Survey Research Methods (2019)
Vol. 13, No. 1, pp. 45-71
doi:10.18148/srm/2019.v13i1.7379

ISSN 1864-3361
http://www.surveymethods.org

Respondent burden in a Mobile App: evidence from a shopping receipt scanning study

Brendan Read
Institute for Social and Economic Research
University of Essex
Colchester, UK

This study considers the burden placed on participants, subjectively and objectively, when asked to use a mobile app to scan shopping receipts. The existing literature on respondent burden is reviewed to present a framework of seven factors that affect burden, and this research demonstrates how these may be used to identify potential predictors of burden. Such an approach, together with the findings of this paper, have potential implications when applied to a number of emerging research contexts involving in-the-moment and repeated data collection. Data from both the *Understanding Society* Spending Study, a shopping receipt scanning study using respondents mobile phones, and the ninth wave of the *Understanding Society* Innovation Panel were used. Evidence was found to suggest that subjective perceptions of burden may not be strongly correlated with the actual objective burden faced. There were no systematic trends in subjective burden throughout the course of the study, though, as respondents completed more of the repeated tasks in the study, the objective burden per task did decrease. In terms of predictors of burden hypothetical willingness to complete the task was predictive of lower subjective burden. Older and female respondents also took longer to complete individual tasks in the study.

Keywords: Subjective, Objective, Cumulative, Fatigue, Expenditure, Measurement of Consumption, Household Panel Survey

1 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, p. 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 images of receipts. However, the concepts considered, and findings presented, in this research are also equally applicable to other research 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 data collected (Ruch, 1941; P. Young & 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, p. 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, p. 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:

Contact information: Brendan Read, ISER, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK (E-Mail: bread@essex.ac.uk).

1. Are subjective and objective measures of burden related?
2. How do subjective and objective burden change over the course of the study?
3. Does objective burden predict breaks in participation?
4. What factors predict subjective and objective burden?

2 Background

2.1 Receipt and UPC scanning

The potential benefits of Universal Product Codes (UPCs), also called barcodes, 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 (Marr, 1971; Sudman, 1964a, 1964b).

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 & Gupta, 1999; Guadagni & Little, 1983; Gupta, Chintagunta, Kaul, & Wittink, 1996; McGloughlin, 1983; Van Heerde, Leeflang, & Wittink, 2000). 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 is the National Consumer Panel in the US (formerly Nielsen HomeScan) from which a number of pieces of research have emerged (e.g. Aguiar & Hurst, 2007; Einav, Leibtag, Nevo, 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 also been used for several pieces of academic research (e.g. Griffith, Leibtag, Leicester, & Nevo, 2009; Leicester & Oldfield, 2009a, 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, Edgar, Geisen, & Stringei, 2012; Inman & Winer, 1998; Inman, Winer, & Ferraro, 2009; Stilley, Inman, & Wakefield, 2010) and health (Appelhans, French, Tangney, Powell, & Wang, 2017; Biediger-Friedman, Sanchez, He, Guan, & Yin, 2016; Chrisinger, DiSantis, Hillier, & Kumanyika, 2018; Cullen et al., 2007; Greenwood, Ransley, Gilthorpe, & Cade, 2006; Martin, Howell, Duan, & Walters, 2006; Rankin et al., 1998; Ransley et al., 2003; Waterlander, de Boer, Schuit, Seidell, & Steenhuis, 2013).

More recently, the potential for using mobile devices to aid the capture of these kind of data sources has been recognised. A body of research conducted by researchers at Nielsen (Scagnelli, Bailey, Link, & Benezra, 2012; Scagnelli & 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 & Cox, 2009) and receipts (Bulusu et al., 2008; Ozarslan & Eren, 2014; Sehgal, Kanhere, & Chou, 2008) have been developed. It is believed that the *Understanding Society* Spending Study, 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.

2.2 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, p. 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 how important the study was, difficulty, whether time and effort was well spent, and 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, Fricker, and Kopp (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.

