Understanding Society

Working Paper Series

No. 2019-01

January 2019

The influence of device characteristics on data collection using a Mobile App.

Brendan Read

University of Essex

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Non-technical summary

Previous research has found differences in outcomes between surveys completed on mobile devices and PCs. This research considers the effect of the large number of different mobile devices that people use to respond to surveys. Not much it known about how the different specifications of these devices might affect the quality of data collected using them.

The data used is from the *Understanding Society* Spending Study One, this was an app-based study asking respondents to take pictures of receipts or submit information about purchases using their mobile devices. Respondents to the *Understanding Society* Innovation Panel were invited to take part in the Spending Study between waves nine and ten of the Innovation Panel. The make, model and operating system of the mobile devices used were captured using the Spending Study app. Additional data on device characteristics was collected using Amazon mTurk and web scraping including: Random-Access Memory (RAM), camera quality and processor performance.

Several survey outcomes were looked at, including: the length of time it took to use the app, the quality of images of receipts, whether data was submitted as a picture of a receipt or input manually, how many shopping items were on the image of the receipt.

It was found that the device used can have a large effect on certain survey outcomes. This was most noticeable for the quality of the photographs of receipts. Additionally, certain characteristics of mobile devices were found to matter more than others in their effect on survey outcomes. For example, whether a mobile device was a tablet or smartphone, whether it was an Apple or Android device, and how much RAM the device had all affected more than one survey outcome. Finally, some of the results that were found seem to be because of different respondents selecting different mobile devices, rather than the effect of the devices themselves.

The influence of device characteristics on data collection using a Mobile App.

Brendan Read – (University of Essex)

Abstract:

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

Keywords: Data Quality, Process Quality, Respondent Behaviour, Expenditure, Household Panel Survey, Amazon mTurk**.**

JEL Classification: C81, C83

Acknowledgments: The author would like to thank Annette Jäckle, Tarek Al Baghal, Thomas Crossley, Mick Couper, and Alexander Wenz for their comments and suggestions on 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 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 Transformative Research scheme and the National Centre for Research Methods (NCRM) awarded to Annette Jäckle [grant number ES/N006534/1].

Contact information: Brendan Read, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex UK, CO4 3SQ email: bread@essex.ac.uk.

1 Introduction

Data collection using mobile devices, whilst offering many new opportunities for innovations in survey methodology, presents its own challenges. One challenge is how to address the diverse range of available makes and models that make up the mobile device market. There were an estimated 1,600 models of mobile device available on the market in 2009 (Zahariev et al. 2009 cited in Callegaro 2010). The number of Android models, the most diverse mobile operating system by number of model types, was reportedly around 24,000 in 2015 (Open Signal 2015).

Such diversity raises questions as to the extent to which the use of different devices influences the data collection process and whether we should be concerned by this influence. These questions are concerning because diversity in the device used comes as a challenge to one of the central tenets of survey design, standardisation. If differences in device result in differences in the survey experience, or in the responses collected, it becomes important to either mitigate those biases, or to be able to correct for them.

When exploring the effect of device differences there are a range of different outcomes that can be examined. One set of outcomes is indicators of the response behaviour of those completing the survey, such as response time durations. A second set of outcomes is data quality indicators such as missing data, or lower quality responses.

It seems likely that differences may be accentuated when data collection makes use of hardware capabilities of devices to collect data besides those collected in a traditional survey. As more features of the device are drawn upon it seems probable that differences in device specifications will have a greater impact. Take for example asking respondents to use the cameras on their mobile device to take pictures for data collection, as was the case in the data collection task examined in this research. Here we might expect that the quality of the camera will affect the data that is collected. However, we might also find that the processor speed and available memory of the device may have an impact on how quickly the respondent can complete the task, or if they can complete the task at all.

Another question that remains unanswered is whether observed differences in outcomes are the result of device characteristics or whether they result from different types of people selecting different types of device. This would mean the selection mechanism through which individuals choose their devices would also account for the observed differences in outcomes. It is therefore necessary to model both device characteristics and respondent characteristics to fully understand the influence of device upon data collection outcomes.

To examine the influence of device characteristics on data collection using a mobile app this paper uses data from the *Understanding Society* Spending Study. The Spending Study tasked respondents with using a mobile app to record their expenditure across a month. Respondents could take a picture of a receipt, manually enter some data about a purchase, or report no spending on a given day. Spending Study data is supplemented with data from wave nine of the *Understanding Society* Innovation Panel and additional data collected on characteristics of the devices used to participate in the Spending Study. These data are used to examine the following research questions:

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

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

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

2 Background

To date, there are no papers that have explicitly examined the model of mobile device used in a survey. Most of the existing literature on device effects has been framed within the mode effects paradigm. One of the drawbacks of this is that it only allows analysis of differences between broad categorisations such as comparing between smartphones, tablets and desktops. To extend upon this and to examine device effects in more detail it is useful to consider device effects as a parallel to interviewer effects. One reason this is useful is that if we consider how the role of the device is like that of an interviewer in a face-to-face survey then it is possible to draw on the extensive interviewer effects literature to establish a conceptual understanding of how and why device effects may occur. In addition, the similar hierarchical structure of survey responses clustered in either devices or interviewers mean that establishing the parallels between the two can help shape the appropriate methodology for studying device effects.

2.1 Mode effects paradigm

The potential effects of the mode used to administer a survey has long been recognised (Deming 1944) and substantial evidence in support of mode effects has been found (e.g. Groves and Kahn 1979, Dillman and Christian 2005, Elliott, Zaslavsky et al. 2009). For a comprehensive discussion of the effects of the mode of data collection the reader is directed to Jäckle, Roberts et al. (2010). In short, the main concern has been the degree to which different modes used to administer surveys contribute to different sources of error, whether the resulting data from different modes are then comparable. With the rise in the number of web administered surveys, research has been conducted to examine the potential for mode effects in web administered surveys, comparing them to face-to-face and telephone surveys (e.g. McCabe, Boyd et al. 2002, Link and Mokdad 2005, Shin, Johnson et al. 2012).

Much of the research into device effects has followed in the tradition of the mode effects literature. Research into device effects has typically made comparisons between surveys completed using a PC (defined as a desktop or laptop computer) and those completed using mobile devices (defined as a mobile phone or tablet) (e.g. De Bruijne and Wijnant 2013, Fernee and Sonck 2013, Mavletova 2013, De Bruijne and Wijnant 2014, Lugtig and Toepoel 2015, Struminskaya, Weyandt et al. 2015, Revilla, Toninelli et al. 2016, Couper and Peterson 2017, Keusch and Yan 2017, Revilla 2017, Revilla and Couper 2018). It should be noted that many of these examples do not refer to PC and mobile survey completion as separate modes; instead, both PC and mobile survey completion are considered as sub-types of the web mode.

Evidence of a number of differences in surveys completion between PCs and mobile devices has been found. Revilla, Toninelli et al. (2016) found that smartphone respondents typically provided longer answers to open-ended questions than those using PCs. This finding was replicated by Antoun, Couper et al. (2017).

Couper and Peterson (2017) found that respondents typically took longer to answer questions on mobile devices, and that much of this could be attributed to increased time spent scrolling. Several other studies have also found evidence that respondents take longer to complete surveys when using a mobile device (De Bruijne and Wijnant 2013, Mavletova 2013, Mavletova and Couper 2013, Cook 2014, Wells, Bailey et al. 2014, Struminskaya, Weyandt et al. 2015). However, some research has found no differences in the average response times between mobile and desktop respondents (Toepoel and Lugtig 2014, Lugtig and Toepoel 2015).

Another finding is that respondents using mobile devices are less likely to straightline than those using a PC (Lugtig and Toepoel 2015, Keusch and Yan 2017). However, conflicting evidence that mobile respondents may in fact be more likely to straightline has also been found (Struminskaya, Weyandt et al. 2015), and it has been suggested that this may be dependent on whether the questions are presented in a grid .

Research into device effects for several other forms of measurement error have found no evidence of such effects, including: disclosure of sensitive information (Mavletova 2013, Revilla, Toninelli et al. 2016, Antoun, Couper et al. 2017); acquiescence (Keusch and Yan 2017); mid-point responding (Keusch and Yan 2017); item nonresponse (Lugtig and Toepoel 2015, Revilla and Couper 2018); and primacy effects (Mavletova 2013, Lugtig and Toepoel 2015).

2.2 Interviewer effects conceptually

Considering device effects in the same fashion as mode effects is useful for establishing differences in broad categorisations such as PCs compared to mobile devices. However, to gain a more granular understanding of device effects, in particular, the clustering effect of a specific model of device, it becomes necessary to turn to a different body of literature to inform our thinking. In this instance, the device effect can be seen to be similar to interviewer effects. To draw this parallel the survey must be seen conceptually as an interaction between the respondent and an agent of the researcher, but instead of the interviewer, the agent is the respondent’s own device. The device takes the place of the interviewer in a face-to-face interview. One key difference (that is returned to in the section on respondent characteristics in the measures section below) is that the respondents have themselves selected which device they are using.

