University of Southampton

Faculty of Social Sciences

School of Economic, Social and Political Sciences

Department of Social Statistics and Demography

Investigating the Impact of Technologies on the Quality of Data Collected through Surveys

DOI 10.5258/SOTON/T0046

by

Carli Lessof

ORCID ID 0000-0002-0553-1623

Thesis for the degree of Doctor of Philosophy

June 2022

University of Southampton

Abstract

Faculty of Social Sciences

School of Economic, Social and Political Sciences

Department of Social Statistics and Demography

Thesis for the degree of Doctor of Philosophy

Investigating the Impact of Technologies on the Quality of Data Collected through Surveys

by

Carli Lessof

Social surveys continue to play an important role in social science and policy making. In addition to providing information about attitudes and behaviours, they act as a vehicle for the collection of many different types such as biomeasures, geographical data, administrative records, social media posts and so on. The resilience and adaptability of the social survey owes much to the way that they have adapted to the enormous changes in the technological environment. Radical changes in telephony, personal computing, the internet, and mobile devices have transformed many aspects of the research process. While these changes have brought many benefits, the application of each new technology in survey data collection needs careful consideration in terms of ethics, burden, cost, and implementation. Moreover, they may introduce representation errors and measurement errors that must be accounted for. In this context, this thesis considers the effects of new technologies on data quality and measurement error. It presents three examples of the use of technologies in social surveys and examines an aspect of data quality in relation to each. The thesis makes specific recommendations and encourages further methodological research in the use of technology in survey data collection.

The focus of the first study is on three biomeasures which are frequently collected in health or multidisciplinary surveys but may be recorded using different equipment. A randomised cross-over trial of 118 healthy adults aged 45-74 years was conducted using two sphygmomanometers to measure blood pressure, four handgrip dynamometers to measure grip strength, and two spirometers to measure lung function. For each of these three measures, multiple readings from each device were combined with information about the individual, drawn from a self-completion questionnaire, to build a pseudo-anonymised analytical dataset. Evidence was found of differences in measurements when assessed using alternative devices. For blood pressure, there is a difference, on average, of 3.85 mm Hg for Systolic Blood Pressure and 1.35 mm Hg for Diastolic Blood Pressure. For grip strength, two electronic dynamometers record measurements

on average 4-5kg higher than either a hydraulic or a spring-gauge dynamometer. For lung function, a difference of 0.47 litres, on average, was found for measures of Forced Vital Capacity, but no difference was found in measures of Forced Expiratory Volume. The primary analysis was conducted using Bland and Altman plots. Sensitivity analyses tested different definitions of each measure and used multilevel regression modelling as an alternative way of estimating device effects. The findings have implications for analysts who may want to test the sensitivity of their findings to the average differences observed with these combinations of devices and may help investigators who are selecting equipment for new studies or changing equipment for longitudinal studies. Further trials are needed to replicate the comparison of these devices and to test different device combinations, both in stand-alone studies and within larger observational surveys. Future analysts may wish to consider using multilevel modelling to assess device effects.

The second paper also considers device effects, this time, exploring whether the device used to complete an online survey (that is a PC, smartphone, or tablet) affects data quality. The study is based on the Wellcome Trust Science Education Tracker, a mobile-optimised, online survey of over 4,000 pupils aged 14-18. It uses the Wellcome Science Education Tracker 2016 dataset, available through the UK Data Archive, with additional survey process variables obtained with the agreement of the Wellcome Trust. The study uses propensity scores (more specifically, Inverse Probability Treatment Weights) to balance the samples, to reduce the possibility that measurement effects are confounded by selection. The analysis draws on linked geographical, administrative and survey process data which provides an opportunity to assess the use of exogenous confounder variables in the matching process. The large sample size makes it possible to test the sensitivity of the finding to the inclusion or exclusion of tablet users. Overall, the study identifies few consistent device effects, and those that are observed are small, providing reassurance for survey practitioners and analysts. After controlling for selection, those who use a mobile device are seen to have higher levels of “don’t know” responses and are more likely to have interruptions during survey completion. Contrary to the findings of some earlier studies, smartphone responders complete the survey more quickly than PC responders. The results for straightlining are mixed and no clear pattern between mobile and PC could be found. The findings encourage the inclusion of a wide range of covariates when controlling for selection, beyond basic demographics, ideally including exogenous variables, and including those which capture topic salience.

