some of the survey questions (e.g. grids) were optimized for smaller screens of mobile devices. The content of the survey covered a range of attitudinal and behavioural questions intended to understand aspirations, expectations and career and education outcomes of students. The survey employed a paging design with one question displayed per page: matrix and grid questions with more than four Likert-scale points were optimized for smartphones and the size of slider scales was increased. When questions were left unanswered, a pop-up screen asked respondents whether they would like to advance in the survey without answering. Respondents were able to pause their survey completion and return to the questionnaire at any time while the survey was in the field.

Our sample is homogeneous when it comes to age, so we are naturally controlling for this effect, allowing us to focus on personality, cognition and other characteristics. Almost all participants in our sample have access to at least a smartphone and a PC (desktop or laptop) for survey completion. Thus, at every wave respondents face a choice between devices to complete a survey. In order to obtain more information on the decision making process of respondents, the core set of cognition and personality covariates such as Big Five traits, need for cognition and probabilistic numeracy was supplemented with behavioural and demographic data from administrative records: students' class attendance, academic performance and age, gender as well as ethnic group.

First, we determine how respondents who consistently use a PC across waves are different in terms of personality and cognition measures from respondents who exclusively use a smartphone or switch to different devices between the waves. Second, we explore whether consistent smartphone respondents are different from those who switch to a different device after taking the survey the first time. Finally, we examine the effects of device choice on survey participation, break-off and item non-response.

We expect that respondents learn the content and length of the survey across waves. Respondents who are willing to complete the survey diligently would choose an optimal device (PC) or switch to it as they get familiar with the content of the survey. Therefore, we expect student with a higher need for cognition to use PC as a more optimal device for a survey that is 45 to 60-minute long. Previously, higher rates of need for cognition were linked with higher conscientiousness and openness (Sadowski & Cogburn, 1997), so we can expect these two traits to also be positively associated with using a PC. The behavioural variables such as higher attendance and higher overall end of the year mark are expected to be linked with a higher student motivation and effort. Again, we expect those students to complete more waves and be more likely to choose the optimal survey device for completion. On the other hand, lower need for cognition, lower probabilistic numeracy and lower openness, conscientiousness, and agreeableness could be associated with the continuous use of a suboptimal device (mobile), completing fewer waves, higher non-response and break-offs.

3.3. Data and Methods

Sample

The initial sampling frame covered 2094 students newly enrolled in a bachelor's degree at the University of Essex in 2015. Of these, 1909 were still enrolled in 2016. All students received invitations to take part in the survey via email. This paper uses data from four waves of approximately hour-long online surveys. The first wave was fielded in November 2015, the next one in March 2016, and the following two waves were conducted in November 2016 and March 2017. Data in each wave were collected over a period of one month. Respondents were provided a £20 cash incentive upon successful completion of each wave. Out of the entire pool of eligible participants, 1540 replied to the initial survey invitation. Only respondents who took two and more waves and provided data on core covariates were included in the analysis hence creating a sample of 1318 individual observations.

The sample consisted of undergraduate students, therefore age was highly clustered around the mean of 21 year (SD=2), males and females were nearly equally represented in the sample (both 45%) and 10% of respondents chose not to disclose their sex. Some administrative records regarding student academic performance were included as covariates: average attendance of classes and lectures for the cohort was 64% (SD=20) with the average mark being 59% out of 100% (SD=12.9).

Device use

User agent strings were recorded for each participant. Based on information from the user agent strings, we used the 'parseuas' module in Stata (Rossmann & Gummer, 2014) to code the devices into three categories – PCs (desktops and laptops), smartphones, tablets. In order to analyse device switching behaviour across waves, device data for all respondents who took more than one wave were grouped according to the devices they were choosing for each of the survey wave. Nearly half of the sample, 765 respondents (48.8 %), preferred to complete all waves of the survey on their PC; 353 participants (22.5%) were switching between PC and smartphone. It was fairly common to switch between PCs and tablets – 180 (11.5%) participants used both devices when completing two or more surveys. Although the online survey lasted 45 to 60 minutes, 130 (8.3%) respondents exclusively used smartphones for survey completion over the course of the longitudinal study. A very small number of 31 respondents completed all surveys on a tablet (2.0%), 26 were switching between smartphone and tablet (1.7%), and 30 used all three devices (1.9%). The sample sizes of the last three groups are too small to analyse, and we will not show results for these groups.

Respondent characteristics

We now turn our attention to the individual differences of respondents and the extent to which device preference can be attributed to them.

Need for cognition was measured by a combined score of 18-items. The need for cognition scale asked students to rate on a 5-point Likert scale the extent to which they agree with each of 18 statements about the satisfaction they gain from thinking. The average score for the sample was 3.3 (SD=.51).

The inter-item reliability Cronbach's $\alpha = .82$ was comparable to previous research that used the short need for cognition scale (Cacioppo, Petty, Feinstein, & Jarvis, 1996).

Probabilistic numeracy was assessed using a short 4-item scale that used questions of easy, average and hard difficulty levels, asking participants to estimate the marginal and joint probabilities of occurring events (Hudomiet et al., 2017). The full wording of questions is documented in Appendix Table 3.6. For research purposes, the probabilistic numeracy scale was used to create a variable that reflected a number of correct responses from 0 to 4. Overall, the majority of the sample answered three out of four questions correctly (52.2%), 28% of respondents answered all four questions correctly, 10% and 3% of respondents gave correct answers in two and one question respectively and 6.8% did not provide correct answers. These numbers are reflective of the study by Hudomiet et al. (2017) supporting the notion that the majority of survey participants understand the marginal probability concept, yet encounter difficulty answering the more difficult questions on joint probabilities.

Next, the study used a short 15-item Big Five personality inventory administered in two grids. Each of the five personality factors was measured with 3 questions and asked respondents to indicate their agreement with each of the statements on a 7-point Likert scale from '*Strongly Disagree'* to '*Strongly Agree'*. We assessed the scale reliability for each factor and produced an average score for each dimension. Average values across the sample for openness were 5.2 out of 7 (Cronbach's $\alpha = .71$), conscientiousness 4.6 (Cronbach's $\alpha = .53$), extraversion 4.5 (Cronbach's $\alpha = .73$), agreeableness 5 (Cronbach's $\alpha = .55$) and, emotional stability 4.2 (Cronbach's $\alpha = .76$). The 15-item scale administered in this survey yielded reliability scores comparable to those tested in larger online surveys such as BHPS and GSOEP (Donnellan & Lucas, 2008; Lang, John, Lüdke, Schupp, & Wagner, 2011).

3.4. Results

Descriptive statistics

Table 3.1 presents descriptive statistics of individual respondent characteristics grouped according to their device switching behaviour. Respondents who switched between PC and tablet across waves had the highest Need for Cognition score of 3.36 (SD=.5), whereas smartphone only respondents reported a noticeably lower score of 3.14 (SD=.6). Average number of correctly answered probability questions was lower for a smartphone group 2.6 (SD=1.3) and the highest for PC respondents 3 (SD=.9).

Average class and lecture attendance over the year was the highest for the PC-tablet group 65.7%, followed by the PC group 64.8%, and was the lowest for the smartphone group 59.1%. PC to tablet switchers had the highest average mark 62.6%, a much lower result was observed in a smartphone group 57.5%.

As age did not differ across the sample, we used the Higher Education Academy classification to split respondents into young (aged under 21) and mature (aged 21 and over) students (Higher Education Statistics Agency, 2018). This measure is not intended to reflect the age of respondent but rather is aimed at controlling for the life experiences of respondents: young respondents are most likely the ones who went to the university straight after completing secondary education, whereas mature students are those who have had a gap between secondary and tertiary education possibly working or studying elsewhere. Respondents 20 years and younger were classified as young and comprised 68% of the sample; 32% of the sample were mature students over 21 years old. We found a small effect of gender: always PC respondents (49% women) and PC-tablet (51% of women) groups showed a nearly even split between men and women using these devices. On the other hand, smartphones only were preferred by more females (58%). Females were also dominating the smartphone to PC (65.7%) group. We also find small effects of ethnicity: white participants were more likely to use only a PC (53.2%) or switch between PC and tablet (52.8%). Other ethnicities were more likely to always use a smartphone (58.5%) or switch between PC and smartphone (52.4%).