From a conceptual standpoint it is therefore useful to consider the framework of interviewer effects outlined by Sudman and Bradburn (1974) which suggests three ways in which the interviewer may bias responses. The first of these they term *interviewer role demands*. This refers to the degree of autonomy given to interviewers when the researcher specifies how they should go about conducting the interviews. They suggest this is a continuum, with a highly standardised approach with zero deviation from the interview script allowed at one end. At the other end of the continuum interviewers are asked to adopt an approach where they are encouraged to adapt their behaviour to best allow them to complete an interview with a given respondent.

In terms of web surveys, the parallel to this would be the degree to which the functionality of the survey website or app is explicitly defined. When designing the survey task, the researcher can choose the degree to which they are explicit in outlining the processes the device goes through as the respondent completes the task. Some, or all, of the background tasks that the device performs may rely on defaults or logic that are provided by the device.

To illustrate this, consider the example of collecting survey data in the form of photographs, taken by respondents using the camera on their mobile device. The researcher, at the design stage, can choose to explicitly make choices such as where on the device the image is stored, the quality of the image that is taken, or whether to use a standardised protocol for transmission of that photograph. Instead of explicitly defining each of these choices, the researcher may instead decide to allow the device to make these choices, either based on defaults set by the device manufacturer, or through logic defined in the software libraries on top of which the app is designed. These decisions, and the differences that they make across different devices may be one source of device effects.

The second source of bias outlined by Sudman and Bradburn is *interviewer role behaviour.* This refers to the degree to which the interviewer carries out the role demands. Even in the case of the highly standardised interview protocol, an interviewer may deviate from the specification. This is perhaps the source of bias where the parallel for device effects is easiest to consider. In terms of device effects, the direct comparison would be where some characteristic of a specific device means that the survey task does not function in the way it was programmed.

By way of illustration consider the example of a mobile device with a lower specification of Random-Access Memory (RAM) that may run out of available memory whilst completing some portion of the survey process. If this were to happen, this may result in the survey app crashing, meaning the data collecting process is halted. Deviations such as this represent a second potential source of device effects.

The final source of bias outlined by Sudman and Bradburn are the *extra-role characteristics of the interviewer.* These are those characteristics that the interviewer possesses which are separate from their role as an interviewer that nevertheless may affect the interviewer-respondent relationship, particularly through the respondent’s perception of the interviewer. Race, social class, educational level, age, and religious, ethnic, political, or other affiliations are offered as examples of such characteristics.

In terms of device effects, the direct comparison would be the indirect effect of device features upon data collection. This would most likely occur as a result of the respondent’s pre-existing relationship with, and perception of, their mobile device. If, for example, a respondent finds it easy to take a picture using their mobile device they may be more motivated to do so, than a respondent who finds this hard to do. Ultimately, just as in the case of the extra-role characteristics of interviewers, the potential for bias only occurs during the interaction with the respondent. It is not the device or the interviewer who is the sole cause of the bias, but the respondent’s reaction to the device or interviewer. Such effects of the respondent’s attitude towards their device therefore makes up a third class of potential device effects.

2.3 Interviewer effects methodology

As well as informing how to conceptually consider device effects, the comparison to interviewer effects can help to inform how to study device effects. The use of multilevel models to account for the clustering effect of respondents within interviewers has long been established (Wiggins, Longford et al. 1990, Pickery, Loosveldt et al. 2001, West and Olson 2010, Jäckle, Lynn et al. 2011, West and Elliott 2014, West, Conrad et al. 2018). In addition to the second-level of analysis to model the clustering effect of interviewers, some research has included a third level, measuring the clustering effect of the area a respondent lives in (Schnell and Kreuter 2005). Attempts have been made to disentangle interviewer effects from area effects by fitting cross-classified models with interviewer and area as higher levels that are not hierarchically nested (O'Muircheartaigh and Campanelli 1998, O'Muircheartaigh and Campanelli 1999, Durrant, Groves et al. 2010, Brunton-Smith, Sturgis et al. 2017).

Most of these studies have used some variation of intra-interviewer correlations (IIC), or interviewer design effects, to assess the clustering effect of interviewers. These measures are related to one another and are derived from the decomposition of variance in the multilevel models. The size of reported intra-interviewer correlations has been quite varied. O'Muircheartaigh and Campanelli (1998) suggest that correlations of larger than 0.10 are rare. In their research they found correlations ranging from 0.06 – 0.17. Jäckle, Lynn et al. (2011) reported IICs ranging from 0.04 – 0.07. West and Olson (2010) cite a wider range of observed IICs ranging from 0.01 – 0.12. It has been suggested that even relatively small interviewer clustering effects can have large impacts when estimating statistics. In both cases, assuming an average of 30 respondents per interviewer, an IIC of 0.01 would result in a twenty-nine percent increase (West and Olson 2010), and an IIC of 0.02 would result in a fifty-four percent increase (West, Conrad et al. 2018) in the variance of an estimated mean.

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

3 Data

3.1 Study designs

Three datasets are used for the analyses in this research. The main data set used is the *Understanding Society* Spending Study, this was supplemented with data from the *Understanding Society* Innovation Panel and some additional data collection capturing characteristics of the devices that were used by respondents in the Spending Study. Details of all three data sets can be found below.

3.1.1 Innovation Panel

The *Understanding Society* Innovation Panel (University of Essex. Institute for Social and Economic Research 2018) is the experimental and methodological research portion of the UK Household Longitudinal Study. The Innovation Panel is an annual household panel survey with a stratified and clustered sample that is representative of the Great British population south of the Caledonian Canal. Data from the ninth wave of the study are used as covariates in the analyses presented in this research. The wave nine (IP9) sample consists of remaining sample members from the original sample along with respondents from two additional refreshment samples who have participated 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 ninth wave had a household response rate of 84.7% and an individual response rate of 85.4% within responding households (Jäckle, Al Baghal et al. 2018).

3.1.2 Spending Study One

The *Understanding Society* Spending Study One (University of Essex. Institute for Social and Economic Research 2018) was an inter-wave data collection task that collected additional information about the expenditure of Innovation Panel members. The first Spending Study took place between waves nine and ten of the Innovation Panel. Data collection used an app that asked respondents in the study to use their mobile devices (smartphones or tablets) to submit information on their purchases as they made them. This app was developed by Kantar Worldpanel, with whom the study was conducted in partnership.

Respondents were asked to submit data about their purchasing behaviour in one of three forms: scanned pictures of receipts, taken using the mobile devices camera; self-reports of purchases, including details of what category of item the purchases were and for how much; or reports of days without spending. More details can be found in the Spending Study One user guide(Jäckle, Burton et al. 2018). Three sets of additional questionnaires were asked of respondents: a registration questionnaire at the start of the study; a series of end of week questionnaires throughout their time in the study; and an end of project questionnaire after they had finished participating.

All adult members (aged 16 and over) of households where at least one person responded at IP9 were included in the issued sample for Spending Study One; except those known to have refused to take part in the Innovation Panel long-term. Incentives were given to respondents in the form of Love2Shop gift vouchers or gift cards. The amount respondents received varied depending on their level of participation in the study. An initial incentive was offered for completing the registration survey and downloading the app, this had two experimental conditions, £2.00 and £6.00. Allocation to the incentive treatments was made at the household level. An additional £5.00 incentive was offered to a random subsample of all members of half of all households where nobody had participated by the third week of the study. Respondents received an incentive of 50p for every day they used the app. Completion of each end of week survey earned respondents an additional 50p and completion of the end of project survey earned £3.00. To maximise compliance with the task across the month, an additional incentive of £10.00 for using the app for 31 consecutive days was offered. When this additional incentive was administered, this criterion was relaxed to participation on 27 out of 31 days.

There were 274 people who used the Spending Study app at least once. This constitutes a response rate of 11.5% amongst the 2,383 Innovation Panel members who were invited to participate. For the purposes of the analyses presented here, this sample was constrained to the 255 respondents for whom IP9 data on all the respondent characteristics used as predictors was available. Those models not including the respondent characteristics were also fitted using all 274 respondents;

there was little to no difference in the results, so only results from the constrained sample are presented here.

3.1.3 Device characteristics data

The make and model of the device used to complete each app use was captured within the main Spending Study One app. There were 97 different makes and model of device amongst all 274 Spending Study respondents, and 90 makes and models used by the analytical sample of 255 respondents. The Spending Study app also captured the Operating System (OS). Whether the device was a tablet or smartphone was then derived during the data cleaning process for the Spending Study.

Additional data collection then took place to capture specific characteristics of each of these mobile devices. This data collection task was completed using the Amazon Mechanical Turk (mTurk) micro-task crowdsourcing platform. The mTurk platform allows the creation of so-called Human Intelligence Tasks (HITs), which harness the labour supplied by the platform's workers to complete them. There has been growing interest in using Amazon mTurk for social science data collection (e.g. Paolacci, Chandler et al. 2010, Buhrmester, Kwang et al. 2011, Berinsky, Huber et al. 2012, Mason and Suri 2012). This has also extended to using mTurk for survey methodology data collection (Antoun, Zhang et al. 2016, Keusch and Yan 2017).