The third research study addresses the potential for app-based research. It is an exploratory study which assesses the quality of data collected using an app-based expenditure diary over a one-month period. A total of 268 members of the *Understanding Society* Innovation Panel agreed to take part. The analysis uses a combination of two datasets from *Understanding Society*: Spending Study 1 (2016-2017) and Wave 9 of the Innovation Panel (2016), both of which are available from

the UK Data Archive. Other analyses have explored initial response rates to this study, noting that just 16.5% of the invited sample completed the registration process and fewer still downloaded the app. In this study, the investigation of data quality involved defining and examining four measures of adherence to protocol, and the extent to which these aspects of adherence were sustained over the duration of the study period. The research identifies a reasonable level of engagement from those who agreed to participate in the app study. For example, the mean number of app use days in the one-month period was 21.7 and the mean number of spending events reported was 27.6. Almost all participants (96.6%) reported at least one spending event and of those, most (95%) used a combination of photographing receipts and making direct entries, or only photographed receipts, with 61% of all spending events reported by photographing receipts. Almost all of those (94.9%) who photographed one or more receipts which had relevant date information did so within, on average, 24 hours of the time of the spending event. Although adherence based on all four measures clearly declines across the study month, it remains reasonably high. This study provides encouragement for further development of the app, and further methodological research and experimentation to increase full and sustained adherence to protocol. If a spending study app is to be embedded successfully in a large-scale study such as *Understanding Society*, future efforts will inevitably focus on ways to raise initial participation rates, but it would be unfortunate if the particular benefits of app-based research, such as capturing detailed spending data from receipts using photographs, were entirely let go in favour of achieving higher initial response rates.

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Carli Lessof

Title of thesis: Investigating the impact of technologies on the quality of data collected through surveys

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

This work was done wholly or mainly while in candidature for a research degree at this University;
Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
Where I have consulted the published work of others, this is always clearly attributed;
Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
I have acknowledged all main sources of help;
Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself *(please see the footnote near the start of each of the three substantive chapters)*;
None of this work has been published before submission;
NB: Related publications covering aspects of the research presented in this thesis are as follows:-

COOPER, R., LESSOF, C., WONG, A. & HARDY, R. 2021. The impact of variation in the device used to measure grip strength on the identification of low muscle strength: Findings from a randomised cross-over study. J Frailty Sarcopenia Falls, 6, 225-230.

LESSOF, C. & STURGIS, P. 2018. New Kinds of Survey Measurements. In: VANNETTE, D. L. & KROSNICK, J. A. (eds.) The Palgrave Handbook of Survey Research. Cham: Springer International Publishing.

JÄCKLE, A., BURTON, B., COUPER, M. P. & LESSOF, C. 2019. Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. Survey Research Methods, 13.

JÄCKLE, A., COUPER, M. P., GAIA, A. & LESSOF, C. 2021. Improving Survey Measurement of Household Finances: A Review of New Data Sources and Technologies. Advances in Longitudinal Survey Methodology.

JÄCKLE, A., GAIA, A., LESSOF, C. & COUPER, M. P. 2019. A review of new technologies and data sources for measuring household finances: implications for total survey error. *Understanding Society* at the Institute for Social and Economic Research.

Signature: Date: 14 June 2022

Published after submission based on this work;

LESSOF, C., COOPER, R., WONG, A., BENDAYAN, R., CALEYACHETTY, R., CHESHIRE, H., et al. 2023. Comparison of devices used to measure blood pressure, grip strength and lung function: A randomised-cross over study. PLoS One 2023 Vol. 18 Issue 12 Pages e0289052 PMCID: PMC10752545 DOI: 10.1371/journal.pone.0289052

Acknowledgements

The analysis carried out for this PhD was supported by an ESRC Doctoral Training Programme grant at the University of Southampton (ES/J500161/1) and included an Overseas Institutional Visit to the University of Michigan, sponsored by Mick Couper.

