OPPORTUNITIES AND CHALLENGES IN NEW SURVEY DATA COLLECTION METHODS USING APPS AND IMAGES

BRENDAN READ

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN SURVEY METHODOLOGY.

INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH

UNIVERSITY OF ESSEX

JANUARY 2020

DECLARATIONS

No part of this thesis has been submitted for another degree and I am the sole author of the whole of this thesis.

Chapter one is published as:

> Read B. (2019) Respondent burden in a Mobile App: evidence from a shopping receipt scanning study. Survey Research Methods 13: 45-71.

An earlier version of chapter one is published as a working paper:

> Read B. (2018) Respondent burden in a mobile app: evidence from a shopping receipt scanning study. Understanding Society Working Paper 2018-04, Colchester: University of Essex.

A version of chapter two is published as a working paper:

> Read B. (2019) The influence of device characteristics on data collection using a Mobile App. Understanding Society Working Paper 2019-01, Colchester: University of Essex.

ACKNOWLEDGEMENTS

This research was supported by a +3 PhD studentship from the Economic and Social Research Council (ESRC) awarded to Brendan Read as an associated studentship funded from a grant awarded to Michaela Benzeval to conduct waves nine through eleven of the UKHLS [grant number ES/N00812X/1]. Data from the ninth wave of the *Understanding Society* Innovation Panel is used, which was funded by a grant awarded to Nick Buck and Michaela Benzeval to conduct waves six through eight of the UKHLS [grant number ES/K005146/1]. In addition, data from the *Understanding Society* Spending Study is used, which was funded by a research grant from the ESRC Transformative Research scheme and the National Centre for Research Methods (NCRM) awarded to Annette Jäckle [grant number ES/N006534/1].

My thanks must first go to my primary supervisor Annette Jäckle. Without her expert guidance I would not have been able to complete this challenge. In addition, her compassion helped to humanise the whole process and was equally as important as her expertise. Her approach to supervision is aspirational and I hope one day to be half the supervisor she is.

I would also like to thank Tarek Al Baghal in his role as my second supervisor. His input at key milestones throughout this process was invaluable. He offered an insightful second opinion and asked challenging questions, the answers to which often greatly improved this work. My thanks extend to Jon Burton, who stepped in as my second supervisor for a period, and whose mixture of humour and infallible knowledge of *Understanding Society* helped make this experience easier.

I feel incredibly fortunate to have been involved in two exciting and interesting research projects during my time at ISER. Aside from those already mentioned, my thanks go to the team members on those projects, including Thomas Crossley, Mick Couper, Paul

Fisher, Alessandra Gaia, Alex Wenz, Sandra Walzenbach and Carli Lessof. Their comments and insightful discussions of my work and the wider projects were extremely useful throughout.

I am grateful to all those in my personal life who have listened (often at great length) to me discussing my work. Whether they smiled and nodded and pretended to care about survey research or took a genuine and passionate interest in what I was working on, the opportunity to discuss my research with those outside the world of academia has been invaluable in helping me shape my thinking. My thanks in particular go to Jenna, who joined this journey in its closing stages, her companionship made this final push a far more enjoyable experience than it might otherwise have been.

Penultimately, I would like to thank my parents, who first instilled in me the thirst for learning and passion for searching for answers to difficult questions that ultimately set me on the path that has led me here. Their support and pride in what I do has been a continual source of strength throughout the process.

Finally, I would like to thank my examiners, Nick Allum, and Vera Toepoel for an interesting and insightful discussion of my thesis during my viva. Their suggestions for corrections have helped refine this work, and in particular the concluding section has benefitted greatly from the changes made based on their input.

~

SUMMARY

Surveys are well established as an effective way of collecting social science data. However, they may lack the detail, or not measure the concepts, necessary to answer a wide array of social science questions. Supplementing survey data with data from other sources offer opportunities to overcome this. The use of mobile technologies offers many such new opportunities for data collection. New types of data might be able to be collected, or it may be possible to collect existing data types in new and innovative ways. As well as these new opportunities, there are new challenges. Again, these can both be unique to mobile data collection, or existing data collection challenges that are altered by using mobile devices to collect the data.

The data used is from a study that makes use of an app for mobile devices to collect data about household spending, the *Understanding Society* Spending Study One. Participants were asked to report their spending by submitting a photo of a receipt, entering information about a purchase manually, or reporting that they had not spent anything that day.

Each substantive chapter offers a piece of research exploring a different challenge posed by this particular research context. Chapter one explores the challenge presented by respondent burden in the context of mobile data collection. Chapter two considers the challenge of device effects. Chapter three examines the challenge of coding large volumes of organic data. The thesis concludes by reflecting on how the lessons learnt throughout might inform survey practice moving forward. Whilst this research focuses on one particular application it is hoped that this serves as a microcosm for contributing to the discussion of the wider opportunities and challenges faced by survey research as a field moving forward.

~

CONTENTS

INTRODUCTION

This thesis is a product of what Robert Groves has called the third era of survey research (Groves, 2011). This is true in terms of the type of data examined, the tools and methods used to collect that data, and the approaches taken to analyse it. Groves describes this third era as being composed of two key developments. The first was a renewed interest in the use of self-completion modes, with the advent of web surveys. One consequence of this was a move towards mixed-mode surveys, combining survey data from different modes, often with distinctly different data generating processes. The second key development was a move towards researchers making use of so-called '*organic data*', data that has been created for non-research purposes. Supplementing data collected through questionnaires with organic data sources presents both new opportunities, as well as new challenges for survey researchers.

This shift in the survey research landscape reflects a wider shift in the past two decades in how societies understand and make use of data. This has led to what has been labelled '*the Age of Big Data*' (Lohr, 2012). Characterised by the increasing abundance and commodification of data, this new age has widely been discussed elsewhere in terms of the new opportunities it presents (e.g. Labrinidis and Jagadish, 2012; Lynch, 2008; Mayer-Schönberger and Cukier, 2013; McAfee et al., 2012; Sagiroglu and Sinanc, 2013). The amount of digital data in the world was estimated to have reached 33 Zettabytes (ZB) in 2018, which is 33 trillion GB of data, or the equivalent of 4,400 GB of data for every person on Earth (Reinsel et al., 2018).

However, Groves (2011) cautions against mistaking data for information. Whilst there are undoubtedly many opportunities presented by the exponential growth in the availability of digital data, there are significant challenges in obtaining useful information from this abundant supply of data.

When reflecting on the impact of this societal shift on the field of survey research, some early discussion went so far as to suggest the demise of social surveys in favour of new and innovative methods of social data collection (Savage and Burrows, 2007). A few years on, the consensus has seemed to settle on a future that harnesses a plurality of data sources, with surveys, administrative data, and Big Data all having a part to play (Groves, 2011; Couper, 2013a; Smith, 2013; Prewitt, 2013). This is perhaps the logical conclusion of the combination of survey research becoming increasingly multi-mode and the abundance of available organic data. In the intervening years, including those during which this thesis was written, survey research as a field has begun to explore the opportunities and challenges that are presented by supplementing data collected through questionnaires with that found in organic data sources.

One particular area of new developments in recent years has been the use of sensors in mobile devices to capture new types of data, or existing data types in new ways. Whilst smartphones are still not ubiquitous, penetration has continued to increase over time, rising from 17% in 2008 to 78% in 2018 (OFCOM, 2018). As smartphones become an increasingly important part of people's lives, the feasibility of making use of them for survey research also increases. As penetration rates continue to rise, concern about the digital divide between those with and without smartphones has waned. However, researchers wanting to make use of mobile devices for research purposes would be well served in being cautious of what has been called the '*second-level digital divide*' Hargittai (2002). As well as being concerned about who has a mobile device or not, it is important to consider the divisions in how different groups of people make use of their mobile devices in different ways.

Outside of survey research, there is a slightly longer history of making use of the sensors available in mobile devices for data collection, particularly within the field of computer science, where this has been explored through the concept of participatory sensing (Burke et al., 2006). Typically, this research has focused on considering the technical feasibility of capturing data using sensors (Bulusu et al., 2008; Sehgal et al., 2008; Deng and Cox, 2009; Ozarslan and Eren, 2014).

Within the field of survey research, many applications using smartphone sensors to collect survey data have been highlighted in a report compiled by the AAPOR Task Force on Emerging Technologies in Public Opinion Research (Link et al., 2014). One of the key survey design questions that emerges when trying to harness mobile devices to capture survey data is whether to equip sample members with devices, or ask them to participate using devices they already possess. There are a number of examples of both of these approaches. Several studies have equipped respondents with devices including wearable accelerometers (Gilbert et al., 2017; Scherpenzeel et al., 2018), and barcode scanners to use in their homes (Leicester, 2012). Examples of research that has asked respondents to use their own devices to participate in the study, include SurveyMotion (Höhne and Schlosser, 2019) and SurveyMaps (Schlosser et al., In progress), which provide general purpose Javascript libraries for capturing motion and geolocation data using smartphone sensors. Respondents have also been asked to use their mobile devices to take photographs of food that were coded to obtain nutritional information (Bruno and Silva Resende, 2017), and to model social networks amongst schoolchildren by making use of the Bluetooth capabilities of smartphones (Stopczynski et al., 2014).

This thesis investigates some of these new opportunities and challenges through the lens of a particular research project, the *Understanding Society* Spending Study 1 (SS1) (University of Essex. Institute for Social and Economic Research, 2018b). All three substantive chapters of this thesis make use of the data from the Spending Study. The study was part of a wider project with an overall aim of developing a better understanding of household finance through better measurement. The study was situated within the context of the *Understanding Society* Innovation Panel (IP). The IP forms part of the UK Household Longitudinal Survey (UKHLS), also known as *Understanding Society*. Both methodological and experimental testing are conducted using the IP, to inform the design of the main *Understanding Society*, and for the purposes of methodological research. The IP has a stratified and geographically clustered sample of households in Great Britain, south of the Caledonian Canal. All household members aged 16 years or older are eligible for annual interview. The IP User Guide (Jäckle et al., 2018a) documents the full details of the sample design for the IP.

Spending Study 1 took place between the ninth and tenth wave of the IP (IP9 & IP10) with all adult members (aged 16 and over) from households where at least one person responded at IP9 being invited to participate. The study consisted of an app in which respondents self-reported their spending, and a series of additional questionnaires asking respondents to reflect on the experience of participating. For a period of a month, sample members were asked to report on their spending by either: submitting pictures of shopping receipts for purchases they made, manually entering information about purchases, or reporting they had not made any purchases that day.

Chapter one explores the challenge presented by respondent burden in the context of mobile data collection. It considers how what we know about burden in a questionnaire context might apply to additional data collection tasks using a mobile device, and how it might differ. Survey research relies on the good grace of respondents and their cooperation when providing us with data. The results suggest that the subjective perception of burden (self-reports of how burdensome the task was) arose separately from the objective burden (the actual burden e.g. time taken to participate) respondents experienced in the Spending Study. These two facets of burden also differed in terms of how they changed throughout the course of the study. There was no systematic change in the subjective burden, but the objective burden decreased as the study continued. There was some limited evidence of the cumulative effect of objective burden on temporary dropout, but this was negligible. Finally, this chapter explores the factors that predict burden. Self-reported willingness to complete this kind of task was associated with shorter app use completion times, older respondents on average took longer to use the app, as did female respondents. Finally, those who reported being willing to use a camera to complete survey tasks were more likely to report their time and effort as being well spent.

Chapter two considers the challenge of device effects. It examines how the hardware specifications of the device used to participate in a mobile data collection task might affect the quality of the data produced. The results from this chapter provide evidence of device effects, with the size of the effects observed being similar in magnitude to those found in much of the interviewer effects literature. The device effects were most prominent when considering the image quality of the photographs of receipts. This perhaps indicates that device effects may be of greater concern in situations harnessing the sensor capabilities now available in mobile devices, as opposed to fielding a questionnaire on a mobile device. The operating system, whether a device was a tablet or smartphone, and the amount of RAM the device had access to were all associated with measures of data quality. There was also some evidence that controlling for respondent characteristics reduced the observed device effects, suggesting that some degree of the observed effects were attributable to selection.

Chapter three examines the challenge of coding large volumes of organic data. This chapter explored a range of automated techniques for transforming data from the scanned pictures of receipts into usable data. Descriptions of individual items purchased contained within receipts were too granular to be useful for most statistical analysis. Therefore, it was necessary to code these descriptions into broader categories. Doing this manually was a time consuming and expensive process. This chapter explores whether an automated approach produced acceptably accurate coding, whilst offering significant savings in terms of cost and time. Having previously manually coded the item descriptions it was possible to validate the automated coding. Different statistical or machine learning techniques were compared to assess their performance. The Random Forest algorithm and an enhanced string-matching algorithm performed best in terms of accuracy and ability to distinguish between categories. These two approaches also produced the smallest biases in terms of misclassifications of items into the wrong spending category. Two methods of improving upon the original automated coding processes were employed. The first was the use of ensemble models, which combined predictions from multiple models to produce one overall prediction; this did not improve performance compared to the best individual models. The second approach used probability thresholds to try to uncover those cases misclassified by the coding algorithms; this produced noticeable improvements in performance, with the two best performing algorithms having the most pronounced improvements.

Finally, this thesis concludes with a discussion of how the lessons learnt throughout the research in the substantive chapters might inform survey practice moving forward. Whilst this research focuses on one particular application it is hoped that this serves as a microcosm for contributing to the discussion of the wider opportunities and challenges faced by survey research as a field moving forward. As such, the discussion also turns to laying the groundwork for future research that may also help contribute to this growing discourse.

~

CHAPTER ONE

RESPONDENT BURDEN IN A MOBILE APP: EVIDENCE FROM A SHOPPING RECEIPT SCANNING STUDY

Abstract

This paper considers the burden placed on participants, subjectively and objectively, when asked to use a mobile app to scan shopping receipts. Using data from both the *Understanding Society* Spending Study, and the ninth wave of the *Understanding Society* Innovation Panel allow measures of burden and related characteristics to be identified. Subjective and objective burden were found to be seemingly unrelated to one another. There is evidence of older respondents facing greater objective burden, however there was some evidence that this did not correspond to an increase in the levels of subjective burden reported. Reported willingness to participate in a task of a similar nature proved to be indicative of both objective and subjective burden.

INTRODUCTION

A number of benefits of using mobile technologies to collect survey data have been highlighted. Chief among these is the ability to collect a range of new data including: '*voice, photography, video, text, email* [and] *GPS*' (Link et al., 2014: 22), to augment survey data. This paper focuses on one such new opportunity: using an app for mobile devices to facilitate the collection of scanned receipts. However, the concepts considered, and findings presented, in this research are also equally applicable to other contexts. This does not just include related tasks involving photography, such as barcode scanning, but also a wider array of event-based supplementary data collection tasks such as time-use diaries, tracking of health behaviours, capture of visual data, and '*in-the-moment*' survey data collection.

Along with the new data collection opportunities offered by these new technologies, it is also important to consider the potential challenges they present. These could be challenges unique to data collection using a mobile device or app, or existing survey data collection challenges altered by the new context. This paper focuses on one such challenge, respondent burden. Historically, there have long been concerns about the demands surveys place upon respondents and how this may affect the quality of the data collected (Ruch, 1941; Young and Schmid, 1956). More recently, such concerns have been conceptualised as respondent burden (Bradburn, 1978).

Burden is expressed as consisting of two dimensions: objective burden, the '*total time and financial resources expended by the survey respondent to generate, maintain, retain, and provide survey information*' (Office of Management and Budget, 2006: 34); and subjective burden, '*the degree to which a survey respondent perceives participation in a survey research project as difficult, time consuming, or emotionally stressful*' (Graf, 2008: 740). Both dimensions, and the relationship between them, are of interest in this paper.

The data collection task that is the focus of this paper is the *Understanding Society* Spending Study One. Participants were asked to use an app every day for one month to scan shopping receipts, submit purchases made without obtaining a receipt, or report days without spending. Data from the app, accompanying debrief questionnaires, and wave nine of the *Understanding Society* Innovation Panel are used to examine the following research questions:

RQ1. Are subjective and objective measures of burden related?

RQ2. How do subjective and objective burden change over the course of the study?

RQ3. Does objective burden predict breaks in participation?

RQ4. What factors predict subjective and objective burden?

BACKGROUND

RECEIPT AND UPC SCANNING

The potential benefits of Universal Product Codes (UPCs) and Electronic Point of Sales (EPOS) systems for the collection of survey data on purchasing behaviours was recognised swiftly following their widespread adoption in the 1980s (McGloughlin, 1983). Both UPCs and till receipts were identified as sources of data on spending which could potentially overcome the underreporting and misreporting that were observed in earlier consumer surveys and diary studies (Sudman, 1964a; Sudman, 1964b; Marr, 1971).

Some of the earliest attempts to capture these new sources of data involved studies situated within supermarket stores, with respondents identifying themselves at the point of purchase to allow the records of their purchases to be attributed to them (Bucklin and Gupta, 1999; Gupta et al., 1996; Guadagni and Little, 1983; Van Heerde et al., 2000; McGloughlin, 1983).

Subsequently, some of these studies evolved to in-home scanning panels, with respondents provided with a device specifically for the purpose of scanning the UPCs on the products they purchased. These panels have typically been formed within the realm of commercial market research. Among the most prominent of these studies are the National Consumer Panel in the US (formerly Nielsen HomeScan) from which a number of pieces of research have emerged (e.g. Aguiar and Hurst, 2007; Einav et al., 2008; Harris, 2005). Similarly, Kantar Worldpanel (formerly TNS Worldpanel) have conducted a number of studies worldwide, including the most prominent example of such a panel in the UK, the data from which has been used for several pieces of academic research (e.g. Griffith et al., 2009; Leicester and Oldfield, 2009a; Leicester and Oldfield, 2009b).

Capturing data from till receipts usually involves respondents collating their receipts and providing them to the research organisation. Respondents are asked to submit them through the mail, or by providing them to an interviewer who would come to their home to collect them. Examples of research making use of till receipts can be found in both economics (Hendershott et al., 2012; Inman et al., 2009; Stilley et al., 2010; Inman and Winer, 1998) and health (Appelhans et al., 2017; Biediger-Friedman et al., 2016; Chrisinger et al., 2018; Cullen et al., 2007; Greenwood et al., 2006; Rankin et al., 1998; Ransley et al., 2003; Martin et al., 2006; Waterlander et al., 2013).

More recently, the potential for using mobile devices to aid the capture of these kinds of data sources has been recognised. A body of research conducted by researchers at Nielsen (Scagnelli et al., 2012; Scagnelli and Bristol, 2014)has examined the feasibility of UPC scanning using a smartphone app. Their study invited millennials (aged 18-29) to participate and provided them with an Android phone with a data plan to participate. Similarly, (Volkova et al., 2016) have developed an app for use in randomized controlled trials, that also makes use of mobile devices for scanning UPCs. In parallel to this, within the field of computer science, the concept of participatory sensing has emerged, which imagines mobile devices as a distributed network of sensors, that through the participation of their users, can be harnessed for large scale data collection (Burke et al., 2006). Much of this emerging literature has focused on the technical feasibility of different use cases for these technologies. As such, working examples of mobile apps to collect both UPCs (Deng and Cox, 2009) and receipts (Bulusu et al., 2008; Sehgal et al., 2008; Ozarslan and Eren, 2014) have been developed. It is believed that the *Understanding Society* Spending Study One, the data collection task analysed in this research, is the first example of a receipts scanning task using a mobile app situated within the context of a nationally representative probability sample.

RESPONDENT BURDEN

Respondent burden has traditionally been examined within the context of traditional survey data collection using questionnaires. The existing body of literature is drawn together here to provide a conceptual account of burden. Throughout an attempt is made to apply these concepts to the kind of task that makes up the *Understanding Society* Spending Study. This conceptual framework of burden can similarly be applied to other new forms of data collection using mobile devices.

The exact relationship between objective (also called actual) and subjective (also called perceived) burden has not always been clearly established. Bradburn, in his seminal discussion of respondent burden, suggested that '*"burdensomeness" is not to be an objective characteristic of the task, but is the product of an interaction between the nature of the task and the way in which it is perceived by the respondent*' (Bradburn, 1978: 49). This acknowledges the importance of the nature of the task, an objective set of features, but suggests its importance comes from how it shapes subjective perception. More recent accounts have made the case for considering both the objective and subjective dimensions of burden (Ampt, 2003; Willeboordse, 1997). By considering both dimensions it is possible to acknowledge the role of objective burden in shaping subjective burden, whilst also considering objective burden in its own right, if for no other reason than the factors determining objective burden are likely to be more easily controllable by the survey practitioner.

Evidence for the relationship between subjective and objective measures of burden has been mixed. Dale and Haraldsen (2005) report a high correlation between subjective and objective measures of burden. However, in this study the objective measure (how long it took to complete the survey) relies on self-reports and therefore it is not surprising that it correlates with other subjective measures.

Sharp and Frankel (1983) examined the relationship between a wider selection of measures of subjective and objective burden. They experimentally varied the objective length of the survey and the level of effort necessary to complete the survey. In addition, measures of objective burden including item refusal and nonresponse rates were collected. Subjective burden was captured through self-reports of willingness to be re-interviewed, willingness to participate for longer, interest in the study, judgement as to important the study was, difficulty, whether time and effort was well spent, belief that the interview was the right length. The evidence suggested that a longer survey resulted in greater reports of subjective burden on the indicators related to length. However, there was little evidence of relationships between the other measures of burden examined.

Yu et al. (2015) attempted to disentangle the subjective from the objective by experimentally varying the actual length of a survey, and the presentation of that length, so as to examine whether separate effects of both increased objective burden and increased subjective burden could be observed. They found that not only did increasing the objective length of the survey increase the levels of reported burden, but presenting the survey as longer and more burdensome also further increased the levels of reported burden.

FACTORS DETERMINING BURDEN

Bradburn (1978) identified four survey characteristics that determine burden: survey length, the amount of effort required to complete the survey, the amount of emotional stress caused, and the frequency of interviewing. Haraldsen (2004) suggested three respondent characteristics as factors determining burden: the respondent's competence/ability, their interest/motivation, and their availability/opportunity to complete the task.

Such a dichotomy into survey and respondent characteristics is somewhat misleading. This is because it suggests that the seven factors identified are solely influenced by either design choices, or the nature of a respondent. Instead the case can be made that each of these seven factors is determined by characteristics of both the survey and the respondent. For example, how long a survey takes to complete is both determined by the amount of content specified, and the variance in the length of time individuals take to respond.

Therefore, in this paper, the approach of combining the list of four factors suggested by Bradburn with those suggested by Haraldsen is taken, resulting in one list of seven factors that contribute to respondent burden. Most of these factors has been discussed in the existing survey methodological literature. Where links to these seven factors have been discussed in the existing literature on receipt/UPC data collection, or mobile data collection more broadly these links are highlighted.

Length. Presenting information that suggests a longer survey to respondents has been found to have a negative impact on response rates in web surveys (Crawford et al., 2001; Galesic and Bosnjak, 2009), telephone surveys (Collins et al., 1988; Roberts et al., 2010), face-to-face surveys (Groves et al., 1999), and postal surveys (Yammarino et al., 1991; Dillman et al., 1993).

However, when it comes to the actual time taken to complete a survey there is some evidence that those with the longest response times may be those individuals who have engaged the most with the topic of the survey, and for whom that topic is particularly relevant (Branden et al., 1995). Similarly, those respondents with the longest response times in a given wave of a panel study have been found to be more likely to respond in subsequent waves (Lynn, 2014). In repeated measures studies it has also been found that respondents perceptions of task durations may not map very well onto the true durations of those tasks (Lee and Waite, 2005; Scagnelli et al., 2012).

Effort. Couper and Nicholls (1998) express concern that the shift from paper or interviewer-based modes to web modes of data collection may result in respondents having to expend more effort to participate. This is because some of the tasks traditionally performed by the data collector are instead coming to be performed by the respondent. This shift, whilst potentially beneficial in terms of reducing costs, or potentially reducing processing errors, comes at the cost of increasing the burden placed upon the respondent. As was noted earlier, data collection involving receipts has typically required the respondent simply to collect their paper receipts, with the data processing being performed by the survey organisation. By asking respondents to take and upload pictures of their receipts, more effort is needed on the part of the respondents in order to participate.

Emotional stress. Typically, research into the emotional stress caused by surveys has looked at the effect of sensitive questions on specific vulnerable populations. For example, emotional stress has been found to make participation harder in surveys on: sexual and physical violence among adults(Walker et al., 1997), bereavement (Dyregrov, 2004), and traumatic injuries (Ruzek and Zatzick, 2000). There has also been some evidence of question sensitivity as a barrier to participation amongst subgroups in general population surveys (Newman et al., 2001; Galea et al., 2005), though the characteristics of the affected subgroups identified have not always been clear. Kreuter et al. (2008) found that questions were more likely to be sensitive for respondents who belonged to groups with a sensitive status related to the concept being measured. This seems to support the idea that the amount of emotional stress caused by a survey instrument is not simply an innate characteristic of that given instrument, but it also shaped by the characteristics of the respondent receiving that instrument. As such, a given survey instrument may potentially be more stressful and thus produce higher burden for some individuals or subgroups of a sample as opposed to others.

It has been suggested that collecting receipts offers a less sensitive form of collection for data on consumption (Martin et al., 2006), with reduced risk of social desirability bias. However, it does not appear that this has been empirically tested.

Frequency. In Bradburn's (1978) original discussion of burden frequency is discussed in terms of the number of surveys by different organisations that any given individual would be invited to participate in. More surveys resulted in a greater burden, with discussion of how this burden may be split amongst a population (for an example of a discussion of how to ensure this distribution of burden in reference to business surveys see Oomens and Timmermans, 2008). However, it is also possible to consider the impact of the frequency of response when considering a study involving a series of repeated measures, as is the case in this research. Here it is possible to draw upon literature regarding the Experience Sampling Method (ESM) (Larson and Csikszentmihalyi, 1983). Csikszentmihalyi and Larson (2014) report that respondents quickly adopted ESM reporting as a habitual behaviour, and frequency of reports did not differ throughout the course of a study. They did however report different frequencies with which different subgroups of the general population would respond, with less educated and lower skilled individuals being less compliant and therefore responding less.

Availability/Opportunity. The finite amount of time available to respondents means that they must make a decision as to whether to spend their time participating. Framing this through the lens of traditional economic thought surrounding issues of resource scarcity (drawing upon Raiklin and Uyar, 1996), participation in the survey comes at the opportunity cost of not using their time for other activities. This cost is most sharply felt where time is a scarce resource. Previous research considering time constraints as a barrier to participation have found evidence to suggest that those who are more likely to have time constraints have a lower propensity to respond (Groves and Couper, 1998; Abraham et al., 2006).

Another important factor when considering the opportunity to participate for studies using mobile devices is whether a sample member has access to a device with which to take part in the study. Where a sample member does not have access to a mobile device, the objective burden of participating is clearly higher, as they must have the opportunity to either borrow or otherwise acquire access to a device to allow participation. The act of having to borrow a device likely increases the level of effort necessary to participate. Whilst a respondent may have the opportunity to gain access to a device, repeatedly acquiring that access may be considered too much effort, meaning the participant chooses either to participate less, or not at all.

Finally, a respondent's opportunity to participate may be broken up by distractions. A number of studies have examined the presence of distractions for respondents completed web questionnaires (Ansolabehere and Schaffner, 2015; Sendelbah et al., 2016; Zwarun and Hall, 2014). However, it has been suggested that the degree to which these distractions impact upon data quality is minimal (Ansolabehere and Schaffner, 2015). There is also some evidence to suggest that distractions are part of deliberate multi-tasking, and therefore may be embedded in respondent's web use behaviour, meaning a certain level of distraction may be necessary for respondents to be comfortable participating (Zwarun and Hall, 2014).

Ability/Competence. Lower cognitive ability has been highlighted as a widely accepted cause of measurement error (Fricker and Tourangeau, 2010). Lower cognitive ability may result in greater difficulty completing a task, thus increasing the burden. Satisficing describes a response strategy where respondents attempt to reduce the burden of participation by producing sub-optimal (in the eyes of the survey practitioner) responses. Lower cognitive ability has been found to increase the likelihood of a respondent satisficing (Krosnick, 1991; Knäuper et al., 1997). Lower device familiarity, or lower ability to complete survey tasks on a mobile device, has also been considered as that may act as a barrier to participation (Jäckle et al., 2019a). This may affect both the subjective burden, as sample members evaluate their ability to perform the task, and the objective burden, how well respondents are able to perform the task.

Motivation/Interest. One factor affecting a respondent's motivation is the topic or subject matter of the survey they are asked to complete. When being approached with a survey request, evidence suggests that if that request is related to a topic in which the respondent has been observed to have an interest, their propensity to respond will be increased (Groves et al., 2004). Conversely, a lack of interest has been found to result in a lower propensity to respond (Couper, 1997). The consensus is that the use of incentives helps to motivate respondents, and improve the rate of participation (Armstrong, 1975; Singer et al., 1999). Typically, unconditional incentives have been found to be better motivators than conditional incentives (Church, 1993; Goyder, 1994; Young et al., 2015). However, there is evidence of a so-called '*ceiling effect*' when using incentives to promote response, with the impact of incentives being diminished when respondents are already motivated to take part in a survey (Groves et al., 2000; Zagorsky and Rhoton, 2008).

For mobile surveys there has been recent interest in increasing motivation to participate through the gamification of surveys (for a summary see Keusch and Yan, 2017). A number of different approaches have been suggested, ranging from gamified question wording (Henning, 2012), borrowing elements of gamified app design, such as achievement badges for use in surveys (Lai et al., 2012; Link et al., 2012), through to games specifically designed for data collection (Adamou, 2013). There is some evidence to suggest that gamified survey designs can reduce burden in mobile surveys, at least amongst a sample of children (Mavletova, 2015).

DYNAMIC BURDEN

Burden has typically been considered as static, either as the perceived burden before beginning a survey, or the total objective burden that is experienced by fully completing a questionnaire. Existing conceptual understandings of drop out of diary studies, or break-off in web-surveys offer insight into how burden may be considered a dynamic concept throughout the duration of a data collection task.

Accounts of break-off in web surveys have suggested participants go through an ongoing decision-making process about whether to continue participating in a survey (Galesic, 2006; Haraldsen, 2004; Peytchev, 2009). Some of these analyses draw upon *decision field theory*, developed by (Busemeyer and Townsend, 1993), which describes a dynamic decision-making process. One of the key aspects of decision field theory is the notion of an inhibitory threshold: '*the point which determines when the difference in the preference for one or the other action is large enough to provoke behaviour*' (Galesic, 2006: 314). When respondents fall below this inhibitory threshold, they shift from making the decision to participate to making a decision to stop participating.