Screenshots of the HIT used to collect the additional device characteristics can be found in Appendix A. Workers were presented with the make and model of a given device[1], and asked to provide values for a series of device characteristics. Workers were paid $0.25 for each HIT they completed. Five device characteristics were collected using the HIT: the device's Random-Access Memory (RAM) (measured in gibibytes or mebibytes), processor speed (measured in hertz), camera quality (measured in megapixels), storage space (in gigabytes or megabytes) and screen size (measured diagonally in inches). Some cleaning was needed to extract the

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

numerical value, and units from the text input by the workers when completing the HITs. However, this was relatively straightforward to complete using Boolean string matching, or regular expressions. Of these five measures, only the device’s RAM and camera quality were ultimately included as measures in the models presented here.

Screen size was not included, as there was limited variation of screen sizes amongst tablets, or amongst smartphones. The device type was the more important distinction, as opposed to the size of the screen, including this as a continuous measure resulted at times in an apparent linear effect, when in fact the important relationship was whether a device was a smartphone or tablet.

The storage space variable that was captured was ultimately excluded from analysis as this was a very imprecise measure. The challenge when capturing storage space is that the same model of device might be available in variants with different default storage capacity; for example, the Apple iPhone 6 is available in 16/32/64/128 GB versions. Whilst it was possible to capture the full range of available storage capacities using mTurk, it was not possible to determine exactly which variant the devices used in the Spending Study were, or whether two devices that were the same model had different storage capacities. This issue was further compounded by the fact that some devices allow the use of additional memory cards to provide extra storage. Finally, even if the full storage capacity of the device could be identified, it is the amount of available storage on the device that would actually affect performance.

The processor speed measure captured was problematic because a number of newer mobile devices use multiple cores in their CPUs. Therefore, a large number of the reported processor speeds only captured the performance of one core, not the total performance of the processor. Consequently, an alternative source of data for the performance of device processors was used, details of this can be found in the measures section below.

3.2 Multi-level structure

Throughout the analyses in this research the data are considered to have a four-level cross-classified structure. This structure is illustrated in the classification diagram in Figure 1.

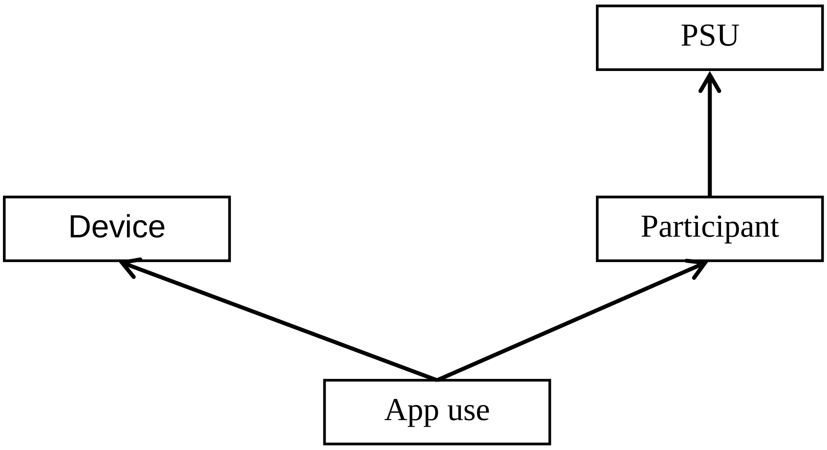


Figure 1. Classification diagram for the four-level cross-classified data structure.

The lowest level considered is the individual app uses. Each individual app use is then nested within two second-level clusters. The first of these is the specific model of device that was used to complete the app use. For example, all app uses completed on Apple iPhone 6s would be in the same cluster. The second is the respondent who completed that app use. Finally, the Primary Sampling Unit (PSU) to which the respondent belongs is also included to account for the complex clustered sample design of the Innovation Panel. There are no variables measured at the PSU level included in the analyses. Variables measured at all three of the other levels are included. Household was also considered as an additional level, but models fitted to include households suggest there was little clustering effects of households, and therefore the more parsimonious four-level structure is presented here.

4 Measures

4.1 Data quality indicators - App use level

Ultimately, what is f interest in this research is the contribution of systematic error or biases that the device used produces in estimates generated from the data collected. Without validation of the true measure (in this instance a total record of true expenditure across the duration of the study) it is necessary to consider the contribution to error indirectly. As the error itself is unobservable it is instead useful to examine the effect of device upon observable measures that are assumed to be correlated with the degree of error in estimates produced using the data. Four data quality indicators have been identified and are outlined below. Descriptive statistics for all four measures can be found in Table 1.

Table 1. *Descriptive statistics of app use outcomes.*

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

4.1.1 App use duration

Response times have previously been examined as a data quality indicator (Malhotra 2008, Yan and Tourangeau 2008, Galesic and Bosnjak 2009). Typically, this has involved the assumption that shorter response times are more likely to be indicative of satisficing, and as a result associated with increased errors. However, as is noted by Malhotra, the relationship between response time and data quality is not easily disentangled. For example, when considering the effect of the device used on app use duration it does not make sense to suggest that faster devices result in lower quality data. In contrast, it seems likely that the opposite may be true, that slower devices may in fact result in poorer quality data. The justification for this is that slower devices may contribute to an increased perception of the time it takes to participate. The negative impact of longer perceptions of time taken to complete a survey on response propensity is well documented (Collins, Sykes et al. 1988, Yammarino, Skinner et al. 1991, Dillman, Sinclair et al. 1993, Groves, Singer et al. 1999, Crawford, Couper et al. 2001, Galesic and Bosnjak 2009, Roberts, Eva et al. 2010). Considering this in the context of the Spending Study, if a device results in app uses taking on average longer, the assumption is that this means that it is less likely the respondent using that device will report all their purchases.

The duration of app uses was measured in seconds. A number of extreme responses were observed, and the possibility that these may be outlying responses was considered. Using the same method as in Read (2018) an adjusted boxplot was used to classify outliers. This method takes into account the skewness of the distribution by using the medcouple (Brys, Hubert et al. 2004), a robust measure of the skewness of the data. This is applied to a boxplot as suggested by Hubert and Vandervieren (2008) by to adjust the interval of the boxplot to take into account the skewness of the data. All data points outside the adjusted interval are then coded as outliers. These outliers are excluded for those models in this research that regress app use duration on predictors. Separate models are then fitted to examine the associations between different predictors and the probability of an app use being an outlier in terms of duration. The mean app use duration was 31 seconds and the percentage of app uses with outlying durations was 3.39 percent.

4.1.2 Type of app use

A second data quality indicator is the type of app use. The three app use types were: taking a picture of a receipt, manually entering data about a purchase, or reporting nothing bought that day. Here, the assumption is that app uses that are reports of purchases made without receipts, or of nothing bought, may be more likely to represent increased error if the “true” response should have been a scanned receipt. One potential issue is that of course both of these categories of app uses can be valid responses and may not represent an increase in error. Therefore, a higher proportion of these types of app use clustered within a given respondent may represent a true difference in purchasing behaviour. In contrast, there is no reason to believe that the device should have a direct effect on the type of app use, therefore any observed effect would suggest a bias towards a certain type of app use. This measure is included as a binary indicator of whether the app use was a scanned receipt, (a value of 0) or one of the other two types of app use (assigned a value of 1). Forty-eight percent of app uses were scanned receipts, and fifty-two percent of app uses were either purchases without receipts, or reports of nothing bought.

4.1.3 Image quality

A third data quality indicator analysed is the quality of the images produced by respondents when scanning their receipts. Here the data quality assumption is more easily understood, namely that poorer quality images increase the potential for error, either because information cannot be collected from them, or because the information collected may be incorrect. This measure is a binary indicator with fully readable receipts being coded as zero. Receipts which could not be fully read, either because part of the receipt was unreadable, the whole receipt was unreadable, or there was not an image captured, were coded as one. Ninety-two percent of receipts were fully readable, and eight percent were either partially readable, unreadable, or missing. For both this measure, and the number of items on receipts (below), the number of respondents and devices is slightly reduced as some respondents never submitted a scanned receipt. The number of respondents and devices for these measures is reported in Table 4 in the Results section.

4.1.4 Number of items on the receipt

The final data quality indicator is the number of items that were on the receipt. Once again, this is a variable that might reasonably be expected to vary as a direct effect of respondent characteristics, as different purchasing behaviour will affect the composition of receipts. However, as is the case with the other data quality indicators, there is no reason to suspect that the characteristics of the device used should have a direct effect on the number of items on a receipt. Therefore, the assumption here is that shorter receipts as a result of device characteristics may represent a downwards bias caused by the device used. The mean number of items on receipts was 7.

4.2 Device Characteristics – Device level

4.2.1 Operating System (OS)

The first of five device characteristics that was identified as possibly affecting data quality in the context of the Spending Study was the Operating System (OS). Descriptive statistics for all five device characteristics are in Table 2. The app was available for iOS and Android, and which OS the respective device was using was captured within the app itself. Differences in the software architecture of the two operating systems were the main reason that it was believed that the OS of the device used may affect data quality. For example, iOS and Android differ in how they handle memory allocation, which can have a significant effect on both app speed and processing performance (Rinaldi 2017, Lee 2018, Brownlee 2019). There are also differences in the demographics of iOS and Android users, with men being found to be slightly more likely to be iOS users than women (Fluent 2016). Amongst the device models used in the Spending Study, 29% were iOS devices and 71% were Android devices.**Mobile device type**

The second device characteristic considered was the type of mobile device used, meaning whether the device was a smartphone or a tablet. Existing research has found differences between smartphone and tablet responses in surveys; it has been suggested that responses to surveys using tablets are at times more similar to PC responses than smartphone responses (Struminskaya, Weyandt et al. 2015). The type of mobile device used to complete the app uses was captured within the app.