The equipment comparison study which is reported in Chapter 2 was carried out within the MRC Lifelong Health & Ageing Unit at UCL as part of the National Survey for Health and Development 2015-16 data collection. The thesis author played a substantial role in the study design, implementation and data collection, carried out all data entry, data cleaning, data linkage, analysis and drafted the report. A journal paper entitled "Comparison of devices used to measure blood pressure, grip strength and lung function: a randomised cross-over study" has been submitted to PLOS ONE and is undergoing final revisions prior to publication. Contributions to the paper were by Rebecca Hardy (Principal Investigator), Di Kuh, Rachel Cooper, Andy Wong, Saif Shaheen, Anna Hansell, Cosetta Minelli, George Kyriakopoulos and Rebecca Bendayan (development of study and sample design), Aradhna Kaushal, Rishi Caleyachetty, Theodore D Cosco, Ahmed Elhakeem, Stella G Muthuri and Andrew Wong (data collection). Analytical advice and comments in preparing the journal paper were provided by Rebecca Hardy, Andrew Wong and Rachel Cooper. An associated paper showing the clinical relevance of the results was published in a paper entitled “The impact of variation in the device used to measure grip strength on the identification of low muscle strength: Findings from a randomised cross-over study” (Cooper and Lessof, et al, 2021, Appendix F).

The Science Education Tracker which is presented in Chapter 3 was funded by the Wellcome Trust and conducted by Kantar Public. Peter Matthews, Luke Taylor, Joel Williams and Becky Hamlyn provided assistance accessing the data and liaising with Wellcome Trust for additional permissions, while Alex Wenz provided analytical advice at the start of the project and Mick Couper provided welcome advice during my Overseas International Visit at the University of Michigan.

The paper reported in Chapter 4 was written as part of a wider project funded by the ESRC UK Economic and Social Research Council (ESRC) Transformative Research Scheme and the National Centre for Research Methods (NCRM) Methodological Research Projects Scheme (ES/N006534/1) carried out at the University of Essex with Annette Jäckle as Principal Investigator. Annette Jäckle, Mick P. Couper and Tom F. Crossley provided analytical guidance at several stages. Annette Jäckle suggested ways of restructuring the paper and provided drafting suggestions. Mick P Couper proposed that continued participation be conceptualised in terms of adherence.

The equipment comparison study received ethical approval from the local UCL ethics committee (ref: 6338/001) and all participants gave written informed consent. Ethical approval for the Spending Study was granted by the University of Essex Ethics Committee. In addition, all three papers were reviewed by the Ethics and Research Governance system at the University of Southampton (ERGOII) (ID 18498, ID 31996 and ID 24482).

I am enormously grateful to my supervisors, Dave Martin, Gabriele Durrant and Patrick Sturgis (now at the LSE) for their patience and encouragement and to Mick Couper for allowing me to spend two extended study periods at the Institute for Social Research in Ann Arbor. My thanks also to Peter Smith, Olga Maslovskaya and Fred Conrad for their insightful comments. I am also grateful to the research teams at the University of Essex and at the MRC Unit for Health and Ageing at UCL for allowing me to build on collaborative projects for my PhD research, in particular Andy Wong, Rebecca Hardy, Rachel Cooper, Annette Jackle, Alessandra Gaia, Alex Wenz, Mick Couper and Tom Crossley. Jane Parsons, Gemma Harris, Glenn Miller, and Alex Frosch at the University have been efficient and kind in equal measure and played an important role in getting me to the finishing line and thanks to Debbie Collins for being a steadfast PhD buddy. Finally, an enormous thank you to my darling Nick, Noa and Gabriel for their love and encouragement, Maurice, Leila, Nick and Suszy for their inspiration and constant support and particularly to Suszy for her endless encouragement, Anita for going above and beyond by providing numerous delicious countryside retreats, Lyla for always being a listening ear, and my truly wonderful friends, colleagues, and community for making everything worthwhile.