In contrast, it has been suggested that drop out in diary studies results from cumulative fatigue (Gillmore et al., 2001). Fatigue builds throughout participation and can therefore only increase as time goes on. Evidence of fatigue, as measured by a decrease in responding throughout the course of a diary study, has been mixed. There are examples of studies in which respondents show evidence of becoming fatigued (Gerstel et al., 1980; Leigh, 1993; Verbrugge, 1980) and some studies in which the effect does not seem to be present (Lemmens et al., 1988; Persky et al., 1981; Searles et al., 1995). Gillmore et al. (2001) suggest that both respondent and design characteristics may play a role in determining whether respondents become fatigued in a diary study. However, their attempts to identify examples of specific characteristics that contribute to fatigue were not able to provide much insight.

Both subjective and objective burden can then be considered in a discrete and cumulative manner. In the case of objective burden, it is felt that this more closely resembles the concept of cumulative fatigue as described in the diary studies literature. Discrete objective burden is the amount of burden each individual task within the study places on the respondent. This may differ from task to task, or even across repeat performances of the same task, due to factors such as the nature of the task, the situational context, or characteristics of the respondent. Cumulative objective burden then consists of the summed total of all episodes of discrete objective burden up to any given point in the study.

In terms of subjective burden, the conceptual model presented here is close to the one offered by decision field theory. When considering subjective burden in a discrete manner this is the disposition of the respondent when considering whether to complete each individual task that makes up a given study. In line with decision field theory, a respondent may be above or below the inhibitory threshold for participating, and this may differ from task to task, different tasks might be perceived as more or less than burdensome, or the same task at different points in the study might produce different perceptions of burden. Cumulative subjective burden then would then not be the summative concept presented by cumulative objective burden. Instead cumulative subjective burden should be considered as the trend in discrete perceptions, this might be a monotonic increase or decrease in perceived burden over time, or it might follow a non-monotonic pattern, with peaks and troughs in the level of perceived burden throughout the study.

DATA

STUDY DESIGNS

This research uses data from both wave nine of the *Understanding Society* Innovation Panel (IP9) and an inter-wave receipt scanning project: the *Understanding Society* Spending Study 1, which took place between waves nine and ten of the Innovation Panel (IP). The main variables of interest are taken from the Spending Study, with variables from IP9 used as covariates for some of the analyses.

Innovation Panel. The Innovation Panel (University of Essex. Institute for Social and Economic Research, 2018a) is one part of the UK Household Longitudinal Study, *Understanding Society*. The IP exists to allow the implementation of experiments and research into issues of data collection procedures within the context of longitudinal surveys. The sample design is a stratified, clustered sample of all households within Great Britain, south of the Caledonian Canal. The ninth wave contains the original sample along with refreshment samples from waves four and seven onwards. All household members aged sixteen and over at the time of interviewing are considered eligible for annual interviews. The data used in this paper come from the ninth wave which had a household response rate of 84.7% and an individual response rate of 85.4% within responding households (Jäckle et al., 2019a).

***Understanding Society* Spending Study.** The *Understanding Society* Spending Study (University of Essex. Institute for Social and Economic Research, 2018b) is part of a project to give a better account of household finances by developing innovative methods of collecting data on this topic. The study was conducted in partnership with Kantar Worldpanel, who developed the app. Respondents were tasked with downloading and using an app on their smartphone or tablet, to provide data about their spending across the span of a month. Spending could be reported by scanning receipts, inputting a purchase without a receipt, or reporting a day in which nothing was spent. Full details of the design of the study, including the full questionnaires and app text, can be found in the User Guide (Jäckle et al., 2018b). Screenshots for the app are documented in the separate Appendix C of the User Guide.

The issued sample for Spending Study 1 consisted of all adult members (aged 16 or over) of households where at least one person in the household responded at IP9. Household members who are known to have refused to participate long-term in the Innovation Panel were not included in the Spending Study sample.

Alongside the data collected via the app, the Spending Study also asked participants to complete several additional questionnaires, with questions regarding the experience of participating and some additional questions about their household expenditure. End of week surveys asked participants to reflect on the previous week's participation. An end of project questionnaire asked participants to reflect on the experience of participating as a whole. The end of project questionnaire was first implemented as an online survey, before a paper follow-up was sent out to those who had not initially responded to the online version.

Different incentive amounts for different forms of participation in the study were offered to participants, with the incentives being made available in the form of either Love2Shop gift vouchers or gift cards. These are redeemable in many high-street stores throughout the UK. There was an initial incentive for completing a registration survey and downloading an app with two randomised conditions (£2 vs £6). All members of a given household received the same incentive treatment. Secondly, in an effort to further increase the rate of response, an additional £5 incentive was sent to members of a random half of all households where no-one had participated by the third week of the study. These first two incentives are included as covariates in the analyses presented here. In addition, participants received a 50p a day incentive for every day in which they used the app. Completion of each end of week survey earned a further 50p, and completing the end of project survey earned £3. Finally, a bonus of £10 was offered if a participant used the app every day for 31 days. Ultimately, this requirement was relaxed so that all participants who used the app on at least 27 days throughout the study received this bonus. Participants were sent an email at the end of each week updating them on how much they had earned in incentives so far.

ANALYTICAL SAMPLE

To allow covariates from IP9 to be used in the analyses in this paper only the 2,112 sample members who completed a full adult interview at IP9 were considered for the analytical sample. Of these IP9 respondents, 270 attempted to use the app, with 268 successfully completing at least one app use, a response rate of 12.7%. This paper focuses only on these participants and does not present analyses examining those who did not participate in the study. Jäckle et al. (2019a) examined participation in the Spending Study, and some of their findings, together with consideration of some of the implications of examining burden amongst participants can be found in the discussion section of this paper.

Of the 268 app users, 238 responded to the end of project survey (88.8%). As the subjective measures of burden were asked in the end of project survey the analytical sample for this paper is constrained to just those participants who completed this survey. Due to an error in the scripting of the web version of the end of project survey, fourteen participants who completed the end of project survey did not receive the subjective burden questions. These fourteen cases were individuals who had not participated in the final week of the study and were allocated to receive questions about why they had dropped out. Instead these participants received a version of the questionnaire intended for non-participants, thus they were not asked any of the questions reflecting back on the experience of participating. This left 224 cases who received the subjective burden questions. Of the 224 cases, a single participant did not answer all of subjective burden questions, and was subsequently dropped from the analyses, leaving a final analytical sample of 223. This constitutes 10.5% of the issued sample and 83.2% of participants in the Spending Study.

The analyses presented here are constrained to the analytical sample, though those analyses which only examined objective measures of burden, were repeated with all 268 app users. The differences between the two specifications were for the most part minimal, with any notable differences highlighted throughout the results section of this paper. Table 1 documents the response rates at different stages of the study, and the analytical sample.

The average number of end of week surveys completed each week was around 136 out of the 223 analytical sample members. This was about 60% of the analytical sample. A breakdown of the number of end of week surveys that participants completed is in Table A1 in Appendix A. That a relatively large portion of participants did not complete the end of week surveys is in line with previous research that found that hypothetical willingness to complete additional questions alongside a data collection task using a mobile device was generally low (Keusch et al., 2017).

Table 1
Breakdown of response rates for different stages of the Understanding Society Spending Study 1.

	n	% of sample	% of participants	% of analytical sample
Issued sample	2112	100.0		
Completed at least one app use	268	12.7	100.0	
Completed end of project survey	238	11.3	88.8	
Received subjective burden questions	224	10.6	83.6	
Analytical sample	223	10.5	83.2	100.0
Completed end of week surveys				
Week one	134	6.3	50.0	60.1
Week two	132	6.2	49.3	59.2
Week three	139	6.6	51.9	62.3
Week four	137	6.5	51.1	61.4

The total number of app uses for the analytical sample of 223 participants was 10,381. There was some concern that a number of extremely long or short app uses may represent outliers. Due to the potential bias these extreme results may have introduced the decision was made to identify potential outliers and remove them from the analytical sample. Outliers were classified as those outside of the interval of a boxplot as defined by (Tukey, 1977). To adjust for the skewed distribution the approach advocated by (Hubert and Vandervieren, 2008) was taken, which uses the medcouple

(Brys et al., 2004), a robust measure of skewness, to adjust the boxplot for skewed distributions. The medcouple was estimated using the Stata package '*medcouple*' (Gelade et al., 2013). App uses that took less than 3 seconds, or more than 173 seconds were classified as outliers. These app uses were then excluded from the analysis leaving 10,029 app uses that were included in the analyses presented here.

Table 2 reports the breakdown of app uses by type of app use, and by type of mobile device used to complete the app use. Nearly half of app uses were scanned receipts, with around thirty percent being purchases submitted without a receipt, and twenty percent being reports of nothing bought. The majority of app uses were completed on smartphones as opposed to tablets (83.7% compared to 16.3%).

Table 2
Number of app uses completed by type of app use and type of mobile device.

	n	% by device type	% of total app uses
Smartphone			
App uses	8395	100.0	83.7
Receipts scanned	4012	47.8	40.0
Purchases without a receipt	2517	30.0	25.1
Nothing bought	1866	22.2	18.6
Tablet			
App uses	1634	100.0	16.3
Receipts scanned	860	52.6	8.6
Purchases without a receipt	424	26.0	4.2
Nothing bought	350	21.4	3.5
All app uses			
App uses	10029		100.0
Receipts scanned	4872		48.6
Purchases without a receipt	2941		29.3
Nothing bought	2216		22.1

MEASURES

MEASURES OF BURDEN

Objective measures of burden. Four measures of objective burden were derived from paradata collected by the app: the number of app uses each participant completed, the total time they spent completing these app uses, their average time per app use, and the durations of the individual app uses. The first two of these measures capture the total cumulative burden of individuals across the course of the whole study. The latter two instead attempt to measures the amount of objective burden per app use. The first three measures are measured at the participant level, the fourth is captured at the app use level. The assumption here is that a longer period of time or more app uses equals a greater objective burden placed upon the participant. Descriptive statistics for these four measures, both broken down by type of app use, and pooled across all types of app use are presented in Table 3.

The mean number of app uses completed by an individual was 45, which is about one or two app uses per day throughout the course of the study. The mean time to complete an individual app use was 31 seconds. The grand mean of the mean time taken by each respondent to complete their app uses was 31 seconds. The mean total time taken by an individual to complete all their app uses was 1,403 seconds, this equates to a little over 23 minutes throughout the course of the study. Descriptive statistics for app use duration for the two types of device used to complete the app use are provided for reference. The impact of device is not considered in the analyses presented here, though some consideration is given as to the impact of device effects in the discussion section.

Table 3

Descriptive statistics for the four measures of objective burden.

	$\bar{x}$	SD	Q_1	Q_2	Q_3
Number of app uses completed by each participant					
All app uses	45	20	33	42	55
Receipts scanned	22	18	8	18	30
Purchases without a receipt	13	12	3	10	19
Nothing bought	10	8	4	8	15
Average duration of app uses for participants (seconds)					
All app uses	31	11	23	30	37
Receipts scanned	45	18	33	42	54
Purchases without a receipt	34	16	23	29	40
Nothing bought	11	7	7	9	13
Total duration of app uses for participants (seconds)					
All app uses	1403	820	812	1266	1884
Receipts scanned	980	684	471	841	1374
Purchases without a receipt	444	347	194	365	619
Nothing bought	100	76	43	85	139
Duration of each app use (seconds)					
All app uses	31	25	14	24	39
Receipts scanned	41	27	23	33	51
Purchases without a receipt	30	20	17	24	36
Nothing bought	9	8	5	7	10
Smartphone	29	24	14	23	37
Tablet	39	30	18	32	51

Subjective measures of burden. Four measures of subjective burden were taken from the end of project survey. All four measures were adapted from measures used by Sharp and Frankel (1983). The distributions for these four subjective measures were skewed towards lower levels of burden. This, combined with the relatively small analytical sample size, means that the number of responses in the categories representing highest burden was typically quite small. The decision was made to recode these variables into four dichotomous measures. Specifications for models using both the original form of these variables and the dichotomised form were considered, however the original form resulted in a number of empty cells at certain levels of the four measures of subjective burden in the multivariate analysis or resulted in estimations being made from a very small number of cases. In most cases this violated the proportional odds assumption of the ordered logistic regression models. Therefore, the dichotomised specifications of models are presented here. The original and recoded responses to these questions can be found in Table 4.

Table 4
Response distributions for four subjective measures of respondent burden (original and recoded).

Original response options			Recoded response options		
	n	%		*n*	%
Likelihood[a]					
Very likely	150	67.3	Higher likelihood	150	67.3
Somewhat likely	57	25.6	Lower likelihood	73	32.7
Somewhat unlikely	11	4.9			
Very unlikely	5	2.2			
Time/effort[b]					
Very well spent	112	50.2	More well spent	112	50.2
Somewhat well spent	106	47.5	Less well spent	111	49.8
Not very well spent	5	2.2			
Interest[c]					
Very interesting	88	39.5	Higher interest	88	39.5
Somewhat interesting	111	49.8	Lower interest	135	60.5
Not interesting	24	10.8			
Difficulty[d]					
Very easy	88	39.5	Lower difficulty	88	39.5
Somewhat easy	95	42.6	Higher difficulty	135	60.5
Somewhat difficult	36	16.1			
Very difficult	4	1.8			

Notes: Original question wordings - **a** '*Imagine you were being asked to do this Spending Study for the first time. Based on your experience, how likely would you be to participate?*' **b** '*Overall do you feel that the time and effort you put into participating in the Spending Study was…*' **c** '*Overall how interesting was participating in the Spending Study?*' **d** '*Overall, how easy or difficult did you find completing the Spending Study?*'

One of these four measures, self-rated ease or difficulty participating in the study, was also asked each week in the end of week surveys, reflecting on the previous week. A week by week breakdown of the response distributions for this variable can be found in Table 5.

Table 5
Response distributions for end of week measure of Spending Study difficulty listed for each week and pooled across all weeks.

	Very easy		Somewhat easy		Somewhat difficult		Very difficult		Missing	
Week	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
1	56	25.1	55	24.7	20	9.0	3	1.4	89	39.9
2	53	23.8	51	22.9	25	11.2	3	1.4	91	40.8
3	58	26.0	53	23.8	23	10.3	5	2.2	84	37.7
4	57	25.6	63	28.3	15	6.7	2	0.9	86	38.6
Pooled	224	25.1	222	24.9	83	9.3	13	1.5	350	39.2

Notes: Original question wording '*How easy or difficult did you find completing the Spending Study this week?'*

PREDICTORS OF BURDEN

To establish predictors of burden from the seven factors affecting burden established earlier in this research two possible approaches could be taken. One approach is to try to uncover a series of direct measures for each of these factors, as was the approach taken by (Fricker, 2016) regarding the four factors originally outlined by Bradburn. An alternative approach, the one advocated here, is to consider the seven factors as conceptually underpinning burden, and then identify indirect measures that may affect each of the factors considered. This may produce a more nuanced understanding of predictors of burden. For example, a general measure of motivation may be informative, but may not provide the in-depth practical insights into how and why a respondent may be motivated or not that would be useful when making survey design choices.

Based on the seven factors determining burden a number of predictors of burden were identified, how these predictors map onto the seven factors is noted throughout. Descriptive statistics for each predictor variable can be found in Table 6.

Mobile device activities - *Ability/Motivation/Emotional stress.* Questions about whether respondents performed a range of activities on their mobile device were asked to respondents who reported access to either a smartphone or tablet. Previous research has used similar questions about tasks completed on mobile devices to attain a measure of device use competence (Fortunati and Taipale, 2014). Respondents were presented with a list of possible activities and asked, '*Do you use your smartphone for the following activities?*' Of those activities three were identified as being related to the Spending Study. The first two of these, '*Taking photos*', and '*Installing new apps (e.g., from iTunes*[1]*, Google Play Store)*', were both necessary skills to participate in the study. Being familiar with performing either of these tasks likely increased the ability of participants to take part in the study, thus decreasing the burden they faced.

The third activity, '*Online banking (e.g., checking account balance, transferring money)*', was a related skill which was included with the idea that those respondents who did this would likely be more comfortable accessing and transmitting their financial information through an app. It was felt that this greater comfort performing the task of transmitting financial information digitally might result in less emotional stress when participating in the study, meaning the burden for those participants used to doing this would be decreased. It was also considered possible that those who checked their finances online

[1] The use of iTunes to refer to what is more commonly known as the Apple App Store is a mistake in the original question wording that is matched here for consistency.

Table 6

Descriptive statistics for predictors of burden.

		n	%
Initial incentive	£2.00	97	43.5
	£6.00	126	56.6
Received unconditional £5 incentive	Yes	39	17.5
	No	184	82.5
Uses device for taking photos	Yes	201	90.1
	No	22	9.9
Uses device for online banking	Yes	158	70.9
	No	65	29.1
Uses device to install apps	Yes	180	80.7
	No	43	19.3
Willing to download app	Not willing	44	19.7
	Willing	179	80.3
Willing to use camera	Not willing	38	17.0
	Willing	185	83.0
Checks bank balance	Less than once a week	43	19.2
	Once a week or more	181	80.8
Keeps a budget	Yes	116	52.0
	No	107	48.0
Poverty threshold	Below the threshold	28	12.6
	Above the threshold	195	87.4
Time constrained	Yes	65	29.1
	No	158	70.9
Disabled/long term illness	Yes	56	25.1
	No	167	74.9
Gender	Male	87	39.0
	Female	136	61.0
Age	$\bar{x}$	44	
	SD	15	
	Q_1	31	
	Q_2	43	
	Q_3	53	
Level of education	Less than a degree	124	55.6
	Degree or higher	99	44.4

may have more interest in the topic of the study, increasing their motivation, thus reducing the subjective burden of participation.

As respondents were asked this set of questions for both mobiles and tablets, each of these activities was coded 1 if the respondent reported performing the activity on either device, or 0 if they did not report performing it on either. As those without access to either device did not receive these questions, these respondents were also coded to 0, with the assumption that without access to a device they could not perform these actions.

Willing to perform survey tasks on mobile device – *Motivation/Ability*. A series of hypothetical questions about willingness to perform different survey activities on mobile devices were asked. Of these, two were felt to be directly related to the tasks performed in the Spending Study, and likely therefore to be indicative of greater motivation to participate. The assumption here is that reporting being willing to perform this task would likely mean that the participant would be more likely to surpass the initial inhibitory threshold for deciding to participate, and as such their subjective perception of burden would be lower from the onset. It is also possible that participant's reported willingness might be indicative of their self-assessment of their ability to complete the task.

Respondents were asked '*How willing would you be to carry out the following tasks on your* [smartphone/tablet] *for a survey?*' Again, this question was asked based on reported possession of a smartphone and/or tablet, so respondents would be asked the question for either smartphone or tablet if they reported having that device, or would be asked for both if they reported having both. The two items included are willingness to '*Download a survey app to complete an online questionnaire*' and '*Use the camera of* *your smartphone to take photos or scan barcodes'*. Both items were measured on a four-point scale of '*not at all willing/a little willing/somewhat willing/very willing*'.

Where the respondent was asked both for tablet and smartphone the higher value of their two answers was taken. This was on the assumption that respondents would choose to use the device they had reported being the most willing to perform the task on. Two alternative specifications were considered, one keeping the original four answer categories, another collapsing these variables into not at all willing vs any of the other levels of willingness. On examination of the alternative specifications, the important distinction seems to be whether the participant was willing or not, as opposed to the degree of willingness; therefore, the dichotomous specification is presented here. Again, these questions were filtered on device access, and subsequently sample members who did not receive these questions were coded to 0.

Existing financial behaviours - *Ability/Motivation*. As with the existing mobile device behaviours, reported participation in certain existing financial behaviours are potential indicators of increased interest in the topic of the Spending Study. In line with existing evidence that interest results in a greater motivation to respond (Groves et al., 2004) it is expected that participants who engage in these financial behaviours will typically report being less burdened.

One measure used was an indicator measuring if respondents kept a budget. Respondents were asked '*Now, thinking about different ways that people have of managing their finances, how, if at all, do you record your budget?*' which was coded 0 if they did not report keeping any form of budget and 1 if they did. Respondents were asked '*How often do you check your bank balance?*' with '*most days/ at least once a week/ a couple of times a month/ at least once a month/ less than once a month/ never*' as response options. The original variable was highly skewed and therefore recoded into a binary indicator of high or low frequency for analysis with '*most days/at least once a week*' being coded as 1, and '*a couple of times a month/at least once a month/less than once a month/never*', coded 0.

As these measures are tied to skills related to tracking your finances (keeping receipts, being aware of how much you have spent, etc.) it also seems likely that those participants who already take part in these activities may have increased ability to complete the task as they already possess a number of associated skills.

Poverty indicator - *Emotional stress*. Given the subject of the Spending Study, it was considered that the topic of the survey may be sensitive for those with the lowest household incomes, and thus cause more emotional stress, making the task more burdensome. As such, an indicator was derived marking the threshold under which individuals were considered to be living in poverty. This was defined as those individuals whose equivalised net household income fell below 60% of the median equivalised net monthly household income. As the Innovation Panel only derives gross income, not net, this figure was first calculated for the seventh wave of the main *Understanding Society* (US7) sample (this wave having occurred for the most part in the same year as IP9). The resulting figure was £922.67. Equivalised gross household income for US7 respondents was then regressed on their equivalised net household income. The resulting regression coefficient was then used to calculate a corresponding gross poverty threshold from the earlier net threshold. The resulting threshold was £1025.38, which was applied to the analytical sample, to derive the final poverty indicator. All individuals whose household equivalised gross income fell below this threshold were considered to be living in poverty.

Time constraints - *Opportunity*. Participants with greater time constraints seem likely to have less opportunities to participate. An indicator of this was derived taking into account a number of factors. This measure was originally derived by Wenz et al. (2019). Participants were considered time constrained if they reported working more than forty hours a week, either in employment or self-employment. Those with a commute of greater than an hour to get to work each day were also coded as time constrained. In addition to this, participants were considered time constrained if they had any children under the age of five living in the household. The final derived variable took the value of 1 if a respondent met any of the criteria for being considered time constrained, or otherwise took a value of 0.

Disability or illness – *Ability*. An indicator for whether an individual had reported to be suffering from any long-standing physical or mental impairment, illness or disability was included as an indicator of participants' ability to participate in the Spending Study. Reporting such a longstanding illness or disability is considered here to reduce ability to participate. This was coded 1 if they reported that they did have a longstanding illness or disability, and 0 if they did not.

Level of education – *Ability*. Level of education was included as a proxy for cognitive ability. Participants' level of education was coded as 1 for a degree or above and 0 if a respondent's highest level of qualification was lower than this. Participants with higher education are expected to find the task easier. This may result in the task taking them less time to complete. It may also result in them reporting finding the task easier, and this may translate to other measures of subjective burden also being lower.

Demographics. Two demographic control variables were included in the analyses. Sex was coded as 0 for male respondents, and 1 for female. Age was included as a continuous variable, and the possibility of a curvilinear relationship was explored, however the introduction of a squared age term did not show evidence of such a relationship, and this squared term was subsequently removed from the analyses presented here.

RESULTS

To address the four research questions in this paper, two different units of analysis are used throughout, either: participants, or the individual app uses, with app uses clustered within participants. All standard errors are calculated adjusting for the complex clustered sample design of the Innovation Panel.

RQ1: Are subjective and objective measures of burden related?

For this first research question the unit of analysis is participants. As the four subjective measures of burden are measured at a participant level, the three objective measures chosen to be introduced in this analysis are those that are calculated at the participant level. To examine the relationship between objective and subjective indicators the matrix of correlations between the seven indicators was initially examined. An exploratory factor analysis was then carried out, examining the underlying structure of the seven indicators.

Polychoric correlations were used due to the potential drawbacks of using other correlation measures: neither Pearson's *r* or Spearman's ρ are appropriate as the subjective measures of burden used here are binary; Kendall's τ is suitable for binary measures, but the resulting correlation matrix cannot be used for factor analysis. The approach of using polychoric correlations to allow both binary correlations, and a subsequent factor analysis has previously been advocated by (Flora and Curran, 2004) and (Holgado-Tello et al., 2010) and is thus adopted here. These correlations were calculated using the user-written '*polychoric*' package written for Stata by Kolenikov (2008) and are presented in Table 7.

Table 7

Correlation matrix of the bivariate relationships between different measures of burden.

	Likelihood	Time/effort	Interest	Difficulty	Average time	Total time	No. of app uses
Likelihood	1.00						
Time/effort	0.66	1.00					
Interest	0.42	0.67	1.00				
Difficulty	0.51	0.62	0.44	1.00			
Average time	0.16	0.00	-0.13	0.19	1.00		
Total time	-0.14	-0.11	-0.22	0.06	0.59	1.00	
No. of app uses	-0.26	-0.12	-0.19	-0.07	0.07	0.81	1.00

Notes: *n* = 223 participants; Correlations between subjective measures are polychoric, correlations between objective measures and subjective measures are polyserial, correlations between objective measures are Pearson's r correlations.

Using established thresholds for interpreting correlations (Hinkle et al., 2003) most of the relationships between each pairing of the four subjective measures fell within the range of moderate positive correlations (0.50 to 0.70). The only exceptions to this were the relationship between interest in the study and difficulty; and between interest and likelihood of participation. Here the correlations were lower, though both were above 0.40, indicating a low positive correlation. The correlations between each of the subjective measures and the objective measures of burden produced coefficients that fell below the threshold for a remarkable relationship, falling within the range of -0.30 to 0.30. This seems to suggest that the subjective measures captured are not associated with any of the three measures of objective burden considered here.

Total time showed a moderate to strong relationship to both the number of app uses, and the average time taken to complete app uses. This is not a surprise as increases in either of these two variables would have been expected to increase the total time taken to complete app uses. The number of app uses did not show a strong association with the average time taken to complete an app use.

Before performing the exploratory factor analysis, a common test for the appropriateness of applying a factor structure to a set of variables was conducted. Bartlett (1951) suggests the test of sphericity to offer validation for one of the assumptions of factor analysis, namely that the variables are not orthogonal from one another. A result of χ^2=1040.56, df = 21, p < 0.001 indicates that the variables are not orthogonal from one another and are therefore suitable for factor analysis.

Having established the appropriateness of using factor analysis on the seven variables, a principal factors factor analysis was conducted, with an orthogonal varimax rotation. This was calculated using the earlier matrix of polychoric correlations. Only those factors that were above the threshold of the Kaiser criterion (Kaiser, 1960), an eigenvalue of 1.0, are presented. This produced a structure with three factors, and the factor loadings for each variable with relation to these factors are presented in Table 8.

Table 8
Factor analysis of the structure of seven indicators of respondent burden.

	Factor one	Factor two	Factor three	Uniqueness	KMO
Likelihood	0.69	-0.20	0.17	0.44	0.77
Time/effort	0.88	-0.06	-0.02	0.22	0.68
Interest	0.68	-0.13	-0.15	0.48	0.77
Difficulty	0.68	0.00	0.17	0.50	0.82
Average time	0.04	0.15	0.90	0.16	0.22
Total time	-0.06	0.85	0.49	0.03	0.39
No. of app uses	-0.09	0.96	-0.06	0.07	0.33
Eigenvalue	2.19	1.72	1.15		
Overall					0.50

Notes: *n* = 223 participants; Factor structure after orthogonal varimax rotation applied; Factors with Eigenvalues greater than 1 presented.

For the first factor each of the four subjective measures of burden produced a factor loading greater than the suggested threshold of 0.60 (Guadagnoli and Velicer, 1988) suggesting strong associations between each of these variables the underlying latent variable. There is very little evidence of an association between the objective measures of burden and this underlying factor, further reinforcing the idea that the subjective measures and the objective measure are capturing different aspects of burden.

The other two factors are largely related to a single variable, either the number of app uses, in the case of factor two, or average time taken to complete app uses for factor three. That total duration strongly loads onto each of these factors is again not surprising as this measure is a product of the other two variables. It is somewhat surprising however that the number of app uses and the average duration to complete app uses were not strongly related to one another. A test for the Kaiser-Mayer-Olkin measure of sampling adequacy (Kaiser, 1970; Kaiser and Rice, 1974) was also conducted with an overall result of 0.50; applying the criteria set out by (Kaiser and Rice, 1974) this value comes at the very lowest end of values considered appropriate for factor analysis. However, examining this for individual variables indicates that the subjective measures of burden have a more evident factor structure than the objective measures. The four subjective measures ranged from 0.68 to 0.82, values that can be considered suitable for factor analysis. This compares to values ranging from 0.22 to 0.39 for the objective measures. This seems to further reinforce the notion that there is a latent structure underlying the four subjective burden measures, whereas the three objective measures are not related in this way.

RQ2: How do subjective and objective burden change over the course of the study?

Subjective burden. To investigate the change in subjective burden across the four weeks of participation the sequence of responses to the weekly difficulty question are examined. These sequences are plotted in Figure 1. Each line in the graph represents the sequence for a single participant. The '*sq*' set of sequence analysis packages written for Stata by (Kohler et al., 2006) were used to produce this plot.

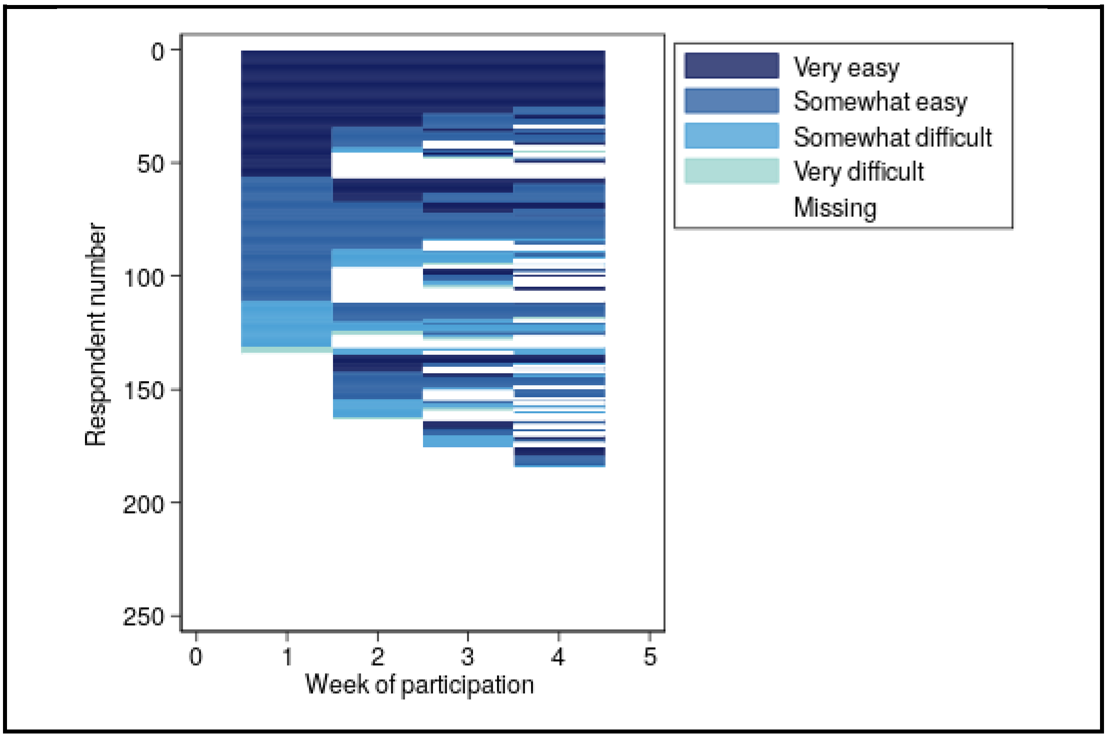


Figure 1
Sequence analysis graph documenting the sequence of weekly reported difficulty participating in the Spending Study

The resulting array of sequences seems to indicate no systematic change in reported burden across the four weeks of participation. One pattern that might have been expected would be that respondents who were not initially burdened accumulate burden, echoing the fatigue observed to occur in some diary studies (Gerstel et al., 1980; Leigh, 1993; Verbrugge, 1980). Conversely, it might be expected that respondents who are initially burdened find themselves adapting to the task, and subsequently their reported levels of burden would decrease. Neither of these patterns is observed in the sequences presented in the graph in Figure 1.