In terms of the Spending Study, the difference in size between tablets and smartphones was considered relevant for two reasons. The first of these is that the increased size of tablets may potentially make it more difficult to take photographs, as they are potentially bulkier and more cumbersome for respondents to use to take the photograph. However, the increased screen size may also have made it easier to see the photograph as it was being taken, potentially resulting in higher quality images. As was noted above, screen size itself was considered as a variable for the models estimated in this paper, however there was little variation in screen size within tablets, or within smartphones. Twenty-two percent of devices were tablets, and seventy-eight percent of devices were smartphones.

Table 2. *Descriptive statistics for the five device characteristics.*

		Device models (n= 90)
Operating system	*Apple*	29%
	Android	71%
Device Type	*Smartphone*	78%
	Tablet	22%
RAM (Gibibytes)	*Mean*	1.79
	SD	0.99
	Min	0.50
	Median	1.50
	Max	4.00
Camera quality (Megapixels)	*Mean*	9.57
	SD	5.01
	Min	0.70
	Median	8.00
	Max	20.70
Processor performance score	*Mean*	2.13
	SD	1.54
	Min	0.21
	Median	1.63
	Max	8.98

4.2.3 Camera quality

The third device characteristic used as predictor of data quality was the quality of the main camera on the mobile device, measured in megapixels. This was coded in the mTurk data collection. Each device was coded by three different workers, and then methodology created for assessing inter-coder reliability was adopted to assess the consensus of the three coders. For 80% of devices the three workers were in perfect agreement as to the value of the quality of the camera. The corresponding kappa statistic of $\kappa = 0.83$ was above the 0.80 threshold describe as “almost perfect” agreement (Landis and Koch 1977). Similarly, the value for Krippendorff’s alpha was above the recommended 0.80 threshold (Krippendorff 2004) at $\alpha = 0.84$. For each device the modal camera quality value for the three coders was selected.[2] The mean camera quality of devices was 9.57 megapixels.

4.2.4 Random-Access Memory (RAM)

The fourth device characteristic was the amount of Random-Access Memory (RAM) available on the device. This is the amount of available immediate storage for software that is running. This was coded in the mTurk data collection. This was measured in gibibytes.[3] For the RAM measure all three coders were in perfect agreement 96% of the time and both the kappa statistic of $\kappa = 0.98$, and Krippendorff’s alpha at$\alpha = 0.95$ suggest there was a high level of agreement amongst coders. Again, for each device the modal RAM value for the three coders was selected.[4] The mean RAM of devices was 1.79 GiB. The available RAM on

[2] For two models of device all three coders were in disagreement about the camera quality value, in these cases the value was manually obtained from the manufacturer’s website: Samsung SM-T210 - http://www.samsung.com/latin_en/consumer/mobile-devices/tablets/galaxy-tab/SM-T2100ZWATPA; Samsung SM-T530" - http://www.samsung.com/uk/tablets/galaxy-tab-4-10-1-t530/SM-T530NYKABTU.

[3] In both the discussion of the HIT, and in much general discussion of RAM the unit measured is typically referred to as gigabytes, however as RAM is measured in multiples of bytes, which is a binary measure, the more technically correct term gibibytes (GiB) is used throughout International Electrotechnical Commission (1999). IEC 60027-2 Amendment 2: Letter symbols to be used in electrical technology - Part 2: Telecommunications and electronics.

[4] For one model of device all three coders were in disagreement about the RAM value, therefore the value was manually obtained from the manufacturer’s website: LGE LG-D855 - http://www.lg.com/uk/mobile-phones/lg-D855.

mobile devices only comes in a select number of values, measured in half or whole gibibyte increments. As a result, alternative specifications of models fitting RAM as a series of categorical variables was considered. These ordinal models produced met the proportional odds assumption, and as RAM is technically a continuous measure the continuous variants of the models are reported.

4.2.5 Processor performance

As was mentioned earlier, the processor performance measure captured in the mTurk data collection did not account for newer mobile devices having multiple cores, therefore it was necessary to obtain an alternative measure of this variable. This was scraped from the Geekbench (2018) database of comparative processor performance scores using an R script. Geekbench provide industry leading benchmarks of processor scores where Intel Core i7-6600U processor is used as the baseline with a score of 4,000 points. Geekbench's database contains multiple records for a given device; the median value for a given device was selected. Double the score represents double the processing performance. The large range of the original measure meant that interpretation of coefficients was difficult, as a one-unit change in processing performance did not really reflect the wide range of scores. Therefore, the decision was made to divide all the processor scores by one thousand to make interpretation easier. The mean processor performance score was 2.13.

4.3 Respondent Characteristics – Respondent level

One of the challenges in examining device effects is disentangling the direct effect of device characteristics from the indirect effects of respondent characteristics as a result of selection. Lugtig and Toepoel (2015) suggest that selection effects accounted for the majority of the observed device effects in their study. It should be noted however, that this finding was based on respondents who had completed successive waves of a survey on different types of device. It is less clear whether this absence of direct device effects might also be observed when respondents are required to complete a study using a mobile device (without a desktop alternative).

As was noted earlier, the device the respondent uses to complete a survey task is not random, and therefore device characteristics are not independent of respondent characteristics. As a result, the potential exists for any observed direct effects of device characteristics to in fact be indirect effects of respondent characteristics, if those respondent characteristics are not adequately controlled for when modelling. Five respondent characteristics have been included in the models presented later in this paper. These have been selected based on a combination of: existing literature that suggests they may be related to device selection and established respondent characteristic controls in a previous paper on device effects by Struminskaya, Weyandt et al. (2015). All five characteristics are taken from the ninth wave of the Innovation Panel. Descriptive statistics for the respondent characteristics can be found in Table 3.

Table 3. *Descriptive statistics for respondent characteristics.*

		Respondents (n= 255)
Sex	*Male*	39%
	Female	61%
Age (years)	*Mean*	43
	SD	15
	Min	16
	Median	42
	Max	86
Equivalised gross monthly household income (£)	*Mean*	£2344
	SD	£1242
	Min	£116
	Median	£2146
	Max	£7921
Employment status	*Management*	36%
	Intermediate	15%
	Routine	18%
	Unemployed	4%
	Retired	15%
	Inactive	11%
Highest level of education	*Degree or higher*	55%
	Lower than a degree	45%

4.3.1 Sex

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

4.3.2 Age

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

4.3.3 Equivalised gross monthly household income

The respondent's level of household income was also included as a relevant respondent characteristic. No previous literature was found that provided evidence to suggest that level of income affects device selection. Price however has been found to be a factor in device selection (Sarker and Wells 2003), so it seems plausible that level of income would influence a respondent's decision about how much they could afford to spend on a device. It also seems likely, given the subject of the Spending Study, that level of income may affect data quality indicators, for example the number of items on receipts. Gross monthly income was equivalised using the modified OECD scale from the ninth wave of the Innovation Panel to account for differences in the number of household members. The mean equivalised gross monthly household income was £2344.

4.3.4 Employment status

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

4.3.5 Level of education:

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

5 Results

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

To decompose the proportion of variance that can be attributed to the device used to participate, a series of five four-level cross-classified regression model were fitted using Markov chain Monte Carlo (MCMC) methods of estimation. These models were estimated using MLwiN (Charlton, Rasbash et al. 2017) using the software's in-built MCMC estimation methods (Browne 2017). All models were fitted with a monitoring chain of 50,000 iterations, a burn in length of 1,000 iterations and with a thinning factor of one. For the two continuous data quality indicators, duration and number of items on the receipt, the equation for the models is as follows:

$$y_{ijkl} = \beta_0 + f_{0l} + v_{0k} + u_{0jl} + e_{ijkl} \qquad (1)$$

where y_{ijkl} is the value of the respective data quality indicator for a given app use i performed by a given respondent j using device model k within PSU l. The coefficient β_0 is then overall mean across all app uses, all respondents, all device models, and all PSUs. The random PSU effect is f_{0l}, the random device effect is v_{0k}, the random effect of the respondent is u_{0jl} and e_{ijkl} is the residual difference of individual app uses. All four of the random terms are assumed to be normally distributed such that:$f_{0l} \sim N\left(0, \sigma^2{}_{f0}\right)$, $v_{0k} \sim N(0, \sigma^2{}_{v0})$, $u_{0jl} \sim N(0, \sigma^2{}_{u0})$ and $e_{ijkl} \sim N(0, \sigma^2{}_{e})$.

For the three other data quality indicators logistic models are fitted with the equational form:

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

where $logit(\pi_{ijkl})$ is the log odds of the occurrence of a value of one for the corresponding data quality indicator. The random parts of the model f_{0l}, v_{0k} and u_{0jl} retain their meaning from equation 1, namely that they are the cluster specific effects. As such, these three terms hold the same assumptions as in equation 2. However, the logistic function, by definition (see Snijders and Bosker 2012 for more details, Hox, Moerbeek et al. 2017), fixes the variance of the lowest level residuals $\sigma^2{}_e$ such that $\sigma^2{}_e = \pi^2/3 \approx 3.29$. Results from all five models that were fitted are presented in Table 4.