2.3 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 chatacteristics 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 variability 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. 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, Couper, & Lamias, 2001; Galesic & Bosnjak, 2009), telephone surveys (Collins, Sykes, Wilson, & Blackshaw, 1988; Roberts, Eva, Allum, & Lynn, 2010), face-to-face surveys (Groves, Singer, & Bowers, 1999), and postal surveys (Dillman, Sinclair, & Clark, 1993; Yammarino, Skinner, & Childers, 1991). 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, Gritz, & Pergamit, 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 & 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, Newman, Koss, & Bernstein, 1997), bereavement (Dyregrov, 2004), and traumatic injuries (Ruzek & Zatzick, 2000). There has also been some evidence of question sensitivity as a barrier to participation amongst subgroups in

general population surveys (Galea et al., 2005; Newman, Willard, Sinclair, & Kaloupek, 2001), though the characteristics of the affected subgroups identified have not always been clear. Kreuter, Presser, and Tourangeau (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 & 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 & 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 & 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 (Abraham, Maitland, & Bianchi, 2006; Groves & Couper, 1998).

Another important factor when considering the opportunity to participate in 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 also 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 & Schaffner, 2015; Sendelbah, Vehovar, Slavec, & Petrovčič, 2016; Zwarun & Hall, 2014). However it has been suggested that the degree to which these distractions impact upon data quality is minimal (Ansolabehere & Schaffner, 2015). There is also some evidence to suggest that distractions are part of deliberate multitasking, 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 & Hall, 2014).

Ability/Competence. Lower cognitive ability has been highlighted as a widely accepted cause of measurement error (S. Fricker & 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 (Knäuper, Belli, Hill, & Herzog, 1997; Krosnick, 1991).

Lower device familiarity, or lower ability to complete survey tasks on a mobile device, has also been considered as a barrier to participation (Jäckle, Burton, Couper, & Lessof, 2019). 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 actually 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, Presser, & Dipko, 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 moti-

vate respondents, and improve the rate of participation (Armstrong, 1975; Singer, van Hoewyk, Gebler, & McGonagle, 1999). Typically, unconditional incentives have been found to be better motivators than conditional incentives (Church, 1993; Goyder, 1994; J. 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, Singer, & Corning, 2000; Zagorsky & Rhoton, 2008).

For mobile surveys there has been recent interest in increasing motivation to participate through the gamification of surveys (for a summary see Florian Keusch & Zhang, 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, Link, & Vanno, 2012; Link, Lai, & Vanno, 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).

2.4 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, p. 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, Harford, & Pautler, 1980; Leigh, 1993; Verbrugge, 1980) and some studies in which the effect does not seem to be present (Lemmens, Knibbe, & Tan, 1988; Persky, Strauss, Lief, Miller, & O'Brien, 1981; Searles, Perrine, Mundt, & Helzer, 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 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 as they choose 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 in contrast to cumulative objective burden is not considered to be summative . 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.

3 Data

3.1 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, 2017) 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, Gaia, Al Baghal, Burton, & Lynn, 2017).

Understanding Society Spending Study. The *Understanding Society* Spending Study (University of Essex. Institute for Social and Economic Research, 2018) 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, Burton, Wenz, & Read, 2018a). Screenshots for the app are documented in the separate Appendix C of the User Guide (Jäckle, Burton, Wenz, & Read, 2018b).

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.

3.2 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 focus only on these participants and does not present analyses examining those who did not participate in the study. Jäckle et al. (2019) 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 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 by the analytical sample each week was 136 out of a possible the 223 . 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 the Appendix. 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, Antoun, Couper, Kreuter, & Struminskaya, 2017).

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, Hubert, & Struyf, 2004), a robust measure of skewness, to adjust the boxplot for skewed distributions. The medcouple was estimated using the Stata package medcouple (Gelade, Verardi, & Vermandele, 2013). Potentially outlying values were identified as those app uses that took less than 3 seconds, or more than 173 seconds. 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 break down 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%).

3.3 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.

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.

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.

3.4 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 iden-

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

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

tify 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 & 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

[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 3
Descriptive statistics for the four measures of objective burden

	M	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 receipts	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 receipts	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 receipts	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 receipts	30	20	17	24	36
Nothing bought	9	8	5	7	10
Smartphone	29	24	14	23	37
Tablet	39	30	18	32	51

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 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 the question for 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

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

Likelihood – *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?*

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 – *Overall do you feel that the time and effort you put into participating in the Spending Study was...*

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 – *Overall how interesting was participating in the Spending Study?*

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 – *Overall, how easy or difficult did you find completing the Spending Study?*

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			

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

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 behaviors—Ability/Motivation. As with the existing mobile device behaviors, reported participation in certain existing financial behaviors are considered to be 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*

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
Willingness to download a survey app	Not willing	44	19.7
	Willing	179	80.3
Willingness to use the camera on device to to take photos or scan barcodes	Not willing	38	17.0
	Willing	185	83.0
Frequency of checking 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	
	s	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

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 at hand 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 measures was originally derived by Wenz, Jäckle, and Couper (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 longstanding 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.