To formally test whether there were any within individual trends in self-reported difficulty a fixed-effects regression model was estimated. This makes it possible to examine the trends within individuals across the course of the study. One challenge that arises in fitting this model is how best to treat the large volume of missing reports that are present in the data. One approach is to treat these as a substantive category, indicative of high levels of burden, with the assumption that a high level of burden would cause a participant to be less likely to complete an end of week survey. A fixed effects regression including missing reports as a substantive category, representing the highest level of burden, produces a coefficient of $\beta = - 0.03$, $p > 0.05$, 95% CI [-0.11, 0.04]. Excluding these missing reports avoids the assumption that these are a substantive category of burden but results in an unbalanced panel. The resulting coefficient for a model excluding missing reports is $\beta = - 0.01$, $p > 0.05$, 95% CI [-0.06, 0.04]. Neither of these specifications of the model produces a result that is indicative of an underlying pattern across time. This is consistent with the lack of a pattern present in the sequence analysis graph.

Objective burden. To examine the change in objective burden across the course of the study trends in the duration of app uses as a participant completes more app uses were modelled. The unit of analysis is app uses clustered within individuals. Fixed-effects models are again fitted to look at the within individual changes. Four separate models were specified, one measuring the change across all app uses and three models measuring the changes within each of the three types of app use. Lines fitted for each of these four models are plotted in Figure 2. The overall trend was a decrease in the time it took to complete app uses with participants typically taking 0.3 seconds less to complete each subsequent app use (β = - 0.29, p < 0.001, 95% CI [- 0.34, -0.24].

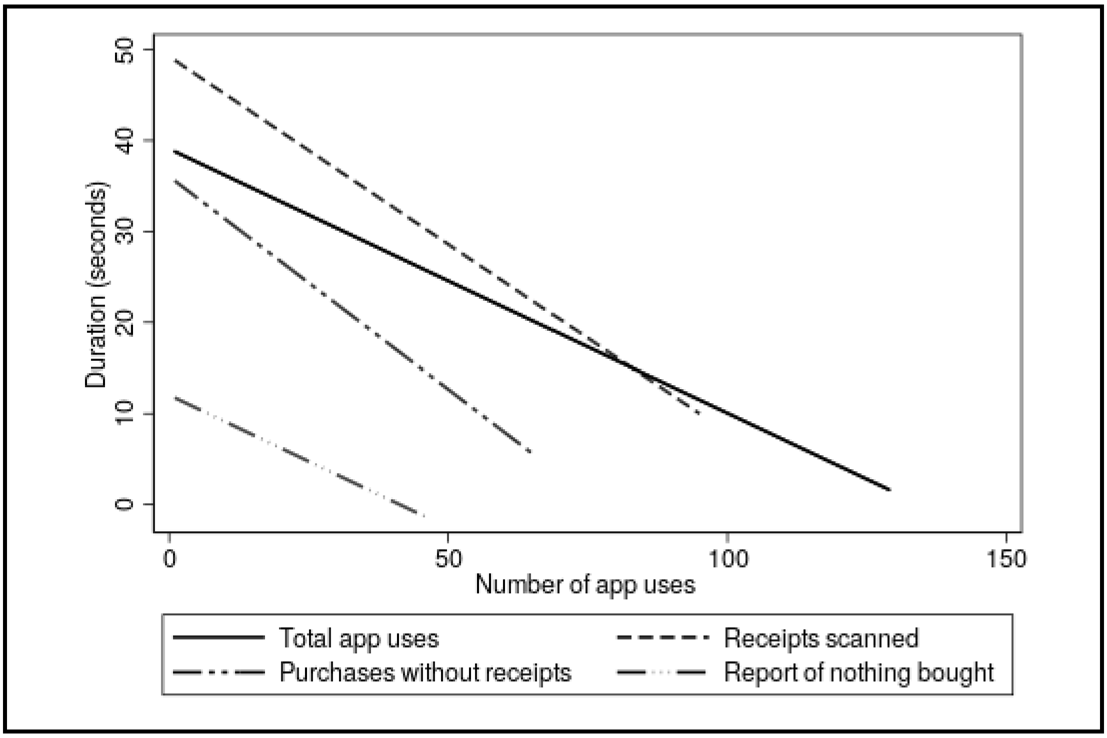


Figure 2
Fixed-effects regression models of changes in app use duration as participation continues split by type of app use

The model was then repeated for each type of app use, with the predictor variable becoming the number of that type of app use that had been completed. The decision was made to run the models separately to test whether the overall trend was truly the product of decreases in time, or whether there was a compositional effect as a result of respondents shifting from the more time-consuming scanning of receipts to the other two less time-consuming methods. The results suggest that participants became between three tenths to half a second quicker with each subsequent app use for all three types of app use: β = - 0.41, β = - 0.47 and β = - 0.29 for receipts scanned, purchases submitted without receipts, and submissions of nothing bought that day, respectively

(95% CIs [-0.51, -0.31], [-0.57, -0.37] and [-0.37, -0.21] respectively, all p-values < 0.001).

It is also possible to consider how patterns in participation inform changes in burden across the course of the study. Jäckle et al. (2019a) report that participation in the study was fairly consistent with 81.5% of participants using the app on at least 29 days. Similarly, they found that the mean number of purchases submitted (either receipts scanned or purchases without receipts) per day per respondent stayed consistent across the study. To expand upon this, the possibility was explored that participants may have shifted in their response behaviour. To test whether participants shifted in their response behaviour within individual fixed effects models of the proportion of each of the three types of app use completed per day were fitted. Throughout the course of the study there was a slight decline in the proportion of receipts scanned (β = -0.0005 , 95% CIs\ [-0.0009,\ -0.0002]) and reports of nothing bought (β = -0.0009 , 95% CIs\ [-0.0013, -0.0005]) both p - values < 0.001. The proportion of purchases without receipts increased across the study (β = 0.0013, 95% CIs [0.0009, 0.0017]). However, the practical effects of these shifts were minimal. From these changes in proportions it is possible to calculate the changes in the percentage share of an individual's app uses that were of each type between the first and last day of the four weeks analysed here. For receipts scanned this was typically a decrease of 1.3 percentage points. Reports of nothing bought typically decreased by 2.4 percentages points. Finally, the share of app uses that were purchases reported without receipts increased by 3.5 percentage points.

RQ3: Does objective burden predict breaks in participation?

Due to the high levels of missingness in the end of week questionnaires it was unfortunately not feasible to model breaks in participation using the weekly subjective measure. The end of project responses were also unsuitable as there were retrospective reports. As such, analyses to predict breaks in participation were only conducted using the objective measures of burden as predictors.

Cox proportional-hazard regression models were fitted to determine whether there was evidence that a higher objective burden resulted in temporary or permanent break-off. Three models were specified, measuring breaks in participation in different ways.

In the first model, the outcome variable is dropout from the Spending Study. Participants were considered to have dropped out (and thus exited from the analysis) after the last day on which they used the app within the 28 days from when they first used the app. There were therefore 223 spells, with one for each participant, running from when they began the study, until the last day on which the app was used.

The second model examined is the time until the first day on which the participant did not use the app. Again, there are 223 spells, this time running from when participants began the study until the first day on which the app was not used. Once the participant missed a day of app use they exit from the analysis.

The third model included repeated spells of participation: when a participant missed a day of app use a new spell began from the day they resumed using the app. Participants remained in the study throughout repeated spells of participation, with the exit condition for this model being dropout, as defined in the first model. This final model consists of 1559 spells. All three models use the Breslow method for handling tied failures (Breslow, 1974). The results of all three models are documented in Table 9.

The main predictor of interest is the average duration of app uses, up to that point in the study. This is a time varying measure, which is recalculated for each day. The proportions of app uses to date that are purchases without receipts and submissions of nothing bought are included as control variables. These are included because the three different types of app use differed in the amount of time taken to complete them. This could lead to a confounding compositional effect if participants have completed different proportions of different types of app uses.

Table 9

Cox regression models examining whether objective burden is predictive of dropout or gaps in participation.

	Dropout		First day missed		All days missed	
	HR	*SE*	*HR*	*SE*	*HR*	*SE*
Average duration	0.98	0.01	1.01*	0.00	1.00	0.00
Prop. purchases without receipts	1.24	0.75	0.97	0.30	1.22	0.24
Prop. Nothing bought	1.19	0.88	2.79**	0.83	1.50	0.42
Wald	4.79		15.21		2.42	
Spells	223		223		1559	

Notes: *n* = 223 participants; * *p* < .05 ** *p* < .01 *** *p* < .001.

For both time until dropout, and time until all missed days the hazard ratio was not statistically significantly different dependent upon the average duration of app uses up until that point (*HR* = 0.98 and *HR* = 1.00 respectively, both p - values > 0.05). In terms of the first missed day of participation, higher average time taken to complete app uses is associated with a higher risk of initially missing a day of participation (*HR* = 1.01, p < 0.05). There was a 1% increase in the expected hazard associated with a one second increase in average time taken to complete app uses. To better understand this result, it has been noted that it can be informative to convert hazard into a corresponding measure of effect size (Azuero, 2016). In this case the value falls below the suggested threshold for a small effect of 1.14, suggesting the observed effect may be inconsequential. Further doubt is cast on whether there is an effect of average duration on initially missing a day when considering the full sample of 268 app users, where this result was not statistically significant (*HR* = 1.00, p > 0.05).

There was also a higher risk of those participants with a higher proportion of reports of nothing bought initially missing a day of using the app (*HR* = 2.79, p<0.05). It is possible that this was due to the task being less salient for these participants, as they were not making purchases as frequently. However, caution should be exercised in interpreting this coefficient directly, as a one-unit change in proportions reflects the entire range of this value. It is therefore more useful to consider a more informative unit shift in proportions, for example the hazard ratio for the difference between the 25th and 75th percentile (Q_1 = 0.07, Q_3 =0.38), which was (*HR* = 1.38). According to \cite(Azuero) this corresponds with a small effect size.

RQ4: What factors predict subjective and objective burden?

Subjective burden. Table A2 in Appendix A shows the bivariate relationship between the predictors of burden and each of the four subjective measures of burden. Multivariate analyses were completed using four logistic regression models, with each of the four measures of subjective burden captured in the end of project survey as the dependent variable in one of the models. Each of the four dependent variables was coded such that 0 meant lower burden, and 1 meant an increased burden. The unit of analysis is the 223 participants. The results of the four models are documented in Table 10.

Throughout, where a statistically significant predictor is observed, this is compared to a series of thresholds for odds ratio values that correspond to recognised thresholds for effect size as measured by Cohen's *d*. The thresholds used are those set out by Cohen (1969) who suggests that *d* = 0.20, *d* = 0.50 and *d* = 0.80 represent a small, medium and large effect size respectively. Formula 1, as set out by Borenstein et al. (2009), allows the conversion of the threshold values of Cohen's *d* to log odds ratios, which can then be converted to odds ratios.

$$LogOddsRatio = d\frac{\pi}{\sqrt{3}} \tag{1}$$

This results in values of *OR* = 1.44, *OR* = 2.48 and *OR* = 4.27 corresponding to small, medium and large effect sizes respectively. To establish thresholds for odds ratios below one the inverse values for these effect size thresholds can be calculated by one over each respective value, resulting in *OR* = 0.69, *OR* = 0.43 and *OR* = 0.23, corresponding to small, medium and large effect sizes respectively. Across all four models the two incentive treatments were not significant predictors of the respective measures of subjective burden. It is possible that this may be a result ceiling effects (Groves et al., 2000) as to the effectiveness of incentives in the presence of other motivating factors. This seems plausible given the seemingly high initial inhibitory threshold to participate (as suggested by the low response rate) together with relatively little variability in the level of self-reported burden. Both perhaps suggest that participants had to be quite highly motivated to participate, so the additional effect of a larger incentive was negligible.

For all four models, downloading apps and online banking were not statistically significantly predictors of any of the four measures of subjective burden. However, using a mobile device to take photos did significantly increase the odds of reporting a lower likelihood of participating in the Spending Study if asked for the first time (*OR* = 5.34, $p < 0.05$), corresponding to a large effect size.

Gender, disability/long term illness, poverty and time constraints were not significant predictors across any of the four models. Participants who reported their highest level of education as a degree or higher had significantly higher odds of reporting that their time and effort was less well spent as compared to those with lower levels of education

(*OR* = 1.87, p < 0.05) though this effect is seemingly small. This perhaps reflects a greater value placed upon their time by these participants.

Table 10
Logistic regression models examining the multivariate relationship between predictors of burden and four measures of subjective burden.

	Likelihood		Time/effort		Interest		Difficulty	
	OR	*SE*	*OR*	*SE*	*OR*	*SE*	*OR*	*SE*
£6 incentive treatment	0.96	0.36	0.99	0.32	1.22	0.38	1.61	0.53
Received additional incentive	1.18	0.54	1.56	0.71	0.95	0.45	0.77	0.30
Uses device for taking photos	5.34*	3.34	1.87	1.04	0.65	0.43	2.04	1.32
Uses device for online banking	0.53	0.19	0.60	0.21	0.80	0.32	0.52	0.28
Uses device to install apps	1.22	0.56	1.08	0.54	2.34	1.26	0.55	0.34
Willing to download app	0.78	0.43	2.45	1.32	1.68	0.75	1.37	0.71
Willing to use camera	0.46	0.28	0.30*	0.16	0.32	0.19	1.09	0.62
Checks balance at least once a week	0.80	0.29	1.03	0.38	0.48	0.21	1.90	0.78
Keeps a budget	0.87	0.31	0.86	0.24	0.84	0.23	1.88	0.55
Below the poverty threshold	2.51	1.36	0.65	0.34	0.59	0.31	2.43	1.55
Time constrained	0.73	0.26	0.91	0.29	0.81	0.3	0.77	0.26
Degree or higher	1.38	0.44	1.87*	0.54	1.86	0.62	1.39	0.39
Disabled/long term illness	0.58	0.23	0.58	0.21	0.64	0.25	0.56	0.21
Female	1.05	0.35	0.76	0.22	1.18	0.35	0.89	0.26
Age	1.00	0.01	1.00	0.01	0.97**	0.01	1.01	0.01

Notes: *n* = 223 participants; * *p* < .05 ** *p* < .01 *** *p* < .001.

Age was a significant predictor of interest, with older respondents reporting finding the study more interesting than younger respondents (*OR* = 0.97, p < 0.01). Though this was a seemingly negligible effect when comparing year to year, the effect was more substantial when comparing across a larger difference in age. For example, when comparing the first and third quartile of age ($Q_1 = 31$, Q_3=53) the odds ratio is *OR* = 0.49, a medium sized effect.

Willingness to download an app to complete survey tasks was not a significant predictor of any of the four measures of subjective burden. Willingness to use a camera to take photos or scan barcodes was a significant predictor of how well participants reported finding their time and effort spent participating. Those who reported being willing to use their camera to take photos for a data collection task had significantly lower odds of reporting lower levels of satisfaction with how well spent their time and effort was (OR = 0.30, ($p < 0.05$) when compared to those who were not willing, again a medium sized effect.

Objective burden. The bivariate relationship between the predictors of burden and the time taken to complete app uses are documented in Table A3 in Appendix A. To understand which factors are predictive of the objective burden experienced by respondents the same covariates that were explored as predictors of subjective burden were included in a model with the duration of individual app uses as the dependent variable. This shifted the unit of analysis from participants down to the level of individual app uses. A mixed effects regression model was used to account for the clustering of app uses within individual participants. The results from the model are presented in Table 11. Type of app use was included to control for the differences in typical durations of each of the three types of app use.

Neither receipt of the higher initial incentive or receipt of the additional incentive proved to be a significant predictor of response times. This is not entirely surprising, it seems more plausible that if an effect of incentives were to be observed it would be found when examining subjective burden, with the assumption that an increased incentive would lead to greater motivation, thus reducing the subjective burden of the task. However, it was considered possible that a larger incentive may have given the impression of greater importance of the task to respondents, thus potentially leading to greater care taken completing the task. These two covariates were retained for this reason, though it turns out there is no evidence of such a relationship.

Table 11
Mixed effects regression model examining the multivariate relationship between predictors of burden and the time taken to complete app uses.

	β	*SE*
£6 incentive treatment	0.93	1.10
Received additional incentive	0.81	1.30
Uses device for taking photos	1.72	2.70
Uses device for online banking	-4.17**	1.42
Uses device to install apps	1.59	1.73
Willing to download app	-4.50*	1.92
Willing to use camera	-0.99	2.05
Checks balance once a week or more	3.98**	1.37
Keeps a budget	-0.84	1.08
Below the poverty threshold	0.19	1.74
Time constrained	-0.87	1.08
Degree or higher	-0.2	1.13
Disabled/long term illness	0.8	0.83
Female	2.09*	0.94
Age	0.33***	0.03
Type of purchase		
Reference: Scanned receipts		
Purchase without receipts	-10.69***	1.03
Nothing bought	-33.46***	1.21
Constant	27.68***	3.99
Wald	1257.50***	

Notes: *n* = 10179 app uses, across 223 participants; * *p* < .05 ** *p* < .01 *** *p* < .001.

Those respondents who reported a long-term illness or disability did not take longer to complete app uses, this perhaps can be explained by the fact that this variable encompasses a wide array of medical conditions, many of which may not be expected to have a direct impact upon participation. Cognitive ability, as measured by level of education, did not have a significant association, though it is unclear whether a better indicator of this characteristic would have revealed an association. Participants whose income fell below the poverty threshold were also not statistically significantly different in how long it took them to complete app uses.

Surprisingly, those participants who reported using their mobile devices for taking photos or installing apps at IP9 were not significantly faster at completing app uses. It was expected that having these existing skills would reflect a greater competency in usage of mobile devices and that this would result in shorter app use durations.

In terms of reported willingness to perform survey tasks on mobile devices, willingness to download an app to complete survey tasks was found to be predictive of app use duration. Respondents who reported being willing were around four and a half seconds faster ($\beta = - 4.50$, $p < 0.05$) than those who reported not being willing to download a survey app. Surprisingly, willingness to use a camera for survey tasks, which is more directly tied to completing app uses, was not found to be a significant predictor of duration.

When it comes to existing financial behaviours, keeping a budget did was not a significant predictor of length of time it took respondents to complete app uses. However, checking one's bank balance more frequently was. Participants who checked their bank account at least once a week took just under 4 seconds longer to complete app uses than those who checked less frequently ($\beta = 3.98$, $p < 0.01$). In contrast, those respondents who reported using their mobile device for online banking were around four seconds faster at completing app uses ($\beta = 3.98$, $p < 0.01$).

Age was found to be a significant predictor of the time taken to complete app uses, with each additional year older a participant was resulting in their app uses typically being around a third of a second longer in duration ($\beta = 0.33$, $p < 0.001$). By again comparing the first and third quartiles of age ($Q_1 = 31$, $Q_3 = 53$) it is possible to get a better understanding of the effect of age on duration within the sample. The predicted duration for an individual at Q_3 compared to one at Q_1 is 7.30 seconds longer. One explanation for this is that it is consistent with evidence of a second-level digital divide in skills, with technical capability being less amongst older individuals (Loges and Jung, 2001).

Finally, gender was a significant predictor with women typically taking around two seconds longer to complete app uses ($\beta = 2.09$, $p < 0.05$).

DISCUSSION

This paper sought to draw together existing literature on respondent burden to establish a conceptual framework, to apply this framework to consider burden in a non-questionnaire survey context, to examine the relationship between subjective and objective burden (RQ1), to consider how burden changes over the course of a study (RQ2 & RQ3), and to illustrate how that conceptual framework might be used to help identify predictors of burden (RQ4). Such an approach could then be adapted to consider burden in an array of different research settings, that involve repeated measures or episode level data collection.

To this end, this paper drew upon the seven factors offered up by (Bradburn, 1978) and (Haraldsen, 2004) and expanded upon these to review much of what has already been established with regards to each of these factors in the existing survey methodological literature. Throughout, the focus was partially on establishing what was known for each of these factors in relation to studies collecting data through receipts or using mobile apps. However, as is expanded upon in the concluding remarks, it is felt that such an approach could be useful when considering other forms of data collection.

The results of RQ1 seem to support the notion that subjective and objective burden arise separately from one another. The four measures of subjective burden were strongly correlated with one another, and also showed strong evidence of mapping onto a latent variable that is seemingly consistent with an underlying concept of subjective burden. This highlights the potential for future use of multi-item scales to capture subjective perceptions of burden. This was not the case for objective burden, where measures were less strongly correlated to one another. This is probably to be expected as these different measures are capturing objective burden in different ways. This highlights the importance of careful consideration when attempting to measure objective burden, as this can be considered either on an event level, or cumulatively across data collection.

The four subjective measures of burden were not strongly correlated with any of the three objective measures. For the three subjective measures not related to time spent participating this is consistent with previous research which has found a lack of correlation between measures of objective burden and subjective measures not explicitly asking about length (Sharp and Frankel, 1983; Oomens and Timmermans, 2008). However, it is surprising that the subjective measure asking about whether time and effort spent participating was well spent is also not strongly correlated with objective measures. Subjective measures asking about survey length have typically been found to have a strong association with objective length (Dale and Haraldsen, 2005; Sharp and Frankel, 1983). It is possible that the lack of correlation here may be a result of asking about effort as well as time (though this is the same as in the case of Sharp and

Frankel); or it could reflect the disconnect between subjective and objective indicators of burden that has at times been observed (Oomens and Timmermans, 2008).

In terms of how burden changes over time (RQ2) the results of the analysis of reported difficulty throughout the course of the study suggest that there is no evidence of systematic changes in subjective burden. It seems likely that in the case of the Spending Study this was because there was a high initial inhibitory threshold that was necessary to surpass to begin participating and that this may have resulted in subjective burden being typically quite low among participants, and indeed, this can be seen in the original distribution of the four subjective measures.

The time taken to participate showed consistent signs of decreasing as participation continued. This is reassuring, as it suggests that the objective burden of each task performed decreased as the number of tasks performed increased. What is less clear is whether this reduction in burden is the result of a learning effect with increases in participant ability, or whether participants were expending less effort to participate in the task, impacting on the quality of the data collected. Examination of indicators of data quality looking for evidence of satisficing behaviour would help to better understand the mechanism driving the reduction in time taken to participate. This result at first glance also seems to contradict the weak correlation between number of app uses and time taken to complete app uses that was found in RQ1. However, this can be explained by considering that these two relationships are subtly different. It seems that whilst an individual who completed more app uses was not necessarily quicker than one who completed less, a given individual tended to complete their app uses faster as they completed more of them.

The possibility that respondents may have changed their response behaviour to manage burden throughout the course of the study was explored. The empirical evidence suggests that whilst this did occur, the effect was minimal throughout the whole of the study, and this did not seem have a practically significant effect.

The effect of cumulative burden on continued participation was small. Respondents who on average took longer to participate had a higher risk of initially missing a day of participation (RQ3). However, this effect was minimal, and was not statistically significant when considering all app users.

It is felt that the framework of seven factors affecting burden was useful for helping to identify predictors of respondent burden. However, when it comes to uncovering which factors predict subjective and objective burden (RQ4) it seems clear that more work is necessary to help better identify these factors. This echoes the difficulties found in uncovering the characteristics which determine whether respondents experience fatigue in a diary study (Gillmore et al., 2001). That said, this paper does begin to find some evidence of the importance of certain factors. Those who reported being willing to download an app to complete survey tasks using a mobile device turned out to be significantly faster at completing app uses. Likewise, those who reported being willing to use a camera to complete survey tasks were more likely to report their time and effort were well spent. This echoes the previous finding that hypothetical willingness is predictive of propensity to respond (Jäckle et al., 2019a), with participants who reported themselves as being very or somewhat willing to download an app to complete survey tasks being eight percentage points more likely to participate. That willingness should prove to be predictive of both participation, together with subjective and objective burden, is a positive argument for making use of hypothetical willingness questions to inform decisions about the use of alternative methods of survey data collection.

Older participants took significantly longer to complete app uses indicative of reduced mobile technology skills amongst older participants (this is consistent with findings in the general population, see Loges and Jung, 2001). It is possible this could also reflect older respondents being more conscientious about responding and taking more time and greater care with their responses. This would echo earlier findings that older individuals are more conscientious survey respondents (Hektner et al., 2007). Similarly, female respondents took significantly longer to respond. This may also be a product of greater care taken responding, as women have also been found to be more conscientious respondents (Hektner et al., 2007).

One important caveat throughout is that the distribution of burden captured in the end of project survey does not fully reflect the full continuum of burden. For those respondents for whom the subjective burden was greatest it seems likely that they never surpassed the initial inhibitory threshold necessary to begin participating in the Spending Study. Jäckle et al. (2019a) examined participation in the Spending Study. They found that certain demographic groups, such as younger participants, and female participants, were overrepresented in the study. They also found differences in financial behaviours between participants and nonparticipants, with those who check their back balance at least once a week, check their bank balance using an app or online, and those who use a spreadsheet or computer document to keep a budget all over represented in the study. Similarly, those who did not keep a budget, used paper statements or cashpoints to check their balance, or did not have store loyalty cards were underrepresented. It is possible that this indicates a greater motivation through greater saliency of the topic of the study for some participants. That a number of these predictors of response biases were related to technology use may also suggest the importance of whether the participant was an active user of mobile technologies, and how this may have shaped both their opportunity and ability to respond. This is also reflected in the response propensity of individuals based on whether they reported owning a mobile device at IP9. Rates of participation were higher for those who reported having a mobile device than those who did not. However, more reassuringly, a number of indicators of the financial situation of participants were not significantly different between participants and nonparticipants, including: personal monthly income, the amount the household spent on food purchases in a month, the amount the household spent each year on fuel, whether the household reported struggling or being behind with paying housing costs or utilities, or the individual's subjective assessment of their financial situation.

Rates of participation were also an issue with regards to the end of week debrief surveys. There are a number of ways these could have been boosted. One barrier to participation in these additional debrief surveys is that they were fielded on a separate web survey platform. If these surveys had been administered each week within the Spending Study app this may have reduced the burden placed on the respondent to participate. As a further extension of this, embedding the feedback questions at the point of participation (e.g. as the participant scanned a receipt) may have helped increase the saliency of these questions, again potentially boosting the response rate. Finally, whilst these additional surveys were incentivised, a conditional incentive based on participating in all of the debrief surveys (as was offered for participating in all days of the study) may have further boosted participation rates.

In addition to not capturing nonparticipants, the analytical sample does not fully capture burden even amongst participants. It seems plausible that those participants in the Spending Study who chose not to complete the additional end of project survey may have been amongst those most burdened by the task. In addition to this, the omission of the small portion of end of project respondents who did not receive the correct questionnaire version further contributes to an inability to account for the full spectrum of burden. Future research into respondent burden may benefit from finding ways of considering burden for both respondents and non-respondents.

There are also a number of potential issues with using retrospective measures of subjective burden. Schwarz (2012) discusses the limitations of having respondents reconstruct subjective measures at some point subsequent to activity about which they are being asked. It is suggested that real-time capture of attitudinal measures may provide more accurate results. Future analyses into burden within repeated measures studies such as the Spending Study may benefit from embedding questions about burden in-situ alongside the main data collection. A further improvement to the subjective measures of burden would have been an inclusion of a measure asking specifically about usability, whilst there was a measure of ease or difficulty, it would have been informative to also have a more nuanced measure of how usable the app was.

Potentially some of the variation in the time it took to complete app uses may be a result of differences in the specifications of the devices used to participate in the app. It is plausible to consider that such differences may be incorporated into the framework presented here, as they may for example decrease the respondent's opportunity to participate. This is explored in paper two.

This paper presents results from only one example of a research context in which burden has been examined. More research is necessary to better understand how burden varies across different types of data collection using mobile apps. It would also be informative for further research to present a comparison between mobile app data collection methods and existing analogue methods. For example, it would be useful to compare the burden between an app scanning task and a study in which respondents submitted paper receipts or kept a paper diary of their spending.

More research is also necessary to better understand the relationship between subjective and objective burden. Qualitative accounts of how objective burden feeds into subjective perceptions of a task may help to shed light on the relationship between experienced burden and subjective perceptions of burden.

CHAPTER TWO

THE INFLUENCE OF DEVICE CHARACTERISTICS ON DATA COLLECTION USING A MOBILE APP.

ABSTRACT

Previous research has found differences in survey outcomes on mobile devices and PCs. A wide variety of mobiles devices are used to respond to surveys. Little is known about how differences in mobile devices may affect data quality. Data is from the *Understanding Society* Spending Study One, an app-based study asking respondents to take pictures of receipts or submit information about purchases. Results suggest some survey outcomes can be strongly affected by the device used. Important device characteristics affecting data quality were whether the device was a tablet or smartphone, the OS, and the amount of RAM.

INTRODUCTION

Using mobile devices for data collection in survey research offers both new opportunities and new challenges. One challenge is the diverse range of available models in the mobile device market. There were an estimated 1,600 models of mobile device available in 2009 (Zahariev et al. 2009 cited in Callegaro, 2010). By 2015 the number of available Android models alone was reportedly around 24,000 (Open Signal, 2015). Such a wide array of devices that respondents could be using to complete surveys comes as a challenge to one of the central tenets of survey research: standardisation. If using different devices results in systematic differences in the survey experience, or in the quality of data collected, this would lead to biases in estimates.

Respondents however are not randomly assigned to their devices. Any observed associations between device characteristics and data quality could therefore also reflect the effects of the respondents themselves.