Table 4. *Results of four-level cross-classified regression models of the data quality indicators with no predictors*.

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

The decomposition of the amount of variance apportioned to each of the levels in the model can be achieved by examining the Variance Partition Coefficient (VPC), which is the proportion of the total variance that is explained by each of the levels. As an example, to calculate the VPC for PSUs the following equation is used:

$$\frac{\sigma^2{}_{f0}}{\sigma^2{}_{f0} + \sigma^2{}_{v0} + \sigma^2{}_{u0} + \sigma^2{}_{e}} \qquad (3)$$

The variance for each level can then be substituted into the numerator of the equation to calculate a level specific VPC. The VPC is similar to an intraclass correlation coefficient (of which intra-interviewer correlations used in the interviewer effects literature is an example). In many circumstances the two are analogous to one another.

However, Leckie (2013) makes the distinction between the two, stating that the VPC reflects the proportion of the response variance the model attributes to each level in the model. In contrast, the ICC measures the expected homogeneity between two of the lowest level units (in this case app uses), based on their membership to each of the higher-level units. In a hierarchically nested model, without random effects for predictor variables, these two measures will be the same (as is the case in much of the interviewer effects literature). However, in a cross-classified model, the different configurations of possible memberships to higher level groups means that the ICC is not equivalent to the VPC, and in fact there will be more possible ICCs than VPCs. For example, the ICC for two app uses that share a device and a respondent will be different than that for two app uses that are completed using the same device model, but by different respondents. To allow comparability to the interviewer effects literature, and to enable device effects to be assessed, the VPC was chosen to be reported throughout.

In terms of durations of the app uses, it was expected that the level of variance that was attributed to the respondent would be quite a bit larger than that which is attributed to the device used. However, this was not the case, with the proportion of the variance attributed to the respondent being 10% and the proportion attributed to the model of the device being used being 8%.

The proportion of variance in whether the duration was an outlier or not was in line with the expected result. Namely, that a greater share (13%) of the variance was attributed to the respondent than to the device (3%).

It was expected that for the type of activity that is performed a far larger share of the variance would be attributed by the model to the respondent; at 35% this was the case. However, that 6% of the variance is attributable to the device used still suggests that the model of device is associated with what type of app use each app use is.

The share of the variance that was attributed to the device model was highest for the quality of the images produced, at 23%. This compares to just 9% of the variance being attributable to the respondent for this indicator. This was unexpected, whilst it was considered that the device used may be associated with the quality of the images produced, it was not expected that almost a quarter of the variance in the measure would be attributable to the device used.

Finally, almost none of the variance (<1%) in the number of items was attributed to the device used, in comparison the share of the variance attributed to the respondent was 10%. This was unexpected, as it was anticipated that some of the variance in this measure would be associated with the model of device used. However, from a data quality perspective this is perhaps reassuring as it suggests there are not device effects downwardly biasing this measure.

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

To examine the effects of specific device characteristics the five characteristics measured at the device level were introduced to each of the five models. The resulting models are the models presented on the left-hand column under each data quality indicator in Table 5.

Table 5. Results of four-level cross-classified regression models of the five data quality indicators with device and respondent characteristics as predictors.

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

Degree or higher			0.52				1.12				0.63**				0.81				-0.11	
			(1.05)				(0.17)				(0.18)				(0.24)				(0.62)	
Income			0.00				1.00				1.00				1.00				0.00	
			(0.00)				(0.00)				(0.00)				(0.00)				(0.00)	
App use type *Ref: Scanned receipt*																				
Purchase without receipt			-11.08***				0.65***													
			(0.53)				(0.13)													
Report of nothing bought			-33.23***				0.52***													
			(0.58)				(0.16)													
Constant	36.72***		36.35***		0.05***		0.06***		0.84		7.11***		0.10***		0.06**		9.08***		6.59***	
	(2.45)		(3.51)		(0.28)		(0.59)		(0.28)		(0.47)		(0.44)		(0.91)		(0.85)		(1.9)	
	σ	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*	*σ*	*VPC*
PSU σ^2_{f0}	5.79	0.01	4.95	0.01	0.02	0.00	0.02	0.00	0.15	0.04	0.10	0.03	0.32	0.07	0.50	0.10	0.36	0.00	0.40	0.00
Device σ^2_{v0}	52.81	0.08	16.96	0.03	0.13	0.03	0.18	0.04	0.23	0.05	0.30	0.07	0.51	0.12	0.80	0.15	0.15	0.00	0.15	0.00
Respondent σ^2_{u0}	68.06	0.10	65.51	0.10	0.45	0.12	0.63	0.14	1.34	0.32	1.22	0.30	0.59	0.14	0.57	0.11	9.99	0.09	9.98	0.09
Residual σ^2_{e}	564.98	0.82	565.48	0.87	3.13	0.84	3.67	0.82	2.40	0.58	2.44	0.60	2.82	0.67	3.43	0.65	96.02	0.90	95.89	0.90
PSUs	90		90		90		90		90		90		89		89		89		89	
Devices	90		90		90		90		90		90		84		84		83		83	
Respondents	255		255		255		255		255		255		233		233		231		231	
App uses	10621		10621		10985		10985		10985		10985		5263		5263		4790		4790	
DIC	97661		94741		3086		3064		12466		12465		2511		2512		35598		35596	

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

As before, the models were fitted in MLwiN, using the same MCMC estimation conditions as those fitted in RQ1. The addition of the device characteristics means that for continuous outcomes equation one becomes:

$$y_{ijkl} = \beta_0 + \mathbf{X}\beta_{\boldsymbol{k}} + f_{0l} + v_{0k} + u_{0jl} + e_{ijkl} \quad (4)$$

and for binary outcomes equation two becomes:

$$logit(\pi_{ijkl}) = \beta_0 + \mathbf{X}\beta_{\boldsymbol{k}} + f_{0l} + v_{0k} + u_{0jl} \quad (5)$$

where in both cases $\mathbf{X}\beta_{\boldsymbol{k}}$ is the five device level predictor variables and their corresponding coefficients. The assumptions about the normality of the random terms, as expressed in equations two and four remain unchanged.

For the logistic models, coefficients and variances have been rescaled for logistic models to allow comparison of nested models as recommended by Hox, Moerbeek et al. (2017) and Snijders and Bosker (2012). This overcomes the issue that because logistic models fix the residual variance at approximately 3.29 the effects of fixed or random effects may be inflated compared to the null model.

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

The time taken to complete app uses was statistically significantly associated with three of the device characteristics included in the model. The first of these was the Operating System, where app uses completed using Android devices were typically associated with taking six seconds longer to complete ($\beta = 6.09,\ p < 0.01,$ $95\%\ CI\ [2.05, 10.13]$). App uses completed on tablets were associated with app use durations that were around seven seconds longer than those completed on smartphones ($\beta = 7.06,\ p < 0.01,\ 95\%\ CI\ [2.16, 11.96]$). Finally, increased RAM was associated with typically shorter app use durations. Each additional gibibyte of RAM a device had was associated with app uses completed on that device typically taking just under five seconds less to complete ($\beta = -4.78,\ p < 0.01,\ 95\%\ CI\ [-7.82, -1.74]$). Processor speed and camera quality were not statistically significantly associated with app use durations. The DIC for the null model of duration was 97656, compared to a DIC of 97661 for the device characteristics model. This suggests the model including device characteristics is potentially a poor fit for the data. To explore whether this increased DIC was just the product of two of the device characteristics seemingly not being predictors of duration, a model retaining just those device characteristics that were statistically significant was fitted. This produced a DIC of 97657. Overall, this suggests that whilst the device used does affect the duration of app uses (as evidenced by the corresponding VPC of 0.08 in RQ1) the device characteristics captured in this study do not seem to account for these device effects very well.

In terms of outlying app use durations, there were two device characteristics that were statistically significantly associated with a lower likelihood of app uses completed using that device being an outlier. The first of these was operating system, with Android devices having 44% lower odds of producing app uses with outlying durations ($OR = 0.56,\ p < 0.01,\ 95\%\ CI\ [0.35,\ 0.90]$). Similarly, increases in processor performance were associated with a decreased likelihood of a device producing outlying durations ($OR = 0.84,\ p < 0.05,\ 95\%\ CI\ [0.69,\ 0.99]$). The other three device characteristics were not statistically significantly associated with the likelihood of app use durations being outlying. The DICs for the null model and the device characteristics model were the same, 3086, indicating that the model with the addition of the device characteristics is not an improvement in terms of how it fits the data. Again, the reduced model with just statistically significant predictors was considered, this produced a DIC of 3085. As with non-outlying durations, this suggests that the device characteristics identified do not explain well the variance in whether an app use was a suspected outlier in terms of duration. This perhaps is less surprising than in the case of the non-outlying durations, as the null model suggested that device only account for 3% of the variance in whether an app uses had an outlying duration.

For the third outcome, the app use type, none of the device characteristics modelled were significant predictors of whether an app use was a scanned receipt, or a manually entered purchase/report of nothing bought.