4 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.

4.1 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 Maydeu-Olivares and D'zurilla (1995), Flora and Curran (2004) and Holgado-Tello, Chacón-Moscoso, Barbero-García, and Vila-Abad (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.

Using established thresholds for interpreting correlations (Hinkle, Wiersma, & Jurs, 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.

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.

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 $\chi^2 = 1040.56$, $df = 21$, $p < 0.001$ is indicative that the variables are not orthogonal from one another, and 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.

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 & 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 & 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.

4.2 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, Luniak, and Brzinsky-Fay (2006) were used to produce this plot.

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; Ver-

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
Avg time	0.04	0.15	0.90	0.16	0.22
Total duration	−0.06	0.85	0.49	0.03	0.39
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.

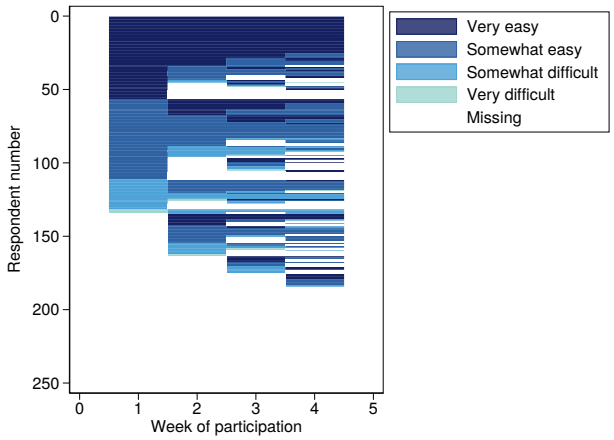

Figure 1. Sequence analysis graph documenting the sequence of weekly reported difficulty participating in the Spending Study

brugge, 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 $\beta = -0.29$, $p < 0.001$, 95% CI $[-0.34, -0.24]$.

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: $\beta = -0.41$, $\beta = -0.47$ and $\beta = -0.29$

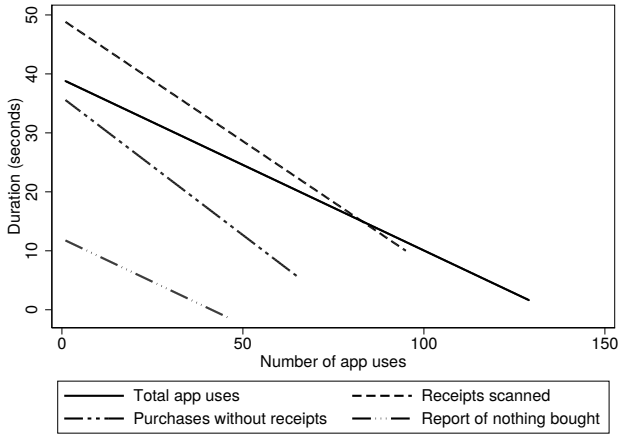

Figure 2. Fixed-effects regression models of changes in app use duration as participation continues split by type of app use

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. (2019) 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 $\beta = -0.0005$, 95% CI $[-0.0009, -0.0002]$ and reports of nothing bought $\beta = -0.0009$, 95% CI $[-0.0013, -0.0005]$ both p-values < 0.001. The proportion of purchases without receipts increased across the study $\beta = 0.0013$, 95% CI $[0.0009, 0.0017]$. 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.

4.3 RQ3: Does objective burden predict breaks in participation?

Due to the high levels of missingness in the end of week questionnaires it was 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 8.

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, that 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.

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

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 χ^2	4.79		15.21		2.42	
Spells	223		223		1559	

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

participation HR = 1.01, $p < 0.05$. There is 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 ratios 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 Azuero (2016) this corresponds with a small effect size.

4.4 RQ4: What factors predict subjective and objective burden?

Subjective burden. Table A2 in the Appendix 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. These thresholds 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. The formula below, as set out by (Borenstein, Hedges, Higgins, & Rothstein, 2009), allows the conversion of the threshold values of Cohen's d to log odds ratios, which can then be converted to odds ratios.