To examine the influence of device characteristics this paper uses data from the *Understanding Society* Spending Study 1 (SS1). Respondents to SS1 used a mobile app to record their purchases across a month. They could take a picture of a receipt, enter data about a purchase, or report no spending that day. These app data were supplemented with data from wave nine of the *Understanding Society* Innovation Panel (IP) and used to examine the following research questions:

RQ1: What proportion of the variance in data quality indicators can be attributed to the device model used to participate, and what proportion to the respondent?

RQ2: Are specific device characteristics associated with data quality indicators?

RQ3: Do any associations between device characteristics and data quality indicators remain after controlling for respondent characteristics?

BACKGROUND

To date, there has been no research explicitly examining the effects of the model of mobile device used to complete a survey. The literature on device effects has followed on from the mode effects literature making comparisons between broad categorisations such as smartphones, tablets and desktops. The assumption is that the data collection process will largely be the same within any one of these device types.

This paper examines whether it is enough to consider device effects by device type, or if it is necessary to consider the more granular effects of specific device models. The clustered structure of survey responses within device models is similar to the structure of survey responses clustered within interviewers in face-to-face surveys.

Mode and device effects

The potential for mode effects has long been recognised (Deming, 1944) and substantial evidence of mode effects has been found (e.g. Groves and Kahn, 1979; Dillman and Christian, 2005; Elliott et al., 2009). For a comprehensive discussion of the effects of the mode of data collection the reader is directed to Jäckle et al. (2010). In short, the main concern has been how different survey modes can contribute to different sources of error and how this affects the comparability of data collected across different modes.

Most of the literature on device effects has made comparisons between large groups such as PCs (defined as a desktop or laptop computer) or mobile devices (defined as a mobile phone or tablet) (e.g. De Bruijne and Wijnant, 2013; De Bruijne and Wijnant, 2014; Couper and Peterson, 2017; Fernee and Sonck, 2013; Keusch and Yan, 2017; Lugtig and Toepoel, 2015; Mavletova, 2013; Revilla, 2017; Revilla and Couper, 2018; Revilla et al., 2016; Struminskaya et al., 2015).

Several studies have found evidence of differences in responses between PCs and mobile devices. Revilla et al. (2016) found that smartphone respondents typically provided shorter answers to open-ended questions than those using PCs, however Antoun et al. (2017) found the opposite. Couper and Peterson (2017) found that respondents typically took longer to answer questions on mobile devices, and that much of this could be attributed to increased time spent scrolling. Several other studies have also found evidence that respondents take longer to complete surveys when using a mobile device (Mavletova, 2013; De Bruijne and Wijnant, 2013; Mavletova and Couper, 2013; Cook, 2014; Wells et al., 2014; Struminskaya et al., 2015). However, some research has found no differences in the average response times between mobile and desktop respondents (Lugtig and Toepoel, 2015; Toepoel and Lugtig, 2014).

Research into the effects of device type on measurement error have found conflicting or no effects. Respondents using mobile devices have been found both to be more likely (Struminskaya et al., 2015) or less likely (Keusch and Yan, 2017; Lugtig and Toepoel, 2015) to straightline than those using a PC. It has been suggested that this may depend on whether the questions are presented in a grid. No effects were found for disclosure of sensitive information (Antoun et al., 2017; Mavletova, 2013; Revilla et al., 2016); acquiescence (Keusch and Yan, 2017); mid-point responding (Keusch and Yan, 2017); item nonresponse (Lugtig and Toepoel, 2015; Revilla and Couper, 2018); and primacy effects (Lugtig and Toepoel, 2015; Mavletova, 2013).

CLUSTERED SURVEY RESPONSES

Multilevel models have widely been used to examine interviewer effects, accounting for the clustering effect of respondents within interviewers (West et al., 2018; West and Elliott, 2014; West and Olson, 2010; Wiggins et al., 1990; Jäckle et al., 2011; Pickery et al., 2001). Cross-classified models have in addition been used to disentangle interviewer effects from area effects (O'Muircheartaigh and Campanelli, 1998; O'Muircheartaigh and Campanelli, 1999; Brunton-Smith et al., 2017; Durrant et al., 2010).

These studies have used either intra-interviewer correlations (IIC) or interviewer design effects to assess the extent of clustering. These measures are related to one another and are derived from the decomposition of variance in the multilevel models. The size of reported intra-interviewer correlations has been varied. O'Muircheartaigh and Campanelli (1998) suggest correlations of larger than 0.10 are rare. They found correlations ranging from 0.06 – 0.17. Jäckle et al. (2011) reported IICs ranging from 0.04 – 0.07. West and Olson (2010) reported IICs ranging from 0.01 – 0.12. It has however been suggested that even relatively small interviewer effects can have large impacts when estimating statistics. Assuming an average of 30 respondents per interviewer, an IIC of 0.01 would result in a twenty-nine percent increase (West and Olson, 2010), and an IIC of 0.02 would result in a fifty-four percent increase (West et al., 2018) in the variance of an estimated mean.

Struminskaya et al. (2015) used multilevel models to examine device effects when comparing surveys responses completed on PCs, tablet and smartphones. Their models did not include the device model as a level. Instead repeated measures were clustered within individual respondents. They reported intra-respondent correlations ranging from 0.16 to 0.62.

DATA

STUDY DESIGNS

Three datasets were used for the analyses in this research. The main data set is the *Understanding Society* Spending Study 1 (SS1), supplemented with data from the *Understanding Society* Innovation Panel and additional coding of characteristics of the device models that were used in SS1.

Innovation Panel Wave 9 (IP9). The *Understanding Society* Innovation Panel (University of Essex. Institute for Social and Economic Research, 2018a) is part of the UK Household Longitudinal Study the and is used for experimental and methodological research. The IP is an annual household panel survey with a stratified and clustered sample that is representative of the Great British population. Data from the ninth wave of the study are used. The wave nine sample consists of the remaining sample members from the first wave, together with two additional refreshment samples from waves four and seven onwards. Household members aged sixteen and over at the time of interviewing are considered eligible for annual interviews. The ninth wave had a household response rate of 84.7% and an individual response rate of 85.4% within responding households (Jäckle et al., 2018a).

Spending Study One (SS1). The *Understanding Society* Spending Study 1 (University of Essex. Institute for Social and Economic Research, 2018b) was an inter-wave data collection task that collected information about expenditure. The SS1 took place between IP9 and IP10, in autumn 2016. Sample members were asked to use an app on their own mobile device to submit information about purchases of goods and services. The app was developed by Kantar Worldpanel, with whom the study was conducted in partnership.

Respondents were asked to submit data in three forms: photographs of till receipts, self-reports of purchases, or reports of days without spending. More details, including the incentive structure, can be found in the SS1 user guide (Jäckle et al., 2018b).

There were 274 people who used the SS1 app at least once. This represents a response rate of 11.5% amongst the 2,112 IP9 respondents who were invited to participate. For the purposes of the analyses presented here, this sample was constrained to the 255 respondents about whom all relevant IP9 data was available.

Device characteristics data. The model of the device used to complete each app use was captured within the main SS1 app. There were 90 different models used by the analytical sample. The Spending Study app also captured the Operating System (OS). Whether the device was a tablet or smartphone was derived during data processing.

Specific characteristics of each model of mobile device were coded using the Amazon Mechanical Turk (mTurk) micro-task crowdsourcing platform. Screenshots of the Human Intelligence Task (HIT) used to collect the additional device characteristics can be found in Appendix B. Workers were presented with a device model[2] (e.g. Google Pixel) and asked to search for and input a series of device characteristics. Workers were paid $0.25 for each HIT they completed. Five device characteristics were collected using the HIT: the device's RAM (gigabytes or megabytes), processor speed (hertz), camera quality (megapixels), storage space (gigabytes or megabytes) and screen size (diagonally in inches). Each device was coded by three workers, and inter-coder reliability was calculated to assess the consensus of the three coders.

Of these five measures, only the device's RAM and camera quality were ultimately included as measures in the models presented here. Screen size was not included, as there was limited variation of screen sizes within tablets and with smartphones. The device type was the more important distinction, as opposed to the size of the screen.

The storage space variable that was captured was ultimately excluded from analysis as this was a very imprecise measure. The challenge is that the same model of device might be available in variants with different default storage capacity; for example, the Apple iPhone 6 is available in 16/32/64/128 GB versions. Whilst it was possible to capture the full range of available storage capacities using mTurk, it was not possible to determine exactly which variant the devices used in SS1 were, or whether two devices with the same model had different storage capacities. This is further complicated by some devices allowing the use of additional memory cards to provide extra storage

[2] The device names captured for iOS devices were the internal machine identifiers used by Apple, these match to the more commonly known product names, for example iPhone7,2 = iPhone 6. These were converted before the HIT was posted to make identification by mTurk workers easier.

The processor speed measure was problematic as some newer mobile devices use multiple cores in their CPUs. Many of the reported processor speeds only captured the performance of one core, not the total performance of the processor. An alternative source of data for the performance of device processors was used, details of this can be found in the measures section below.

MULTI-LEVEL STRUCTURE

Throughout the analyses in this research the data are considered to have a four-level cross-classified structure. This structure is illustrated in the classification diagram in Figure 3.

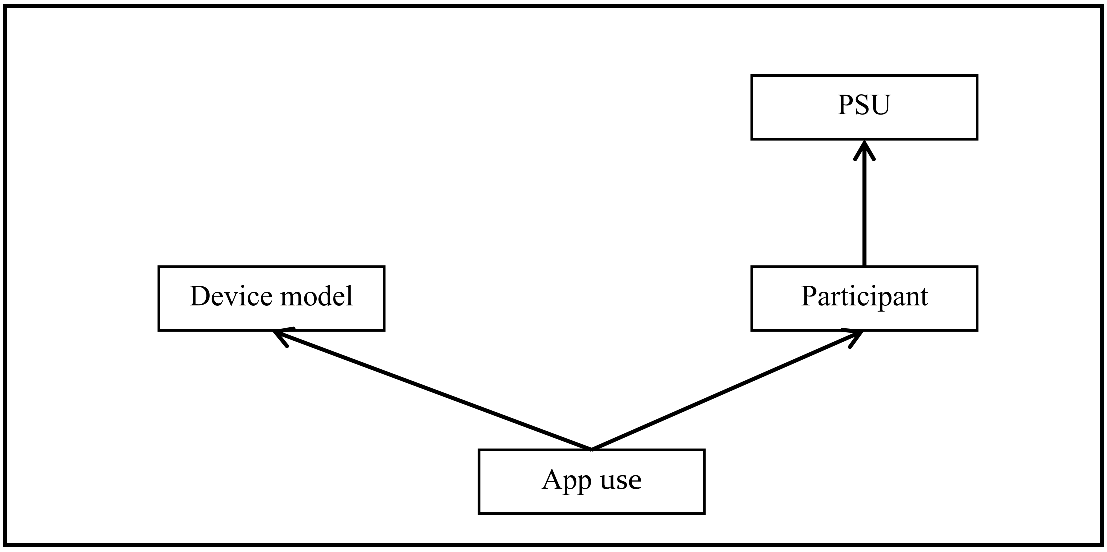


Figure 3
Classification diagram for the four-level cross-classified data structure.

The lowest level considered is individual app uses. Each app use is nested within two second-level clusters: the device model (e.g. all app uses completed on Apple iPhone 6s form one cluster) and the respondent who completed the app use. Finally, the Primary Sampling Unit (PSU) of the participant is included to account for the complex sample design of the Innovation Panel. Household was also considered as an additional level, but models fitted to include households suggest there was little clustering effect within households, so the more parsimonious four-level structure is presented.

MEASURES

DATA QUALITY INDICATORS - *APP USE LEVEL*

Without validation of the true measure of expenditure it is not possible to quantify measurement error directly. It is instead useful to examine the effect of device upon observable measures that are assumed to be correlated with measurement error. Four data quality indicators have been identified and are outlined below. Descriptive statistics for all four measures can be found in Table 12.

Table 12

Descriptive statistics of app use outcomes.

App use duration (seconds) (*n=10621*)	Mean	31
	SD	26
	Min	3
	Median	24
	Max	172
Classified as outlier in terms of app use duration (*n=10985*)	Yes	3%
	No	97%
Type of app use (*n=10985*)	Receipt scanned	48%
	Purchase without receipt	30%
	Report of nothing bought	22%
Was the receipt fully readable (*n=5263*)	Yes	92%
	No	8%
Number of items on the receipt (*n=4790*)	Mean	7
	SD	10
	Min	1
	Median	3
	Max	129

For all four measures, variance attributed to the respondent level might represent genuine variation in the outcome. However, the assumption is made that variance attributed to the device model represents systematic biases, and that these would be correlated with increased measurement error.

App use duration. Response times have previously been examined as a data quality indicator (Galesic and Bosnjak, 2009; Yan and Tourangeau, 2008; Malhotra, 2008). Typically, this has involved the assumption that shorter response times are indicative of satisficing, increasing measurement error. However, as is noted by Malhotra, the relationship between response time and data quality is not easily disentangled. When considering the effect of the device used on app use duration it does not make sense to suggest that faster devices result in lower quality data. In contrast, it seems likely that the opposite may be true, that slower devices result in poorer quality data. The justification for this is that slower devices may contribute to an increased perception of the time it takes to participate. The negative impact of longer perceptions of time taken to complete a survey on response propensity is well documented (Crawford et al., 2001; Galesic and Bosnjak, 2009; Collins et al., 1988; Roberts et al., 2010; Groves et al., 1999; Yammarino et al., 1991; Dillman et al., 1993). Respondents may be less motivated to accurately record all their purchases if their perception of how long it will take them to participate is increased by using a slower model of device.

The duration of app uses was measured in seconds. A number of extreme responses were again observed, and classified as outliers, using an adjusted boxplot. As suggested by Hubert and Vandervieren (2008) the medcouple (Brys et al., 2004), a robust measure of skewness, is applied to a boxplot to calculate an interval adjusted for skewed data. All data points outside the adjusted interval were coded as outliers. These outliers were excluded for models that regress app use duration on predictors. Separate models were then fitted to examine the associations between predictors and outlying durations. The mean app use duration was 31 seconds and the percentage of app uses with outlying durations was three percent.

Type of app use. A second data quality indicator is the type of app use. The three app use types were: taking a photograph of a receipt, manually entering data about a purchase, or reporting nothing bought that day. Here, the assumption is that app uses that are reports of purchases made without receipts, or of nothing bought, may be more likely to represent increased error if the 'true' response should have been a photographed receipt. This measure is included as a binary indicator of whether the app use was a scanned receipt, (coded as zero) or one of the other two types of app use (coded as one). Forty-eight percent of app uses were scanned receipts, and fifty-two percent of app uses were either purchases without receipts, or reports of nothing bought.

Image quality. A third data quality indicator analysed is the quality of the images of receipts. Here the data quality assumption is that poorer quality images increase the potential for error. Either because information cannot be coded from them or because the information coded may be incorrect. This measure was a binary indicator with fully readable receipts coded as zero. Receipts that were partially or completely unreadable or missing were coded as one. Ninety-two percent of receipts were fully readable, and eight percent were either partially readable, unreadable, or missing. For both this measure, and the number of items on receipts (below), the number of respondents and devices is slightly reduced as some respondents never scanned a receipt.

Number of items on the receipt. The final data quality indicator is the number of items that were on the receipt. The assumption here is that shorter receipts as a result of device characteristics may represent a downwards bias caused by the device model. The mean number of items on receipts was seven.

DEVICE CHARACTERISTICS – *DEVICE LEVEL*

Operating System (OS). Five device characteristics were identified as possibly affecting data quality, descriptive statistics for all five can be found in Table 13. The first was the Operating System (OS). The app was available for iOS and Android, and the OS was captured within the app. Differences in the software architecture of the two operating systems were the main reason for suspecting that the OS of the device used may affect data quality. For example, iOS and Android differ in how they handle memory allocation, which can have a significant effect on app speed and processing performance (Rinaldi, 2017; Brownlee, 2019; Lee, 2018). Amongst the device models used in the SS1, 29% were iOS devices and 71% were Android devices.

Mobile device type. The second device characteristic considered was device type: whether it was a smartphone or a tablet. Existing research has found differences between smartphone and tablet responses in surveys. Some evidence has suggested that responses to surveys using tablets are more similar to PC responses than smartphone responses (Struminskaya et al., 2015). The device type was derived during data processing.

The difference in size between tablets and smartphones was considered relevant for two reasons. The first of these is that the increased size of tablets may potentially make it more difficult to take photographs, as they are potentially bulkier and more cumbersome for respondents to use to take the photograph. However, the increased screen size may also have made it easier to see the photograph as it was being taken, potentially resulting in higher quality images. Twenty-two percent of devices were tablets, and seventy-eight percent of devices were smartphones.

Table 13.
Descriptive statistics for the device characteristics.

Operating system	Apple	29%
	Android	71%
Device Type	Smartphone	78%
	Tablet	22%
RAM (Gigabytes)	Mean	1.8
	SD	1.0
	Min	0.5
	Median	1.5
	Max	4.0
Camera quality (Megapixels)	Mean	9.6
	SD	5.0
	Min	0.7
	Median	8.0
	Max	20.7
Processor performance score	Mean	2.1
	SD	1.5
	Min	0.2
	Median	1.6
	Max	9.0

Notes: *n* = 97 devices.

Camera quality. The third device characteristic was the quality of the main camera on the mobile device, measured in megapixels. This was coded in the mTurk data collection. For 80% of devices the three workers were in perfect agreement as to the value of the quality of the camera. The corresponding kappa statistic of $\kappa = 0.83$ was above the 0.80 threshold describe as '*almost perfect*' agreement (Landis and Koch, 1977). Similarly, the value for Krippendorff's alpha was above the recommended 0.80

threshold (Krippendorff, 2004) at $\alpha = 0.84$. For each device the modal camera quality value for the three coders was selected. In two cases where the coders were in disagreement, the author obtained the value from the manufacturer's website. The mean camera quality of devices was 9.57 megapixels.

RAM. The fourth device characteristic was the amount of RAM available on the device. This is the amount of available immediate storage for software that is running. This was coded in the mTurk data collection and measured in gigabytes. For this measure all three coders were in perfect agreement 96% of the time and both the kappa statistic of $\kappa = 0.98$, and Krippendorff's alpha at $\alpha = 0.95$ suggest a high level of agreement amongst coders. Again, the modal RAM across coders for each model was selected. In one case the coders were in disagreement, the author again obtained the value from the manufacturer's website. The mean RAM of devices was 1.79 GB.

The available RAM on mobile devices only comes in a select number of values, measured in half or whole gigabyte increments. Alternative specifications of models fitting RAM as an ordinal measure were considered. These models met the proportional odds assumption, and as RAM is technically a continuous measure, the continuous variants of the models are reported.

Processor performance. As was mentioned earlier, the processor performance measure captured in the mTurk data collection did not account for some devices having multiple cores, meaning an alternative measure was needed. This was scraped from the Geekbench (2018) database of processor performance scores. Geekbench's database contains multiple records for each device model, so the median value was selected. Double the score represents double the processing performance. The wide range of the original measure meant that interpretation of coefficients was difficult, so all processor scores were divided by one thousand to make interpretation easier. The mean processor performance score was 2.13.

RESPONDENT CHARACTERISTICS – RESPONDENT LEVEL

One of the challenges in examining device effects is disentangling the direct effect of device characteristics from the indirect effects of selection. Lugtig and Toepoel (2015) suggested that selection accounted for most of the observed device effects in their study. This finding was based on respondents who had completed successive waves of a survey on device types such as mobiles devices, laptops, or PCs. Instead, this research focuses on the more granular effects of specific device models.

To account for selection five respondent characteristics have been included in the models presented in this research. These have been selected based on a combination of existing literature suggesting they are related to device selection and a previous paper by Struminskaya et al. (2015) that documented respondent characteristic controls. All five characteristics are taken from IP9 and descriptive statistics can be found in Table 14.

Sex. The first of these respondent characteristics was the respondent's sex. This has previously been found to be related to device selection (Karjaluoto et al., 2005)**.** Sex was also one of the respondent characteristics controlled for by Struminskaya et al. (2015). Male respondents were coded as zero and female respondents were coded as one. Amongst respondents in the analytical sample 39% were male, and 61% were female.

Age. The second respondent characteristics was age. Age has previously been found to be a predictor of technical ability using a mobile device (Loges and Jung, 2001). Struminskaya et al. (2015) found age to be a significant predictor of all the data quality indicators they examined. This was a continuous variable measured in years and the mean age of respondents in the Spending Study was 43.

Table 14.
Descriptive statistics for respondent characteristics.

Sex	Male	39%
	Female	61%
Age (years)	Mean	43
	SD	15
	Min	16
	Median	42
	Max	86
Equivalised gross monthly household income (£)	Mean	2344
	SD	1242
	Min	116
	Median	2146
	Max	7921
Employment status	Management	36%
	Intermediate	15%
	Routine	18%
	Unemployed	4%
	Retired	15%
	Inactive	11%
Highest level of education	Degree or higher	55%
	Lower than a degree	45%

Notes: *n* = 255 respondents.

Equivalised gross monthly household income. The respondent's level of household income was also included as a relevant respondent characteristic. No previous literature was found that provided evidence that level of income affects device selection. Price however has been found to be a factor in device selection (Sarker and Wells, 2003), and it was considered plausible that level of income would be correlated with how much a respondent was willing or able to pay. Gross monthly income was equivalised using the modified OECD scale to account for differences in household composition. The mean equivalised gross monthly household income was £2,344.

Employment status. Social class has previously been found to be related to device selection, with different factors being important to white-collar and blue-collar workers when making device selection decisions (Karjaluoto et al., 2005). Struminskaya et al. (2015) found differences in data quality indicators in a mobile survey, based on whether a respondent was in paid employment. Employment status was measured using the three category NSSEC classification, which classifies those in paid employment into management (36%), intermediate (15%) and routine (18%) plus categories for respondents who were unemployed (4%), retired (15%) and inactive (11%).

Level of education. The final respondent characteristic was the respondent's level of education. This was also previously found to be a significant predictor of data quality indicators in a mobile survey (Struminskaya et al., 2015). This was categorised into those whose highest level of qualification obtained was a degree or higher (55%), and those whose highest level of qualification was less than a degree (45%).

RESULTS

RQ1: What proportion of the variance in data quality indicators can be attributed to the device model used to participate, and what proportion to the respondent?

To decompose the proportion of variance that can be attributed to the device model used to participate, a series of five four-level cross-classified regression models were fitted using Markov chain Monte Carlo (MCMC) methods of estimation. These models were estimated using MLwiN (Charlton et al., 2017) using the software's in-built MCMC estimation methods (Browne, 2017). All models were fitted with a monitoring chain of

50,000 iterations, a burn in length of 1,000 iterations and with a thinning factor of one. For the two continuous data quality indicators, duration and number of items on the receipt, the equation for the models is as follows:

$$y_{ijkl} = \beta_0 + f_{0l} + v_{0k} + u_{0jl} + e_{ijkl} \qquad (2)$$

where y_{ijkl} is the value of the respective data quality indicator for a given app use i performed by a given respondent j using device model k within PSU l. The coefficient β_0 is then overall mean across all app uses, all respondents, all device models, and all PSUs. The random PSU effect is f_{0l}, the random device effect is v_{0k}, the random effect of the respondent is u_{0jl} and e_{ijkl} is the residual difference of individual app uses. All four random terms are assumed to be normally distributed. For the three binary outcomes logistic variations of this model were fitted. The logistic link function, by definition, fixes the variance of the lowest level residuals such that $\sigma^2{}_e = \pi^2/3 \approx 3.29$ (for more details see Snijders and Bosker, 2012; Hox et al., 2017). Results from all five models that were fitted are presented in Table 15.

The Variance Partition Coefficient (VPC) measures the proportion of the total variance that is explained by each of the levels. The VPC is similar to an intraclass correlation coefficient (of which intra-interviewer correlations are an example). In many circumstances the two measures will be the same. However, in cross-classified models the VPC reflects the proportion of the variance attributed to each level in the model; whereas, the ICC measures the expected homogeneity between two lowest level units, based on their membership to all higher-level units (Leckie, 2013). The VPC is then the more useful in this circumstance, as it allows comparison to the clustering effects observed in the interviewer effects literature, whilst also allowing the size of any device effects to be estimated.

Table 15

Results of four-level cross-classified regression models of data quality indicators with no predictors.

	Duration		Duration outlier		Other activity types		Low quality image		Number of items	
	σ	*VPC*	σ	*VPC*	σ	*VPC*	σ	*VPC*	σ	*VPC*
PSU $\boldsymbol{\sigma^2}_{\boldsymbol{f0}}$	5.79	0.01	0.02	0.01	0.03	0.01	0.26	0.05	0.36	0.00
Device $\boldsymbol{\sigma^2}_{\boldsymbol{v0}}$	52.81	0.08	0.13	0.03	0.32	0.06	1.20	0.22	0.23	0.00
Respondent $\boldsymbol{\sigma^2}_{\boldsymbol{u0}}$	68.06	0.10	0.53	0.13	1.95	0.35	0.64	0.12	10.49	0.10
Residual $\boldsymbol{\sigma^2}_{\boldsymbol{e}}$	564.98	0.82	3.29	0.83	3.29	0.59	3.29	0.61	96.02	0.90
PSUs	90		90		90		89		89	
Devices	90		90		90		84		83	
Respondents	255		255		255		233		231	
App uses	10621		10985		10985		5263		4790	
DIC	97656		3086		12463		2519		35599	

For the app use durations, it was expected that the level of variance that was attributed to the respondent would be quite a bit larger than that which is attributed to the device. However, this was not the case, the proportion attributed to the respondent was 10% compared to the 8% attributed to the device model. The proportion of variance in whether the duration was an outlier or not was in line with the expected result. A greater share (13%) of the variance was attributed to the respondent than to the device (3%).

It was expected that for activity type a larger share of the variance would be attributed by the model to the respondent; at 35% compared to 6% this was the case. The share of the variance that was attributed to the device model was highest for image quality, at 23%. This compares to just 9% for the respondent. This was unexpected, whilst it was considered that the device used may be associated with image quality, it was not expected that almost a quarter of the variance in this measure would be attributable to the device used. Less than 1% of the variance in the number of items was attributed to the device used, in comparison the variance attributed to the respondent was 10%.

RQ2: Are specific device characteristics associated with data quality indicators?

To examine the effects of specific device characteristics the same models as in RQ1 were fitted, with the addition of the five device characteristics. These models are presented on the left-hand column under each data quality indicator in Table 5.

The models were again fitted in MLwiN, using the same MCMC estimation conditions. The addition of the device characteristics means that equation one becomes:

$$y_{ijkl} = \beta_0 + \mathbf{X\beta}_{\boldsymbol{k}} + f_{0l} + v_{0k} + u_{0jl} + e_{ijkl} \tag{3}$$

where $\mathbf{X}\beta_{\boldsymbol{k}}$ is the vector of the five device level predictor variables and their corresponding coefficients. The assumptions about the normality of the random terms remain unchanged. For the models of duration and outlying durations, the type of app use was introduced as a control, as this was highly predictive of duration.

For the logistic models, coefficients and variances have been rescaled to allow comparison of nested models, as recommended by Hox et al. (2017) and Snijders and Bosker (2012). This overcomes the issue of potentially inflated fixed or random effects when comparing to the null model, as a result of the residual variance being fixed in logistic models.

The Deviance Information Criterion (Spiegelhalter et al., 2002) is used as a diagnostic tool for assessing model fit. This balances the likelihood of the model with the number of estimators. A lower DIC indicates a better fitting model. The comparison made here is between the DIC of the device characteristics models (the left-hand models for each outcome in Table 5), and the DIC of the null models (presented in Table 4).

App use duration was statistically significantly associated with three device characteristics. The first was the OS, with app uses completed using on average six seconds longer ($\beta = 6.09,\ p < 0.01,\ 95\%\ CI\ [2.05, 10.13]$). App uses completed on tablets were on average seven seconds longer than those completed on smartphones ($\beta = 7.06,\ p < 0.01,\ 95\%\ CI\ [2.16, 11.96]$). Finally, increased RAM was associated with typically shorter app use durations. Each additional gigabyte of RAM was associated with durations just under five seconds shorter ($\beta = -4.78,\ p < 0.01,\ 95\%\ CI\ [-7.82, -1.74]$). Processor speed and camera quality were not statistically significantly associated with duration. The DIC for the null model of duration was 97656, compared to a DIC of 97661 for the device characteristics model. This suggests the model including device characteristics is potentially a poor fit for the data.

In terms of outlying app use durations, there were two device characteristics that were statistically significantly associated with a lower likelihood of an outlying duration. The first was OS, with Android devices having 44% lower odds of producing app uses with outlying durations ($OR = 0.56,\ p < 0.01,\ 95\%\ CI\ [0.35,\ 0.90]$). Increases in processor performance were significantly associated with a decreased likelihood of an outlying duration($OR = 0.84,\ p < 0.05,\ 95\%\ CI\ [0.69,\ 0.99]$). The other three device characteristics were not statistically significantly associated with the likelihood of app use durations being outlying. The DICs for the null model and the device characteristics model were the same, 3086, indicating that the model with the addition of the device characteristics is not an improvement in terms of how it fits the data. This perhaps is not surprising as the null model suggested that device only account for 3% of the variance in whether an app uses had an outlying duration. None of the device characteristics modelled were significant predictors of the type of app use completed.

Three device characteristics were significantly associated with image quality. Android devices ($OR = 3.14,\ p < 0.001, 95\%\ CI\ [1.61,\ 6.11]$), and tablets ($OR = 2.25,\ p < 0.05,\ 95\%\ CI\ [1.01,\ 5.03]$) were associated with an increase in the odds of producing a low quality image. Higher RAM was associated with lower odds of producing low quality images ($OR = 0.49,\ p < 0.05, 95\%\ CI\ [0.27,\ 0.90]$). This effect of RAM is likely to have one of two causes. Firstly, devices with less RAM available might be expected to be more likely to run out of available memory. If this happens, one likely outcome is the app would fail to capture an image at all. In addition, devices with lower RAM may produce lower quality images because the camera software restricts the amount of memory allocated, and therefore the quality of the images captured, due to the limited hardware resources available. The DIC of the null model of image quality was 2519, compared to a smaller DIC of 2511 for the device characteristics model. This suggests that the addition of the device characteristics improved the model fit.

For the number of lines, the only statistically significant association was the device type, with receipts scanned on tablets typically having one less item on them than those scanned on smartphones ($\beta = -1.50,\ p < 0.05,\ 95\%\ CI\ [-2.83,\ -0.17]$). Upon further investigation it was discovered that the average image size of receipts scanned on tablets was smaller than those on smartphones. It is believed that this was caused by different software libraries typically being used by tablets and smartphones for handling photography due to differences in camera hardware. The DIC for the null model was 35599, and the DIC for the device characteristics model was 35598. Again, this suggests that the inclusion of the device characteristics did not produce a better fitting model. This is not particularly surprising as the VPC for the null model for this outcome suggested that device accounted for less than one percent of the variance in the number of lines on a scanned receipt.