Definitions and Abbreviations

ADMIN	A specified set of variables drawn from administrative records used in Chapter 3
ATS	American Thoracic Society
BMI	Body Mass Index
CAPI	Computer Assisted Personal Interviewing, in-person interviewer administered
CATI	Computer Assisted Telephone Interviewing, a voice call interview method
CE marking	The item fulfils the requirements of the relevant European product directives European Conformity standard
CLOSER	The Cohort and Longitudinal Studies Enhancement Resources
DBP	Diastolic Blood Pressure
DEM	A set of demographic variables drawn from survey data and used in analysis in Chapter 3 (these are gender, school year group, ethnic group)
ERS	European Respiratory Society
ESRC	Economic and Social Research Council
FEV_1	Maximum forced expiratory volume in 1 second
FSM	Free School Meals
FSM6	Free School Meals within the last 6 years
FVC	Forced vital capacity
GOR	Government Office Region
IDACI	Income Deprivation Affecting Children Index
IPTW	Inverse Probability of Treatment Weighting
KS2	Key Stage 2: four years of school when pupils are 7-11 years old
KS4	Key Stage 4: two years of school when pupils are 14-16 years old and includes General Certificate of School Education (GCSE) examinations
LCF	Living Costs and Food Survey
LOA	Limits of agreement
mmHg	Millimetres of mercury
MCS	Millennium Cohort Study
MRC	Medical Research Council
NHANES	National Health and Nutrition Examination Survey

NHANES III	National Health and Nutrition Examination Survey III (1988-1994)
NSHD	The National Survey of Health and Development (the 1946 birth cohort study)
NIH	National Institute of Health
PC	Personal computer – used to refer to desktop, laptop, or netbook
PRE	A set of geographically based variables used in analysis in Chapter 3 (this includes Income Deprivation Affecting Children Index, Government Office Region and, Rural/Urban)
RAM	Random-Access Memory
RDD	Random Digit Dialling
SBP	Systolic Blood Pressure
SEN	Special Educational Needs
SET	Science Education Tracker
SUR	A set of miscellaneous variables drawn from survey data and used in analysis in Chapter 3 (studies or intends to study maths or science at Level 3, aspires to higher education qualification, lives with parent who attended university)
UCL	University College London

Chapter 1 Introduction

Social surveys have a long history. This introduction sets out a definition of what constitutes a social survey for the purpose of this thesis (Section 1.1.) and suggests that this approach to data collection has remained useful over time by evolving to provide the information necessary to address the needs of contemporary society (Section 1.2). In part, social surveys have proven resilient by responding to, and adopting new technologies that have emerged in wider society. Some of these technologies have provided operational improvements, while others have changed the way that data are being collected or have extended the kinds of data that can be gathered through them (Section 1.3). While many of the technological changes that have affected social surveys have brought benefits, they have also presented challenges, in areas such as ethics and data security, participant burden, cost and logistics. Furthermore, technological change does not necessarily result in improvements in the quality of research data being collected, and careful consideration needs to be given to ways that technological change may present problems for data quality, for example by leading to errors of representation or errors of measurement (Section 1.4).

Against this context, this thesis asks the question: do technologies compromise the quality of data collected in social surveys?

It will focus specifically on the effects of these new technologies on data quality and measurement error. It presents three research projects that provide examples of the use of technologies in social surveys and examines an aspect of data quality in relation to each (Section 1.5). These themes are reconsidered in the conclusions (0).

1.1 A note on the definition of the social survey in this thesis

Although in common parlance, what constitutes a social survey is uncontroversial, there is some discussion about the boundaries of what should and should not be included. A good working definition is that a survey is “a systematic method for gathering information from (a sample of) entities for the purpose of constructing quantitative descriptors of the attributes of the larger population of which the entities are members. The word “systematic” distinguishes surveys from other ways of gathering information. The phrase “(a sample of)” appears in the definition because sometimes surveys attempt to measure everyone in a population and sometimes just a sample.” (Groves et al., 2011, p.2). The entities described are most often people (whether individuals, couples, families, or households) but can also be other units such as establishments. Groves et al. (2011) correctly state that survey “information is gathered primarily by asking people questions …

collected either by having interviewers ask questions and record answers or by having people read or hear questions and record their own answers" (p.3). However, De Vaus (2013) usefully broadens this focus: "The data … could be collected by other means such as interviewing or observing each case, by extracting information from records we have on each person or by many other means" (p.4). Perhaps deliberately controversially he states that "There is no necessary connection between questionnaires and survey research" (p.4). Instead, he places emphasis on the form of the data that a survey produces. "Surveys are characterised by a structured or systematic set of data which I will call a 'variable by case data grid'. All this means is that we collect information about the same variables or characteristics from at least two (normally far more) cases and end up with a data grid" (De Vaus, 2013).

In this thesis the focus is on surveys of people, and a conventional view is taken of the essential core of a survey, which asks questions of the target population or their proxies, either using an agent such as an interviewer or survey nurse or through a form of self-completion. However, a wide range of additional types of data are included, which relate to that individual or their behaviours, and which may be collected from a variety of sources. From this perspective, the survey can be seen as an effective vehicle for gathering a wide range of data in a variety of ways.