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

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 of so called "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 con-

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 once a week or more	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	0.30	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; * $p < .05$ ** $p < .01$ *** $p < .001$

straints 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.

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 the Appendix. 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.

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 participa-

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

	β	SE
Six pounds 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 poverty threshold	0.19	1.74
Time constrained	−0.87	1.08
Degree or higher	−0.20	1.13
Disabled/ long term illness	0.80	0.83
Female	2.09^{*}	0.94
Age	0.33^{***}	0.03
Type of purchase		
Reference: Scanned receipts		
Purchase without receipt	-10.69^{***}	1.03
Nothing bought	-33.46^{***}	1.21
Constant	27.68^{***}	3.99
Wald χ	1257.50^{***}	

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

tion. 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 behaviors keeping a budget 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 & Jung, 2001).

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

5 Discussion and conclusion

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 consis-

tent 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 (Oomens & Timmermans, 2008; Sharp & Frankel, 1983). 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 & Haraldsen, 2005; Sharp & 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 & 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., 2019), 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, Loges & 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, Schmidt, & Csikszentmihalyi, 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 con-

scientious 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. (2019) 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 reporting 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.

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 nonrespondents.

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. A separate analysis of the effects of device characteristics is currently in progress (Read, 2019).

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.

Acknowledgements

The author would like to thank Annette Jäckle, Tarek Al Baghal, Thomas Crossley, Mick Couper, Jonathon Burton, Paul Fisher, Carli Lessof, and Alexander Wenz for their comments and suggestions on earlier drafts of this paper. This research was supported by a +3 PhD studentship from the Economic and Social Research Council (ESRC) awarded to Brendan Read as part of an associated studentship funded from a grant awarded to Michaela Benzeval to conduct waves nine through eleven of Understanding Society [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 Understanding Society [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 Transforma-

tive Research scheme and the National Centre for Research Methods (NCRM) awarded to Annette Jäckle [grant number ES/N006534/1].

References

Abraham, K., Maitland, A., & Bianchi, S. (2006). *Nonresponse in the American Time Use Survey: Who is missing from the data and how much does it matter?* NBER Technical Working Paper Series 328.

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

Aguiar, M. & Hurst, E. (2007). Life-cycle prices and production. *American Economic Review*, *97*(5), 1533–1559.

Ampt, E. (2003). Respondent burden. In P. Stopher & P. Jones (Eds.), *Transport survey quality and innovation* (pp. 507–521). Bingley, West Yorkshire, UK.: Emerald Group Publishing.

Ansolabehere, S. & Schaffner, B. F. (2015). Distractions: The incidence and consequences of interruptions for survey respondents. *Journal of Survey Statistics and Methodology*, *3*(2), 216–239.

Appelhans, B. M., French, S. A., Tangney, C. C., Powell, L. M., & Wang, Y. (2017). To what extent do food purchases reflect shoppers' diet quality and nutrient intake? *International Journal of Behavioral Nutrition and Physical Activity*, *14*(1), 46.

Armstrong, J. S. (1975). Monetary incentives in mail surveys. *Public Opinion Quarterly*, *39*(1), 111–116.

Azuero, A. (2016). A note on the magnitude of hazard ratios. *Cancer*, *122*(8), 1298–1299.

Bartlett, M. (1951). The effect of standardization on a $\chi 2$ approximation in factor analysis. *Biometrika*, *38*(3-4), 337–344.

Biediger-Friedman, L., Sanchez, B., He, M., Guan, J., & Yin, Z. (2016). Food purchasing behaviors and food insecurity among college students at the University of Texas at San Antonio. *Journal of Food Security*, *4*(3), 52–57.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Chicago, IL: John Wiley & Sons.

Bradburn, N. (1978). Respondent burden. In *Proceedings of the section on survey research methods* (pp. 35–40). Alexandria, VA: American Statistical Association.

Branden, L., Gritz, R., & Pergamit, M. (1995). The effect of interview length on nonresponse in the National Longitudinal Survey of Youth. In *Proceedings of the 1995 Census Bureau Annual Research Conference* (pp. 129–154). Arlington, VA.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, *30*(1), 89–99.

Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, *13*(4), 996–1017.

Bucklin, R. E. & Gupta, S. [Sunil]. (1999). Commercial use of upc scanner data: Industry and academic perspectives. *Marketing Science*, *18*(3), 247–273.