RQ3: Do any associations between device characteristics and data quality indicators remain after controlling for respondent characteristics?

To examine the potential effects of selection, respondent characteristics were introduced to each of the five models. The resulting models are the models presented on the right-hand column under each data quality indicator in Table 16.

The addition of the respondent characteristics means that for continuous outcomes equation two becomes:

$$y_{ijkl} = \beta_0 + \mathrm{X}\beta_k + \mathbf{X}\beta_j + f_{0l} + v_{0k} + u_{0jl} + e_{ijkl} \tag{4}$$

where $\mathbf{X}\beta_j$ is the vector of the respondent characteristics variables and their corresponding coefficients. The assumptions about the normality of the random terms remain the same.

All three associations between device characteristics and duration were diminished when controlling for selection. App uses completed using Android devices on average took four seconds longer to complete ($\beta = 4.12,\ p < 0.05,\ 95\%\ CI\ [0.69,\ 7.55]$) when controlling for respondent characteristics, compared to six seconds longer in the model without respondent characteristics. App uses completed on tablets were no longer statistically significantly different in terms of their duration when controlling for respondent characteristics. Finally, each additional gigabyte of RAM a device had was associated with app use durations that were a little under three seconds shorter ($\beta = -2.70,\ p < 0.05,\ 95\%\ CI\ [-5.21,\ -0.19]$), compared to just under five seconds shorter when not controlling for selection. This suggests that some of the observed device effects may be the result of selection.

Table 16.

Results of four-level cross-classified regression models of the five data quality indicators with device and respondent characteristics as predictors.

	Duration		Duration outlier†		Other activity types†		Low quality image†		Number of items	
	β	*β*	*OR*	*OR*	*OR*	*OR*	*OR*	*OR*	*β*	*β*
Android	6.09**	4.12*	0.56**	0.54**	0.87	0.92	3.14***	2.91**	0.60	0.65
	(2.06)	(1.75)	(0.24)	(0.27)	(0.22)	(0.27)	(0.34)	(0.36)	(0.73)	(0.75)
Tablet	7.06**	3.47	0.84	0.72	1.11	1.19	2.25*	2.16*	-1.50*	-1.61*
	(2.50)	(2.09)	(0.28)	(0.30)	(0.30)	(0.31)	(0.41)	(0.44)	(0.68)	(0.80)
Camera quality	0.11	-0.14	1.00	1.00	1.01	0.99	1.01	1.01	0.00	0.01
	(0.27)	(0.23)	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.05)	(0.11)	(0.11)
RAM	-4.78**	-2.70*	1.09	1.16	1.28	1.26	0.49**	0.50*	-0.65	-0.48
	(1.55)	(1.28)	(0.19)	(0.21)	(0.19)	(0.18)	(0.31)	(0.32)	(0.63)	(0.65)
Processor	-1.01	-0.74	0.84*	0.84	1.00	0.94	0.94	0.94	-0.34	-0.21
	(0.71)	(0.62)	(0.10)	(0.11)	(0.08)	(0.08)	(0.13)	(0.14)	(0.27)	(0.28)
Female		0.20		1.04		0.73*		1.14		1.86***
		(0.99)		(0.18)		(0.17)		(0.21)		(0.60)
Age (years)		0.20***		1.00		0.97***		1.02		0.04
		(0.05)		(0.01)		(0.01)		(0.01)		(0.03)
Employment status (Ref: Management)										
Intermediate		1.49		1.09		0.64*		0.56		0.01
		(1.45)		(0.24)		(0.24)		(0.35)		(0.84)
Routine		1.71		1.11		0.77		0.76		-0.56
		(1.44)		(0.25)		(0.24)		(0.36)		(0.87)
Unemployed		0.85		1.43		1.36		1.11		-1.96
		(2.87)		(0.47)		(0.49)		(0.74)		(1.81)
Retired		7.39***		1.85*		1.12		0.61		-0.57
		(2.02)		(0.32)		(0.31)		(0.46)		(1.13)
Inactive		4.32**		0.97		0.82		1.03		-0.60
		(1.79)		(0.32)		(0.29)		(0.43)		(1.10)

Continues from previous page

	Duration				Duration outlier†				Other activity types†				Low quality image†				Number of items			
	β		*β*		*OR*		*OR*		*OR*		*OR*		*OR*		*OR*		*β*		*β*	
			(1.05)				(0.17)				(0.18)				(0.24)				(0.62)	
Income			0.00				1.00				1.00				1.00				0.00	
			(0.00)				(0.00)				(0.00)				(0.00)				(0.00)	
App use type (Ref: Scanned receipt)																				
Purchase without receipt			-11.08***				0.65***													
			(0.53)				(0.13)													
Report of nothing bought			-33.23***				0.52***													
			(0.58)				(0.16)													
Constant	36.72***		36.35***		0.05***		0.06***		0.84		7.11***		0.10***		0.06**		9.08***		6.59***	
	(2.45)		(3.51)		(0.28)		(0.59)		(0.28)		(0.47)		(0.44)		(0.91)		(0.85)		(1.9)	
	σ	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*
PSU $\boldsymbol{\sigma^2}_{\boldsymbol{f0}}$	5.79	0.01	4.95	0.01	0.02	0.00	0.02	0.00	0.15	0.04	0.10	0.03	0.32	0.07	0.50	0.10	0.36	0.00	0.40	0.00
Device $\boldsymbol{\sigma^2}_{\boldsymbol{v0}}$	52.81	0.08	16.96	0.03	0.13	0.03	0.18	0.04	0.23	0.05	0.30	0.07	0.51	0.12	0.80	0.15	0.15	0.00	0.15	0.00
Respondent $\boldsymbol{\sigma^2}_{\boldsymbol{u0}}$	68.06	0.10	65.51	0.10	0.45	0.12	0.63	0.14	1.34	0.32	1.22	0.30	0.59	0.14	0.57	0.11	9.99	0.09	9.98	0.09
Residual $\boldsymbol{\sigma^2}_{\boldsymbol{e}}$	564.98	0.82	565.48	0.87	3.13	0.84	3.67	0.82	2.40	0.58	2.44	0.60	2.82	0.67	3.43	0.65	96.02	0.90	95.89	0.90
PSUs	90		90		90		90		90		90		89		89		89		89	
Devices	90		90		90		90		90		90		84		84		83		83	
Respondents	255		255		255		255		255		255		233		233		231		231	
App uses	10621		10621		10985		10985		10985		10985		5263		5263		4790		4790	
DIC	97661		94741		3086		3064		12466		12465		2511		2512		35598		35596	

Notes: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; † Coefficients and variances rescaled for logistic models to allow comparison of nested models as recommended by (Hox et al., 2017; Snijders and Bosker, 2012); Standard errors in parentheses.

Three respondent characteristics were significantly associated with app use duration: age ($\beta = 0.20,\ p < 0.001,\ 95\%\ CI\ [0.10,\ 0.30]$), being retired ($\beta = 7.39,\ p < 0.001,$ $95\%\ CI\ [3.43,\ 11.35]$) and being otherwise inactive in terms of employment ($\beta = 4.32,$ $p < 0.01, 95\%\ CI\ [0.81,\ 7.83]$). The DIC for the model including respondent characteristics dropped quite significantly, from 97661 to 94741 suggesting the addition of respondent characteristics substantially improved the goodness of the fit of the model.

For outlying app use durations, the device's OS remained significantly associated, with very little change in the magnitude of the effect ($OR = 0.54,\ p < 0.01,$ $95\%\ CI\ [0.32, 0.92]$). The processor performance of the device had been significant, however the addition of the respondent characteristics resulted in a nonsignificant result. The only statistically significant respondent characteristic was that retired respondents had a higher likelihood of having an outlying app use duration ($OR\ = 1.85,$ $p\ < 0.05, 95\%\ CI\ [1.34,\ 2.46]$). Again, the decrease in the DIC (3086 to 3064) suggests that the addition of the respondent characteristics improved the fit of the model.

Whilst none of the device characteristics included in the RQ2 model for activity type were found to be significant, the possibility was considered that a relationship might be seen when controlling for respondent characteristics. Therefore, the respondent characteristics model was also fitted for this outcome. However, the device characteristics all remained not statistically significantly associated with activity type.

All three device characteristics that were statistically significantly associated with image quality remained significant when controlling for respondent characteristics. The first two of these had slight reductions in the size of their odds ratios: $OR\ = 2.91,\ p < 0.01,$ $95\%\ CI\ [1.44,\ 5.89]$ down from an odds ratio of 3.14 for the OS; and $OR = 2.16,\ p <$

$0.05,\ 95\%\ CI\ [1.03,\ 4.55]$ down from an odds ratio of 2.25 for tablets compared to smartphones. However, these reductions were relatively small, and this stability of estimates between models supports that there are some direct effects of these device characteristics. The coefficient for the third significant predictor of image quality, the device's RAM, changed very little $OR = 0.50,\ p < 0.05,\ 95\%\ CI\ [0.27,\ 0.94]$ compared to a value of 0.49 in the device characteristics only model. None of the respondent characteristics were significantly associated with image quality.

Finally, the association between number of items and device type remained. The coefficient for this relationship changed little when controlling for selection $\beta = -1.61$, $p < 0.05,\ 95\%\ CI\ [-3.18, -0.04]$ (compared to $\beta = -1.50$ previously). Gender was significantly associated with the number of items on scanned receipts, with female respondents typically submitting receipts that were nearly two lines longer $\beta = 1.86$, $p < 0.001,\ 95\%\ CI\ [0.68,\ 3.04]$. The slight decrease in the DIC (35596 compared to 35598) suggests the model with both sets of characteristics was a better fit for the data.

DISCUSSION

This paper expands upon the existing device effects literature by moving beyond comparing the broad categorisations of smartphone, tablet and PCs. This paper is the first, to date, to consider device effects at the more granular level of device model. To achieve this, multilevel models were fitted that consider the clustering effect of survey app reports that were completed using the same model of mobile device.

This research also explored which characteristics of mobile devices might be contributing to any observed device effects. Some device characteristics were captured in the data collection task (the *Understanding Society* Spending Study) itself. To supplement this selected device characteristics were coded using workers from Amazon mTurk to complete data collection. To the best of the author's knowledge, this paper is the first example of using mTurk to search for additional data. It may be possible to harness mTurk to collect other types of paradata, or perform other data processing tasks, such as coding of textual responses. Amazon mTurk is a fast and inexpensive way of achieving these types of tasks.

The results of RQ1 suggest that there were device effects in the Spending Study 1. The device level VPCs ranged from <0.00 to 0.22, which is of a similar magnitude to those reported within the interviewer effects literature (e.g. O'Muircheartaigh and Campanelli, 1998; Jäckle et al., 2011; West and Olson, 2010). Based on these results, further investigation into the potential for device effects seems to be warranted. It would be useful to examine whether mobile device model clustering effects are found when considering the kinds of data quality indicators traditionally examined in questionnaire-based surveys: e.g. straightlining, acquiescence, mid-point responding, item nonresponse, and primacy effects.

One of the results from RQ1 stands out, nearly a quarter (0.22) of the variance in the quality of the image was a result of the device model used. Whilst this measure if very specific to the context of the SS1, it does suggest that device effects may be more of a concern when mobile devices are being harnessed for enhanced data collection: e.g. asking respondents to take photographs, collecting GPS data, collecting data from wearables. This is potentially problematic, and also warrants further study, as the ability to collect these kind of data has widely been regarded as an important part of the future of role of surveys (Link et al., 2014; Couper, 2013a).

From a survey design perspective, potentially having to take into consideration the wide variety of mobile devices available to respondents is daunting. This is without taking into consideration the variety of models of desktops and laptops that might also be used to respond to web surveys. The 90 devices used by the 255 respondents in the Spending Study suggest that even an approach of testing for the most commonly used devices may not be sufficient. Attempting to test a survey app or website on physical versions of this many devices is unlikely to be feasible, therefore alternative approaches may be needed. One approach could be to use services such as Amazon's AWS Device Farm, or Google's Firebase Test Lab that allow testing of apps or websites across many digital emulations of physical devices.

With regards to RQ2, two of the most important device characteristics across the five measures were the OS and whether the device was a tablet or smartphone. This perhaps suggests that comparisons between categories, as has previously been the case in the device effects literature, may suffice. However, a third characteristic, the amount of RAM a device possesses, was also related to more than one data quality indicator. This is more problematic, though perhaps could be overcome through careful consideration at the survey design stage.

As in RQ1, the quality of the images produced when scanning receipts was the only outcome where there was particularly convincing evidence of device effects. One surprising finding was that the quality of the camera was not a significant predictor of image quality, whilst the RAM was.

In terms of selection effects, the evidence from RQ3 is consistent with some of the observed device effects being the result of selection. The image quality outcome was the main indicator where the device effects did not seem to substantially disappear when controlling for respondent effects. This seems to further support the idea that device effects are most problematic for outcomes that specifically rely on smartphone capabilities to perform tasks beyond those in a traditional survey.

It is important to acknowledge that this study is not without its limitations. Just as in Struminskaya et al. (2015) and Lugtig and Toepoel (2015) it is not possible to fully disentangle device effects from selection effects. Both studies attempted this by looking at transitions in the devices used, however this was not possible in SS1, meaning the only way to try to disentangle these two mechanisms was by using statistical controls. The success of identifying and controlling for relevant respondent characteristics is likely to always be limited. The challenge comes in identifying characteristics for which measures can be obtained, and that make good statistical controls, for example needing to be measured pre-selection (Gelman and Hill, 2006). Preferably, the solution to this issue would be an experimental design, allocating respondents to specific models of devices, however this is likely to prove prohibitive in terms of cost.

Secondly, without some form of validation for the data collected in the study, it is necessary to use indirect measures to look at data quality. A validation study that examined the effects of device models on sources of error would be a useful addition to the emerging literature on device effects.

CHAPTER THREE

AUTOMATED CODING OF DATA FROM SHOPPING RECEIPTS FOR SURVEY RESEARCH.

ABSTRACT

Shopping receipts offer a rich source of additional information to that obtained through traditional survey methods. However, the challenge of organic data such as receipts is extracting relevant information and curating this into a useful format. With receipts, one such useful format is the categorisation of item descriptions. Manual classification of data is a time consuming and costly process that is not easily scalable. Automating this process may help to reduce the costs in terms of time and effort, improving scalability.

INTRODUCTION

There is increasing interest in supplementing questionnaire data with other forms of data, for example, administrative (e.g. Calderwood and Lessof, 2009), or '*Big Data*' (e.g. Baker, 2017). Shopping receipts provide a rich source of data for studies of diet and health (e.g. Appelhans et al., 2017; Biediger-Friedman et al., 2016; Chrisinger et al., 2018; Cullen et al., 2007; Greenwood et al., 2006; Martin et al., 2006; Rankin et al., 1998; Ransley et al., 2003; Waterlander et al., 2013) and consumption or purchasing behaviour (e.g. Hendershott et al., 2012; Inman and Winer, 1998; Inman et al., 2009; Stilley et al., 2010).

Data extracted from receipts constitute what Groves (2011) has called '*organic data*'. Data of this type present challenges for research because the data generating processes are not under the control of the survey researcher. As a result, they may not always perfectly match the needs of the researcher, both in terms of their content and their format. The additional data processing needed to obtain the desired measure from the original data can come with significant costs, both in terms of time and resources. Automation may offer cost savings by reducing these costs.

Advances in information technology mean that automated data processing has increasingly become a feasible option. However, as things currently stand, automated approaches typically do not yet match the level of accuracy achieved in manual data processing. As has been noted in the discussion of automated coding of open-ended survey responses (Schonlau and Couper, 2016), a semi-automated approach may offer the right balance between cost-saving and accuracy. Using automated coding methods that provide probability thresholds for the processed results may allow most of the processing to be completed automatically, whilst identifying cases that would benefit from being reviewed manually.

For manual coding it has long been considered best practice to task multiple coders with independently coding the items, then combining the sets of codes (Saldaña, 2015). Similarly, the potential for combining multiple automated processes to produce a more accurate final code has been raised. In the field of machine learning such combinations of individual classifiers have been termed ensemble methods. The largest gains in accuracy of using an ensemble have been found when combining models that are both accurate and diverse (Dietterich, 2000).

Whilst it is important to try to maximise the overall accuracy of the data processing, it is also important to consider how different data processing algorithms may introduce biases to the data. An automated approach that produces the highest overall accuracy may not be preferred if a less accurate model produces results with smaller biases.

This study uses images of shopping receipts collected as part of the *Understanding Society* Spending Study 1 (University of Essex. Institute for Social and Economic

Research, 2018b). Descriptions of purchased shopping items from this data set have previously been categorised, providing validation to allow different automatic approaches to be tested for their performance. This research uses this as an opportunity to examine the following research questions:

RQ1: How accurately can automated approaches code data from shopping receipts?

RQ2: Does a thresholding approach, that identifies cases for manual checking, offer an improvement on the fully automated model?

RQ3: Can an ensemble approach provide greater accuracy than single-method approaches?

RQ4: To what extent do these automated processes introduce biases to the data produced?

BACKGROUND

Survey research as a field is increasingly recognising the value of supplementing questionnaire-based surveys with additional forms of data collection (Groves, 2011; Couper, 2013b; Smith, 2013). Couper (2013b) has used the metaphor of an expanding toolbox to describe this. Surveys are but one tool amongst many that can be used to uncover information about the social world. Similarly, Breiman (2001b) makes the argument for the need for both stochastic and algorithmic data modelling for obtaining information from data. In both instances the argument is made for greater diversity in both the sources of data (and corresponding data generating processes) and in the methods used for processing and analysing that data.

In his discussion of the new types of data available to survey researchers, Groves (2011) makes an important distinction between what he calls designed and organic data.

Designed data he defines as being generated with a predefined research purpose in mind. This has been the mainstay of survey research throughout most of the field's history. Organic data, in contrast, emerges as a product of non-research processes within society. The purpose for which this data is generated is not the purpose for which the researcher intends to use the data. This difference in purposes introduces new challenges for obtaining useful information from the data.

Three main types of organic data have been identified: administrative data, collected for regulatory or other governmental purposes; social media data, created and curated by users as a presentation/expression of themselves; and transactional data, generated as an automatic by-product of an individual's transactions and activities. Receipts fall under the last of these three groups.

To provide a framework for understanding the process of obtaining information from organic data sources survey researchers have turned to the Extract, Transform and Load (ETL) process framework, well established in database management (Baker, 2017; Biemer, 2016). Extraction involves identifying the range of sources from which the data originates, validating the data, and collating it together into a common data set. Transformation then involves tasks that alter the data to allow it to be used for research purposes. This might involve coding or recoding, producing aggregated or disaggregated measures or otherwise editing the data to make it fit for purpose. Loading refers to the storage and database management of the final curated data set. This research focuses on the second of these three phases, the transformation of organic data into a format usable for research purposes.

TRANSFORMATION

Transformation is a crucial step in allowing organic data to be repurposed for research purposes. A number of examples can be found of transforming various different types of organic data for integration into research contexts. These include: open-ended answers to survey questions (Fielding et al., 2013; Giorgetti and Sebastiani, 2003; Schonlau and Couper, 2016), social media data (Ceron et al., 2014; Murthy, 2015; Resch et al., 2018; Schwartz and Ungar, 2015), administrative data (Brignone et al., 2018; Dehghani et al., 2015; Gundlapalli et al., 2014), and transactional data (Huang et al., 2005; Patel et al., 2015).

In the case of shopping receipts, transformation involves taking an item description coding it to an expenditure category. For example, a receipt might list '*Fresh Milk*' amongst the items purchased, whereas the desired code for this item might be '*Food and groceries*'. The computational complexity of the method for completing this transformation can vary depending on the technique used, ranging from less computationally complex methods such as simple string matching, through to statistical learning models, or even more computationally complex techniques, such as deep learning models.

A direct parallel for coding item descriptions from receipts can be found in attempts to automate the coding of occupational and industrial codes from open-text survey responses. This parallel offers insights into the appropriate methodology for gaining usable data from item descriptions on shopping receipts.

Perhaps the simplest approaches involve exact string matches between the uncoded description and a dictionary of strings that have been coded. An example of this approach can be found in the work of Ossiander and Milham (2006). Their approach counted the number of strings in a description that matched dictionary terms, and then applied the code for which the most strings matched. They implemented some simple rules for dealing with ties, and descriptions that were still tied after applying these rules were referred for manual inspection.

The Office for National Statistics in the UK have developed a web-based tool for occupation coding (Office for National Statistics, 2018). The tool uses three edit distance measures to suggest a code based on an occupation description. One is the standard Levenshtein distance (Levenshtein, 1966), the other two are custom string distance metrics that capture the difference in spelling, and difference in phonetic sound between a search term and entries in a database of occupation codes. (Kirby et al., 2015) have also previously used the Levenshtein distance for occupation coding.

The Computer Assisted Structured Coding Tool (CASCOT) developed at the Institute for Employment Research at the University of Warwick (Jones and Elias, 2004) uses a rule-based system to suggest a code. In addition to the suggested output code the software provides a certainty score (ranging from 0-100) to express the probability that the suggested code is correct. Recommendations for thresholds based on this certainty score have been offered, with the suggestion that scores above 80 should be accepted, scores between 60-80 can be reasonably sure of correct coding, scores between 40-60 suggest some ambiguity that would benefit from additional information, and scores below 40 being inconclusive (Ellison, 2010). It has been suggested that the optimum single cut-off for the certainty score is 64 (Institute for Employment Research, 2015).

Statistical learning models require the transformation of textual data into numeric data to allow the use of a stochastic or algorithmic model to predict a suggested code as an outcome. Typically, a series of binary indicators are constructed, representing an n-gram, which is a unit of lexical meaning. An n-gram consisting of one word is called a unigram, two words make a bigram, and three words make a trigram. For each record the corresponding n-gram is coded as one if that n-gram appears within the description, or zero if it does not.

Traditional statistical models such as a multinomial logistic regression have been used to predict occupational codes (Nahoomi, 2018; Thompson et al., 2012; Kirby et al., 2015). In addition, a variety of statistical learning (sometimes called machine learning) techniques have been applied to automating this process. These range from k-nearest neighbour approaches (Creecy et al., 1992; Jung et al., 2008; Russ et al., 2014), through to support vector machines (Gweon et al., 2017; Nahoomi, 2018), Naïve Bayes classifiers (Burstyn et al., 2014; Nahoomi, 2018; Schierholz, 2014; Kirby et al., 2015) and Gradient Boosting Machines (Schierholz, 2014). One recent effort has seen the application of so called '*deep-learning*' techniques to this task, using a Convolutional Neural Network to classify occupations (Nahoomi, 2018).

In addition to these individual classifiers, there has been increasing interest in the possibility of combining classifiers to produce an ensemble classifier. Such a classifier takes the predictions from a set of classifiers and combines them through either weighted or unweighted vote system to produce the final classification (Dietterich, 2000). Traditionally, this has resulted in what have been termed homogenous ensembles, classifiers that fit multiple models of the same type and combine the results of those models. Gradient Boosting Machines offer one example of a homogenous ensemble. In contrast, heterogenous ensembles combine classifiers based around different algorithms to fit an overall classifier that has been found to have improved performance, reduced bias, and to produce better estimates of probabilities of being in each class (Large et al., 2017). Heterogenous ensembles have previously found some success when applied to occupation coding (Russ et al., 2016; Kirby et al., 2015).

DATA

The main data set used for this research is the *Understanding Society* Spending Study 1 (University of Essex. Institute for Social and Economic Research, 2018b), this was a supplementary data collection task situated between waves nine and ten of the *Understanding Society* Innovation Panel (University of Essex. Institute for Social and Economic Research, 2018a). Some covariates for the analyses presented here are taken from the Innovation Panel annual interview.

INNOVATION PANEL

The *Understanding Society* Innovation Panel (IP) forms part of the UK Household Longitudinal Study, more commonly known as *Understanding Society*. The IP is used for methodological and experimental research to inform the design and content of the main *Understanding Society*. The same stratified and clustered sample design is used for the IP as in the main *Understanding Society,* and the IP sample is representative of the population of Great Britain. The ninth wave of the study (IP9), the wave prior to the Spending Study 1, consists of the original IP sample together with two refreshment samples, introduced at IP4 and IP7. All household members aged sixteen and over at the time of interviewing are considered eligible for annual interviews. The IP9 household response rate was 84.7%, and there was an individual response rate of 85.4% within responding households (Jäckle et al., 2018a).

SPENDING STUDY ONE

The aim of the *Understanding Society* Spending Study 1 (SS1) was to try to provide a full account of household expenditure for IP members across a period of a month. Full details about the study design can be found in the SS1 user guide (Jäckle et al., 2018b).

Data collection involved an app that was developed by Kantar Worldpanel, with whom the study was conducted in partnership.

Respondents were asked to use the app to provide data about purchases they made during the study period. This could be provided by either photographing shopping receipts, self-reports of purchases, or reports of days without spending. It is the first of these three types of app use that is the focus of this research. There were 5,541 receipts submitted by SS1 respondents. There were 274 participants who completed at least one app use. This is a response rate of 11.5% amongst the 2,383 sample members who were invited to participate.

There were several incentives offered to participants in the study; these were in the form of either Love2Shop gift vouchers or gift cards. The initial incentive for completing the registration survey and downloading the app was either £2.00 or £6.00 depending on the experimental treatment group to which the household was allocated. At the end of the third week of the study an additional £5.00 conditional incentive was offered to all members of a random sample of half of all households where nobody had yet participated. For each day in which a participant used the app they received another 50p incentive. In addition, respondents were promised an additional £10.00 for using the app consecutively for 31 days.

The extraction of data from the raw images of the receipts was completed manually. This produced two datasets, one containing receipt level information, and another containing item level records. The 5,541 receipts contained 37,259 purchased items.

The item descriptions were then transformed by coding to match a set of eleven spending categories. These categories were derived from work carried out by the Institute for Fiscal Studies (d’Ardenne and Blake, 2012) and were also used in the app when respondents self-reported purchases instead of uploading images of receipts. The eleven categories were: Food and groceries; Clothes and footwear; Transport costs (e.g. petrol, car maintenance, public transport costs); Child costs (e.g. childcare, school equipment and fees); Home improvements and household goods (e.g. DIY, gardening, furniture, white goods or electrical goods); Health expenses (e.g. glasses, dental care, prescriptions, social care); Socialising and hobbies (e.g. going out (restaurants, pub, cinema, theatre, concert), gym or club membership, arts and crafts, children's activities); Other goods and services (e.g. books, magazines, DVDs, CDs, games, toys, beauty products, haircuts, manicures, massages); Holidays; Giving money or gifts to other people (e.g. money for children, gifts or money for relatives, donations to charity); and Other.

When the receipts were manually input, items were labelled as either a physical item, or another type of item commonly found on a receipt, such as a promotion, VAT, or gratuity. These other items were all manually coded. For those items that were labelled physical items the first step of the coding process was automated.

This automated process used Volume D: Expenditure codes 2015-16 of the UK Living Cost and Food Survey (LCF) User Guide (Office for National Statistics, 2017) to create a dictionary of words categorised within each of the eleven categories. This table is a list of item descriptions from the LCF that have been classified using the Classification of Individual Consumption by Purpose (COICOP). The first stage in constructing the dictionary involved matching these COICOP codes to the eleven categories used in the Spending Study 1. Each word that appeared in the LCF Expenditure code item descriptions was then matched to one of our 11 categories. Where a word matched more than one category, it was assigned to the category in which it appeared most frequently. Common stop-words (e.g. the, and, it) were removed from the dictionaries.

Once this dictionary of terms was created the item descriptions from Spending Study 1 were classified by calculating the number of words in each item description that matched with each of the categories. The category for which the most words in the item description matched was then assigned as a suggested code. Where there was a tie between two categories, no code was assigned. These automated codes were then manually checked, with a first coder suggesting alternative codes for all items they felt were incorrectly coded in the automatic coding. Where the first coder was unsure of a suggested code this was flagged for review by a second coder, who assigned a final code.

MEASURES

The main variable of interest is the item description that was extracted from the images of the receipts when the data from the receipts were manually input. Coders were instructed to enter these verbatim as they appeared on the receipts. These descriptions were then parsed and a series of binary indicators representing unigrams of the words in the item descriptions were derived. Table 17 illustrates the data structure for some example cases. The training dataset refers to the LCFS expenditure codes, which were used for the original semi-automated coding applied to the SS1 data, as well as the training dataset for the automated coding attempted in this research. The test dataset refers to the Spending Study 1 item descriptions.

Examples from both the training and test dataset are included for illustrative purposes; however, the two datasets were kept separate during the analysis. From the table it is clear that both datasets had a similar structure, with the only difference being the lack of an assigned category for the test data set. To simulate automated coding in the field the test dataset was treated as though manual codes were not available. The codes assigned for the Spending Study 1 data by the semi-automated approach outlined above were held in reserve as validation for the automated approaches applied.

Table 17.

Example of data structure and binary indicators for unigrams.

Dataset	Description	Category	Fresh	Milk	Double	Cream	Paint	Whole	Nuts	Bolts
LCF	'Fresh milk'	*Food and groceries*	1	1	0	0	0	0	0	0
LCF	'Double cream'	*Food and groceries*	1	0	1	1	0	0	0	0
LCF	'Cream paint'	*Home*	0	0	0	1	1	0	0	0
LCF	'Whole milk'	*Food and groceries*	0	1	0	0	0	1	0	0
LCF	'Whole nuts'	*Food and groceries*	0	0	0	0	0	1	1	0
LCF	'Nuts and bolts'	*Home*	0	0	0	0	0	0	1	1
SS1	'Fresh milk'		1	1	0	0	0	0	0	0
SS1	'Double cream'		1	0	1	1	0	0	0	0
SS1	'Cream paint'		0	0	0	1	1	0	0	0
SS1	'Whole milk'		0	1	0	0	0	1	0	0

METHODS

This section outlines the methods that were applied to the task of classifying physical items from the receipts obtained in SS1 to the relevant expenditure category. The first set of approaches consist of attempts to match strings in the SS1 item descriptions with a dictionary of terms for each expenditure category, derived from the LCFS item descriptions. The second set of approaches use statistical learning models to classify the item descriptions, with the LCFS data being used as the training dataset for these models. Throughout the descriptions of these methods the appropriate method for estimating the probability of class membership for use when applying thresholds to the predicted classifications are outlined. Finally, the different approaches to constructing a heterogenous ensemble model are outlined.

STRING-MATCHING APPROACHES

Two string-matching approaches were applied to the data to provide predicted classifications. In both cases, the dictionary of terms from the LCF Expenditure Codes that was generated for the original automated coding of the data was used. Similarly, the item descriptions from the SS1 shopping receipts were parsed into individual words.