For the fourth outcome, the image quality, Android devices $(OR = 3.14,\ p < 0.001, 95\%\ CI\ [1.61,\ 6.11])$, and tablets $(OR = 2.25,\ p < 0.05,\ 95\%\ CI\ [1.01,\ 5.03])$ were associated with an increase in the odds of receipt scan producing a low quality image. Higher RAM was associated with lower odds of producing low quality images $(OR = 0.49,\ p < 0.05, 95\%\ CI\ [0.27,\ 0.90])$. The DIC of the null model of image quality was 2519, compared to a smaller DIC of 2511 for the corresponding device characteristics model. This suggests that the addition of the device characteristics to the model produced a better fitting model.

For the final outcome, the number of lines, the only statistically significant association was the device type, with receipts scanned on tablets typically having one less item on them than those scanned on smartphones $(\beta = -1.50,\ p < 0.05,$ $95\%\ CI\ [-2.83,\ -0.17])$. The DIC for the null model was 35599, and the DIC for the device characteristics model was 35598. Again, this suggests that the inclusion of the device characteristics did not produce a better fitting model. This is not particularly surprising as the VPC for the null model for this outcome suggested that device accounted for less than one percent of the variance in the number of lines on a scanned receipt.

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

To examine the potential effects of selection, respondent characteristics were introduced to each of the five models. The resulting models are the models presented on the right-hand column under each data quality indicator in Table 5.

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

$$y_{ijkl} = \beta_0 + \mathbf{X}\beta_{\boldsymbol{k}} + \mathbf{X}\beta_{\boldsymbol{j}} + f_{0l} + v_{0k} + u_{0jl} + e_{ijkl} \tag{6}$$

and for binary outcomes equation five becomes:

$$logit(\pi_{ijkl}) = \beta_0 + \mathbf{X}\beta_{\boldsymbol{k}} + \mathbf{X}\beta_{\boldsymbol{j}} + f_{0l} + v_{0k} + u_{0jl} \tag{7}$$

where in both cases $\mathbf{X}\beta_{\boldsymbol{j}}$ is the respondent characteristics variables and their corresponding coefficients. The assumptions about the normality of the random terms, as expressed in equations two and four remain unchanged. In addition to this, for the models for duration and outlying durations, an additional control, measured at the app use level, was introduced. This was what the type the app use was, included as this was highly predictive of duration.

For all three device characteristics that were statistically significant predictors of app use duration, the effects were diminished when controlling for respondent characteristics. The first of these was the Operating System, where app uses completed using Android devices were typically associated with taking four seconds longer to complete ($\beta = 4.12,\ p < 0.05,\ 95\%\ CI\ [0.69,\ 7.55]$) when controlling for respondent characteristics, as opposed to six seconds longer when not. App uses completed on tablets were associated with durations that were around three and a half seconds longer than those completed on smartphones ($\beta = 3.47,\ p > 0.05,$ $95\%\ CI\ [-0.63,\ 7.57]$), down from seven seconds when not controlling for respondent characteristics. This measure was also no longer statistically significant. Finally, each additional gibibyte of RAM a device had was associated with app use durations that were a little under three seconds shorter ($\beta = -2.70,\ p < 0.05,$ $95\%\ CI\ [-5.21,\ -0.19]$), compared to just under five seconds shorter when not controlling for respondent characteristics. This perhaps suggests that some of the observed device effects may in fact be the result of selection. Three respondent characteristics were significant predictors of app use duration: age ($\beta = 0.20,$ $p < 0.001,\ 95\%\ CI\ [0.10,\ 0.30]$), being retired ($\beta = 7.39,\ p < 0.001, 95\%\ CI\ [3.43,$ $11.35]$) and being otherwise inactive in terms of employment ($\beta = 4.32,\ p <$ $0.01, 95\%\ CI\ [0.81,\ 7.83]$). The DIC for the model including respondent characteristics dropped quite significantly, from 97661 to 94741. This suggests that the addition of these respondent characteristics quite substantially improved the goodness of the fit of the model.

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

Whilst none of the device characteristics included in the device characteristics models for activity types were found to statistically significant predictors the possibility was considered that a relationship may be seen when controlling for respondent characteristics. Therefore, the respondent characteristics model was fitted for this outcome. However, the device characteristics all remained statistically nonsignificant predictors of activity type in this model.

Three device characteristics were significant predictors of image quality in the device characteristics models. All three remained statistically significant when controlling for respondent characteristics. The first two of these had slight reductions in the size of their odds ratios: $OR = 2.91,\ p < 0.01,\ 95\%\ CI\ [1.44,\ 5.89]$ down from an odds ratio of 3.14 for the OS; and $OR = 2.16,\ p < 0.05,\ 95\%\ CI\ [1.03,\ 4.55]$ down from an odds ratio of 2.25 for tablets compared to smartphones. However, these reductions were relatively small, and this stability of estimates between models supports that there are some direct effects of these device characteristics. The coefficient for the third significant predictor of image quality, the device's RAM, changed very little $OR = 0.50,\ p < 0.05,\ 95\%\ CI\ [0.27,\ 0.94]$ compared to a value of 0.49 in the device characteristics only model. None of the respondent characteristics were significantly associated with image quality.

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

6 Discussion

This paper expands upon the existing device effects literature by moving beyond comparing the broad categorisations of smartphone, tablet and PCs. Instead, this paper is the first, to date, to consider the effects of the models of mobile devices used for survey tasks. To achieve this, models were fitted that consider the potential for homogeneity amongst survey responses that were completed using the same model of mobile device.

This research also sought to explore what characteristics of mobile devices might be contributing to any observed device effects. Some device characteristics were captured in the data collection task (the *Understanding Society* Spending Study) itself. However, to supplement this selected device characteristics were coded using workers from Amazon mTurk to complete data collection. To the best of the author's knowledge this paper is the first example of using mTurk to collect paradata after the main stage of data collection has been completed. It may be possible to harness mTurk to collect other types of paradata, or perform other data processing tasks, such as coding of textual responses. One of the major advantages of this would be that mTurk represents a fast and inexpensive way of achieving this.

The results of RQ1 suggest that there were device effects in the Spending Study. The device level VPCs ranged from <0.00 to 0.22, which is of a similar magnitude to those reported within the interviewer effects literature (e.g. O'Muircheartaigh and Campanelli 1998, West and Olson 2010, Jäckle, Lynn et al. 2011). The evidence is not strong enough to suggest that survey researchers should be as concerned about device effects as they are about interviewer effects. However, based on these results, it seems that further investigation into the potential for device effects is warranted. For example, examining whether mobile device model clustering effects are found when considering the kinds of data quality indicators traditionally examined in questionnaire-based surveys, for example straightlining, acquiescence, mid-point responding, item nonresponse, and primacy effects.

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

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

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

As in RQ1, the quality of the images produced when scanning receipts was the only outcome where there was particularly convincing evidence of device effects. That the amount of RAM a device has was a significant predictor when the quality of the camera was not was an unexpected result. The earlier consideration of how Sudman and Bradburn’s (1974) conceptual framework for interviewer effects can be applied to device effects perhaps sheds some light on this result. It seems possible this finding is consistent with either the first or second source of bias. The first explanation may be that some portion of the allocation of memory in the photography process was assigned to the device to manage, and subsequently differences across devices resulted in either poorer quality images being captured, or in some cases no images at all. Alternatively, even if the allocation of RAM was adequately accounted for when the app was programmed, it is possible that circumstances beyond the control of the programmers resulted in devices running out of RAM. Ultimately, as this study used a pre-existing app developed by a commercial partner, it is not possible to examine the underlying software of the app to attempt to uncover which of these accounts best explains this finding. However, it is felt that making the parallel to interviewer effects is useful for considering how to think about device effects both conceptually and methodologically. Further consideration of how the device used to complete a web survey acts in the place of an interviewer may have implications for best practices for designing web surveys.

To illustrate the potential magnitude of the combined effects of the device’s characteristics of a specific device model it is useful to examine the change in odds between devices that had a high and low likelihood of producing a low-quality image. The device model with the highest odds of producing a low-quality image was the Motorola Moto E, an entry-level budget smartphone targeted at first-time smartphone buyers (Gibbs 2014). In contrast, the device model with the lowest odds of producing a low-quality image was the Google Pixel XL, a device that was optimised for its photography capabilities (Goodwin 2017). Both devices were Android smartphones, the Moto E had a 5MP camera, 1GiB of RAM and a processor score of 0.63 and the Pixel XL had a 12.3MP camera, 4GiB of RAM and a processor score of 4.08.

Assuming all else (that is, the respondent) is kept constant, the difference in the likelihood of producing a low-quality image between the two devices produces an odds ratio of 48.75. This suggests the odds of a Moto E producing a low-quality image are 4775% higher than on the Google Pixel XL.

This perhaps offers an overly extreme comparison. As a more conservative comparison we can look at the devices with the lowest and highest likelihood of producing a low-quality image amongst those devices that were used by more than one respondent. Amongst these devices, the device model with the highest likelihood was the Apple iPad 2 with Wi-Fi only capabilities, this was an iOS Tablet, with a 0.70 MP camera, 0.5 GiB of RAM, and a processor score of 0.59. The lowest was the Samsung Galaxy S7 edge, an Android smartphone, with a 12MP camera, 4GiB of RAM, and a processor score of 3.77. The device characteristic values for these two devices are also documented in Table C1. The difference in likelihood between these two devices produced an odds ratio of 20.62, which means the odds of the iPad producing a low-quality image were 1962% higher.