Using the term *social* survey reflects a long tradition of using this research approach for public benefit as explained by Shelby in 1931:

> "The social survey is a cooperative undertaking which applies scientific method to the study and treatment of current related social problems and conditions having definite geographical limits and bearings, plus such a spreading of its facts, conclusions, and recommendations as will make them, as far as possible, the common knowledge of the community and a force for intelligent coordinated action" (Shelby, 1931, p.20)

1.2 Adaptability and continued importance of the social survey

The social survey has provided an important source of information for public debate since its earliest manifestations in England and the United States in the 1880s. The approach became more recognisable as the modern social survey with the development of sampling theory, inferential statistics and attitude measurement in the 1930s (Bulmer et al., 1991, Converse, 2017). Its popularity grew in response to the rapid social changes that accompanied industrialisation, urbanisation and war (Bulmer, 2001). Periodically, concerns about the future of the social survey have been expressed and pushed away (Biemer, 2018, Bogart et al., 1987, Couper, 2013a, Couper, 2013b), for example in response to rising costs and declining response rates (Boyle, 2020), or increased concerns about data disclosure risk (Lambert, 1993), or loss of confidence in survey

responses based on insights from social and cognitive psychology (see, for example, Beniger in Bogart et al., 1987) or in response to widely reported failures, such as the inability of polls to accurately predict election results (Sturgis et al., 2018). More recently there have been debates about how other data sources, whether from administrative records or Big Data, could entirely, or partially, replace the need for survey research (Biemer, 2018).

Although the question "Is the sky falling?" may continue to be asked intermittently (Couper, 2013a), in practice, surveys have proven to be remarkably resilient. One of the reasons for this is their extraordinary adaptability. Social surveys can be used to gather a wide range of social science research data about people's circumstances, experiences, attitudes, and behaviours, and provide a vehicle for questions on an unlimited range of topics. They have been adapted to meet varied and changing data needs, both for academic research and planning purposes. In Martin Bulmer's words: "The social survey developed in close relationship with public policy and social reform," involving an interplay of "both social scientists and social reformers" (Bulmer, 2001, p.14469).

Over time, they have been used by the government in the UK to address an array of policy issues, ranging from enumerating poverty to understanding the effect of wartime rationing, to measuring the health needs of the population. Alongside the census, surveys provide evidence for almost every area of government decision-making, including housing, taxation and social security benefits, pensions, and education. In academia, surveys have been tailored to the needs of different disciplines such as economics, epidemiology, sociology, psychology and geography. Surveys can be used to gather data from the general population or can be targeted at specific population sub-groups. They may cover subjects which are relatively dry, for example by collecting household incomes from multiple sources to model take-up of means-tested benefits (Department for Work and Pensions, 2022, Pudney et al., 2006) but also those which are highly sensitive, for example surveying the sexual attitudes and lifestyles of the general population to understand illicit drug use and sexual risk behaviours (Erens et al., 2014, Paquette et al., 2017). They can also be used to investigate challenging populations, for example to understand the role of childhood trauma among adult sex offenders (Levenson and Grady, 2016). They can provide ad hoc information at a point in time, provide repeated cross-sections, for example tracking social attitudes in Britain and Europe (Jowell et al., 2007, Park et al., 2013) or can be used to understand longitudinal change, for example through the British birth cohort studies (Connelly and Platt, 2014, Elliott and Shepherd, 2006, Power and Elliott, 2006, Wadsworth et al., 2006). In summary, social surveys are incredibly diverse in scale and topic, and stretch across sectors.

The building blocks or scaffolding of almost any social survey are questions and answers. By maintaining consistency over time, these can allow governments to track trends (for example

using the Crime Survey of England and Wales, the Family Resources Survey or the Labour Force Survey) or by developing new question sets to address emerging policy concerns (Campbell-Hall et al., 2010). In fields such as economics, a range of different approaches have, by necessity, been used to measure key concepts such as income and wealth and to adapt to the opportunities provided by single focus surveys and the constraints of multidisciplinary studies (Crossley and Winter, 2016) while in other disciplines there is a strong focus on standardised question sets.