Table 3.1. Mean values (SD) for respondent characteristics grouped by device switching behaviour.

	Always PC	**Always Smartphone**	**PC-Smartphone**	**PC-Tablet**
Need for Cognition	3.3 (.5)	3.14 (.6)	3.18 (.5)	3.36 (.5)
Probabilistic numeracy	3 (.9)	2.6 (1.3)	2.8 (1.1)	2.9 (1.00.)
Openness	5.2 (1.1)	5 (1.2)	5.1 (100)	5.2 (100.)
Conscientiousness	4.6 (.9)	4.7 (.9)	4.6 (.9)	4.6 (.9)
Extraversion	4.5 (.1.3)	4.6 (1.4)	4.6 (1.2)	4.6 (1.2)
Agreeableness	4.9 (1.00)	4.9 (1.2)	5 (1.00)	5 (1.00)
Emotional Stability	4.2 (1.4)	4.1 (1.4)	4.2 (1.3)	4.2 (1.3)
Average Attendance	64.8 (20.1)	59.1 (21.8)	61.3 (20.3)	65.7 (19.8)
Average year mark	61.6 (12.8)	57.5 (13.6)	58.9 (12.6)	62.6 (11.4)
Sample size	765	130	353	180

Individual differences as determinants of the device switching behaviour

Multinomial logistic regression (MLR) was used to determine the significance of individual characteristics in determining device switching behaviour in a multivariate model. The dependent variable – device switching groups – was captured as every possible combination of the three devices respondents could have used to complete up to four survey waves, therefore producing six exhaustive and mutually exclusive categories. We chose not to show results for the groups that used always tablet, smartphone and tablet or all three devices due to small sizes of these groups. Independent variables include measure of probabilistic numeracy, need for cognition score, average scores for the Big Five personality traits satisfied linearity assumption and were included in the analysis as continuous measures. Additionally, academic records for performance and attendance as well as respondent's age group, gender, ethnic group were included as predictors. Table 3.2 presents the results from the MLR model where 'always PC' was used as a reference category. We can conclude that need for cognition is significant in distinguishing Smartphone and PC-Smartphone users from respondents who used exclusively PCs for the survey. For each unit increase in the need for cognition score, the odds of being in the Smartphone or PC-Smartphone group decreased by 4% and 2% respectively. Similarly, respondents who performed better on the probabilistic numeracy test were less likely to always use smartphones or smartphones and PCs to participate in the survey. For each unit increase in the probabilistic numeracy score, the odds of being in the Smartphone or PC-Smartphone group decreased by 37% and 11% respectively. Males were less likely to represent Smartphone and PC users group (OR=.43). Those with the higher attendance (OR=.99) and a higher average year mark (OR=.98) were slightly less likely to be in the PC-Smartphone switching group than in the PC. These findings are in line with our hypotheses that higher cognitive indicators and more responsible behavioural traits will be associated with selecting a more convenient device for the survey, whereas females are a more likely demographic group to participate in a survey from a mobile device.

The odds of being in the PC-Smartphone group were higher for respondents who reported higher scores on agreeableness scale (OR=1.13). Those who scored higher on agreeableness scale (OR=1.3) were also more likely to respond from PC and tablets, whereas those who scored higher on extraversion were less likely to use PC and tablets for responding than always use PC: for each unit increase on the extraversion scale the odds of switching between PC and tablet decreased by 16%. Overall it appears that probabilistic numeracy, need for cognition, attendance, sex and some of the personality traits provided a substantive contribution towards understanding of the device switching behaviour.

Table 3.2. Multinomial logistic regression model showing results for need for cognition, probabilistic numeracy score, Big Five personality traits and administrative data as determinants of device switching pattern.

	Always Smartphone	**PC-Smartphone**	**PC-Tablet**
Need for Cognition	.96**	.98**	1.01
	(.02)	(.01)	(.01)
Probabilistic numeracy	.63***	.89**	.84*
	(.1)	(.07)	(.09)
Openness	.83	.89	.99
	(.12)	(.07)	(.09)
Conscientiousness	1.2	1.07	.87
	(.15)	(.09)	(.1)
Extraversion	1.1	1.04	.84*
	(.1)	(.06)	(.07)
Agreeableness	1.1	1.13*	1.3**
	(.12)	(.07)	(.09)
Emotional Stability	.85	.9	.9
	(.1)	(.06)	(.07)
Average Attendance	.99	.99**	1.00
	(.007)	(.004)	(.005)
Average year mark	1.00	.98*	1.00
	(.01)	(.01)	(.009)
Sex (male)	.62	.43***	.96
	(.25)	(.16)	(.19)
Age (young)	.96	1.01	1.17
	(.26)	(.17)	(.21)
Ethnicity (white)	.6	.65	.88
	(.36)	(.23)	(.29)
Sample size	89	312	168

**p*<.05, ***p*<.01, ****p*<.001.

Note. Odds ratios and standard errors (in parentheses) are reported. Reference group: always PC users (n=675). Between 12-35% of missing values for age, ethnicity, average lecture attendance, average final mark were imputed using linear imputation commands.

Individual differences as determinants of smartphone users' behaviour

The regression analysis reported in Table 3.3 will look specifically at the smartphone respondents group. The question we would like to answer is for which students the survey experience (one-hour survey on a mobile device) at wave one resulted in a switch to a different device in the next longest wave. The multinomial regression included respondent characteristics as covariates in order to determine whether smartphone respondent characteristics are associated with the likelihood of switching to a different device or dropping out after taking the first wave of the survey. The first survey took an hour, leading some respondents to choose a different device onwards. Respondents who continued using a smartphone in the next longest survey were used as a reference category in the multinomial logistic regression model. We observed that respondents from the young age group were less likely to drop out after using smartphone in the first wave (OR=.23) and continue responding from a smartphone in subsequent waves while the other predictors were not significant.

Table 3.3. Multinomial logistic regression model for the device switch behaviour of wave one smartphone respondents.

	Switch to PC or tablet	**Dropout**
Need for Cognition	1.04	1.02
	(.02)	(.04)
Probabilistic	1.12	.93
numeracy	(.12)	(.23)
Openness	1.14	.64
	(.15)	(.28)
Conscientiousness	1.01	1.07
	(.19)	(.38)
Extraversion	.81	.85
	(.13)	(.26)
Agreeableness	.96	1.06
	(.15)	(.31)

(continued)

	Switch to PC or tablet	Dropout
Emotional Stability	.84	.91
	(.11)	(.24)
Average	1.00	1.02
Attendance	(.01)	(.02)
Sex (male)	1.05	1.65
	(.32)	(.59)
Average year mark	1.00	1.02
	(.02)	(.03)
Age (young)	.74	.23**
	(.36)	(.58)
Ethnicity (white)	1.6	2.43
	(.30)	(.6)
Sample size	105	17

*p<.05, **p<.01, ***p<.001.

Note. Odds ratios and standard errors (in parentheses) are reported. Reference group: Smartphone users in both wave one and wave three (n=116).

Implications of device use over the waves on response quality

In the final analysis, a series of multivariate regression models were fitted to predict whether respondents completed all four waves of the survey, dropped out or provided more missing items while controlling for self-selection effects and including respondent characteristics as covariates (Table 3.4).

A logistic regression model was fitted to model completion of all four waves or less for different device user groups. Compared to always PC users, smartphone (OR=.34) and PC-smartphone switchers (OR=.22) were less likely to complete all four waves. The results suggest that using a device such as smartphone or a combination of PC and a smartphone could lead to a lower rate of participation however there seems to be no such effect for tablet switchers.

The next logistic model looked at the number of surveys that respondents started but failed to complete. We created a binary variable indicating that a survey was successfully completed or respondents started but did not complete. Smartphone users (OR=.2) were significantly less likely than PC users to break off before submitting a survey.