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

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

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

Finally, the Spending Study is a particular use of mobile devices for data collection. The question remains how generalisable the findings presented here are to survey research more broadly. In response to this, in the first instance, it seems likely that the results will be generalisable both to studies that very closely resemble the Spending Study (e.g. making use of cameras on mobile devices) but also for other studies that make use mobile device features to collect data beyond that which is traditionally captured in surveys, for example: tracking of health behaviours, collecting, GPS data, or administering “in-the-moment" surveys. In addition to this, hopefully the approach of using the literature on interviewer effects to inform how to think both conceptually and methodologically about device effects may also be relevant in more traditional survey settings.

7 References:

Antoun, C., et al. (2017). "Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel." Public Opinion Quarterly **81**(S1): 280-306.

Antoun, C., et al. (2016). "Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk." Field methods **28**(3): 231-246.

Berinsky, A. J., et al. (2012). "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." Political analysis **20**(3): 351-368.

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

Brownlee, J. (2019). "iOS is twice as memory-efficient as Android. Here's why. | Cult of Mac." Retrieved 13/01/2019, from https://www.cultofmac.com/303223/ios-twice-memory-efficient-android-heres/.

Brunton-Smith, I., et al. (2017). "Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model." Journal of the Royal Statistical Society: Series A (Statistics in Society) **180**(2): 551-568.

Brys, G., et al. (2004). "A robust measure of skewness." Journal of Computational and Graphical Statistics **13**(4): 996-1017.

Buhrmester, M., et al. (2011). "Amazon's Mechanical Turk:A New Source of Inexpensive, Yet High-Quality, Data?" Perspectives on Psychological Science **6**(1): 3-5.

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

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

Collins, M., et al. (1988). Diffusion of technological innovation: Computer assisted data collection in the U.K. Computer assisted survey information collection. R. M. Groves, P. P. Biemer, L. E. Lyberg et al., John Wiley & Sons.

Cook, W. A. (2014). "Is mobile a reliable platform for survey taking? Defining quality in online surveys from mobile respondents." Journal of Advertising Research **54**(2): 141-148.

Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Survey Research Methods.

Couper, M. P. and G. J. Peterson (2017). "Why do web surveys take longer on smartphones?" Social Science Computer Review **35**(3): 357-377.

Crawford, S. D., et al. (2001). "Web Surveys:Perceptions of Burden." Social Science Computer Review **19**(2): 146-162.

De Bruijne, M. and A. Wijnant (2013). "Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey." Social Science Computer Review **31**(4): 482-504.

De Bruijne, M. and A. Wijnant (2014). "Mobile response in web panels." Social Science Computer Review **32**(6): 728-742.

Deming, W. E. (1944). "On errors in surveys." American Sociological Review **9**(4): 359-369.

Dillman, D. A. and L. M. Christian (2005). "Survey mode as a source of instability in responses across surveys." Field methods **17**(1): 30-52.

Dillman, D. A., et al. (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.

Durrant, G. B., et al. (2010). "Effects of interviewer attitudes and behaviors on refusal in household surveys." Public Opinion Quarterly **74**(1): 1-36.

Elliott, M. N., et al. (2009). "Effects of survey mode, patient mix, and nonresponse on CAHPS® hospital survey scores." Health services research **44**(2p1): 501-518.

Fernee, H. and N. Sonck (2013). "Is everyone able to use a smartphone in survey research? Tests with a Time-use App with Experienced and Inexperienced Users." Survey Practice **6**(4): 2884.

Fluent (2016). Devices and Demographics 2016. Devices and Demographics.

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

Geekbench (2018). "Geekbench 4." Retrieved 15/06/2018, from https://www.geekbench.com/.

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

Gibbs, S. (2014). "Motorola launches £89 Moto E – its bargain basement smartphone." Retrieved 13/01/2018, from https://www.theguardian.com/technology/2014/may/13/motorola-moto-e-android-smartphone.

Goodwin, R. (2017). "This is Why The Google Pixel's Camera is Almost Unbeatable." Retrieved 13/01/2019, from https://www.knowyourmobile.com/mobile-phones/google-pixel/24147/google-pixel-camera-review-specs-features-detailed-vs-iphone-7-plus.

Groves, R. M. and R. L. Kahn (1979). Surveys by Telephone; A national comparison with personal interviews. New York, Academic Press.

Groves, R. M., et al. (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.

Hox, J. J., et al. (2017). Multilevel analysis: Techniques and applications, Routledge.

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

International Electrotechnical Commission (1999). IEC 60027-2 Amendment 2: Letter symbols to be used in electrical technology - Part 2: Telecommunications and electronics.

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

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

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

Jäckle, A., et al. (2010). "Assessing the effect of data collection mode on measurement." International Statistical Review **78**(1): 3-20.

Karjaluoto, H., et al. (2005). "Factors affecting consumer choice of mobile phones: Two studies from Finland." Journal of Euromarketing **14**(3): 59-82.

Keusch, F. and T. Yan (2017). "Web versus mobile web: An experimental study of device effects and self-selection effects." Social Science Computer Review **35**(6): 751-769.

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

Landis, J. R. and G. G. Koch (1977). "The measurement of observer agreement for categorical data." Biometrics: 159-174.

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

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

Link, M. W. and A. H. Mokdad (2005). "Effects of survey mode on self-reports of adult alcohol consumption: a comparison of mail, web and telephone approaches." Journal of Studies on Alcohol **66**(2): 239-245.

Link, M. W., et al. (2014). "Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the AAPOR task force on emerging technologies in public opinion research." Public Opinion Quarterly **78**(4): 779-787.

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

Lugtig, P. and V. Toepoel (2015). "The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error." Social Science Computer Review **34**(1): 78-94.

Malhotra, N. (2008). "Completion Time and Response Order Effects in Web Surveys." Public Opinion Quarterly **72**(5): 914-934.

Mason, W. and S. Suri (2012). "Conducting behavioral research on Amazon’s Mechanical Turk." Behavior research methods **44**(1): 1-23.

Mavletova, A. (2013). "Data quality in PC and mobile web surveys." Social Science Computer Review **31**(6): 725-743.

Mavletova, A. and M. P. Couper (2013). Sensitive topics in PC web and mobile web surveys: Is there a difference? Survey Research Methods.

McCabe, S. E., et al. (2002). "Mode effects for collecting alcohol and other drug use data: Web and US mail." Journal of Studies on Alcohol **63**(6): 755-761.

O'Muircheartaigh, C. and P. Campanelli (1999). "A multilevel exploration of the role of interviewers in survey non-response." Journal of the Royal Statistical Society: Series A (Statistics in Society) **162**(3): 437-446.

O’Muircheartaigh, C. and P. Campanelli (1998). "The relative impact of interviewer effects and sample design effects on survey precision." Journal of the Royal Statistical Society: Series A (Statistics in Society) **161**(1): 63-77.

Open Signal (2015). Android Fragmentation 2015.

Paolacci, G., et al. (2010). "Running experiments on amazon mechanical turk."

Pickery, J., et al. (2001). "The Effects of Interviewer and Respondent Characteristics on Response Behavior in Panel Surveys:A Multilevel Approach." Sociological Methods & Research **29**(4): 509-523.

Read, B. (2018). Respondent burden in a Mobile App: evidence from a shopping receipt scanning study, Understanding Society at the Institute for Social and Economic Research.

Revilla, M. (2017). "Are there differences depending on the device used to complete a web survey (PC or smartphone) for order-by-click questions?" Field methods **29**(3): 266-280.

Revilla, M. and M. P. Couper (2018). "Comparing grids with vertical and horizontal item-by-item formats for PCs and smartphones." Social Science Computer Review **36**(3): 349-368.

Revilla, M., et al. (2016). "PCs versus Smartphones in answering web surveys: Does the device make a difference?" Survey Practice **9**(4): 1-6.

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

Roberts, C., et al. (2010). "Diffusion of technological innovation: Computer assisted data collection in the U.K." ISER Working Paper Series **2010**(36).

Sarker, S. and J. D. Wells (2003). "Understanding mobile handheld device use and adoption." Communications of the ACM **46**(12): 35-40.

Schnell, R. and F. Kreuter (2005). "Separating interviewer and sampling-point effects." Journal of Official Statistics **21**(3): 389-410.

Shin, E., et al. (2012). "Survey mode effects on data quality: Comparison of web and mail modes in a US national panel survey." Social Science Computer Review **30**(2): 212-228.

Snijders, T. A. B. and R. J. Bosker (2012). Multilevel analysis : an introduction to basic and advanced multilevel modeling.

Spiegelhalter, D. J., et al. (2002). "Bayesian measures of model complexity and fit." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(4): 583-639.

Struminskaya, B., et al. (2015). "The effects of questionnaire completion using mobile devices on data quality. Evidence from a probability-based general population panel." methods, data, analyses **9**(2): 32.

Sudman, S. and N. M. Bradburn (1974). Response Effects in Surveys: A Review and Synthesis, Aldine Publishing Company.