Bulusu, N., Chou, C. T., Kanhere, S., Dong, Y., Sehgal, S., Sullivan, D., & Blazeski, L. (2008). Participatory sensing in commerce: Using mobile camera phones to track market price dispersion. In *Proceedings of the international workshop on urban, community, and social applications of networked sensing systems (UrbanSense 2008)*.

Burke, J. A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., & Srivastava, M. B. (2006). *Participatory sensing*. Paper presented to the 4th ACM Conference on Embedded Networked Sensor Systems, Boulder, CO.

Busemeyer, J. & Townsend, J. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432–459.

Chrisinger, B. W., DiSantis, K. I., Hillier, A. E., & Kumanyika, S. K. (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, A. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, *57*(1), 62–79.

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

Collins, M., Sykes, W., Wilson, P., & Blackshaw, N. (1988). Diffusion of technological innovation: Computer assisted data collection in the U.K. In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls II, & J. Waksberg (Eds.), *Computer assisted survey information collection* (Chap. Nonresponse: The UK experience). New York, NY: John Wiley & Sons.

Couper, M. (1997). Survey introductions and data quality. *Public Opinion Quarterly*, *61*(2), 317–338.

Couper, M. & Nicholls, W., II. (1998). The history and development of computer assisted survey information collection methods. In M. Couper, R. Baker, J. Bethlehem, C. F. Clark, J. Martin, W. Nicholls II, & J. O'Reilly (Eds.), *Computer assisted survey information collection*. New York, NY: John Wiley & Sons.

Crawford, S., Couper, M., & Lamias, M. (2001). Web surveys. *Social Science Computer Review*, *19*(2), 146–162.

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

Cullen, K., Baranowski, T., Watson, K., Nicklas, T., Fisher, J., O'Donnell, S., … Missaghian, M. (2007). Food category purchases vary by household education and race/ethnicity: Results from grocery receipts. *Journal of the American Dietetic Association*, *107*(10), 1747–1752.

Dale, T. & Haraldsen, G. (2005). Embedded evaluation of perceived and actual response burden in business surveys. In D. Hedlin, T. Dale, G. Haraldsen, & J. Jones (Eds.), *Developing methods for assessing perceived response burden* (pp. 112–125). Luxembourg: Eurostat.

Deng, L. & Cox, L. P. (2009). LiveCompare: Grocery bargain hunting through participatory sensing. In *Proceedings of the 10th workshop on mobile computing systems and applications*. ACM.

Dillman, D., Sinclair, M., & Clark, J. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *Public Opinion Quarterly*, *57*(3), 289–304.

Dyregrov, K. (2004). Bereaved parents' experience of research participation. *Social science & medicine*, *58*(2), 391–400.

Einav, L., Leibtag, E., Nevo, A., et al. (2008). *On the accuracy of Nielsen homescan data*.

Flora, D. & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466–491.

Fortunati, L. & Taipale, S. (2014). The advanced use of mobile phones in five European countries. *The British Journal of Sociology*, *65*(2), 317–337.

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

Fricker, S. [S.] & Tourangeau, R. (2010). Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Public Opinion Quarterly*, *74*(5), 934–955.

Galea, S., Nandi, A., Stuber, J., Gold, J., Acierno, R., Best, C., … Resnick, H. (2005). Participant reactions to survey research in the general population after terrorist attacks. *Journal of Traumatic Stress*, *18*(5), 461–465.

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313–328.

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.

Gelade, W., Verardi, V., & Vermandele, C. (2013). Medcouple. [Stata package]. Retrieved from https://ideas.repec.org/c/boc/bocode/s457699.html

Gerstel, E., Harford, T., & Pautler, C. (1980). The reliability of drinking estimates obtained with two data collection methods. *Journal of Studies on Alcohol*, *41*(1), 89–94.

Gillmore, M., Gaylord, J., Hartway, J., Hoppe, M., Morrison, D., Leigh, B., & Rainey, D. (2001). Daily data collection of sexual and other health-related behaviors. *Journal of Sex Research*, *38*(1), 35–42.

Goyder, J. (1994). An experiment with cash incentives on a personal interview survey. *Journal of the Market Research Society*, *34*(4), 1–7.

Graf, I. (2008). Respondent burden. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods* (p. 740). Thousand Oaks, CA: Sage Publications.