Finally, an OLS regression was fitted to predict item non-response while controlling for selection effects, participation in the waves and drop-outs. The item missing values were coded only for questions that were seen and not answered by the respondents; therefore, there should be no direct positive association of missing responses and dropping out of the survey. Smartphone use is predicting on average 3.6 more missing answers than PC only use whereas PC-Smartphone switching causes slightly less missing items – an average of 2.4. Respondents who used tablet and PC did not report higher rate of missing items compared to PC only users. Respondents who used PC in all or at least in some of the waves produced less missing responses compared to those who used smartphone. These results evince some adverse effects on data quality posed by less web survey friendly devices such as smartphones.

Table 3.4. Factors predicting completion, drop-outs and item non-response across four waves of the survey controlling for personal characteristics.

	Completed 4 waves (binary) OR	**Break-off (binary) OR**	**Item non-response B**
Always Smartphone	.34*	.2**	3.6*
	(.51)	(.59)	(1.6)
PC-Smartphone	.22*	.18	2.4**
	(.69)	(1.02)	(1.00)
PC-Tablet	.56	.48	.94
	(.52)	(.58)	(1.2)
Need for Cognition	1.01	1.02	-.01
	(.01)	(.01)	(.05)
Probabilistic numeracy	1.03	.45***	-.84*
	(.06)	(.09)	(.41)
Openness	1.02	103	.26
	(.07)	(.11)	(.4)
Conscientiousness	1.12	.79	-.01
	(.08)	(.1)	(.48)
Extraversion	1.05	.99	-.29
	(.06)	(.1)	(.33)
Agreeableness	.96	.99	-.44
	(.06)	(.11)	(.39)
Emotional Stability	1.05	.86	-.25
	(.05)	(.09)	(.31)
Average Attendance	1.02***	1.00	.02
	(.004)	(.01)	(.02)
Average year mark	1.00	.99	.001
	(.01)	(.01)	(.04)
Sex (male)	.9	.7	.11
	(.14)	(.24)	(.84)
Age (young)	5.7***	.19***	3.1***
	(.14)	(.24)	(.84)
Ethnicity (white)	.9	1.28	-.76
	(.13)	(.23)	(.82)
Drop-out	--	--	-10.85***
			(1.5)
Waves completed	--	--	8.6***
			(.58)
Model	-2Log=1498	-2Log=604.7	R^2=.23

*p<.05, **p<.01, ***p<.001.

Note: Standard errors are reported in parentheses. Reference group: always PC respondents (n=675). Total sample size 1318.

3.5. Discussion

This study explored the association between respondent characteristics and device preference in a longitudinal web survey of university students. We tested whether certain cognitive and non-cognitive traits are more characteristic of PC respondents, smartphone respondents or those who switch devices over the course of the study. Furthermore, having that information we looked at the effect of device preference on data quality.

As we have been studying a homogenous sample of university students, we examined cognitive determinants of device use. We found that respondents with a lower need for cognition were more likely to respond from a smartphone or use both smartphone and PC to participate in the waves. If surveys are as long as the one used in this study (lasting between 45-60 minutes), having lower need for cognition may lead to a lower data quality on mobile devices. One way to counter this is to break the survey down into several shorter sections allowing respondents to complete it in parts. Next, as this survey was only partially mobile optimized, a mobile first design would be a solution to further reduce cognitive burden on smartphone respondents.

Next, we found that respondents with a higher probabilistic numeracy were more likely to always use PC for completing surveys than smartphone, tablet or switch between devices. Probabilistic numeracy has been linked to respondents' expectation measures in surveys and highly predictive of probability-related individual decision making. In case respondents who have higher probabilistic numeracy tend to rely on smartphone and tablets less often then these different respondent groups will provide different responses in terms of beliefs about the future and consequently make different future probability-related decisions. Future research could look at the substantive data provided by different device groups to evaluate whether there is a difference in life decisions between these groups of respondents.

In terms of personality traits, extravert respondents were less likely to use PC and tablet during the survey, whereas higher agreeableness scores were linked to higher likelihood of participating from PC and smartphone or PC and tablet than just PC. Previous studies found that agreeable respondents are more likely to respond to the mobile survey (Antoun, 2015). The interpretation within the current research suggests that agreeable respondents made more effort to participate in all waves that were advertised to them which required relying on less convenient but readily available mobile devices. In case agreeableness is linked to more cooperation with the survey requests survey designers could take advantage of this trait to improve survey response. The initial survey invitation could include instructions advising respondents to best take the survey on a PC while not strictly forbidding mobile access.

Behavioural information showed that respondents with higher class attendance are more likely to respond from PC only, whereas lower attendance was linked to switching between PC and smartphone. Similarly, lower overall mark was associated with using PC and smartphone rather than responding exclusively from PC. Completing a long only partially optimized survey via a PC rather than a smartphone would benefit quality of responses given. Both findings could indicate that more responsible behaviour in terms of academic performance and attendance was associated with a preference for a PC (more optimal device) as opposed to mobile devices. Among respondents who first participated in a survey via smartphone respondents from a young group were less likely to drop-out in later waves. Again, when dealing with the long and effortful survey that might be additionally difficult to complete via smartphone researchers could encourage respondents to use PC rather than risk increased drop-out rates.

When controlling for personal differences we observed that devices alone influence participation and break-off in a cohort study. Compared to PC-only users, those who used smartphones were less likely to participate in all four waves of the survey. Compared to PC, smartphone respondents were producing the highest level of missing items. However, those who used exclusively smartphone throughout the survey were less likely to break-off before submitting their responses. Nonetheless, even though smartphone users are less likely to break off they are also less likely to participate in all surveys and while they do so they are more likely to skip answering some items. Collectively, the results strongly suggest that using exclusively smartphone survey leads to a lower data quality. Respondents who use smartphones in addition to PC are less likely to participate in all waves and overall produce more missing items than PC only group. Using both smartphone and PC did not reduce break-off rates – this group is just as likely to break off as PC users. Overall, the results show a certain trade-off in smartphone participation boosting response rates yet sacrificing some data quality. Data quality provided by users who responded from tablet and PC showed virtually no difference to the PC group once again supporting the notion that tablet use is comparable to the PC. Overall, the results show a certain trade-off in smartphone participation boosting response rates yet sacrificing some data quality.

The findings are based on four waves administered over two years in a study following the same cohort of undergraduate students. We expect this group to be IT proficient, with high levels smartphone ownership and availability of IT facilities at the university. This assumption is supported by a significantly higher smartphone response compared to the general population research panels. That allows us to have more confidence that the use of the device was more of a deliberate choice than the issue of the ownership. In the general population the pattern of switching might differ due to differences in demographics, smartphone penetration levels as well as the situational and social background factors. While the psychometric scales are highly associated with gender, age and education, having a homogenous sample allowed us to observe the effect of personality differences in a more controlled environment. Furthermore, the online survey waves that the results were based on were significantly longer than in studies previously used to study device switching – each online survey lasted 45 to 60 minutes which would require a larger investment of time and effort from participants. We observed respondents' behaviour under conditions where they became familiar with the content and duration of the survey and using that information made a choice of the device they responded from next.

Differences in respondent characteristics can be used to improve response quality. First, pilot web surveys should be tested on a wide range of respondents with different demographics, cognitive ability and personality traits. The process of determining a final survey design should follow a 'principle of the lowest common denominator' – the final survey design should perform well even among respondents with the lowest cognitive ability and those more susceptible to the cognitive fatigue. If the selected design has demonstrated an improved performance among more reluctant respondents, we can expect it to perform equally well with the rest of the survey population. Furthermore, survey research involving respondent characteristics have shown that respondent cognition and personality traits can determine participation in surveys and the final survey sample might differ from the population on a range of personality traits. Future research could test the plausibility of constructing weights using respondent characteristics for post collection adjustments.

3.6. Appendix

Table 3.5. Questions assessing the need for cognition.