So far, the focus has been on surveys that include traditional question and answer formats, but an important aspect of the adaptability of the survey has been its ability to measure phenomena that cannot be collected by asking a participant to respond to a set off direct survey questions and instead require some form of assessment, normally administered by interviewers or survey nurses, following a defined protocol. In all these cases, the core survey questions provide a data record about the respondent, and additional data is collected and then linked to these records.

Health studies such as the Health Survey for England (Mindell et al., 2012) and the National Health and Nutrition Examination Survey (NHANES), or multi-disciplinary studies which include a health component such as the English Longitudinal Study of Ageing (Steptoe et al., 2012), the Health and Retirement Study (HRS; Sonnega et al., 2014) or Understanding Society (McFall et al., 2014), provide numerous examples. These surveys are used to collect objective assessment of an individual's cognitive function (Crimmins et al., 2011, Formanek et al., 2019, Langa et al., 2020, Steel et al., 2003), anthropometric measurements such as height, weight, waist and hip circumference (Cobiac and Scarborough, 2021, Hirani et al., 2010, Power and Elliott, 2006), physical performance such as gait speed, chair rises and balance assessments (Melzer et al., 2006, Ofstedal et al., 2013, Zhang et al., 2019) and innovative measures such as digit ratio (Al Baghal et al., 2014, Hand, 2020).

Another important area of data collection where supplementary data is linked to the basic survey record, is in biomarkers taken from biological samples, for example from blood, saliva, or urine (Sakshaug et al., 2014, Woodhall et al., 2016). Once again, this provides objective evidence that is unknown to respondents so could not be reported by them. For example, while a study participant may be able to report whether their doctor has told them that they are pre-diabetic or have diabetes, only an objective, contemporaneous fasting blood glucose test will make it possible to accurately identify undiagnosed diabetes (Pierce et al., 2009).

Other fields of study have used surveys as a vehicle to gather specialist data, linked to the individual, which goes beyond the simple format of questions and answers. For example, surveys have been extended through the placement of "leave behind" diaries to measure expenditure (Ralph and Manclossi, 2016), time use (Chatzitheochari et al., 2018) and travel (King et al., 2019). The National Travel Survey is an example of several studies that have experimented with

capturing GPS data as a supplement to survey data collection (Bricka et al., 2009, Rofique et al., 2012). More controversially, the Millennium Cohort Study incorporated interviewer observations of the physical environment of the home and parental interactions with the child during the assessments (Chaplin Gray et al., 2009).

In all the examples above, the core survey questions generate a data record about the respondent and supplementary data is collected and linked to those records. Another approach to achieving this goal, which also extends the scope and value of social surveys, is to link the survey information about the individual, with the consent of the survey participant, to non-survey data sources such as administrative records and, in some cases, data from social media sources (Gibson et al., 2016, Kim et al., 2016). In the case of administrative data, these external records may provide detailed information about the person's health or health service use, education, or benefit and pension entitlements at the level of the individual, the benefit unit or household. In recent years it has become commonplace for large scale academic and government surveys to seek permission to link survey responses to an array of detailed administrative data held by, for example, the Department of Work and Pensions (DWP), HM Revenue and Customs (HMRC) and Hospital Episode Statistics (HES) (Blom and Korbmacher, 2018, Calderwood and Lessof, 2009, Jones et al., 2019). Education research is particularly powerful when attitudinal and behavioural data is collected from parents and school pupils, and is then combined with administrative records from the National Pupil Database (NPD) or Independent Learner Record (ILR) held by the Department for Education, which provide objective information about, for example, the student's Free School Meal status, the qualities of their school and their individual academic attainment (Lessof et al., 2019, Lessof et al., 2016).

These examples show some of the diverse ways in which the utility of surveys has been increased by combining basic survey questions with a wide range of additional data collection methodologies and sources. Crucially, the core element of the survey provides information about the circumstances, attitudes, or behaviours of the respondents, which provides a mainstay for subsequent analyses. Arguably, therefore, the resilience of surveys can be attributed to both the endless adaptability of survey questions to gather the populations attitudes, behaviours, and experiences, but at the same time the ability to take a much wider set of measurements into people's homes or to draw complex data from other sources to that population sample.