Strict string-matching. This first approach matches the automated approach that was originally applied to the SS1 data to aid the manual coding. The probability of class membership was estimated as the proportion of words within the item description that were matched with dictionary terms, multiplied by the proportion of match terms that fell in the final predicted category. Where a code could not be assigned, the most frequent category, food and groceries, was assigned.

Approximate string matching. The second approach extends on the first by attempting to match those unigrams within the item descriptions that did not strictly match with a dictionary term. Approximate string matches were made for all dictionary terms with a Levenshtein edit distance (Levenshtein, 1966) of less than three. These approximate matches were then incorporated into the category count from the first string-matching method. However, where a unigram from the item description produced more than one approximate string match the count was downweighted by dividing by the number of matched terms to capture the uncertainty of the match. The probability of class membership was estimated in the same fashion as with the strict string-matching algorithm, with the uncertainty in the approximate matches incorporated into the predicted probability through the downweighted counts.

STATISTICAL LEARNING MODELS

There were seven types of statistical learning models that were used to automate the process of coding item descriptions, though some of these model types had more than one variation, resulting in ten statistical learning models in total. Many of these models require additional hyperparameters to be set, these are noted throughout the descriptions of the individual models. Where hyperparameters were required, these were set using a combination of a tuning grid of plausible values, and ten-fold cross validation when training the models. The best performing hyperparameter set were then used when applying the models to make predictions on the SS1 data.

Multinomial logistic regression. Multinomial logistic regression extends the standard logistic regression for use with a dependent variable with more than two classes. The multinomial logistic regression model used in this analysis, together with the models implementing the lasso and ridge regularisations (outlined below) are all fit using the glmnet package in R (Hastie and Qian, 2014).

Multinomial logistic regression – LASSO. LASSO regression (Tibshirani, 1996) uses L1 regularisation to shrink the absolute magnitude of coefficients to avoid overfitting, high-valued regression coefficients are penalised to reduce the risk of overfitting. One effect of this form of regularisation is that it also performs variable selection, as some of the coefficients shrink to zero and are excluded from the final model. The λ hyperparameter was set, which establishes the level of shrinkage in the model.

Multinomial logistic regression – Ridge. Ridge regression uses another form of regularisation, L2 regularisation, which introduces a penalty equalling the square of the magnitude of all the coefficients. All coefficients are therefore shrunk by the same factor. In the case of L2 regularisation, variable selection does not take place, as the shrinkage does not result in any coefficients shrinking to zero. Once again, the λ hyperparameter sets the level of shrinkage in the model.

k-nearest Neighbours. The k-nearest neighbours (knn) algorithm is a non-parametric classification algorithm, that assigns a class based on a vote of the classes of the k nearest points based on a distance metric (Altman, 1992). The knn algorithm is often implemented using the Euclidean distance metric, however this can result in a large number of ties when used for text classification. Therefore, two alternative distance metrics were implemented, that have previously been applied to knn classification of texts: the cosine similarity distance (Manning et al., 2010) and the Jaccard similarity distance (Ouyang, 2016). The dbscan (Hahsler et al., 2015) package in R was used to calculate the predicted class based on the corresponding distance metrics. The probability of the assigned class is then estimated from the proportion of the k-nearest neighbours from which the predicted class is assigned that fell within the assigned class. The value of k, that is, the number of points upon which the classification is based is a hyperparameter, set through cross-validation.

Support Vector Machines. Support Vector Machines (SVMs) (Vapnik, 1998) are non-probabilistic binary linear classifiers that fit a hyperplane to the data that divides the data into two separate classes of the outcome variable, whilst maximising the distance between the hyperplane and the support vectors, those data points at the edge of the class closest to the hyperplane. The SVMs applied to the data were fitted using the e1071 package in R (Meyer et al., 2019), which provides an interface for the libsvm C++ package (Chang and Lin, 2011).

In the case of an outcome with more than two classes multiple SVMs are fitted, the libsvm package uses the one-versus-one multiclass classification where $k(k-1)/2$ models are fitted (with k being the number of classes, in this case 11), each of which involves only two classes. A voting mechanism is then incorporated to assign the final suggested class, using the method outlined by Friedman (1996).

The performance of SVMs can be improved through the use of kernels (Hofmann et al., 2008). Where the classes of a dependent variable cannot be linearly separated in the dimensional space of the original independent variables, often transforming to a higher-dimensional space can allow linear separation to be achieved. For example, squaring the independent variables may separate the classes in this new higher dimensional space, where they were not previously separable. Transforming all of the independent variables can quickly become computationally intensive, therefore the use of a kernel allows the original vector of independent variables to be transformed into the dot product of the transformed variables. Calculating one value that represents the higher dimensional space is less computationally intensive than transforming all of the independent variables. In addition to the linear kernel, models with radial and polynomial kernels were fitted.

As SVMs are non-probabilistic, it is not possible to directly estimate class probabilities from the model. However, libsvm uses Platt scaling (Platt, 1999) to estimate the probability as a logistic transformation of the classifier scores, after the SVM is fitted. For the multiclass case libsvm implements the approach outlined by (Wu et al., 2004).

Random Forest. The Random Forest algorithm (Breiman, 2001a) is a homogenous ensemble, which fits a series of decision trees (Breiman et al., 1984) to assign a class, and then takes the modal classification across the fitted models. The Random Forest algorithm incorporates the technique of bootstrap aggregation (Breiman, 1996), whereby each decision tree is fitted using a random sample with replacement of the training data set. In addition, each decision node within any given tree selects a variable to split on from a random sample with replacement subset of the predictor variables. By applying bootstrap aggregation to both the sample and feature selection for each tree the correlation between trees is reduced, producing better estimates than an individual decision tree (Breiman, 2001a). The probability of the assigned class is then estimated by the proportion of the total trees fitted that predicted the modal predicted class, which is the final class assigned by the model. The ranger package in R (Wright and Ziegler, 2015) was used to fit the random forest model in this analysis.

Gradient Boosting Machine. Gradient Boosting Machines (Friedman, 2002; Friedman, 2001) are another example of homogenous ensembles, which fits decision trees iteratively, rather than independently, as the Random Forests algorithm does. In the case of classification, the initial decision tree returns the probability of class membership. A pseudo-residual is then calculated, that is, the difference between the predicted probability, and the observed class membership (either 0 or 1). This pseudo-residual is then the output of a subsequent model, predicted by the same set of predictor variables as the initial model. This process is repeated iteratively to minimise the size of the pseudo-residuals. In the case of multiclass classification, at each step a decision tree is fitted for each class, and the softmax function is used to produce k probabilities for class membership, that is one probability per output class. A variant of the standard GBM, extreme gradient boosting or xgboost (Chen et al., 2015) was used in these analyses. This adapts the GBM algorithm, as described above, to incorporate bootstrap aggregation of the sample for each iterative tree, and the set of features for nodes within a tree, as is the case in the Random Forests algorithm.

HETEROGENOUS ENSEMBLES

Heterogeneous ensembles offer the possibility of overcoming the potential shortcomings of any given prediction algorithm by combining them. Six different methods of combining the models described above are implemented. All six methods use some form of vote to assign the final category, which can result in ties between models. Throughout these ties are resolved at random.

Majority vote – Unweighted. For this ensemble the final predicted category is assigned by using the modal predicted category across all of the models. In effect, each model gets one vote for the final predicted class across all models.

Majority vote – Weighted. This second heterogeneous ensemble takes the principle of the above majority vote but weights the votes of each model according to the predicted probability of the final class selected by that model. Each vote is therefore downweighted with predictions of less certainty carrying less weight in the final vote.

Theoretically grouped majority – Unweighted. The twelve algorithms can be broadly categorised into five theoretical groups: the first contains the two string matching algorithms, the second contains the three regression models, the third contains the two k-nearest neighbours algorithms, the fourth contains the three support vector machines, and the final group contains the two tree based ensemble methods. Within each group the modal value was first selected, resulting in five predictions the modal value of which was then taken to provide the final prediction.

Theoretically grouped majority – Weighted. For this ensemble the models are again collapsed down into five groups based on type of model as described above. The final votes of these five predicted categories are downweighted by the average probability amongst the models in the group for the modal category.

Empirically grouped majority – Unweighted. Ensembles perform best when the different classifiers are uncorrelated with one another (Dietterich, 2000). If different models excel at modelling different types of items, then combining them should produce greater accuracy. By combining the more highly correlated models the heterogeneity of the predictions contributing to the final ensemble should be increased. The expectation is that this would improve the accuracy of the classifications made. The individual approaches were collapsed down into six groups where the predictions of each approach within a group had correlations that were greater than 0.75 with all other approaches in the same group.

Empirically grouped majority – Weighted. The same process of using correlations to collapse the models down into six groups is applied here as is described above. The final votes of these six predicted categories are downweighted by the average probability amongst the models in the group for the modal category.

THRESHOLDING

Thresholding makes use of the predicted probability of category membership that the model produces to determine whether the assigned category should be accepted or whether manual checking is required. This results in a semi-automated approach (as advocated by Schonlau and Couper, 2016). The desired effect is that this should help distinguish between correctly and incorrectly categorised items.

Three levels of probability threshold were selected. These are based on the CASCOT probability thresholds, where it is suggested that scores above 80 should be accepted, there is a reasonable level of confidence for scores between 60-80, scores between 40-60 are somewhat ambiguous, and scores below 40 are considered inconclusive (Ellison, 2010). Rather than ranges, three of these values were selected, with probabilities of 0.40, 0.60, and 0.80 taken as plausible thresholds. Two additional thresholds (0.90 and 0.95) were also examined to assess the effectiveness of stricter thresholds.

RESULTS

RQ1: How accurately can automated approaches code data from shopping receipts?

Two measures are used to assess the performance of the different automated coding approaches. The first of these is the overall accuracy measure. This is the percentage of item descriptions that were correctly coded. The second measure is the area under the curve of the Receiver Operating Characteristics curve (AUROC). This is a measure of separability, that is informative for understanding how well the model performs in distinguishing between classes. A higher AUROC suggests greater model performance (Hastie et al., 2009).

Table 18 documents the accuracy and AUROC for each of the automated methods used. These can be compared to a baseline, which assigns all item descriptions to the dominant class, '*Food and groceries*'. This baseline has an accuracy 72.7%, 95% CI [72.5%, 73.5%] and an AUROC of 0.50.

Table 18

Coding performance in terms of accuracy and separability of twelve automatic coding approaches.

	Accuracy	AUROC
Baseline	72.7%	0.50
Strict string match	77.4%	0.69
Approx. string match	78.4%	0.70
Logistic	76.8%	0.58
Lasso	77.6%	0.58
Ridge	79.2%	0.59
knn Cosine	74.7%	0.59
knn Jaccard	73.2%	0.58
LSVM	77.1%	0.59
PSVM	78.4%	0.57
RSVM	71.9%	0.66
Random Forest	80.4%	0.78
xGBoost	75.2%	0.57

Almost all of the automated methods applied were more accurate than the baseline, with the exceptions of the k-nearest neighbours algorithm using the Jaccard distance measure, and the Support Vector Machine using the radial kernel. All of the algorithms produced an AUROC that was higher than the baseline separability of 0.50.

The strict string-matching algorithm produced an accuracy of 77.4%, 95% CI [77.0%, 77.8%] and an AUROC of 0.69. The approximate string-matching algorithm improved on this performance slightly with an accuracy of 78.4%, 95% CI [78.0%, 78.8%] and an AUROC of 0.70.

The logistic regression model produced an accuracy of 76.8%, 95% CI [76.4%, 77.2%], the LASSO regression performed better with an accuracy of 77.6%, 95% CI [77.2%, 78.0%], both with an AUROC of 0.58. The ridge regression algorithm again performed better, with an accuracy of 79.2%, 95% CI [78.8%, 79.6%] and an AUROC of 0.59.

Of the two knn algorithms the implementation using the Cosine distance measure performed the best. This produced an accuracy of 74.7%, 95% CI [74.3%, 75.1%] and an AUROC of 0.59. The Jaccard distance implementation of knn was no more accurate than the baseline, at 73.2%, 95% CI [72.8%, 73.6%], however the separability of the model was slightly improved, with an AUROC of 0.58.

With regards to the Support Vector Machines, the linear and polynomial kernel performed similarly with accuracies of 77.1%, 95% CI [76.7%, 77.5%] and 78.4%, 95% CI [78.0%, 78.8%] respectively. The linear kernel produced an AUROC of 0.59, and the polynomial kernel produced an AUROC of 0.57. The SVM with the radial kernel produced a lower accuracy, at 71.9%, 95% CI [71.4%, 72.4%], but performed better than the other two SVM algorithms in terms of the AUROC, at 0.66.

Of the two tree-based algorithms, the Random Forest algorithm performed better than the Extreme Gradient Boosting algorithm. The accuracy for the Random Forest algorithm was 80.4%, 95% CI [80.0%, 80.8%], the highest of any of the algorithms used, and the AUROC was 0.78, also the highest of the algorithms used. The xGBoost algorithm produced an accuracy of 75.2%, 95% CI [74.8%, 75.6%] and an AUROC of 0.57.

RQ2: Does a thresholding approach, that identifies cases for manual checking, offer an improvement on the fully automated model?

Table 19 documents the effects on accuracy of only accepting the predicted codes of those items for which the probability of being in that category was higher than the probability threshold. The assumption here is that those items below the threshold would be categorised manually, and therefore are not considered for assessing the performance of the automatic coding methods. The change in accuracy column for each threshold documents the difference in percentage points accuracy between the coding without a threshold, and the accuracy amongst those items for which the probability was above the respective threshold The percentage of items that are uncoded and would require manual coding is also documented. In addition to the thresholds based on those found in the CASCOT coding two stricter thresholds were applied (0.90 and 0.95). Across all the models assessed these stricter thresholds either did not improve accuracy or even, in some instances, reduced accuracy (with one exception to this being a slight increase in accuracy when applying the strictest threshold to the polynomial support vector machine). The results of these stricter thresholds are therefore also reported in Table 19, but not discussed below. The percentage points change in accuracy at these different thresholds is reported in Table 19, the accuracies themselves, along with 95% confidence intervals are reported in Table C1 in Appendix C.

The 0.40 probability threshold did little to improve the performance across any of the different approaches used. This is because the percentage of cases falling below the threshold was very low, ranging from <0.01% to 0.02% of cases across all of the models.

The strict and approximate string-matching algorithms, Lasso regression, the k-nearest neighbours algorithm using the Cosine distance, polynomial support vector machine all saw increases of less than one percentage point in accuracy. Ridge regression and the Random Forest approaches had the greatest increases in accuracy at this level of the threshold, with increases of 1.2 and 1.0 percentage point respectively. Logistic regression and the radial support vector machine actually performed slightly worse when the threshold was applied, with decreases in accuracy of 0.7 and 1.4 percentage points respectively.

The 0.60 threshold excluded a distinctly higher percentage of cases than the 0.40 threshold across all the algorithms, ranging from 1.3% to 2.1% of cases.

Both of the string-matching algorithms saw a further percentage points increase in accuracy when applying the 0.60 probability threshold, rising to 1.7 for the strict match, and 1.6 for the approximate match. Likewise, the Random Forest algorithm had a further increase to 1.9 percentage points greater accuracy with the stricter threshold. The accuracy of the ridge regression was approximately the same with both the 0.40 and 0.60 probability thresholds. The 0.60 threshold also improved the accuracy of the linear support vector machine (1.1 percentage points) and the extreme gradient boosting algorithm (0.7 percentage points).

Table 19

Changes in accuracy after applying different levels of probability thresholds to predicted codes.

		0.40 threshold		0.60 threshold		0.80 threshold		0.90 threshold		0.95 threshold	
	Original accuracy	Δ accuracy	Not coded	Δ Accuracy	Not coded	Δ accuracy	Not coded	Δ accuracy	Not coded	Δ accuracy	Not coded
Strict match	77.4%	+0.8	0.1%	+1.7	2.0%	+2.8	3.5%	+2.9	8.1%	+0.5	12.2%
Approx. match	78.4%	+0.8	0.1%	+1.6	1.8%	+2.5	2.7%	+2.6	5.6%	+0.8	8.9%
Logistic	76.8%	-0.7	0.2%	+0.1	1.3%	+0.3	3.1%	+0.8	6.1%	-0.1	6.8%
Lasso	77.6%	+0.7	0.0%	+0.2	1.6%	+0.9	3.1%	+1.5	4.1%	+0.8	4.3%
Ridge	79.2%	+1.2	0.1%	+1.1	1.3%	+0.7	2.8%	+0.5	4.4%	+0.7	4.6%
knn Cosine	74.7%	+0.5	0.1%	-0.4	1.4%	-0.8	2.5%	-0.7	5.5%	-1.4	5.5%
knn Jaccard	73.2%	-0.3	0.1%	0.0	2.1%	+0.1	4.3%	+0.1	4.7%	-0.1	5.3%
LSVM	77.1%	0.0	0.1%	+1.1	1.3%	+0.3	3.6%	+0.1	5.2%	-0.3	5.6%
PSVM	78.4%	+0.7	0.1%	-0.1	1.9%	+0.6	4.2%	+0.8	4.8%	+1.7	4.8%
RSVM	71.9%	-1.4	0.2%	-0.5	1.7%	-0.8	2.4%	-0.8	7.7%	-3.3	12.7%
Random Forest	80.4%	+1.0	0.1%	+1.9	1.3%	+3.9	3.2%	+3.0	3.6%	+1.9	9.6%
xGBoost	75.2%	0.0	0.1%	+0.7	1.2%	+1.1	4.0%	+1.2	4.2%	+0.7	4.5%

The logistic regression, lasso regression and the k-nearest neighbours algorithm using the Jaccard distance measure all produced little or no increases in accuracy at this threshold (0.1, 0.2 and <0.1 respectively). The polynomial support vector machine had produced a slight increase in accuracy using the 0.40 threshold, however this disappeared when using the stricter threshold (from an increase of 0.7 to a decrease of 0.1 percentage points).

The k-nearest neighbours algorithm using the Cosine distance measure saw a decrease in accuracy of 0.4 percentage points when using the 0.60 threshold. The radial support vector machine produced a decrease in accuracy of 0.5 percentage point thresholds at this threshold.

Finally, the strictest threshold was probabilities greater than 0.80. The percentage of cases falling below this threshold ranged from 2.4% to 4.3% of cases across all of the models.

This stricter threshold resulted in an increase of 2.8 percentage points for the strict string-matching algorithm and 2.5 for the approximate match, once again producing a greater increase than when implementing a lower threshold. The same was true for the Random Forest algorithm, the accuracy of which increased by 3.9 percentage points. This was the largest increase in accuracy of any of the three thresholds.

This strictest threshold also produced an increase in accuracy of 0.9 percentage points for the lasso regression. As with the 0.60 threshold, both the ridge regression and extreme gradient boosting machine approaches increases in accuracy, by 0.7 and 1.1 percentage points, respectively. The logistic regression (0.3 percentage points), linear support vector machine (0.3 percentage points) and polynomial support vector machine (0.6 percentage points) all produced modest increases in accuracy at this strictest threshold. The k-nearest neighbours algorithm using the Jaccard distance measure showed little difference in accuracy at this threshold, an increase of only 0.1 percentage points.

Both the radial support vector machine and the k-nearest neighbours algorithm using the Cosine distance measure produced a decrease in accuracy at the strictest threshold, both decreasing by 0.8 percentage points.

RQ3: Can an ensemble approach provide greater accuracy than single-method approaches?

Six heterogenous ensemble methods were tested to examine their effect on the accuracy and separability, compared to the individual approaches examined in RQ1. Table 20 documents the performance of these ensembles, in terms of accuracy and AUROC. The best performing of the individual approaches, Random Forest, is also included in the table for comparison.

Table 20

Coding accuracy and separability of six heterogenous ensembles of individual approaches.

	Accuracy	AUROC
Random Forest	80.4%	0.78
Majority vote – Unweighted	77.4%	0.58
Majority vote – Weighted	80.8%	0.63
Theoretically grouped majority – Unweighted	76.3%	0.57
Theoretically grouped majority – Weighted	78.7%	0.60
Empirically grouped majority – Unweighted	77.2%	0.60
Empirically grouped majority – Weighted	79.1%	0.62

Of the six ensemble methods, only the weighted majority vote was more accurate than the Random Forest algorithm, with 80.8%, 95% CI [80.4%, 81.2%] and 80.4%, 95% CI [80.0%, 80.8%] accuracy respectively. The next most accurate heterogenous ensembles were the weighted empirically grouped (79.1%, 95% CI [78.7%, 79.5%]) and weighted theoretically grouped ensemble (78.7%, 95% CI [78.3%, 79.1%]). The unweighted empirically grouped ensemble was accurate in 77.2% (95% CI [76.8%, 77.6%]) of cases, and the unweighted theoretically grouped ensemble in 76.3% of cases [95% CI [75.9%, 76.7%].

However, in terms of separability, all of the ensembles performed to a lower standard than the Random Forest algorithm, which had an AUROC of 0.78. Again, the best performing of the ensembles was the weighted majority vote, with an AUROC of 0.63. This was followed by the weighted empirically grouped ensemble, with an AUROC of 0.62. The weighted theoretically grouped and unweighted empirically grouped ensembles produced an AUROC of 0.60. The unweighted majority vote produced an AUROC of 0.58, and the unweighted theoretically grouped ensemble performed the worst in terms of separability, with an AUROC of 0.57.

RQ4: To what extent do these automated processes introduce biases to the data produced?

The first assessment of biases introduced by the automated processes examines the percentage of items classified that fell within each of the spending categories. Table 21 documents the percentage point differences between the true percentage of items in each of the spending categories, and the categorisations given by each of the automatic approaches.

Table 21

The size of biases introduced by different automated approaches across spending categories.

Category	*True %*	*Strict match*	*Approx. match*	*Logistic*	*Lasso*	*Ridge*	*knn Cosine*	*knn Jacard*	*LSVM*	*PSVM*	*RSVM*	*Random Forest*	*xGBoost*
Food	73	+8	+6	+10	+10	+10	+11	+12	+8	+10	-5	+5	+11
Clothes	3	-1	-1	-2	-2	-2	-2	-2	-2	-2	-2	-1	-2
Travel	1	0	0	0	0	0	-1	-1	0	-1	-1	0	-1
Child	1	0	0	0	0	0	0	0	0	0	0	0	0
Home	8	-4	-2	-3	-3	-3	-3	-5	-3	-2	+12	-1	-3
Health	1	0	0	0	0	0	0	0	0	-1	0	0	0
Social	3	-1	-1	0	0	1	-1	-1	-1	-1	0	-1	-1
Misc. goods	8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Holidays	<1	0	0	0	0	0	0	0	0	0	0	0	0
Gifts	<1	+1	+1	0	0	0	0	0	0	0	0	1	0
Other	<1	0	0	0	0	0	0	0	0	0	0	0	0
Absolute bias		16	12	16	16	17	19	22	15	18	22	10	19

Consistently, the automated approaches overestimated the percentage of items that were classified as Food and groceries. This was true for all the approaches, with the exception of the radial support vector machine, which underestimated the share of items in this category. The overestimates in this category ranged from 12 percentage points for the k-nearest neighbours algorithm using the Jaccard distance measure to five percentage points for Random Forest. The average overestimate was just over nine percentage points.

These overestimates typically came at the expense of underestimating the second and fourth most frequently occurring spending categories: home improvements and household goods and clothes and footwear, respectively. Home improvements was underestimated by all the approaches, apart from the radial support vector machine, which overestimated this category by twelve percentage points. The underestimates ranged from five percentage points for the k-nearest neighbours algorithm using the Jaccard distance measure to one percentage point for Random Forest. The average underestimate was just under three percentage points. Clothes and footwear were underestimated by all twelve approaches. Random forest and the two string matching approaches underestimated this by one percentage point, all the other approaches underestimated by two percentage points.

To assess the overall biases of the respective algorithm across all spending categories the absolute percentage point difference between the predictions and the actual categorisations was calculated. This is the absolute sum of the percentage points difference for each of the categories. The radial support vector machine and the k-nearest neighbours using the Jaccard distance measure were the two algorithms that introduced the greatest biases, with an absolute percentage points difference of 22. The

Random Forest and approximate string-matching approaches resulted in the smallest biases, with an absolute percentage points difference of 10 and 12 respectively.

DISCUSSION

The contribution of this research is to draw together a range of different techniques from the fields of data science and statistical learning and apply them to a specific survey data management task. Whilst the majority of the individual techniques have been explored individually elsewhere, as highlighted in the particular context of occupation coding, the aim is to present a more holistic overview of these different techniques to inform not just the particular task at hand, but best practice for processing textual data more broadly for survey research.

The comparison of different algorithmic approaches to transforming textual data from receipts (RQ1) offers insight into the comparative performance of these algorithms. The conventional wisdom is that more computationally complex algorithms, in particular, the homogenous ensembles, such as the Random Forests and Extreme Gradient Boosting algorithms, will have greater predictive power. Whilst the Random Forest algorithm was the approach with the best performance, both in terms of accuracy and separability, there was no clear gradient between computational complexity and performance. In fact, the approximate string matching, one of the least computationally complex approaches, was the second highest performing approach. This highlights the importance of considering a variety of different approaches as the choice of an appropriate algorithm is likely to be context specific.

In terms of the effects of applying different thresholds, the evidence suggests that applying probability thresholds did not uniformly improve the performance across all the automated approaches used in this research. In fact, in some instances the thresholds decreased the performance of certain approaches. However, there was a consistent pattern that those algorithms that performed the best when unconstrained by a threshold, such as the Random Forest, and approximate string matching, improved the most in their accuracy when applying the thresholds. The threshold of 0.80 was able to produce around a three to four percentage point increase in accuracy for these approaches at the cost of around three or four percent of cases falling below the threshold. This effect had diminished returns beyond this threshold, with no instances of further improvements, and large numbers of uncategorised cases.

There was little evidence of improvements offered by implementing a heterogenous ensemble to combine the individual approaches (RQ3). Only one of the six ensembles, the weighted majority vote, was more accurate than the best single approach and all six performed worse in terms of separability. The results suggest that the ensembles that weighted the vote based on the probabilities of class membership performed better than their unweighted equivalents. This is likely to be because these produced fewer tied votes, which introduce additional noise as the ties are resolved at random. Of the two methods of improving accuracy, using probability thresholds seemed to perform better than using ensembles.

With regards to the biases the algorithm introduced to the data via their predictions (RQ4), the single largest bias was that the vast majority of the algorithms overestimated the number of cases in the most frequent category, food and groceries, at the expense of the second and fourth most frequent categories, household goods and clothes and footwear. With regards to the overall biases of these measures across all categories, it was again the Random Forest and approximate string-matching approaches that resulted in the smallest biases being introduced.

One further extension of this research would be to explore ensembles of models with probability thresholds applied, to consider whether the combination of these two methods of improving accuracy may yield greater improvements than either method individually. It may also be beneficial to explore whether more advanced ensembles, for example fitting a second stage model using the training data that takes the predicted classes of each of the individual models as predictors of the observed class.

This research is limited in that it covers only the second stage of the ETL framework. The initial extraction of the data would also benefit from automation. In the research presented here this was not feasible as the app did not optimise the process of capturing images of receipts sufficiently to allow Optical Character Recognition to obtain text from the images.

Ideally this research would have used training data that exactly matched the receipt data that was being coded. However, in the absence of coded receipt data, the LCF data from spending diaries was a sufficiently functional equivalent. Whilst the format of the two datasets are similar, the different data-generating processes that produced the two datasets means that it is likely that a training set of coded receipts may have performed better than the training dataset used. That said, the results achieved using a training dataset generated through a different process than the data that is intended to be processed suggests that this may be a feasible alternative, in cases where an appropriate training dataset is not available.

Finally, the level of specificity in the final coded outcomes was relatively high level. Realistically, the level of specificity that would be desirable for this task would be more granular, with items being classified according to a standardised classification scheme such as the COICOP coding scheme (United Nations, 2018). It is likely that the same approaches adopted here could be applied to coding to a greater level of specificity, however, it is likely that this would be a more challenging task, with the potential for errors increasing as specificity increased. One solution to this may be to make use of the hierarchical structure of a framework like COICOP to initially assign codes at higher levels and then increase the specificity with subsequent models within these initial high-level classifications. Such an approach may then benefit from using different models at different stages of this coding process.

~

CONCLUSION

This thesis explored some of the opportunities and challenges presented by new methods of survey data collection using mobile devices. Each of the substantive chapters in this thesis examined a particular challenge presented by the *Understanding Society* Spending Study 1. The key findings from each chapter are presented below, followed in turn by a discussion of some of the implications of these findings for survey research designs. This is followed by a discussion of the emerging landscape of opportunities and challenges for using mobile apps for survey research.

SUMMARY OF FINDINGS

Chapter one examined respondent burden within the context of a mobile data collection task. The first main finding was that subjective burden appeared to form independently of the objective burden. Secondly, subjective and objective burden differed in terms of how they changed over the course of the study. There was no evidence that subjective burden changed systematically across the study. However, there was equally no evidence that a respondent's level of subjective burden was constant over time. Instead, changes occurred without any discernible pattern. In contrast, objective burden systematically decreased over the course of the study, but it was not clear whether a learning effect or satisficing drove this. There was some evidence that higher cumulative objective burden resulted in a greater number of breaks in participation, however the observed effects were small.

Some of the implications of these findings for informing implementation of survey research designs are as follows. The first is to stress the importance of considering the subjective perception of burden in addition to the objective burden. The finding that these arose separately from one another indicates careful consideration must be given to both in the design process. The second is to highlight the need for a better understanding of burden as a dynamic concept across the time in which a sample member is participating.

Chapter two examined the potential impacts of device effects as a source of error. Here evidence was found for device effects of a similar magnitude to those previously found in studies of interviewer effects. The device effects were largest when examining outcomes that required the use of hardware features of a smartphone beyond displaying and entering text. Android devices, tablets and devices with lower RAM were all associated with lower data quality. There was some evidence that device effects are driven by selection, but this is less apparent for outcomes that make greater use of device hardware.

These findings have important implications for developing mobile data collection tasks. They further support the notion that device effects are a non-ignorable source of error. As such, when designing a study careful consideration should be given to maximising the standardisation of the delivery of the survey instrument across devices with different specifications. The evidence supports the idea that this could be most pertinent when the survey instrument makes use of additional hardware features of the device.