Question wording
For each of the statements below please indicate to what extent the statement is characteristic of you.
Extremely uncharacteristic; Somewhat uncharacteristic; Uncertain; Somewhat characteristic; Extremely uncharacteristic
I would prefer complex to simple problems
I like to have the responsibility of handling a situation that requires a lot of thinking
Thinking is not my idea of fun
I would rather do something that requires little thought than something that is sure to challenge my thinking abilities
I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something
I find satisfaction in deliberating hard and for long hours
I only think as hard as I have to
I prefer to think about small, daily projects than long-term ones
I like tasks that require little thought once I've learned them
The idea of relying on thought to make my way to the top appeals to me
I really enjoy a task that involves coming up with new solutions to problems
Learning new ways to think doesn't excite me very much
I prefer my life to be filled with puzzles that I must solve

	Question wording
	The notion of thinking abstractly is appealing to me
	I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought
	I feel relief rather than satisfaction after completing a task that required a lot of mental effort
	It's enough for me that something gets the job done; I don't care how or why it works
	I usually end up deliberating about issues even when they do not affect me personally

Table 3.6. Questions assessing the probabilistic numeracy.

	Question wording
	Now we would like to ask you some questions about the chances of some events. We would like you to give a number from 0 to 100, where "0" means that you think there is absolutely no chance, and "100" means that you think the event is absolutely sure to happen. Consider a bowl that holds 10 balls. Some of the balls may be white and some red. You will be asked to take one ball from the bowl without looking.
Q1	First, suppose this bowl has 10 white balls and no red balls. On a scale from 0 percent to 100 percent, what is the percent chance that the ball you take is red? *[Textbox]*
Q2	Now suppose that the bowl holds 7 white balls and 3 red balls. What is the percent chance that the ball you take is white? *[Textbox]*

	Question wording
Q3	Assume that the weather forecast accurately reports the chance of rain. Suppose the weather forecast tells you that the chance it will rain tomorrow is 70%. What is the chance it will NOT rain tomorrow? *[Textbox]*
Q4	Suppose that whether it rains in your town and whether it rains in Paris are unrelated. The chance that it will rain in your town tomorrow is 50%. The chance that it will rain in Paris is also 50%. What is the chance that it will rain both in your town AND in Paris tomorrow? *[Textbox]*

Table 3.7. Fifteen-item Big Five personality measurement scale.

	Question wording
	To what extent do you agree with the following statements. I see myself as someone who... *Strongly disagree; Disagree; Neither agree nor disagree; Somewhat agree; Agree; Strongly agree* Worries a lot Gets nervous easily Remains calm in tense situations Is talkative Is outgoing, sociable

	Question wording
	Is reserved
	Is original, comes up with new ideas
	Values artistic, aesthetic experiences
	Has an active imagination
	Is sometimes rude to others
	Has a forgiving nature
	Is considerate and kind to almost everyone
	Does a thorough job
	Tends to be lazy
	Does things efficiently

Table 3.8. Description of the sampling frame and process of selecting a sample used for the analysis

	Sample
Initial sampling frame. Students newly enrolled in a Bachelor degree at the University of Essex in 2015	2094
Students still enrolled in a Bachelor degree at the University of Essex in 2016	1909
Participants who replied to the initial survey invitation	1540
Final sample. Respondents who participated in two and more waves and provided data on core covariates	1318

Conclusion

Summary

In the three chapters of this thesis, I have investigated how survey data quality is affected by the design of the response scale, the device used to complete the questionnaire, and respondent personality and cognition. When evaluating web survey data quality, researchers should take into account the collective contribution of the device, survey design and sample characteristics. Two chapters of this thesis examine the effects of different scale designs and conclude that even a small change in design might contribute to changes in response processing and selection. The third chapter demonstrates the influence of respondent characteristics on selecting a certain device for survey completion and of how the type of device can affect data quality.

Chapter one has extended earlier research on how the design of response scales affects measurement error by studying the implications of scale design on data quality for smartphone and tablet users. A vertical Likert scale design administered on a smartphone revealed primacy effects: options that were initially presented in a drop-down list were selected more frequently, while the effect was not observed on tablets. This primacy effect indicates that smartphone respondents are more susceptible to variations of the response scale design than tablet users. A consistent scale design (radio-button grids) was shown to reduce the level of measurement error between PC, smartphone and tablet users. Tablet users show little variation in data quality depending on the response scale design. When using designs adapted for smartphones, researchers should be aware of the unique effects that scale designs can have on smartphone response distributions. The findings suggest that researchers ought to adjust for differences in response distribution occurring due to the use of different devices.

Chapter two has explored the response process behind vertical item-specific Likert-type response scales presented in the 'forward' and 'reversed' order. The results presented in the chapter have contributed to the field by extending earlier research to examine if respondent characteristics could explain the satisficing and anchoring mechanisms that have been attributed to scale direction effects. The experiment did not observe a strong scale direction effect on selection of responses – the 'forward'-ordered scale induced only a slightly higher selection of responses from the high/positive end of the scale but no difference was observed compared to the 'reversed'-ordered scale. Respondents' tendency to select low/negative responses was predicted by indicators that traditionally are attributed to the satisficing response mechanism – higher age, lower educational level. However, variables indicative of satisficing (age, education, cognitive reflection) that should predict respondents' likelihood to select earlier response options did not interact with the scale direction to predict the selection of responses. This lack of a significant interaction strengthens the support for the anchoring mechanism taking place in scale effects. The results lend additional support towards the anchoring heuristic in the selection of options from the high/positive end of the scale. Respondents' tendency to select high/positive responses was explained by the lower extraversion – a personality trait that is attributed to the use of anchoring. What is more a significant interaction between one of the Big Five traits and a scale direction strengthens the case of anchoring taking place. Conscientious respondents' responses were affected by the scale direction: more conscientious respondents who were assigned to 'forward'-ordered scales selected less high/positive answers. This result suggests that using 'forward'-ordered scales induced more acquiescent responding among less conscientious respondents. Presenting 'forward'-ordered scales resulted in more conscientious respondents giving less high/positive responses as compared to the less conscientious respondents. The results from Chapter two were mixed as there were no strong advantages and disadvantages to each scale direction design; these results could be due to the sample composition of motivated respondents and due to a user-friendly questionnaire design. Administering a 'reverse-ordered' survey scale could reduce the possibility of the higher acquiescent responding among less conscientious respondents and have otherwise no detrimental effects on response quality of more conscientious respondents. However, if survey developers can motivate respondents to provide high-quality responses and provide clear and unambiguous questions and response categories, then the choice of the scale direction may not be a crucial factor in designing a scale.

The findings from chapter three provided insights into the role of respondent characteristics in affecting response behaviour in a multi-device online survey. The results indicate that respondent characteristics are linked to the choice of devices. In addition, the choice of device to respond to an online survey has been shown to affect data quality. Respondents with lower cognition and motivation are more likely to use a smartphone for survey completion. Overall, PC and tablet data are comparable and of a higher quality than smartphone responses. When participant characteristics are controlled for smartphone response is linked to a lower break-off but also a lower likelihood of participating in all waves and a higher item non-response. The suggestion of this chapter is that we should allow participation from different devices, however, perhaps, more radical changes to smartphone optimised designs should be implemented: feeding a survey in several shorter sections to reduce burden, use additional incentives to increase motivation, explore alternative questionnaire designs to reduce cognitive burden.

Limitations

The studies presented in this thesis are subject to certain limitations that could be addressed in future research. The studies were based on probability surveys of higher education students or on surveys implemented in a non-probability online panel. The probability samples were homogenous and consisted of respondents who were proficient in IT and likely to have had access to a mobile device. The downside of using these samples is that they differ from the general population – participants in the general population are less experienced with online surveys and less mobile affluent. In the near future the issue of mobile response will not only concern younger respondent groups. Over time, a larger share of the general population is likely to use a mobile device and will become more familiar using mobile devices for daily tasks including taking surveys on a smartphone.

The findings in chapter one and three are based on self-selected mobile device users. The self-selection into device reflects natural conditions where respondents can access online surveys from a device of their choice. The drawback of allowing respondents to choose their own device lies in the difficulty of separating device effects from self-selection effects on data quality. The findings from the chapters indicating that several respondent characteristics are significantly associated with device choice raise questions about the representativeness of the mobile samples used for methodological research in mixed-device surveys. The findings from the chapters are applicable to surveys with self-selected into device participants. Previous findings that allocated participants to mobile device surveys reported high drop-out rates or non-compliance with the instructions indicating that the results of experimental allocation might be biased and not representative of the general mobile user population (Keusch & Yan, 2017). The benefit of allowing self-selection into device as opposed to device allocation is that I have avoided respondent non-compliance with the assigned device and subsequently increased drop-out rates.