Greenwood, D., Ransley, J., Gilthorpe, M., & Cade, J. (2006). Use of itemized Till receipts to adjust for correlated dietary measurement error. *American Journal of Epidemiology*, *164*(10), 1012–1018.

Griffith, R., Leibtag, E., Leicester, A., & Nevo, A. (2009). Consumer shopping behavior: How much do consumers save? *Journal of Economic Perspectives*, *23*(2), 99–120.

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

Groves, R., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, *68*(1), 2–31.

Groves, R., Singer, E., & 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*(2), 251–268.

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

Guadagni, P. M. & Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. *Marketing science*, *2*(3), 203–238.

Guadagnoli, E. & Velicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–75.

Gupta, S. [Sachin], Chintagunta, P., Kaul, A., & Wittink, D. R. (1996). Do household scanner data provide representative inferences from brand choices: A comparison with store data. *Journal of Marketing Research*, 383–398.

Haraldsen, G. (2004). Identifying and reducing response burdens in internet business surveys. *Journal of Official Statistics*, *20*(2), 393–410.

Harris, J. M. (2005). Using homescan data and complex survey design techniques to estimate convenience food expenditures. American Agricultural Economics Association Conference Paper.

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Sage.

Hendershott, A., Edgar, J., Geisen, E., & Stringei, C. (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*. Paper presented to the 67th Annual Conference of the American Association for Public Opinion Research, Orlando, FL.

Hinkle, D., Wiersma, W., & Jurs, S. (2003). *Applied statistics for the behavioral sciences*. Boston, MA: Houghton Mifflin.

Holgado-Tello, F., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153–166.

Hubert, M. & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, *52*(12), 5186–5201.

Inman, J. J. & Winer, R. S. (1998). Where the rubber meets the road: A model of in-store consumer decision making.

Inman, J. J., Winer, R. S., & Ferraro, R. (2009). The interplay among category characteristics, customer characteristics, and customer activities on in-store decision making. *Journal of Marketing*, *73*(5), 19–29.

Jäckle, A., Burton, J., Couper, M., & Lessof, C. (2019). Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: Response rates and response biases. *Survey Research Methods*, *13*(1), 23–45.

Jäckle, A., Burton, J., Wenz, A., & Read, B. (2018a). Understanding Society: The UK Household Longitudinal Study. Spending Study 1, User Guide. Report. Retrieved from https://www.iser.essex.ac.uk/files/projects/household-finance/Spending_User_Guide.pdf

Jäckle, A., Burton, J., Wenz, A., & Read, B. (2018b). Understanding Society: The UK Household Longitudinal Study. Spending Study 1, User Guide. Appendix C: App Screenshots. Report. Retrieved from https://www.iser.essex.ac.uk/files/projects/household-finance/Spending_User_Guide_AppC.pdf

Jäckle, A., Gaia, A., Al Baghal, T., Burton, J., & Lynn, P. (2017). Understanding Society: The UK Household Longitudinal Study Innovation Panel, Waves 1–9, User Manual. Retrieved from https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/innovation-panel/user-guides/6849_ip_waves1-9_user_manual_June_2017.pdf

Kaiser, H. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151.

Kaiser, H. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401–415.

Kaiser, H. & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117.

Keusch, F., Antoun, C., Couper, M., Kreuter, F., & Struminskaya, B. (2017). Willingness to participate in passive mobile data collection. In *Annual meeting of the American Association for Public Opinion Research*. New Orleans, LA: American Association for Public Opinion Research.

Keusch, F. [Florian] & Zhang, C. (2017). A review of issues in gamified surveys. *Social Science Computer Review*, *35*(2), 147–166.

Knäuper, B., Belli, R., Hill, D. H., & Herzog, A. (1997). Question difficulty and respondents' cognitive ability: The effect on data quality. *Journal of Official Statistics*, *13*(2), 181–199.

Kohler, U., Luniak, M., & Brzinsky-Fay, C. (2006). Sq. [Stata package]. Retrieved from http://www.stata-journal.com/software/sj6-4

Kolenikov, S. (2008). Polychoric. [Stata package]. Retrieved from http://www.komkon.org/~tacik/stata/

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys. the effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865.

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

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

Larson, R. & Csikszentmihalyi, M. (1983). The experience sampling method. *New directions for methodology of social & behavioral science*, *15*, 41–56.