Chapter three examined the application of statistical learning techniques to automate the process of coding textual data. The coding method that performed the best, both in terms of accuracy and separability, was the Random Forest algorithm. Applying probability thresholds improved the accuracy of the best performing model, however, as the strictness of the threshold was increased the returns in terms of gains in accuracy diminished. The benefits of applying probability thresholds were found to be greatest when applied to the best performing coding methods. Combining different coding methods into heterogenous ensembles was found to offer little improvement over the best performing single model. All models consistently overestimated the occurrence of the most frequent coding category, though reassuringly this bias was smallest in the most accurate models. Similarly, the biases in terms of which types of respondents for whom receipt items were correctly coded were smallest in the most accurate models.

The key practical implications of these findings are as follows. It is feasible to use automated coding techniques to reduce the time necessary to process organic data. However, as advocated by (Schonlau and Couper, 2016), a semi-automated approach, making use of thresholds to flag those cases in need of manual inspection would be beneficial. In addition, the overestimation of the dominant class, at least partially driven by the overrepresentation of this class in the training data set, highlights the importance of careful consideration when selecting an appropriate training data set.

OPPORTUNITIES AND CHALLENGES FOR SURVEY RESEARCH USING APPS

Finally, this thesis considers the broader context of the emerging landscape of opportunities and challenges for using apps for survey research. As a means of uncovering some of the topography of this emerging landscape the discussion below highlights seven opportunities and seven challenges the field of survey research faces when embracing mobile technologies, these are presented in Table 22 below. Whilst this list of considerations is presented as a dichotomy it is important to stress that many of the challenges presented here offer opportunities to overcome them, and likewise many of the opportunities present their own challenges to overcome.

Table 22

Summary of opportunities and challenges of using apps for survey research.

Opportunities	*Challenges*
New types of data	Low response rates
Greater granularity	Second-level digital divide
Real-time data collection	Privacy and informed consent
Repeated measurements	Volume of data
Gamification	Complexity of data
Reaching hard to reach populations	New skillsets for survey researchers
Reduced burden	Competition with commercial products

OPPORTUNITIES

New types of data. As was highlighted in Chapter 1 of this thesis, one opportunity of using mobile apps for data collection is the potential to collect new types of data, such as: '*voice, photography, video, text, email* [and] *GPS*' (Link et al., 2014: 22). The authors of this report also highlight that new types of data not only enhance existing research contexts, but also present research opportunities that would not be possible using questionnaires.

Greater granularity. A second opportunity presented by mobile apps for survey research is the opportunity to obtain data for certain survey outcomes with greater granularity. This may be particularly pertinent for behavioural measures, which are known to be problematic due to recall error (Tourangeau et al., 2000). To better illustrate this, we can look to the example of the Spending Study 1. It seems extremely unlikely that participants could be expected to recall their spending with the same level of accuracy and granularity as was obtained through scanned images of receipts.

Real-time data collection. A third opportunity is presented in the ability to use mobile apps to enable collection of data in real-time. This may be particularly useful for attitudinal measures or subjective reflections on experiences that may be hard for participants to reconstruct after the fact (Schwarz, 2012; Tourangeau et al., 2000). One example of using mobile apps to collect real-time data can be found in the context of Ecological Momentary Assessments (Moskowitz and Young, 2006).

Repeated measurements. A fourth opportunity where mobile apps excel as a form of data collection is when the research context requires repeated measurements. Once the initial hurdle of having the participant install the app is surpassed (this is not an insignificant hurdle, as evidenced by the section on low response rates below) an app offers a good interface to which respondents can easily return to in order to continue participating. There is some evidence to suggest that in the case of collecting repeated measures an app-based survey instrument will have lower drop-out than a browser based alternative, and it is hypothesised this is driven by the app being easier to return to (Jäckle et al., 2019b).

Gamification. Gamification involves using elements of game design in the hope of increasing the engagement and reducing the burden of participating in a survey. As was discussed in Chapter 1, mobile apps as an interface are well suited to implementing aspects of gamification. A mobile app as an instrument allows for a more polished user experience that can incorporate aspects like achievement badges (Lai et al., 2012; Link et al., 2012) feedback (Wenz et al., 2020), or fully implemented games (Adamou, 2013).

Reaching hard to reach populations. The fifth opportunity presented by using mobile apps for data collection is that they may provide the means of surveying populations who are otherwise difficult to gain access to. For example, it has been highlighted that harnessing mobile technologies for social research may be beneficial in countries where there are not established networks of interviewers to conduct face-to-face research (Pfeffermann, 2019). Similarly, mobile apps have been harnessed for research into migration (Gillespie et al., 2016), where traditional survey methods may not be effective due to the high levels of mobility of the population.

Reduced burden. It has been suggested that replacing traditional survey questions with data collected through a mobile app may give us one opportunity to reduce respondent burden (Keusch et al., 2019). Whilst replacing actively answering questions with passively collecting data may be an attractive proposition for reducing burden, there is a danger that this may be a false economy. Whilst passive data collection might be decreasing objective burden, this may come at the cost of increasing subjective burden. Evidence suggests that participants are less willing to take part in passive mobile data collection methods that offer participants reduced agency (Wenz et al., 2019). Where mobile data collection requires active participation, it is not yet clear whether burden is reduced compared to fielding a questionnaire. Further comparative research is needed to ascertain this.

CHALLENGES

Low response rates. The first challenge is maximising sample member's cooperation with survey requests involving mobile devices. For the Spending Study 1 13% of the sample used the app at least once (Jäckle et al., 2019a). Other mobile data collection tasks using apps have found similar response rates: 16% (Kreuter et al., 2018), 19% and 22% (Scherpenzeel, 2017). However, as is noted by Groves (2006) nonresponse does not inherently result in nonresponse biases. More work is necessary to fully understand to what extent the low levels of response observed in the studies listed above result in nonresponse biases. We can gain some insight from the findings of Jäckle et al. (2019), as discussed earlier in this thesis, they examined participation biases in the context of the Spending Study 1. To summarise, they found some evidence of differences in participation between different demographic groups such as age and gender, between sample members who reported engaging in financial behaviours such as budgeting and tracking their finances, and between those who reported having access to a mobile device or not. However, crucially, they found little evidence of differences in participation rates between those with different financial situations, which would directly impact the outcomes being measured in the study.

Second-level digital divide. A second challenge can be found in what has been termed the '*second-level digital divide*' Hargittai (2002). Discussed earlier in the thesis, this raises the potential issue of how different groups of people make use of their mobile devices in different ways. Placing this in the context of the Total Survey Error framework, such differences may contribute to both nonresponse and measurement error.

Privacy and informed consent. The third challenge are concerns about privacy and informed consent when harnessing an app for survey research. In part this is a product of the enhanced requirements placed on those wishing to collect digital data by the recently instated European General Data Protection Regulation (GDPR). Emphasis on the duty of care those handling data have to the subjects of that data have resulted in increased efforts to safeguard participants, for example providing extensive participant materials offering explanations regarding data collection and data usage. However, Kreuter et al. (2018) found evidence that respondents often did not engage with or read such explanations when they are offered. Perhaps this lack of engagement with participant materials is reflective of wider patterns of behaviour in relation to privacy protections. Evidence suggests many people routinely choose to ignore, or not engage with privacy policies and terms of service policies (Obar and Oeldorf-Hirsch, 2020). Drawing on the work of Kahneman (2011), Kreuter et al. (2018) suggest participants often use heuristics to '*shortcut*' consent decisions, rather than giving them full consideration.

It has been highlighted that privacy concerns may be particularly prevalent when using mobile technologies to collect data not typically collected through surveys, for example photographs. In one example, van Heerden et al. (2020) discuss the ethical concern of the potential for survey data collection that captures images potentially capturing images of nudity or of a sexual nature. In this context the potential for causing harm to participants is very apparent. The authors suggest a negotiated ethical approach that emphasises agency for participants to be involved in the framing of what is ethically acceptable in terms of data to collect. Such an approach is however not without its own challenges, for example, the potential for increased social desirability bias when asking respondents to play an active role in defining the ethical boundaries of research.

Volume of data. A fourth challenge is the volume of data that is typical of the types of organic data that may be captured using mobile devices. In their pilot study examining the feasibility of capturing accelerometery data through wearable mobile devices Scherpenzeel et al. (2018) report 17,000,000 records were generated per respondent. Such volumes of data require survey researchers to draw on techniques developed in the fields of data science and computational social science to enable analysis to be conducted.

Complexity of data. The fifth challenge is the complexity of different types of organic data. Japec et al. (2015) argue that it is the complexity of organic data that should be of most concern to social scientists rather than the volume. They argue that the often unstructured nature of the data means that drawing useful insights presents a significant challenge. In part, this echoes Groves' (2011) caution about mistaking an abundance of data for an abundance of information. In the field of data science, the term data mining has come to represent this goal of extracting useful insight from raw data. This evocative metaphor of mining for insight in data also speaks to the increasing need for mechanisation and automation, in the form of machine learning, to achieve this goal. Chapter three explored one attempt at facing this challenge, however the diversity of forms in which organic data can be found suggest it is unlikely a universal solution to this challenge will be found. Instead, further developments across disciplines will be necessary to make the most of the opportunities mobile devices present for capturing organic data.

New skillsets for survey researchers. The sixth challenge is the need for survey researchers to develop new skills to be able to harness the new opportunities presented by using mobile apps for survey research. Japec et al. (2015) discuss the adoption of new skillsets in the context of making use of Big Data for survey research, but many of the same principles apply when considering the use of mobile apps. They highlight the need for the following skills: domain expertise, research skills, computer science skills, and systems administration skills. We can take these skillsets and apply them to data collection using mobile apps. Domain expertise might both refer to the substantive expertise related to the topic of the research at hand, but also to expertise on any norms and conventions that shape how the participant may use the app. Research skills refers to the traditional survey research skills that are required for fielding a questionnaire, though as has hopefully been shown across this thesis, it is important to continually assess how this body of knowledge is to be applied to this different context. Computer science skills are important not just for programming the app, researchers having an understanding of principles of computer science can help ensure that pitfalls causing issues such as missing data are avoided. System administration skills are important for understanding the infrastructure underlying the software making up the app. As was demonstrated in Chapter 2, the underlying software of an app can have impacts upon data quality. Of course, a survey researcher does not need to be an expert in all of these disparate skillsets, but having some understanding of all of them, and access to experts in each will help to improve survey app design.

Competition with commercial products. The mobile app market is highly saturated, there are a large number of commercial apps available covering an incredibly diverse array of use cases. There are two key ways in which this presents challenges for survey research. The first is the potential for a research app to get "lost in the crowd", if participants cannot find the app, they cannot participate in the study. Secondly, researchers may find participants implicitly or explicitly making comparisons between the research app and a related commercial app. Such comparisons may often not be favourable for the research app due to limitations such as budget. Careful design choices may be necessary to ensure participants are not put off from participating by an app that deviates too far from expectations of what an app should look like, or how it should function.

With these considerations in mind, this thesis concludes by briefly considering the broader outlook for the field of survey research. Mobile data collection, be that the collection of traditional designed survey data or of new forms of organic data using sensors and other device hardware, seems now to be firmly established as part of the survey researcher's toolbox. Returning to where this thesis began, characterising itself as a product of the third era of survey research, gives an interesting insight into how the innovative use of mobile data collection methods could shape the field of survey research. Innovations, such as those necessary to face the opportunities and challenges offered by mobile data collection are not new to survey research. Indeed, Groves' (2011) historical narrative account of the field suggests that innovation has been embedded into survey research from the very beginning. It is with this in mind that it seems certain that the ongoing innovations necessary to explore new opportunities and overcome new challenges that present themselves will continue to characterise this third era of survey research, and the future beyond it.

~

REFERENCES

Abraham KG, Maitland A and Bianchi SM. (2006) Non-response in the American Time Use Survey: Who Is Missing from the Data and How Much Does It Matter? *NBER Technical Working Paper Series*.

Adamou B. (2013) ResearchGames as a methodology: The impact of online ResearchGames upon participant engagement and future ResearchGame participation. *Association for Survey Computing Conference.* Winchester, UK.

Aguiar M and Hurst E. (2007) Life-cycle prices and production. *American Economic Review* 97: 1533-1559.

Altman NS. (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46: 175-185.

Ampt E. (2003) Respondent Burden. In: Stopher P and Jones P (eds) *Transport Survey Quality and Innovation.* Emerald Group Publishing, 507-521.

Ansolabehere S and Schaffner BF. (2015) Distractions: The incidence and consequences of interruptions for survey respondents. *Journal of Survey Statistics and Methodology* 3: 216-239.

Antoun C, Couper MP and Conrad FG. (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.

Appelhans BM, French SA, Tangney CC, et al. (2017) To what extent do food purchases reflect shoppers' diet quality and nutrient intake? *International journal of behavioral nutrition and physical activity* 14: 46.

Armstrong JS. (1975) Monetary incentives in mail surveys. *Public Opinion Quarterly* 39: 111-116.

Azuero A. (2016) A note on the magnitude of hazard ratios. *Cancer* 122: 1298-1299.

Baker R. (2017) Big Data: A Survey Research Perspective. In: Biemer PP, de Leeuw E, Eckman S, et al. (eds) *Total Survey Error in Practice.* John Wiley & Sons, 47-69.

Bartlett MS. (1951) The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika* 38: 337-344.

Biediger-Friedman L, Sanchez B, He M, et al. (2016) Food purchasing behaviors and food insecurity among college students at The University of Texas at San Antonio. *Journal of Food Security* 4: 52-57.

Biemer PP. (2016) Data Quality and Inference Errors. In: Foster I, Ghani R, Jarmin RS, et al. (eds) *Big Data and Social Science: A Practical Guide to Methods and Tools.* Chapman and Hall.

Borenstein M, Hedges LV, Higgins JPT, et al. (2009) *Introduction to Meta-Analysis*: John Wiley & Sons.

Bradburn N. (1978) Respondent burden. *Proceedings of the Survey Research Methods Section of the American Statistical Association.* American Statistical Association Alexandria, VA, 40.

Branden L, Gritz RM and Pergamit MR. (1995) The Effect of Interview Length on Nonresponse in the National Longitudinal Survey of Youth. *1995 Census Bureau Annual Research Conference.* Arlington, VA, 129-154.

Breiman L. (1996) Bagging predictors. *Machine learning* 24: 123-140.

Breiman L. (2001a) Random forests. *Machine learning* 45: 5-32.

Breiman L. (2001b) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16: 199-231.

Breiman L, Friedman JH, Olshen R, et al. (1984) Classification and Regression Trees (Belmont, CA: Wadsworth International Group). *Biometrics* 40: 17-23.

Breslow N. (1974) Covariance Analysis of Censored Survival Data. *Biometrics* 30: 89-99.

Brignone E, Fargo JD, Blais RK, et al. (2018) Applying Machine Learning to Linked Administrative and Clinical Data to Enhance the Detection of Homelessness among Vulnerable Veterans. *AMIA Annual Symposium Proceedings.* American Medical Informatics Association, 305.

Browne WJ. (2017) MCMC Estimation in MLwiN v3.00. Centre for Multilevel Modelling, University of Bristol.

Brownlee J. (2019) *iOS is twice as memory-efficient as Android. Here's why. | Cult of Mac*. Available at: https://www.cultofmac.com/303223/ios-twice-memory-efficient-android-heres/.

Brunton-Smith I, Sturgis P and Leckie G. (2017) Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180: 551-568.

Brys G, Hubert M and Struyf A. (2004) A robust measure of skewness. *Journal of Computational and Graphical Statistics* 13: 996-1017.

Bucklin RE and Gupta S. (1999) Commercial use of UPC scanner data: Industry and academic perspectives. *Marketing Science* 18: 247-273.

Bulusu N, Chou CT, Kanhere S, et al. (2008) Participatory sensing in commerce: Using mobile camera phones to track market price dispersion. *Proceedings of the international workshop on urban, community, and social applications of networked sensing systems (UrbanSense 2008).* 6-10.

Burke JA, Estrin D, Hansen M, et al. (2006) Participatory sensing. *4th ACM Conference on Embedded Networked Sensor Systems.* Boulder, CO.

Burstyn I, Slutsky A, Lee DG, et al. (2014) Beyond crosswalks: reliability of exposure assessment following automated coding of free-text job descriptions for occupational epidemiology. *Annals of occupational hygiene* 58: 482-492.

Busemeyer JR and Townsend JT. (1993) Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review* 100: 432-459.

Calderwood L and Lessof C. (2009) Enhancing longitudinal surveys by linking to administrative data In: Lynn P (ed) *Methodology of longitudinal surveys.* Chichester: John Wiley & Sons, Ltd, 55-72.

Callegaro M. (2010) Do you know which device your respondent has used to take your online survey.

Ceron A, Curini L, Iacus SM, et al. (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16: 340-358.

Chang C-C and Lin C-J. (2011) LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2: 27.

Charlton C, Rasbash J, Browne WJ, et al. (2017) MLwiN Version 3.00. Centre for Multilevel Modelling, University of Bristol.

Chen T, He T, Benesty M, et al. (2015) Xgboost: extreme gradient boosting. *R package version 0.4-2*: 1-4.

Chrisinger BW, DiSantis KI, Hillier AE, et al. (2018) Family food purchases of high-and low-calorie foods in full-service supermarkets and other food retailers by Black women in an urban US setting. *Preventive medicine reports* 10: 136-143.

Church AH. (1993) Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly* 57: 62-79.

Cohen J. (1969) *Statistical power analysis for the behavioral sciences*: Academic Press.

Collins M, Sykes W, Wilson P, et al. (1988) Diffusion of technological innovation: Computer assisted data collection in the U.K. In: Groves RM, Biemer PP, Lyberg LE, et al. (eds) *Computer assisted survey information collection.* John Wiley & Sons.

Cook WA. (2014) Is mobile a reliable platform for survey taking? Defining quality in online surveys from mobile respondents. *Journal of Advertising Research* 54: 141-148.

Couper MP. (1997) Survey introductions and data quality. *Public Opinion Quarterly* 61: 317-338.

Couper MP. (2013a) Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods.* 145-156.

Couper MP. (2013b) Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods.* 145-156.

Couper MP and Nicholls IWL. (1998) The history and development of computer assisted survey information collection methods. In: Couper MP, Baker RP, Bethlehem J, et al. (eds) *Computer assisted survey information collection.* John Wiley & Sons.

Couper MP and Peterson GJ. (2017) Why do web surveys take longer on smartphones? *Social Science Computer Review* 35: 357-377.

Crawford SD, Couper MP and Lamias MJ. (2001) Web Surveys:Perceptions of Burden. *Social Science Computer Review* 19: 146-162.

Creecy RH, Masand BM, Smith SJ, et al. (1992) Trading MIPS and Memory for Knowledge Engineering. *Communications of the ACM* 35: 48-64.

Csikszentmihalyi M and Larson R. (2014) Validity and reliability of the experience-sampling method. *Flow and the foundations of positive psychology.* Springer, 35-54.

Cullen K, Baranowski T, Watson K, et al. (2007) Food category purchases vary by household education and race/ethnicity: results from grocery receipts. *Journal of the American Dietetic Association* 107: 1747-1752.

d'Ardenne J and Blake M. (2012) Developing expenditure questions: Findings from focus groups. *IFS Working Paper* W12/18.

Dale T and Haraldsen G. (2005) Embedded evaluation of perceived and actual response burden in business surveys. In: Hedlin D, Dale T, Haraldsen G, et al. (eds) *Developing methods for assessing perceived response burden.* Eurostat, 112-125.

De Bruijne M and 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 and Wijnant A. (2014) Mobile response in web panels. *Social Science Computer Review* 32: 728-742.

Dehghani M, Azarbonyad H, Marx M, et al. (2015) Sources of evidence for automatic indexing of political texts. *European Conference on Information Retrieval.* Springer, 568-573.

Deming WE. (1944) On errors in surveys. *American Sociological Review* 9: 359-369.

Deng L and Cox LP. (2009) Livecompare: grocery bargain hunting through participatory sensing. *10th workshop on Mobile Computing Systems and Applications.* Santa Cruz, CA, 1-6.

Dietterich TG. (2000) Ensemble methods in machine learning. *International workshop on multiple classifier systems.* Springer, 1-15.

Dillman DA and Christian LM. (2005) Survey mode as a source of instability in responses across surveys. *Field methods* 17: 30-52.

Dillman DA, Sinclair MD and Clark JR. (1993) Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly* 57: 289-304.

Durrant GB, Groves RM, Staetsky L, et al. (2010) Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly* 74: 1-36.

Dyregrov K. (2004) Bereaved parents' experience of research participation. *Social science & medicine* 58: 391-400.

Einav L, Leibtag ES and Nevo A. (2008) On the accuracy of Nielsen Homescan data.

Elliott MN, Zaslavsky AM, Goldstein E, et al. (2009) Effects of survey mode, patient mix, and nonresponse on CAHPS® hospital survey scores. *Health services research* 44: 501-518.

Ellison R. (2010) *Classifications and CASCOT.* Available at: https://warwick.ac.uk/fac/soc/ier/software/cascot/cascot_soc2010_demo_for_web.pptx.

Fernee H and Sonck N. (2013) Is everyone able to use a smartphone in survey research? Tests with a Time-use App with Experienced and Inexperienced Users. *Survey Practice* 6: 2884.

Fielding J, Fielding N and Hughes G. (2013) Opening up open-ended survey data using qualitative software. *Quality & Quantity* 47: 3261-3276.

Flora DB and Curran PJ. (2004) An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods* 9: 466-491.

Fortunati L and Taipale S. (2014) The advanced use of mobile phones in five E uropean countries. *The British journal of sociology* 65: 317-337.

Fricker S. (2016) Defining, measuring, and mitigating respondent burden. *Workshop on Respondent Burden in the American Community Survey*. Washington, DC.

Fricker S and Tourangeau R. (2010) Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly* 74: 934-955.

Friedman JH. (1996) *Another Approach to Polychotomous Classification*. Available at: http://statweb.stanford.edu/~jhf/ftp/poly.pdf.

Friedman JH. (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics*: 1189-1232.

Friedman JH. (2002) Stochastic gradient boosting. *Computational statistics & data analysis* 38: 367-378.

Galea S, Nandi A, Stuber J, et al. (2005) Participant reactions to survey research in the general population after terrorist attacks. *Journal of Traumatic Stress* 18: 461-465.

Galesic M. (2006) Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics* 22: 313-328.

Galesic M and Bosnjak M. (2009) Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly* 73: 349-360.

Geekbench. (2018) *Geekbench 4*. Available at: https://www.geekbench.com/.

Gelade W, Verardi V and Vermandele C. (2013) Medcouple [Stata package].

Gelman A and Hill J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*: Cambridge University Press.

Gerstel EK, Harford TC and Pautler C. (1980) The reliability of drinking estimates obtained with two data collection methods. *Journal of Studies on Alcohol* 41: 89-94.

Gillespie M, Ampofo L, Cheesman M, et al. (2016) Mapping refugee media journeys: Smartphones and social media networks. Open University.

Gillmore MR, Gaylord J, Hartway J, et al. (2001) Daily data collection of sexual and other health-related behaviors. *Journal of Sex Research* 38: 35-42.

Giorgetti D and Sebastiani F. (2003) Automating survey coding by multiclass text categorization techniques. *Journal of the American Society for Information Science and Technology* 54: 1269-1277.

Goyder J. (1994) An experiment with cash incentives on a personal interview survey. *Journal of the Market Research Society* 34: 1-7.

Graf I. (2008) Respondent Burden. In: Lavrakas P (ed) *Encyclopedia of survey research methods.* Sage Publications, 740-740.

Greenwood D, Ransley J, Gilthorpe M, et al. (2006) Use of itemized till receipts to adjust for correlated dietary measurement error. *American journal of epidemiology* 164: 1012-1018.

Griffith R, Leibtag E, Leicester A, et al. (2009) Consumer shopping behavior: how much do consumers save? *Journal of Economic Perspectives* 23: 99-120.

Groves RM. (2006) Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly* 70: 646-675.

Groves RM. (2011) Three eras of survey research. *Public Opinion Quarterly* 75: 861-871.

Groves RM and Couper MP. (1998) *Nonresponse in household interview surveys*: John Wiley & Sons.

Groves RM and Kahn RL. (1979) *Surveys by Telephone; A national comparison with personal interviews,* New York: Academic Press.

Groves RM, Presser S and Dipko S. (2004) The role of topic interest in survey participation decisions. *Public Opinion Quarterly* 68: 2-31.

Groves RM, Singer E and Bowers A. (1999) A laboratory approach to measuring the effects on survey participation of interview length, incentives, differential incentives, and refusal conversion. *Journal of Official Statistics* 15: 251-268.

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

Guadagni PM and Little JD. (1983) A logit model of brand choice calibrated on scanner data. *Marketing Science* 2: 203-238.

Guadagnoli E and Velicer W. (1988) Relation of sample size to the stability of component patterns. *Psychological Bulletin* 103: 265-275.

Gundlapalli AV, Carter ME, Divita G, et al. (2014) Extracting concepts related to homelessness from the free text of VA electronic medical records. *AMIA Annual Symposium Proceedings.* American Medical Informatics Association, 589.

Gupta S, Chintagunta P, Kaul A, et al. (1996) Do household scanner data provide representative inferences from brand choices: A comparison with store data. *Journal of Marketing Research* 33: 383-398.

Gweon H, Schonlau M, Kaczmirek L, et al. (2017) Three methods for occupation coding based on statistical learning. *Journal of Official Statistics* 33: 101-122.

Hahsler M, Piekenbrock M and Doran D. (2015) *dbscan: Fast Density-based Clustering with R.* Available at: https://cran.r-project.org/web/packages/dbscan/vignettes/dbscan.pdf.

Haraldsen G. (2004) Identifying and reducing response burdens in internet business surveys. *Journal of Official Statistics* 20: 393–410-393–410.

Harris JM. (2005) Using homescan data and complex survey design techniques to estimate convenience food expenditures. *American Agricultural Economics Association Annual Meeting.* Providence, RI.

Hastie T and Qian J. (2014) Glmnet vignette. *Retrieve from http://www. web. stanford. edu/~ hastie/Papers/Glmnet_Vignette. pdf. Accessed September* 20: 2016.

Hastie T, Tibshirani R and Friedman J. (2009) *The elements of statistical learning: data mining, inference, and prediction*: Springer Science & Business Media.

Hektner JM, Schmidt JA and Csikszentmihalyi M. (2007) *Experience sampling method: Measuring the quality of everyday life,* Thousand Oaks, CA: Sage Pub. Inc.

Hendershott A, Edgar J, Geisen E, et al. (2012) Would You Like a Receipt With That? Availability of Respondent Records When Collecting Expenditure Information

Henning J. (2012) King me! How anyone can easily gamify their next survey. *67th Annual Conference of the American Association for Public Opinion Research.* Orlando, FL.

Hinkle DE, Wiersma W and Jurs SG. (2003) *Applied statistics for the behavioral sciences*: Houghton Mifflin.

Hofmann T, Schölkopf B and Smola AJ. (2008) Kernel methods in machine learning. *The annals of statistics*: 1171-1220.

Holgado-Tello FP, Chacón–Moscoso S, Barbero–García I, et al. (2010) Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity* 44: 153-166.

Hox JJ, Moerbeek M and Van de Schoot R. (2017) *Multilevel analysis: Techniques and applications*: Routledge.

Huang W, Nakamori Y and Wang S-Y. (2005) Forecasting stock market movement direction with support vector machine. *Computers & operations research* 32: 2513-2522.

Hubert M and Vandervieren E. (2008) An adjusted boxplot for skewed distributions. *Computational statistics & data analysis* 52: 5186-5201.

Inman JJ and Winer RS. (1998) Where the rubber meets the road: A model of in-store consumer decision making.

Inman JJ, Winer RS and Ferraro R. (2009) The interplay among category characteristics, customer characteristics, and customer activities on in-store decision making. *Journal of Marketing* 73: 19-29.

Institute for Employment Research. (2015) *CASCOT FAQ*. Available at: https://warwick.ac.uk/fac/soc/ier/software/cascot/faq.

Jäckle A, Al Baghal T, Burton J, et al. (2018a) *Understanding Society: The UK Household Longitudinal Study Innovation Panel, Waves 1-10, User Manual* Colchester: Institute for Social and Economic Research, University of Essex.

Jäckle A, Burton J, Couper MP, et al. (2019a) Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. *Survey Research Methods* 13: 23-44.

Jäckle A, Burton J, Wenz A, et al. (2018b) *Understanding Society: The UK Household Longitudinal Study. Spending Study 1, User Guide.* , Colchester: Institute for Social and Economic Research, University of Essex.

Jäckle A, Lynn P, Sinibaldi J, et al. (2011) The effect of interviewer personality, skills and attitudes on respondent co-operation with face-to-face surveys. ISER Working Paper Series.

Jäckle A, Roberts C and Lynn P. (2010) Assessing the effect of data collection mode on measurement. *International Statistical Review* 78: 3-20.

Jäckle A, Wenz A, Burton J, et al. (2019b) Increasing participation in a mobile app study: the effects of a sequential mixed-mode design and in-interview invitation. *Understanding Society Working Paper Series*.

Japec L, Kreuter F, Berg M, et al. (2015) Big Data in Survey Research: AAPOR Task Force Report. *Public Opinion Quarterly* 79: 839-880.

Jones R and Elias P. (2004) CASCOT: Computer-Assisted Structured Coding Tool. Coventry, University of Warwick: Institute for Employment Research.

Jung Y, Yoo J, Myaeng S-H, et al. (2008) A web-based automated system for industry and occupation coding. *International Conference on Web Information Systems Engineering.* Springer, 443-457.

Kahneman D. (2011) *Thinking, fast and slow*: Macmillan.

Kaiser HF. (1960) The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20: 141-151.

Kaiser HF. (1970) A second generation little jiffy. *Psychometrika* 35: 401-415.

Kaiser HF and Rice J. (1974) Little jiffy, Mark IV. *Educational and Psychological Measurement* 34: 111-117.

Karjaluoto H, Karvonen J, Kesti M, et al. (2005) Factors affecting consumer choice of mobile phones: Two studies from Finland. *Journal of Euromarketing* 14: 59-82.

Keusch F, Antoun C, Couper MP, et al. (2017) Willingness to participate in passive mobile data collection. *72nd Annual Conference of the American Association for Public Opinion Research,.* New Orleans, LA.

Keusch F, Struminskaya B, Antoun C, et al. (2019) Willingness to Participate in Passive Mobile Data Collection. *Public Opinion Quarterly* 83: 210-235.

Keusch F and Yan T. (2017) Web versus mobile web: An experimental study of device effects and self-selection effects. *Social Science Computer Review* 35: 751-769.

Kirby G, Carson J, Dunlop F, et al. (2015) Automatic methods for coding historical occupation descriptions to standard classifications. In: Bloothooft G, Christen P, Mandemakers K, et al. (eds) *Population Reconstruction.* Springer, 43-60.

Knäuper B, Belli RF, Hill DH, et al. (1997) Question difficulty and respondents' cognitive ability: The effect on data quality. *Journal of Official Statistics* 13: 181-199.