Next, chapters two and three that explored the role of respondent characteristics in data quality assessed personality (Big Five) and need for cognition with the main questionnaire. There is a possibility that the instruments evaluating the personality and cognition might be affected by the same measurement error as the main questionnaire (e.g. extreme responding, mid-point responding, and straightlining) and not be representative of the true estimates of respondent personality and cognition (Vaerenbergh & Thomas, 2013). A way to deal with the issue of the reliability of personality and cognition measurements obtained from the questionnaire would be to use alternative methods of measuring personality and cognitive capacity such as orally administered ability tests or previously recorded performance not based on self-reported measurements. Chapter three has attempted to overcome these limitations by supplementing the personality measurements with behavioural (indirect) indicators of respondent cognition and motivation such as academic performance scores and class attendance.

Final discussion

Information technology is constantly evolving introducing new challenges and possibilities for survey design. Questionnaire design elements that performed well in one survey setting might need adjustments depending on the characteristics of the target population, the design of the survey and new devices being introduced on the market. In the modern world of web surveys, smartphones and tablets are expected to receive a great deal of attention form survey designers. It is established that allowing smartphone access is necessary as participation from different devices increases response rates. However, devices affect data quality differently and researchers are facing a dilemma of the extent to which the design can be mobile optimized with minimal trade-offs in data comparability.

Measurement error related to device and survey design

Mobile optimization should be developing with the focus on how to reduce device effects on measurement error and produce a survey design with the most comparable data across modes. Several design solutions are currently used when designing web surveys: the unimode design approach, the adaptive mobile design and the device-agnostic design. The unimode design approach suggests keeping the design identical across modes (Dillman, 2007). In a web survey setting, a unimode approach could mean presenting non-optimised surveys on mobile devices. Previous research has found higher break-off rates, longer completion times and lower satisfaction with the survey for non-mobile-optimised surveys. On the other hand, this design avoids differences in measurement occurring due to design variations. The generalized mode design (de Leeuw, 2005) proposes utilizing the best features of each mode offering respondents the same stimuli but not necessary the same questionnaire design. In web surveys, the generalized mode design would suggest creating a mobile optimized survey alongside the main PC design. The downside of this approach is that differences in design may introduce different types of measurement error to each device, thereby reducing data comparability across devices. Finally, the device-agnostic approach suggests creating a universal web survey design focusing on the limitations of mobile devices (Alexandre, Carre, Leavy, & Leonick, 2016). The approach is similar to the unimode framework but it stresses the importance of creating a mobile-first design. The device-agnostic design should perform equally well on PCs, smartphones and tablets. Potentially, the device-agnostic design can avoid both the detrimental effects of non-optimised surveys on data quality of the unimode approach and differences in measurement error of the adaptive approach. However, there is scope for future web survey research to establish design features that perform comparably across all devices. The findings from chapter one suggest that the unimode approach for scale design indeed yields similar levels of measurement error across PC, smartphone and tablet responses. The caveat of such approach lies in the fact that while this design causes similar measurement error, it may not explore the best features of each device and produce the lowest level of measurement error possible. The experimental manipulation of the scale design demonstrated that certain vertical response scale designs may reduce straightlining on smartphones. These findings open a discussion on the best scale design that could satisfy both PC and mobile device requirements and lower the level of measurement error. More research is necessary testing the response scale design that would reduce the measurement error while remaining user-friendly on mobile devices and PCs.

Measurement error and respondents

In order to identify questionnaire design that produces the least measurement error, depending on the survey design, the questionnaire should be tested with different parts of the population, across different modes and on different devices.

To reduce respondent-induced measurement error, we need to understand how respondents process information and respond to design manipulations. The differences between respondents occur at every stage of data collection: willingness to respond, selection of the device, response strategies while taking a survey. Next, while responses are collected, we are still facing limitations to data quality imposed by cognitive limitations resulting in the adoption of response strategies such as satisficing and anchoring-and-adjustment.

Chapter two and three assessed the role of respondent personality and cognition on response process, participation and data quality. Each of the studies have relied on different psychological and cognitive scales to operationalize these characteristics. The second chapter used a short ten-item Big Five scale developed by Gosling et al. (2005) to assess respondent personality traits. This scale has performed comparably to other shorter versions of the original scale as well as the original Big Five scale and showed a good test-retest reliability. The third chapter used a fifteen-item scale developed by Lang et al. (2011). This scale was developed as a shorter version of a full scale and captured the personality traits similarly to the full Big Five scale. Both shorter versions of the Big Five scales are designed to assess the personality traits with the highest precision in the shortest time. Both scales perform better than their shorter equivalents as they are able to control for acquiescent responding and errors by implementing reverse-scored items. As both BIG Five scales are a reduced version of an extended scale they may fail to capture the finer personality facets, yet overall they have been proven to capture the core Big Five attributes in a comparable way to the main Big Five scale (Chiorri et al., 2014; Gosling et al., 2005; Lang et al., 2011)

Cognition measurement in the second chapter was operatonalised using a Cognitive Reflection test (Frederick, 2005). This scale involves basic calculation skills however the most difficult part of this scale is that it assesses users' self-reflection – the questions intended to capture whether respondents notice that a seemingly easy answer is not a correct one and come back to think about it more. The purpose of the CRT is to differentiate between more impulsive and more reflective decision-making. The CRT consists of three questions designed so that the intuitive spontaneous answer is wrong. Correct answers indicate a higher degree of reflective and deliberate thinking. This scale is a good measure caputring whether respondents were willing to spend more time on the questions ensuring the answers they give are a true reflection of their opinions.

The third chapter assesses cognition by using a combination of self-reported need for cognition scales and probabilistic literacy assessment. Individuals who score high on the need for cognition scale are more inclined to consider answers at a deeper level and reflect upon ideas than those who score low. Those with a higher need for cognition scores are more likely to excel at tasks that involve effortful thought (Cacioppo & Petty, 1982; Cacioppo, Petty, Feinstein, & Jarvis, 1996). The self-assessed need for cognition used as a standalone measure of cognition might be prone to bias arising either due to the social desirability effect or simply inattentive responding. Therefore, the cognition measure was further supplemented by the probabilistic numeracy scores. Hudomiet et al. (2017) have demonstrated that this scale is positively correlated with education and math skills and addresses respondent basic knowledge on probability and mathematical operations. In combination these two scales can be used for a more comprehensive assessment of the two distinct components of one's cognition – the innate curiosity and drive for complex thinking as well as ability for analytical thinking.

The experiment in chapter two demonstrated that Big Five personality traits are linked to the selection of particular responses. Presenting a 'forward'-ordered scale design could reduce the occurrence of high/positive responding among more conscientious respondents. Overall, the interaction of scale direction with respondent characteristics was relatively small and the conclusion was that if a survey is following good survey design practices, then the scale direction is not a source of poor data quality. The study did not observe a significant relationship between respondent CRT scores and scale direction effects. Future studies could capture more direct cognition measurements such as the results of cognitive tests (numeric and verbal reasoning) or exam scores in order to disentangle motivation and cognition components from the cognition assessment. On the other hand, similar testing should be performed with other questionnaire designs employing longer response scales, different types of response scales (agree/disagree or item-specific), different response scale orientation and layout (vertical or horizontal Likert-scales or question grids) or different types of questions (attitudinal, behavioural) to check the robustness of findings across different questionnaire design elements.

Chapter three demonstrated that respondents who are less motivated or with a lower need for cognition are more likely to use smartphones. This indicates that respondents who are not willing to participate and give best data quality may be more likely to use smartphones. As smartphone is a less optimal device to fill in longer or only partially-optimized web surveys it further reduces response quality. In order to negate some detrimental effects of mobile devices researchers should adjust the mobile survey design – present survey in several shorter sections, allow returning to the survey at a later time, making sure the questions use the least cognitively demanding design.