Toepoel, V. and P. Lugtig (2014). "What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users." Social Science Computer Review **32**(4): 544-560.

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

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

Wells, T., et al. (2014). "Comparison of smartphone and online computer survey administration." Social Science Computer Review **32**(2): 238-255.

West, B. T., et al. (2018). "Can conversational interviewing improve survey response quality without increasing interviewer effects?" Journal of the Royal Statistical Society: Series A (Statistics in Society) **181**(1): 181-203.

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

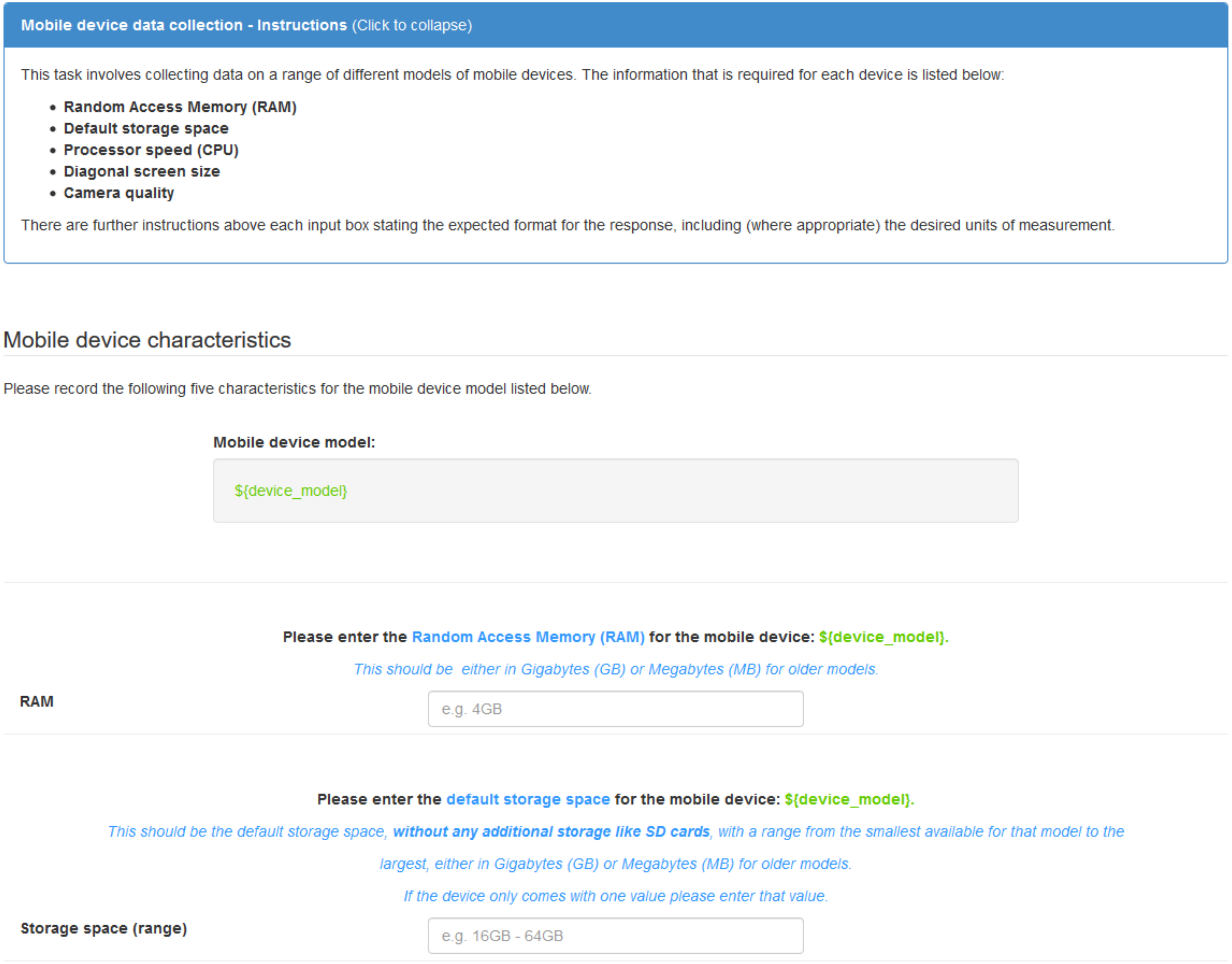
West, B. T. and K. Olson (2010). "How much of interviewer variance is really nonresponse error variance?" Public Opinion Quarterly **74**(5): 1004-1026.

Wiggins, R. D., et al. (1990). A Variance Components Approach to Interviewer Effects, Joint Centre for Survey Methods.

Yammarino, F. J., et al. (1991). "Understanding mail survey response behavior a meta-analysis." Public Opinion Quarterly **55**(4): 613-639.

Yan, T. and R. Tourangeau (2008). "Fast times and easy questions: The effects of age, experience and question complexity on web survey response times." Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition **22**(1): 51-68.

Appendix A: Amazon mTurk Human Intelligence Task Screenshots:



Mobile device data collection - Instructions (Click to collapse)

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

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

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

Mobile device characteristics

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

Mobile device model:

${device_model}

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

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

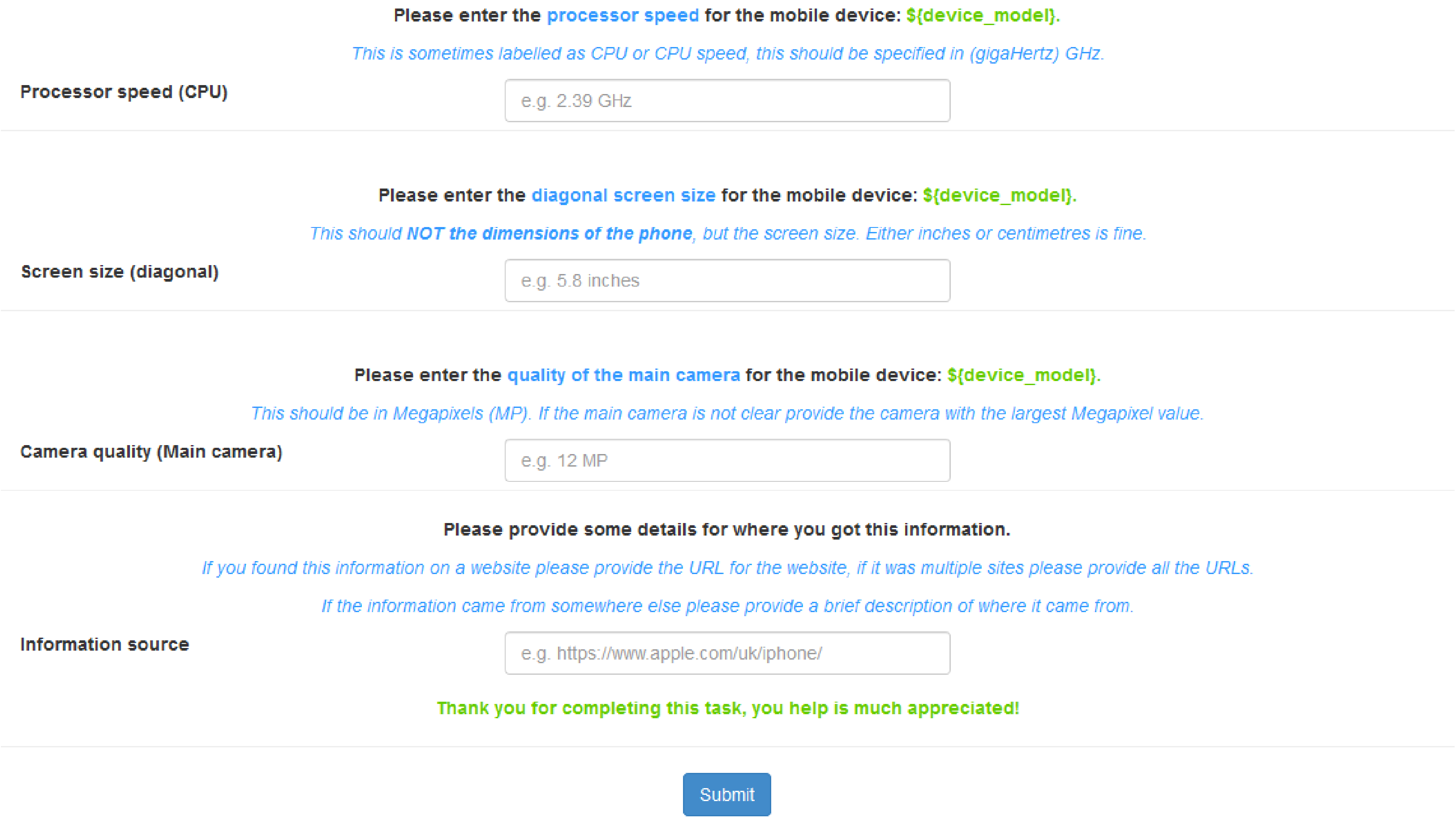
RAM e.g. 4GB

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

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

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

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



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

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

Processor speed (CPU) e.g. 2.39 GHz

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

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

Screen size (diagonal) e.g. 5.8 inches

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

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

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

Please provide some details for where you got this information.

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

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

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

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