Lee, Y.-S. & Waite, L. J. (2005). Husbands' and wives' time spent on housework: A comparison of measures. *Journal of Marriage and Family*, *67*(2), 328–336.

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

Leicester, A. & Oldfield, Z. (2009b). Using scanner technology to collect expenditure data. *Fiscal Studies*, *30*(3–4), 309–337.

Leigh, B. (1993). Alcohol consumption and sexual activity as reported with a diary technique. *Journal of Abnormal Psychology*, *102*(3), 490–493.

Lemmens, P. H., Knibbe, R., & Tan, F. (1988). Weekly recall and diary estimates of alcohol consumption in a general population survey. *Journal of studies on alcohol*, *49*(2), 131–135.

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

Link, M., Murphy, J., Schober, M., Buskirk, T., Hunter Childs, J., & Langer Tesfaye, C. (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*(4), 779–787.

Loges, W. & Jung, J. (2001). Exploring the digital divide: Internet connectedness and age. *Communication research*, *28*(4), 536–562.

Lynn, P. (2014). Longer interviews may not affect subsequent survey participation propensity. *Public Opinion Quarterly*, *78*(2), 500–509.

Marr, J. (1971). Individual dietary surveys: Purposes and methods. In *World review of nutrition and dietetics* (Vol. 13, pp. 105–164). Karger Publishers.

Martin, S. L., Howell, T., Duan, Y., & Walters, M. (2006). The feasibility and utility of grocery receipt analyses for dietary assessment. *Nutrition Journal*, *5*(1), 10.

Mavletova, A. (2015). Web surveys among children and adolescents: Is there a gamification effect? *Social Science Computer Review*, *33*(3), 372–398.

Maydeu-Olivares, A. & D'zurilla, T. (1995). A factor analysis of the social problem-solving inventory using polychoric correlations. *European Journal of Psychological Assessment*, *11*(2), 98–107.

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

Newman, E., Willard, T., Sinclair, R., & Kaloupek, D. (2001). Empirically supported ethical research practice: The costs and benefits of research from the participants' view. *Accountability in Research*, *8*(4), 309–329.

Office of Management and Budget. (2006). Standards and guidelines for statistical surveys. Retrieved from https://unstats.un.org/unsd/dnss/docs-nqaf/USA_standards_stat_surveys.pdf

Oomens, P. & Timmermans, G. (2008). *The dutch approach to reducing the real and perceived administrative burden on enterprises caused by statistics*. Paper presented to the 94th DGINS Conference, Vilnius, Lithuania.

Ozarslan, S. & Eren, P. E. (2014). Text recognition and correction for automated data collection by mobile devices. In *Imaging and multimedia analytics in a web and mobile world 2014* (Vol. 9027). International Society for Optics and Photonics.

Persky, H., Strauss, D., Lief, H., Miller, W., & O'Brien, C. (1981). Effect of the research process on human sexual behavior. *Journal of Psychiatric Research*, *16*(1), 41–52.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74–97.

Raiklin, E. & Uyar, B. (1996). On the relativity of the concepts of needs, wants, scarcity and opportunity cost. *International Journal of Social Economics*, *23*(7), 49–56.

Rankin, J. W., Winett, R. A., Anderson, E. S., Bickley, P. G., Moore, J. F., Leahy, M., ... Gerkin, R. E. (1998). Food purchase patterns at the supermarket and their relationship to family characteristics. *Journal of Nutrition Education*, *30*(2), 81–88.

Ransley, J., Donnelly, J., Botham, H., Khara, T., Greenwood, D., & Cade, J. (2003). Use of supermarket receipts to estimate energy and fat content of food purchased by lean and overweight families. *Appetite*, *41*(2), 141–148.

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.

Roberts, C., Eva, G., Allum, N., & Lynn, P. (2010). *Diffusion of technological innovation: Computer assisted data collection in the U.K.* ISER Working Paper Series 36.

Ruch, F. (1941). Effects of repeated interviewing on the respondent's answers. *Journal of Consulting Psychology*, *5*(4), 179–182.

Ruzek, J. & Zatzick, D. (2000). Ethical considerations in research participation among acutely injured trauma survivors: An empirical investigation. *General Hospital Psychiatry*, *22*(1), 27–36.