Future research should test the performance of several PC and mobile design solutions (question design, response scale design, and modular design) and observe whether data quality can be improved by varying the design elements even among less motivated, more cognitively burdened respondents. Understanding the role of respondent personality and cognition characteristics in respondent behaviour could inform the design solutions that would be most user-friendly and do not pose additional cognitive burden. While we cannot regulate characteristics of our respondents, we could work on the survey design that would be the least burdensome for a diverse group of respondents.

Smartphone data collection

Smartphone functionality as means of data collection will keep on changing and developing. Beyond administering web questionnaires, the additional functionalities of smartphones can offer innovative methods in data collection. Examples of data that could be obtained via mobile data collection are GPS location and tracking, accelerometer measurements, screen activity, physical motion and communication activity. Additionally, the possibility of constant connection to the internet, camera and portability expand the potential of mobile data collection. The last few years have seen a rapid development of mobile devices as additional data collection tools – surveys via text messages or WhatsApp, additional GPS data collection, mobile apps to record time diaries, tracking consumer behaviour or real-time response collection. Smartphones offer a vast array of additional data collection methods that should be explored in the near future.

When the ownership and proficiency with mobile devices become less of an issue respondents' trust in the institution collecting their data and their control over the data being collected might remain an important limiting factor for mobile participation. For example, respondents might need to be willing to install data collection app and consent to further sharing of data such as GPS location or uploading photographs. The imminent challenge of innovative data collection methods in the foreseeable future is the need to establish a trusting relationship between researchers and respondents to obtain potentially sensitive data.

Survey methodology is a stimulating discipline that is constantly adapting and evolving. Inclusion of mobile devices in the web mode has offered a range of new data collection possibilities. The findings presented in this thesis contribute to the practice of mixed-device web design and question answering process. The next steps in web survey design are to embrace smartphones in data collection process and reduce the measurement error by adapting the questionnaire to devices and respondents.

Future research

The constant evolution of survey design makes it an exciting field to work in. Since the early days of survey data collection methods, this field has been constantly developing and embracing technical advancements. I believe there are several directions that the field of web surveys will be developing in - further design improvements for mixed-device web surveys and the adoption of smartphones as a standalone tool for data collection. The fact that most of the web survey methodology is discussed in the context of mixed-device surveys makes a compelling case that smartphones are here to stay. Further support comes from the Dillman (2018) discussion on the future of mixed-mode research stressing the importance of smartphones in data collection in the 2020's. Currently, the general population is very mixed in their proficiency and desire to use mobile devices and higher smartphone use is linked to younger segments of the population. However, in the upcoming years, the share of general population proficient in internet and mobile device technology will keep on growing. I anticipate that survey researchers will see a decrease in the detrimental effects of coverage and mobile device proficiency on measurement error and non-response. Instead, greater attention will be paid to further integration of smartphones as a data collection tool.

The results from the three chapters can be summarized to provide the following recommendations for survey practitioners.

Differences in respondent characteristics can be used to improve response quality. First, pilot surveys should be tested on a wide range of respondents of different age, cognitive ability and personality traits. The process of determining a final survey design should follow a ‘principle of the lowest common denominator’ – the final survey design should perform well even among respondents with the lowest cognitive ability and those more susceptible to the cognitive fatigue. If the selected design has demonstrated an improved performance among more reluctant respondents, we can expect it to perform equally well with the rest of the survey population.

Studies involving respondent characteristics have demonstrated that characteristics such as cognition and personality can be a valuable if not a more important predictor of participation in surveys and quality of provided data than a basic set of demographics (i.e. gender, age, education, and ethnicity). Due to the differences between those who respond to the survey and those who do not, the final survey sample might differ from the general population on a range of personality traits. My suggestion is that once surveys are collected researchers might use data on respondent characteristics to construct weights for post collection adjustments.

Next, research on mixed-device surveys demonstrated that smartphones attract different type or respondents and smartphone surveys respond differently to the instrument design. The issue of comparability between data collected from mobile devices and PCs is resolved if surveys take a ‘device agnostic (mobile first)’ approach. As smartphones are anticipated to become a significant part of web survey data collection it is worth investing research in a design that performs equally well across devices. Yet I would like to stress the importance of targeting the smartphone device limitations first. Once identifying mobile-friendly designs they should be tested on PC and the design that produces comparable responses should be selected as a final device agnostic design. Next steps in survey design development will be developing and testing a survey design that performs well on smartphones in the first place and produces comparable responses on PCs.

References

Alexandre, F., Carre, C., Leavy, A., Leonick, J. (2016). Device agnostic: Why you need to transform now. Retrieved on November 27, 2018 from https://www.ipsos.com/sites/default/files/2016-11/Ipsos-Views-Device-Agnostic.pdf

Andreadis, I. (2015). Web surveys optimized for smartphones: Are there differences between computer and smartphone users? *Methods, data, analysis 9*(2), pp. 213-228.

Antoun, C. (2015). *Mobile web surveys: A first look at measurement, nonresponse, and coverage errors* (Doctoral dissertation). Retrieved from http: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/116722/antoun_1.pdf?sequence=1&isAllowed=y

Antoun, C., Couper, M. (2013). Mobile-mostly internet users and selection bias in traditional web surveys. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. Tucker, B. West (Eds.), *Total survey error in practice (1st Ed.)* (pp. 133 – 154). Hoboken, NJ: John Wiley & Sons, Inc.

Antoun, C., Couper, M., Conrad, F. (2017). Effects of mobile versus PC web on survey response quality. A crossover experiment in a probability web panel. *Public opinion quarterly*, *81*, 280-306.

Arn, B., Klug, S., Kołodziejski, J. (2015). Evaluation of an adapted design in a multi-device online Panel. *Methods, data, analyses*, *9*(2).

Barge, S., Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in higher education*, *53*(2), 182–200.

Bergman, O., Ellingsen, T., Johannesson, M., Svensson, C. (2010). Anchoring and cognitive ability. *Economics letters, 107*(1), 66–68.

Biemer, P. P., Lyberg, L. E. (2003). *Introduction to survey quality*. John Wiley & Sons.

Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L., Tucker, N., West, B. (2017). *Total survey error in practice* (1st ed). Hoboken, NJ: John Wiley & Sons, Inc.

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., Sudman, S. (1991). *Measurement error in surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., Couper, M. P. (2013). Sample composition discrepancies in different stages of a probability-based online panel. *Field methods*, *25*(4), 339-360.

Bosnjak, M., Metzger, G., Gräf, L. (2010). Understanding the willingness to participate in mobile surveys: Exploring the role of utilitarian, affective, hedonic, social, self-expressive, and trust-related factors. *Social science computer review, 28*(3), 350–370.

Buskirk, T.D., Andrus, C. (2012). Smart surveys for smart phones: exploring various approaches for conducting online mobile Surveys via smartphones. *Survey practice 5*(1), 1-11.

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.

Butt, S., Phillips, J. G. (2008). Personality and self-reported mobile phone use. *Computers in Human behavior, 24*, 346-360.

Cacioppo, J., Petty, R. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116—131.

Cacioppo, J., Petty, R. (1983). Effects of Need for Cognition on message evaluation recall and persuasion. *Journal of personality and social psychology, 45*(4), 805-818.

Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological bulletin*, *119*(2), 197.

Callegaro, M., Manfreda, K., Vehovar, V. (2015). *Web Survey Methodology*. London: Sage.

Chapman, G., Johnson, E. (1999). Anchoring, activation, and the construction of values. *Organizational behavior and human decision processes, 79*(2), 115–153.

Chiorri, C., Bracco, F., Piccinno, T., Modafferi, C., Battini, V. (2014). Psychometric Properties of a Revised Version of the Ten Item Personality Inventory. *European Journal of Psychological Assessment*, *1,* 1-11.

Christian, L., Dillman, D. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public opinion quarterly, 68*(1), 57–80.

Christian, L., Dillman, D., Smyth, D. (2007). Helping respondents get it right the first fime: the influence of words, symbols, and graphics in web surveys. *Public opinion quarterly*, *71*, 113 – 125.