Kohler U, Luniak M and Brzinsky-Fay C. (2006) sq. [Stata Package].

Kolenikov S. (2008) polychoric. [Stata Package].

Kreuter F, Haas G-C, Keusch F, et al. (2018) Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*: 0894439318816389.

Kreuter F, Presser S and Tourangeau R. (2008) Social Desirability Bias in CATI, IVR, and Web Surveys The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly* 72: 847-865.

Krippendorff K. (2004) *Content Analysis: An Introduction to Its Methodology*: Sage.

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

Lai JW, Link M and Vanno L. (2012) Emerging techniques of respondent engagement: Leveraging game and social mechanics for mobile application research. *67th Annual Conference of the American Association for Public Opinion Research.* Orlando, FL.

Landis JR and Koch GG. (1977) The measurement of observer agreement for categorical data. *Biometrics*: 159-174.

Large J, Lines J and Bagnall A. (2017) The heterogeneous ensembles of standard classification algorithms (HESCA): the whole is greater than the sum of its parts. *arXiv preprint arXiv:1710.09220*.

Larson R and Csikszentmihalyi M. (1983) The Experience Sampling Method. *New Directions for Methodology of Social & Behavioral Science*.

Leckie G. (2013) Cross-Classified Multilevel Models - Concepts. LEMMA VLE Module 12, 1-60 http://www.bristol.ac.uk/cmm/learning/course.html.

Lee J. (2018) This Is Why iOS Devices Use Less RAM Than Android Devices.

Lee YS and Waite LJ. (2005) Husbands' and wives' time spent on housework: A comparison of measures. *Journal of Marriage and Family* 67: 328-336.

Leicester A and Oldfield Z. (2009a) An analysis of consumer panel data.

Leicester A and Oldfield Z. (2009b) Using scanner technology to collect expenditure data. *Fiscal Studies* 30: 309-337.

Leigh BC. (1993) Alcohol consumption and sexual activity as reported with a diary technique. *Journal of Abnormal Psychology* 102: 490-493.

Lemmens PHHM, Knibbe RA and Tan F. (1988) Weekly recall and diary estimates of alcohol consumption in a general population survey. *Journal of Studies on Alcohol* 49: 131-135.

Levenshtein VI. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady* 10: 707.

Link M, Lai JW and Vanno L. (2012) Smartphone applications: The next (and most important?) evolution in data collection. *67th Annual Conference of the American Association for Public Opinion Research.* Orlando, FL.

Link MW, Murphy J, Schober MF, et al. (2014) Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly* 78: 779-787.

Loges WE and Jung J. (2001) Exploring the digital divide: Internet connectedness and age. *Communication research* 28: 536-562.

Lugtig P and 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: 78-94.

Lynn P. (2014) Longer interviews may not affect subsequent survey participation propensity. *Public Opinion Quarterly* 78: 500-509.

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

Manning C, Raghavan P and Schütze H. (2010) Introduction to information retrieval. *Natural Language Engineering* 16: 100-103.

Marr JW. (1971) Individual dietary surveys: purposes and methods. *World review of nutrition and dietetics.* Karger Publishers, 105-164.

Martin SL, Howell T, Duan Y, et al. (2006) The feasibility and utility of grocery receipt analyses for dietary assessment. *Nutrition journal* 5: 10.

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

Mavletova A. (2015) Web surveys among children and adolescents: is there a gamification effect? *Social Science Computer Review* 33: 372-398.

Mavletova A and Couper MP. (2013) Sensitive topics in PC web and mobile web surveys: Is there a difference? *Survey Research Methods.* 191-205.

McGloughlin I. (1983) The scanner revolution—collection of purchasing data from consumer panel households. *Section on Survey Research Methods at the Joint Statistical Meeting.* Toronto, Canada.

Meyer D, Dimitriadou E, Hornik K, et al. (2019) Package 'e1071'. *R Package version 1.7-3.*

Moskowitz DS and Young SN. (2006) Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience* 31: 13.

Murthy D. (2015) Twitter and elections: are tweets, predictive, reactive, or a form of buzz? *Information, Communication & Society* 18: 816-831.

Nahoomi N. (2018) Automatically Coding Occupation Titles to a Standard Occupation Classification.

Newman E, Willard T, Sinclair R, et al. (2001) Empirically supported ethical research practice: The costs and benefits of research from the participants' view. *Accountability in Research* 8: 309-329.

O'Muircheartaigh C and Campanelli P. (1999) A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162: 437-446.

O'Muircheartaigh C and Campanelli P. (1998) The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 161: 63-77.

Obar JA and Oeldorf-Hirsch A. (2020) The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23: 128-147.

Office for National Statistics. (2017) *Living Costs and Food Survey. User Guide. Volume D: Expenditure codes 2015-16.* Available at:

http://doc.ukdataservice.ac.uk/doc/8210/mrdoc/excel/8210_volume_d_expenditure_codes_2015-16.xls.

Office for National Statistics. (2018) *ONS Occupation Coding Tool*. Available at: https://onsdigital.github.io/dp-classification-tools/standard-occupational-classification/ONS_SOC_occupation_coding_tool.html.

Office of Management and Budget. (2006) Standards and guidelines for statistical surveys.

Oomens P and Timmermans G. (2008) The Dutch approach to reducing the real and perceived administrative burden on enterprises caused by statistics. Paper present to the 94th DGINS Conference.

Open Signal. (2015) *Android Fragmentation 2015*.

Ossiander EM and Milham S. (2006) A computer system for coding occupation. *American journal of industrial medicine* 49: 854-857.

Ouyang M. (2016) KNN in the Jaccard space. *2016 IEEE High Performance Extreme Computing Conference (HPEC).* IEEE, 1-7.

Ozarslan S and Eren PE. (2014) Text recognition and correction for automated data collection by mobile devices. *Imaging and Multimedia Analytics in a Web and Mobile World 2014.* International Society for Optics and Photonics, 902706.

Patel J, Shah S, Thakkar P, et al. (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications* 42: 259-268.

Persky H, Strauss D, Lief HI, et al. (1981) Effect of the research process on human sexual behavior. *Journal of Psychiatric Research* 16: 41-52.

Peytchev A. (2009) Survey Breakoff. *Public Opinion Quarterly* 73: 74-97.

Pfeffermann D. (2019) Benefits and Issues in the Use of Internet-Based Surveys-Experience from Israel. *The future of online data collection in social surveys.* Southampton, UK.

Pickery J, Loosveldt G and Carton A. (2001) The Effects of Interviewer and Respondent Characteristics on Response Behavior in Panel Surveys:A Multilevel Approach. *Sociological Methods & Research* 29: 509-523.

Platt J. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.

Raiklin E and Uyar B. (1996) On the relativity of the concepts of needs, wants, scarcity and opportunity cost. *International Journal of Social Economics* 23: 49-56.

Rankin JW, Winett RA, Anderson ES, et al. (1998) Food purchase patterns at the supermarket and their relationship to family characteristics. *Journal of Nutrition Education* 30: 81-88.

Ransley J, Donnelly J, Botham H, et al. (2003) Use of supermarket receipts to estimate energy and fat content of food purchased by lean and overweight families. *Appetite* 41: 141-148.

Resch B, Usländer F and Havas C. (2018) Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science* 45: 362-376.

Revilla M. (2017) Are there differences depending on the device used to complete a web survey (PC or smartphone) for order-by-click questions? *Field methods* 29: 266-280.

Revilla M and Couper MP. (2018) Comparing grids with vertical and horizontal item-by-item formats for PCs and smartphones. *Social Science Computer Review* 36: 349-368.

Revilla M, Toninelli D and Ochoa C. (2016) PCs versus Smartphones in answering web surveys: Does the device make a difference? *Survey Practice* 9: 1-6.

Rinaldi C. (2017) Android vs iOS: how different is their RAM management? | AndroidPIT.

Roberts C, Eva G, Allum N, et al. (2010) Diffusion of technological innovation: Computer assisted data collection in the U.K. *ISER Working Paper Series* 2010.

Ruch FL. (1941) Effects of repeated interviewing on the respondent's answers. *Journal of Consulting Psychology* 5: 179-182.

Russ DE, Ho K-Y, Colt JS, et al. (2016) Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occup Environ Med* 73: 417-424.

Russ DE, Ho K-Y, Johnson CA, et al. (2014) Computer-based coding of occupation codes for epidemiological analyses. *2014 IEEE 27th International Symposium on Computer-Based Medical Systems.* IEEE, 347-350.

Ruzek JI and Zatzick DF. (2000) Ethical considerations in research participation among acutely injured trauma survivors: An empirical investigation. *General Hospital Psychiatry* 22: 27-36.

Saldaña J. (2015) *The coding manual for qualitative researchers*: Sage.

Sarker S and Wells JD. (2003) Understanding mobile handheld device use and adoption. *Communications of the ACM* 46: 35-40.

Scagnelli J, Bailey J, Link M, et al. (2012) On the run: In the moment smartphone data collection. *67th Annual Conference of the American Association for Public Opinion Research.* Orlando, FL.

Scagnelli J and Bristol K. (2014) Scan all: Smartphones for measuring household purchases in developing markets. *69th Annual Conference of the American Association for Public Opinion Research.* Anaheim, CA.

Scherpenzeel A. (2017) Mixing online panel data collection with innovative methods. *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung.* Springer, 27-49.

Scherpenzeel A, Angleys N and Weiss L. (2018) Testing the logistics of the accelerometry project in SHARE. *BigSurv 18.* Barcelona, Spain.

Schierholz M. (2014) Automating survey coding for occupation.

Schonlau M and Couper MP. (2016) Semi-automated categorization of open-ended questions. *Survey Research Methods.* 143-152.

Schwartz HA and Ungar LH. (2015) Data-Driven Content Analysis of Social Media:A Systematic Overview of Automated Methods. *The ANNALS of the American Academy of Political and Social Science* 659: 78-94.

Schwarz N. (2012) Retrospective and concurrent self-reports: The rationale for real-time data capture. In: Stone AA, Shiffman SS, Atienza A, et al. (eds) *The science of real-time data capture: Self-reports in health research.* New York, NY: Oxford University Press.

Searles JS, Perrine MW, Mundt JC, et al. (1995) Self-report of drinking using touch-tone telephone: extending the limits of reliable daily contact. *Journal of Studies on Alcohol* 56: 375-382.

Sehgal S, Kanhere SS and Chou CT. (2008) Mobishop: Using mobile phones for sharing consumer pricing information. *Demo Session of the Intl. Conference on Distributed Computing in Sensor Systems.*

Sendelbah A, Vehovar V, Slavec A, et al. (2016) Investigating respondent multitasking in web surveys using paradata. *Computers in Human Behavior* 55: 777-787.

Sharp LM and Frankel J. (1983) Respondent burden: A test of some common assumptions. *Public Opinion Quarterly* 47: 36-53.

Singer E, van Hoewyk J, Gebler N, et al. (1999) The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics* 15: 217-230.

Smith TW. (2013) Survey-research paradigms old and new. *International Journal of Public Opinion Research* 25: 218-229.

Snijders TAB and Bosker RJ. (2012) Multilevel analysis : an introduction to basic and advanced multilevel modeling.

Spiegelhalter DJ, Best NG, Carlin BP, et al. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64: 583-639.

Stilley KM, Inman JJ and Wakefield KL. (2010) Planning to make unplanned purchases? The role of in-store slack in budget deviation. *Journal of consumer research* 37: 264-278.

Struminskaya B, Weyandt K and 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: 32.

Sudman S. (1964a) On the accuracy of recording of consumer panels: I. *Journal of Marketing Research* 1: 14-20.

Sudman S. (1964b) On the accuracy of recording of consumer panels: II. *Journal of Marketing Research* 1: 69-83.

Thompson M, Kornbau ME and Vesely J. (2012) Creating an automated industry and occupation coding process for the American Community Survey. United States Census Bureau.

Tibshirani R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58: 267-288.

Toepoel V and Lugtig P. (2014) What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users. *Social Science Computer Review* 32: 544-560.

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

Tukey JW. (1977) *Exploratory data analysis,* Reading, MA: Addison-Wesley Pub. Co.

United Nations. (2018) Classification of Individual Consumption According to Purpose (COICOP) 2018. New York: United Nations Department of Economic and Social Affairs.

University of Essex. Institute for Social and Economic Research. (2018a) Understanding Society: Innovation Panel, Waves 1-10, 2008-2017 [data collection]. 9th Edition. UK Data Service. SN: 6849 http://doi.org/10.5255/UKDA-SN-6849-10

University of Essex. Institute for Social and Economic Research. (2018b) Understanding Society: Spending Study 1, 2016. [data collection]. UK Data Service SN: 8348 http://doi.org/10.5255/UKDA-SN-8348.

Van Heerde HJ, Leeflang PS and Wittink DR. (2000) The estimation of pre-and postpromotion dips with store-level scanner data. *Journal of Marketing Research* 37: 383-395.

van Heerden A, Wassenaar D, Essack Z, et al. (2020) In-home passive sensor data collection and its implications for social media research: perspectives of community women in rural South Africa. *Journal of Empirical Research on Human Research Ethics* 15: 97-107.

Vapnik V. (1998) The support vector method of function estimation. *Nonlinear Modeling.* Springer, 55-85.

Verbrugge LM. (1980) Health diaries. *Medical Care* 18: 73-95.

Volkova E, Li N, Dunford E, et al. (2016) "Smart" RCTs: development of a smartphone app for fully automated nutrition-labeling intervention trials. *JMIR mHealth and uHealth* 4: e23.

Walker EA, Newman E, Koss M, et al. (1997) Does the study of victimization revictimize the victims? *General Hospital Psychiatry* 19: 403-410.

Waterlander WE, de Boer MR, Schuit AJ, et al. (2013) Price discounts significantly enhance fruit and vegetable purchases when combined with nutrition education: a randomized controlled supermarket trial. *The American journal of clinical nutrition* 97: 886-895.

Wells T, Bailey JT and Link MW. (2014) Comparison of smartphone and online computer survey administration. *Social Science Computer Review* 32: 238-255.

Wenz A, Jäckle A, Burton J, et al. (2020) The Effects of Personalized Feedback on Participation and Reporting in Mobile App Data Collection. *Social Science Computer Review* 0: 0894439320914261.

Wenz A, Jackle A and Couper MP. (2019) Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods.* European Survey Research Association, 1-22.

West B, Conrad FG, Kreuter F, et al. (2018) Can conversational interviewing improve survey response quality without increasing interviewer effects? *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181: 181-203.

West B and Elliott MR. (2014) Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers. *Surv. Methodol* 40: 163-188.

West B and Olson K. (2010) How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly* 74: 1004-1026.

Wiggins RD, Longford N and O'Muircheartaigh CA. (1990) *A Variance Components Approach to Interviewer Effects*: Joint Centre for Survey Methods.

Willeboordse A. (1997) Minimizing response burden. In: Willeboordse A (ed) *Handbook on design and implementation of business surveys.* Eurostat, 111-118.

Wright MN and Ziegler A. (2015) ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Wu T-F, Lin C-J and Weng RC. (2004) Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5: 975-1005.

Yammarino FJ, Skinner SJ and Childers TL. (1991) Understanding mail survey response behavior a meta-analysis. *Public Opinion Quarterly* 55: 613-639.

Yan T and Tourangeau R. (2008) Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 22: 51-68.

Young JM, O'Halloran A, McAulay C, et al. (2015) Unconditional and conditional incentives differentially improved general practitioners' participation in an online survey: randomized controlled trial. *Journal of Clinical Epidemiology* 68: 693-697.

Young PV and Schmid CF. (1956) *Scientific social surveys and research : an introduction to the background, content, methods, principles, and analysis of social studies*: Prentice-Hall.

Yu E, Fricker S and Kopp B. (2015) Can Survey Instructions Relieve Respondent Burden? : Paper presented to the 70th Annual Conference of the American Association for Public Opinion Research.

Zagorsky JL and Rhoton P. (2008) The effects of promised monetary incentives on attrition in a long-term panel survey. *Public Opinion Quarterly* 72: 502-513.

Zwarun L and Hall A. (2014) What's going on? Age, distraction, and multitasking during online survey taking. *Computers in Human Behavior* 41: 236-244.

APPENDICES

Appendix A

Additional material for Chapter One.

Table A1

Summary of how many participants completed which number of end of week surveys

Number of end of week surveys completed	*n*	*%*
Zero	*39*	*17*
One	*34*	*15*
Two	*31*	*14*
Three	*30*	*13*
Four	*89*	*40*

Table A2

Pearson χ^2 tests examining the bivariate relationship between predictors of burden and four measures of subjective burden

	Likelihood		Time/effort		Interest		Difficulty	
	χ^2	*F*	χ^2	*F*	χ^2	*F*	χ^2	*F*
£6 incentive treatment	0.36	0.10	1.16	0.50	1.16	0.50	5.11	1.65
Received additional incentive	1.99	0.61	2.2	0.95	0.46	0.20	2.10	0.70
Uses device for taking photos	1.97	0.64	0.66	0.35	0.29	0.17	1.23	0.43
Uses device for online banking	4.11	1.44	0.79	0.42	0.58	0.29	3.72	1.20
Uses device to install apps	1.23	0.41	0.04	0.02	1.96	1.08	3.75	1.23
Willing to download app	11.55	1.36	3.30	0.54	2.76	0.49	12.17	1.38
Willing to use camera	14.72	1.71	6.21	0.99	3.08	0.52	15.16	1.69
Checks balance at least once a week	2.94	1.00	1.51	0.79	1.30	0.65	3.52	1.26
Keeps a budget	3.22	1.00	0.20	0.10	1.44	0.69	5.17	1.84
Below the poverty threshold	11.20*	3.03	1.88	0.86	0.70	0.29	5.60	1.47
Time constrained	8.76*	3.32	0.28	0.13	0.91	0.38	1.10	0.36
Degree or higher	2.87	1.03	4.49	2.52	6.94*	3.20	1.50	0.55
Disabled/long term illness	3.78	1.19	3.30	1.48	2.59	3.51	4.02	1.41
Female	1.13	0.36	1.04	0.51	3.51	1.78	2.72	0.94

Notes: n = 223 participants; * $p < .05$ ** $p < .01$ *** $p < .001$.

Table A3

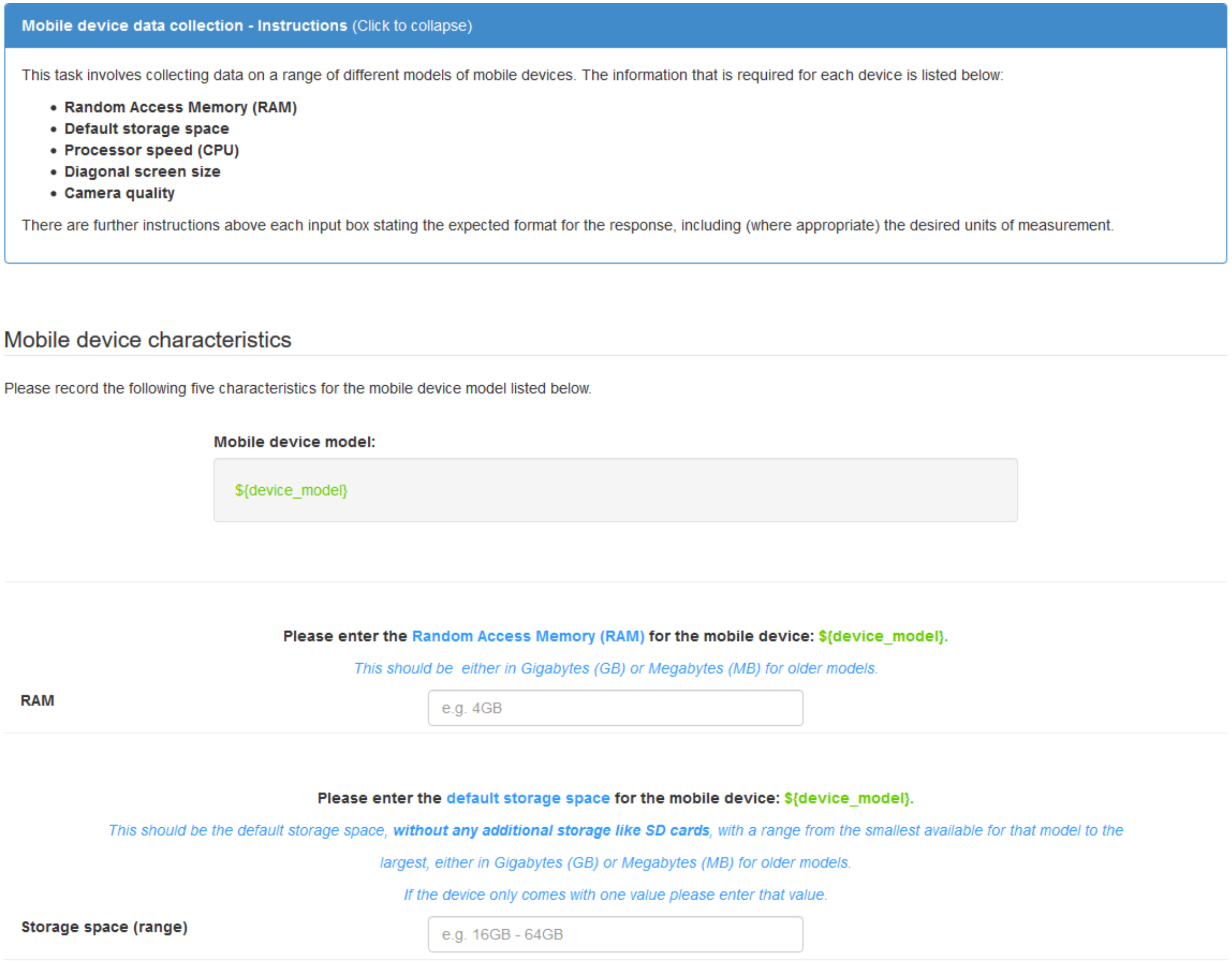
Two-tailed t-tests examining the bivariate relationship between predictors of burden and a measure of objective burden, the time taken to complete app uses

	$\bar{x}_1 - \bar{x}_2$	*SE*	*t*
£6 incentive treatment	-0.60	1.65	-0.36
Received additional incentive	-0.96	1.65	-0.58
Uses device for taking photos	3.44	3.02	1.14
Uses device for online banking	6.51**	1.61	4.05
Uses device to install apps	4.67*	2.02	2.31
Willing to download app	4.78*	2.10	2.28
Willing to use camera	2.85	2.29	1.25
Checks balance once a week or more	0.35	1.76	0.20
Keeps a budget	1.42	1.76	0.81
Below the poverty threshold	0.87	2.63	0.33
Time constrained	3.74*	1.79	2.10
Degree or higher	-0.51	1.56	-0.33
Disabled/long term illness	-0.95	1.65	-0.57
Female	-2.37	1.25	-1.89

Notes: *n* = 10179 app uses, across 223 participants; * *p* < .05 ** *p* < .01 *** *p* < .001.

Appendix B

Amazon mTurk Human Intelligence Task Screenshots.



Mobile device data collection - Instructions (Click to collapse)

This task involves collecting data on a range of different models of mobile devices. The information that is required for each device is listed below:

- **Random Access Memory (RAM)**
- **Default storage space**
- **Processor speed (CPU)**
- **Diagonal screen size**
- **Camera quality**

There are further instructions above each input box stating the expected format for the response, including (where appropriate) the desired units of measurement.

Mobile device characteristics

Please record the following five characteristics for the mobile device model listed below.

Mobile device model:

${device_model}

Please enter the Random Access Memory (RAM) for the mobile device: ${device_model}.

This should be either in Gigabytes (GB) or Megabytes (MB) for older models.

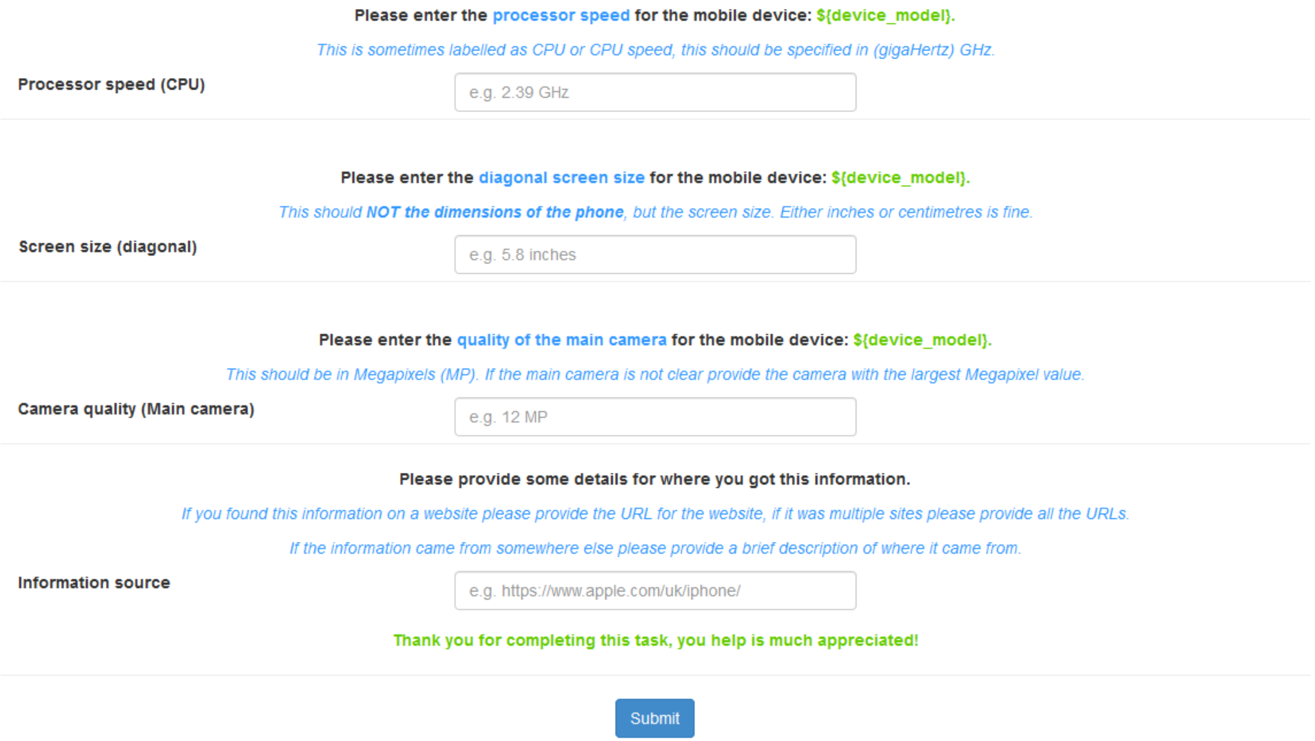
RAM e.g. 4GB

Please enter the default storage space for the mobile device: ${device_model}.

*This should be the default storage space, **without any additional storage like SD cards**, with a range from the smallest available for that model to the largest, either in Gigabytes (GB) or Megabytes (MB) for older models.*

If the device only comes with one value please enter that value.

Storage space (range) e.g. 16GB - 64GB



Please enter the processor speed for the mobile device: ${device_model}.

This is sometimes labelled as CPU or CPU speed, this should be specified in (gigaHertz) GHz.

Processor speed (CPU) e.g. 2.39 GHz

Please enter the diagonal screen size for the mobile device: ${device_model}.

*This should **NOT the dimensions of the phone**, but the screen size. Either inches or centimetres is fine.*

Screen size (diagonal) e.g. 5.8 inches

Please enter the quality of the main camera for the mobile device: ${device_model}.

This should be in Megapixels (MP). If the main camera is not clear provide the camera with the largest Megapixel value.

Camera quality (Main camera) e.g. 12 MP

Please provide some details for where you got this information.

If you found this information on a website please provide the URL for the website, if it was multiple sites please provide all the URLs.

If the information came from somewhere else please provide a brief description of where it came from.

Information source e.g. https://www.apple.com/uk/iphone/

Thank you for completing this task, you help is much appreciated!

Appendix C

Additional material for Chapter Three.

Table C1

Levels of accuracy after applying different levels of probability thresholds to predicted codes.

		0.40 threshold		0.60 threshold		0.80 threshold		0.90 threshold		0.95 threshold	
	Original % Acc.	% Acc.	95% CI	% Acc.	95% CI	% Acc.	95% CI	% Acc.	95% CI	% Acc.	95% CI
Strict match	77.4	78.2	[77.8, 78.6]	79.1	[78.7, 79.5]	80.2	[79.8, 80.6]	80.3	[79.9, 80.7]	77.9	[77.5, 78.3]
Approx. match	78.4	79.2	[78.8, 79.6]	80.0	[79.6, 80.4]	80.9	[80.5, 81.3]	81.0	[80.6, 81.4]	79.2	[78.8, 79.6]
Logistic	76.8	76.1	[75.7, 76.5]	76.9	[76.5, 77.3]	77.1	[76.7, 77.5]	77.6	[77.2, 78.0]	76.7	[76.3, 77.1]
Lasso	77.6	78.3	[77.9, 78.7]	77.8	[77.4, 78.2]	78.5	[78.1, 78.9]	79.1	[78.7, 79.5]	78.4	[78.0, 78.8]
Ridge	79.2	80.4	[80.0, 80.8]	80.3	[79.7, 80.5]	79.9	[79.5, 80.3]	79.7	[79.3, 80.1]	79.9	[79.5, 80.3]
knn Cosine	74.7	75.2	[74.8, 75.6]	74.3	[73.9, 74.7]	73.9	[73.5, 74.3]	74.0	[73.6, 74.4]	73.3	[72.9, 73.7]
knn Jaccard	73.2	72.9	[72.4, 73.4]	73.2	[72.8, 73.6]	73.3	[72.9, 73.7]	73.3	[72.9, 73.7]	73.1	[72.6, 73.6]
LSVM	77.1	77.1	[76.7, 77.5]	78.2	[77.8, 78.6]	77.4	[77.0, 77.8]	77.2	[76.8, 77.6]	76.8	[76.4, 77.2]
PSVM	78.4	79.1	[78.7, 79.5]	78.3	[77.9, 78.7]	79.0	[78.6, 79.4]	79.2	[78.8, 79.6]	80.1	[79.7, 80.5]
RSVM	71.9	70.5	[70.0, 71.0]	71.4	[70.9, 71.9]	71.1	[70.6, 71.6]	71.1	[70.6, 71.6]	68.6	[68.1, 69.1]
Random Forest	80.4	81.4	[81.0, 81.8]	82.3	[81.9, 82.7]	84.3	[83.9, 84.7]	83.4	[83.0, 83.8]	82.3	[81.9, 82.7]
xGBoost	75.2	75.2	[74.8, 75.6]	75.9	[75.5, 76.3]	76.3	[75.9, 76.7]	76.4	[76.0, 76.8]	75.9	[75.5, 76.3]