Scagnelli, J., Bailey, J., Link, H., M.W. Moakowska, & Benezra, K. (2012). *On the run: In the moment smartphone data collection*. Paper presented to the 67th Annual Conference of the American Association for Public Opinion Research, Orlando, FL.

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

Schwarz, N. (2012). Retrospective and concurrent self-reports: The rationale for real-time data capture. In *The science of real-time data capture: Self-reports in health research*. Oxford University Press, New York.

Searles, J., Perrine, M., Mundt, J., & Helzer, J. (1995). Self-report of drinking using touch-tone telephone: Extending the limits of reliable daily contact. *Journal of Studies on Alcohol*, *56*(4), 375–382.

Sehgal, S., Kanhere, S. S., & Chou, C. T. (2008). Mobishop: Using mobile phones for sharing consumer pricing information. In *International conference on distributed computing in sensor systems*.

Sendelbah, A., Vehovar, V., Slavec, A., & Petrovčič, A. (2016). Investigating respondent multitasking in web surveys using paradata. *Computers in Human Behavior*, *55*, 777–787.

Sharp, L. M. & Frankel, J. (1983). Respondent burden: A test of some common assumptions. *Public Opinion Quarterly*, *47*(1), 36–53.

Singer, E., van Hoewyk, J., Gebler, N., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, *15*(2), 217–230.

Stilley, K. M., Inman, J. J., & Wakefield, K. L. (2010). Planning to make unplanned purchases? The role of in-store slack in budget deviation. *Journal of consumer research*, *37*(2), 264–278.

Sudman, S. (1964a). On the accuracy of recording of consumer panels: I. *Journal of Marketing Research*, *1*(2), 14–20.

Sudman, S. (1964b). On the accuracy of recording of consumer panels: II. *Journal of Marketing Research*, *1*(3), 69–83.

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

University of Essex. Institute for Social and Economic Research. (2017). Understanding Society: Innovation Panel, Waves 1–9, 2008–2016. [data collection]. 8th Edition. UK Data Service. SN: 6849. Dataset. doi:10.5255/UKDA-SN-6849-9

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

Van Heerde, H. J., Leeflang, P. S., & Wittink, D. R. (2000). The estimation of pre- and postpromotion dips with store-level scanner data. *Journal of Marketing Research*, *37*(3), 383–395.

Verbrugge, L. (1980). Health diaries. *Medical Care*, *18*(1), 73–95.

Volkova, E., Li, N., Dunford, E., Eyles, H., Crino, M., Michie, J., & Mhurchu, C. N. (2016). ''smart" RCTs: Development of a smartphone app for fully automated nutrition-labeling intervention trials. *JMIR mHealth and uHealth*, *4*(1).

Walker, E., Newman, E., Koss, M., & Bernstein, D. (1997). Does the study of victimization revictimize the victims? *General Hospital Psychiatry*, *19*(6), 403–410.

Waterlander, W. E., de Boer, M. R., Schuit, A. J., Seidell, J. C., & Steenhuis, I. H. (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*(4), 886–895.

Wenz, A., Jäckle, A., & Couper, M. (2019). Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods*, *13*(8), 1–22.

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

Yammarino, F., Skinner, S., & Childers, T. (1991). Understanding mail survey response behavior a meta-analysis. *Public Opinion Quarterly*, *55*(4), 613–639.

Young, J., O'Halloran, A., McAulay, C., Pirotta, M., Forsdike, K., Stacey, I., & Currow, D. (2015). Unconditional and conditional incentives differentially improved general practitioners' participation in an online survey: Randomized controlled trial. *Journal of Clinical Epidemiology*, *68*(6), 693–697.

Young, P. & Schmid, C. (1956). *Scientific social surveys and research: An introduction to the background, content, methods, principles, and analysis of social studies*. Englewood Cliffs, NJ: Prentice-Hall.

Yu, E., Fricker, S., & Kopp, B. (2015). *Can survey instructions relieve respondent burden*. Paper presented to the 70th Annual Conference of the American Association for Public Opinion Research, Hollywood, FL.

Zagorsky, J. & Rhoton, P. (2008). The effects of promised monetary incentives on attrition in a long-term panel survey. *Public Opinion Quarterly*, *72*(3), 502–513.

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

Appendix
Tables

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	17.49
One	34	15.25
Two	31	13.90
Three	30	13.45
Four	89	39.91

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	2.25	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 once a week or more	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;
* $p < .05$ ** $p < .01$ *** $p < .001$