Christian, L. M., Parsons, N. L., Dillman, D. A. (2009). *Designing scalar questions for web surveys.* Sociological methods & research, *37*(3), 393-425.

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

Couper, M. (2017). New developments in survey data collection. *Annual review of sociology, 43*, 121-145.

Couper, M., Antoun, C., Mavletova, A. (2017). Mobile web surveys. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. Tucker, B. West (Eds.)

Total survey error in practice (1st Ed.). (pp. 133 – 154). Hoboken, NJ: John Wiley & Sons, Inc.

Couper, M., Bosnjak, M. (2010). Internet surveys. In J. Wright & P. Madsen (Eds.), *Handbook of survey research* (2nd ed) (pp. 527-550). Bingley: Emerald Group Publishing Limited.

Couper, M., Kennedy, C., Conrad, F., Tourangeau, R. (2011). Designing input fields for non-narrative open-ended responses in web surveys. *Journal of official statistics*, *27*(1), 65-85.

Couper, M., Nicholls, W. (1998). The History and development of computer assisted survey information collection methods. In M. Couper, R. Baker, J. Martin, W. Nicholls, J. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 1-22). New York: John Wiley & Sons.

Couper, M., Peterson, G. (2016). Why do web surveys take longer on smartphones? *Social science computer review*, *35*(3), 357 – 377.

Couper, M., Tourangeau, R., Conrad, F., Crawford, F. (2004). What they see is what we get - response objects for web surveys. *Social science computer review*, *22*(1), 111 – 127.

Couper, M., Traugott, M., Lamias, M. (2000). Experiments on the design of Web surveys. *Paper presented at the Fifth International Conference on Social Science Methodology, Cologne, Germany, October 3–6, 2000.*

Courtright, M., Pashupati, K., Pettit, A., Knowles, R. (2013). Multimode, global scale usage: Understanding respondent scale usage across borders and devices. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. Tucker, B. West (Eds.)

Total survey error in practice (1st Ed.). (pp. 133 – 154). Hoboken, NJ: John Wiley & Sons, Inc.

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.

De Bruijne, M., Wijnant, A. (2014). Mobile response in web panels. *Social science computer review*, *32*(6), 728 – 742.

De Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of official statistics*, *21*, 233-255.

De Leeuw, E. (2008). Choosing the method of data collection. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 113–135). European Association of Methodology Series. New York: Lawrence Erlbaum Associates.

De Leeuw, E., Hox, J. (2011). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies (pp. 45-76.)*. Oxford: Taylor and Francis.

Dillman, D. A. (2007). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley & Sons

Dillman, D. (2018). *How web-push surveys are changing survey methodology.* Presented at NatCen/ESS/City University methods seminar, Oct. 2, London. City University of London

Dillman, D. A., Smyth, J. D., Christian, L. M. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method (3rd ed.).* John Wiley & Sons.

Donnellan, M. B., Lucas, R. E. (2008). Age differences in the Big Five across the life span: evidence from two national samples. *Psychology and aging*, *23*(3), 558-66.

Duggan, M., Smith, A. (2013). Cell internet use 2013. Retrieved in June 2017 from http://www.pewinternet.org/2013/09/16/cell-internet-use-2013/

Epley, N., Gilovich, T. (2006). The anchoring-and-adjustment heuristic: why the adjustments are insufficient. *Psychological science, 17*(4), 311-318.

Eroglu, C., Croxton, K. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International journal of forecasting, 26*(1), 116–133.

Eurostat. (2016). Retrieved July 9th, 2017 from http://ec.europa.eu/eurostat/documents/2995521/7771139/9-20122016-BP-EN.pdf/f023d81a-dce2-4959-93e3-8cc7082b6edd

Frederick, S. (2005). Cognitive reflection and decision making. *journal of economic perspectives, 19*(4), 25-42.

Gehlbach, H., Barge, S. (2012). Anchoring and adjusting in questionnaire responses. *Basic and applied social psychology, 34*(5), 417-433.

Gosling, S. D., Rentfrow, P. J., Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of research in personality, 37*, 504-528.

Groves, R., M. (1991). Measurement error across the disciplines. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, S. Sudman (Eds.) *Measurement error in surveys (pp. 1-25).* Hoboken, NJ: John Wiley & Sons, Inc.

Guidry, K.R. (2012). *Response quality and demographic characteristics of respondents using a mobile device on a web-based survey.* Paper presented at the AAPOR Annual

Conference, May 17-20, Orlando, USA. Retrieved May 20th, 2017 from: http://cpr.indiana.edu/uploads/Response%20Quality%20and%20Demographic%20Characteristics%20of%20Respondents.pdf

He, L., Yan, T., Keusch, F., Han, S. (2014). *The impact of question and scale characteristics on scale direction effect*. In Proccedings of the AAPOR 69th Annual Conference, Anaheim, California, May 15–18. Accessed August 2018 at: http://www.aapor.org/AAPOR_Main/media/AnnualMeetingProceedings/2014/Session-J-6-2-He-Y.pdf

Healey, B. (2007). Drop downs and scroll mice: The effect of response option format and input mechanism employed on data quality in web surveys. *Social science computer review*, *25*, 111–128.

Heerwegh, D., Looseveldt, G. (2002). An evaluation of the effect of response formats on data quality in web surveys. *Social science computer review*, *20*, 471-484.

Hern, A. (2015). *Smartphones now most popular way to browse Internet OfCom Report. The Guardian*. Retrieved from http://www.theguardian.com/technology/2015/aug/06/smartphones-most-popularway-tobrowse-internet-ofcom

Higher Education Funding Council for England. (2012). *National Student Survey: Analysis of National Results for 2011.* Retrieved on 5th of August 2014 from https://www.heacademy.ac.uk/sites/default/files/NSS_report_2011_FINAL.pdf

Higher Education Funding Council for England. (2014). *UK review of the provision of information about higher education: National Student Survey results and trends analysis 2005-2013.* Retrieved on 14th of June 2018 from http://dera.ioe.ac.uk/20480/1/HEFCE2014_13.pdf

Higher Education Statistics Agency. (2018). *Widening participation summary: UK Performance Indicators 2016/17*. Retrieved on November 27, 2018 from https://www.hesa.ac.uk/news/01-02-2018/widening-participation-summary

Höhne, J., K., Krebs, D. (2018). Scale direction effects in agree/disagree and item-specific questions: a comparison of question formats. *International journal of social research methodology, 21*(1), 91-103.

Höhne, J. K., Lenzner, T. (2015). Investigating response order effects in web surveys using eye tracking. *Psihologija*, *48*, 361–77.

Höhne, J. K., Revilla, M., Lenzner, T. (2018). Comparing the performance of agree/disagree and item-specific questions across PCs and smartphones. *Methodology: European journal of research methods for the behavioral and social sciences, 14*(3), 109-118.

Höhne, J. K., Schlosser, S., Krebs, D. (2017). Investigating cognitive effort and response quality of question formats in web surveys using paradata. *Field methods*, *29*(4), 365–382.

Holbrook, L., Krosnick, J., Moore, D., Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public opinion quarterly*, *71*, 325–48.

Hudomiet, P., Hurd, M., Rohwedder, S. (2017, January). *Probability numeracy: Measurement of heterogeneity in the quality of subjective probability data*. Paper presented at International Pension Workshop, Leiden, Netherlands.

Kaminska, O., McCutcheon, A., Billiet, J. (2010). Satisficing Among Reluctant Respondents in a Cross-National Context. *Public Opinion Quarterly*, *74*(5), 956–984.

Keusch, F., Yan, T. (2017). Web versus mobile web: an experimental study of device effects and self-selection effects. *Social science computer review*, *35*(6), 751–769.

Keusch, F., Yan, T., Han, S., and He, L. (2014, March). *Impact of response scale direction on survey responses in web and mobile web surveys.* Paper presented at the General Online Research conference, Cologne, Germany.

King, G., Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137–163.

Krebs, D., Hoffmeyer-Zlotnik, J. (2010). Positive first or negative first? effects of the order of answering categories on response behavior. *Methodology: european journal of research methods for the behavioral and social sciences, 6,* 118–127.

Krosnick, J. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology, 5*(3), 213-236.

Krosnick, J. (1999). Survey research. *Annual review of psychology*, *50*, 537-567.

Lang, F. R., John, D., Lüdtke, O., Schupp, J., Wagner, G. G. (2011). Short assessment of the Big Five: robust across survey methods except telephone interviewing. *Behavior research methods*, *43*(2), 548-67.

Levy, E., Cober, T., Norris-Watts, C. (2004). Specific personality measures. In M. Hersen (Ed.) *Comprehensive handbook of psychological assessment*, 4, (pp. 130-148). Hoboken, NJ: John Wiley & Sons, Inc.

Liu, M., Keusch, F. (2017). Effects of scale direction on response style of ordinal rating scales. *Journal of official statistics*, *33*, 137-154.

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*, *34*(1), 78 – 94.

Lugtig, P., Toepoel, V., Amin, A. (2016). Mobile-only web survey respondents. *Survey practice*, *9* (3).

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public opinion* quarterly, *72*(5), 914-934.

Maloshonok, N., Terentev, E. (2016). The impact of visual design and response formats on data quality in a web survey of MOOC students. *Computers in human behaviour, 62*, 506-515.

Mavletova, A. (2013). Data quality in pc and mobile web surveys. *Social science computer review, 31*(6), 725–743.

Mavletova, A., Couper, M. (2013). Sensitive topics in pc web and mobile web surveys: Is there a difference? *Survey research methods, 7*(3), 191–205.

Mavletova, A., Couper, M. (2014). Mobile web survey design: Scrolling versus paging, SMS versus e-mail invitations. *Journal of survey statistics and methodology, 2*(4), 498–518.

Mavletova, A., Couper, M. (2015). A meta-analysis of breakoff rates in mobile web surveys. In: Toninelli, D, Pinter, R & de Pedraza, P (eds.) *Mobile research methods: Opportunities and challenges of mobile research methodologies*, 81–98. London: Ubiquity Press.

Mavletova, A., Couper, M. (2016a). Device use in web surveys: the effect of differential incentives. *International journal of market research*, *58*(4), 523-544.

Mavletova, A., Couper, M. (2016b). Grouping of items in mobile web questionnaires. *Field methods, 28*(2), 170-193. First published online: July 26, 2015.

McElroy, T., Dowd, K. (2007). Susceptibility to anchoring effects: How openness-to-experience influences responses to anchoring cues. *Judgment and decision making, 2*, 48–53.

Oechssler, J., Roider, A., Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of economic behavior & organization, 72*(1), 147–152.

Ofcom (2017). *The communications market report*. Retrieved February 26, 2018 from https://www.ofcom.org.uk/__data/assets/pdf_file/0017/105074/cmr-2017-uk.pdf

Office for National Statistics (2012). *Statistical Bulletin: Internet Access - Households and Individuals*. Retrieved July 30, 2018, from http://webarchive.nationalarchives.gov.uk/20160109012432/http://www.ons.gov.uk/ons/dcp171778_275775.pdf

Office for National Statistics (2018). *Statistical Bulletin: Internet users, UK: 2018*. Retrieved July 30, 2018, from https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2018

Oppenheimer, D., Meyvis, T., Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology, 45*(4), 867-872.

Peterson, G., Mechling, J., LaFrance, J., Swinehart, J., Ham, G. (2013, March). *Solving the unintentional mobile challenge.* Paper presented at the CASRO Online Research Conference, San Francisco, USA.

Pew Research Center. (2015). *Technology device ownership: 2015.* Retrieved September 26, 2017, from http://www.pewinternet.org/files/2015/10/PI_2015-10-29_device-ownership_FINAL.pdf

Pew Research Center. (2016). *Smartphone ownership and internet usage continues to climb in emerging countries.* Retrieved June 25, 2017 from http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/

Pew Research Center. (2018). *Mobile Fact Sheet*. Retrieved May 7, 2018, from http://www.pewinternet.org/fact-sheet/mobile/Peytchev, A., and Hill, C.A. (2010). Experiments in Mobile Web Survey *Design. Social Science Computer Review, 28*(3): 319-335.

Peytchev, A., Hill, C. A. (2010). Experiments in mobile web survey design: Similarities to other modes and unique considerations. *Social science computer review, 28*, 319–335.

Rammstedt, B., Goldberg, L. R., Borg, I. (2010). The measurement equivalence of Big Five factor markers for persons with different levels of education. *Journal of research in personality*, *44*(4), 53–61.

Rammstedt, B., John, O. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *journal of research in personality, 41*(1), 203-212.

Rammstedt, B., Krebs, D. (2007). Does response scale format affect the answering of personality scales? Assessing the Big Five dimensions of personality with different response scales in a dependent sample. *European journal of psychological assessment*, *23*, 32–38.

Robinson, J., Shaver, P., Wrightsman, L. (Eds.) (1991). *Measures of personality and social psychological attitudes.* Vol. 1 of measures of social psychological attitudes series. San Diego, New York: Academic Press.

Robinson, J., Shaver, P., Wrightsman, L. (Eds.). (1999). *Measures of political attitudes.* Vol. 2 of measures of social psychological attitudes series. San Diego, London: Academic Press.

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

Sadowski, C. J., Cogburn, H. E. (1997). Need for cognition in the Big Five factor structure. *Journal of psychology*, *131*(3), 307-312.

Smyth, J., Dillman, D., Christian, L., Stern, M. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public opinion quarterly, 70*(1), 66–77.

Smyth, J., Dillman, D., Christian, L. (2012). Context effects in Internet surveys: New issues and evidence. *Oxford handbooks online.* retrieved on 10th of May 2016.

Sommer, J., Diedenhofen, B., Musch, J. (2017). Not to be considered harmful: Mobile-device users do not spoil data quality in web surveys. *Social science computer review*, *35*(3), 378–387.

Stanovich, K., West, R. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*, 672-695.

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

Struminskaya, B., Weyandt, K., Bosnjak, M. (2015). The effects of questionnaire completion using mobile devices on data quality. Evidence from a probability-based general population panel. *Methods, data, analyses*, *9,* 261–292.

Toepoel, V., Das, M., Soest, A. (2009). Design of web questionnaires: The effect of layout in rating scales. *Journal of official statistics, 25,* 509-528.

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.

Toepoel, V., Lugtig, P. (2015). Online surveys are mixed-device surveys. *Methods, data, analyses*, *9*(2).

Toepoel, V., Vis, C., Das, M., van Soest, A. (2009). Design of Web Questionnaires: An Information-Processing Perspective for the Effect of Response Categories. *Sociological Methods & Research*, *37*(3), 371–392.

Tourangeau, R., Couper, M., Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public opinion quarterly, 68*(3), 368-383.

Tourangeau, R., Couper, M., Conrad, F. (2007). Color, labels, and interpretive heuristics for response scales. *Public opinion quarterly 71*(1), *91–112.*

Tournageau, R., Maitland, A., Rivero, G., Sun, H., Williams, D., Yan, T. (2017). Web surveys by smartphone and tablets. Effects on survey responses. *Public opinion quarterly, 81*(4), 896–929.

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

Tversky, A., Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*, 1124-1130.

University of Bristol. (2016). *NSS Questions 2005-2016.* Retrieved on 14th of June, 2018 www.bristol.ac.uk/academic-quality/ug/nss/nssqs05-16.html/

Vaerenbergh, Y., Thomas, T. (2013). Response styles in survey research: A literature review of antecedents, consequences and remedies. *International journal of public opinion research*, *25*(2), 195-2017.

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.

Wells, T., Bailey, J., Link, M. (2012). Filling the void: Gaining a better understanding of tablet-based surveys. *Survey practice*, *6*(1), 1–9.

Weng, L., Cheng, C. (2000). Effects of response order on likert-type scales. *Educational and psychological measurement*, *60*(6), 908–24.

Yan, T., Keusch, F. (2015). The effects of the direction of rating scales on survey responses in a telephone survey. *Public opinion quarterly, 79*(1), 145-165.

Yan, T., Keusch, F., He, L. (2018). The impact of question and scale characteristics on scale direction effects. *Survey practice*, *11*(2), 1–10.