1.3 Adaptation to technological change

Many of the developments discussed in the previous section rely on the adoption of new technologies. Consequently, an important and inter-related aspect of the resilience of the social survey is the way that practitioners have, over time, adapted surveys in response to the changing

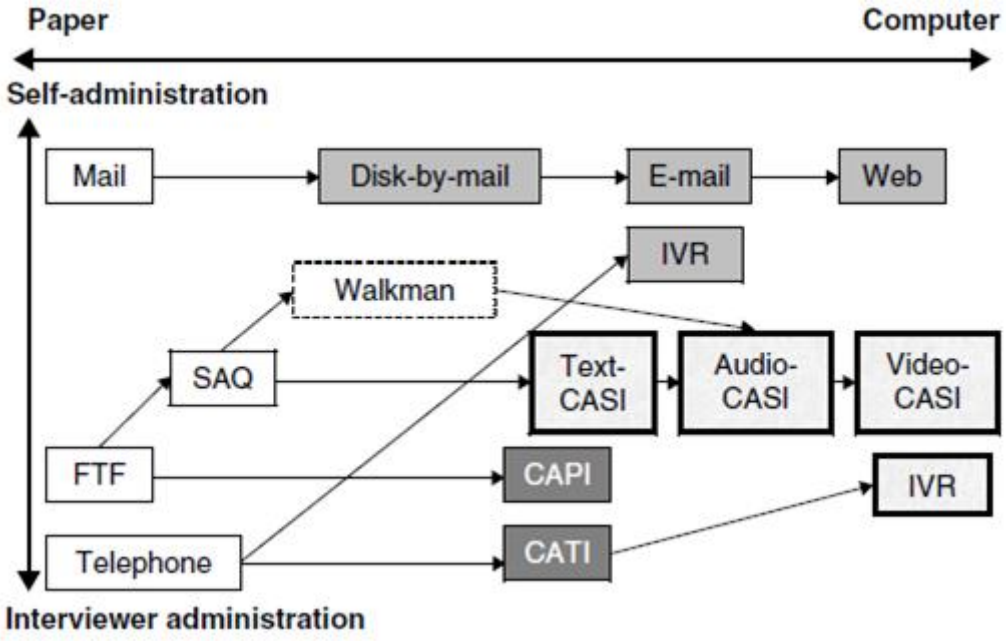
technological environment and to emerging technological opportunities. Adaptations have been seen both in the communication technologies used for survey delivery in different modes, and in the diverse range of technological developments that have delivered supplementary measures and data sources.

The first of these, the development of communication technologies used to deliver survey questions, is well documented (Schober and Conrad, 2008). Before the advent of the first recognisable modern social surveys, trained social investigators used pen and paper to record the socio-economic characteristics of neighbourhoods in 19th century Britain and in the United States (Shelby, 1931). As the modern survey emerged, with the scientific developments of probability sampling and social measurement, surveys continued to rely on face-to-face interviews administered using paper and pen, but the speed and scale of data entry and analysis was transformed through the 1920s and 1930s with the development of punch cards, the antecedent of digital technologies (Armstrong, 2019). The transformative changes in surveys that followed all occurred in response to massive shifts in the technological environment and the ingenuity of survey professionals in responding after each disruptive technology emerged (for example by Couper, 2008). For example, survey research was changed radically by the invention and penetration of telephony (from landline through to mobile and smartphone) which spawned the development of techniques such as Random Digit Dialling (RDD), Interactive Voice Response (IVR), text messaging and text interviews (Conrad et al., 2014, Conrad et al., 2017). In parallel, the emergence of mainframe and then personal computing facilitated the development of computer assisted interviewing in person (CAPI) and by telephone (CATI). And the invention of the internet facilitated new forms of surveys, most notably web or online[1].

It is possible to conceptualise these changes along two dimensions, from paper to computer-based interviewing and from interviewer administration to self-administration as illustrated in Figure 1 below (Couper, 2008), including the growth of self-administered paper and computer-assisted self-interviewing (CASI). Couper's illustration shows the interplay of these two dimensions of development and the many transitional technologies that were employed such as disk-by-mail and email, which preceded the emergence of web or online surveys as a substantial force. Understandably, it does not show the more recent technological developments particularly those based on mobile device technology.

[1] The use of the phrase computer-assisted interviewing is no longer seen as helpful given the ubiquitous presence of computers and digital technology. Nevertheless, this describes an important part of the development of the survey landscape.

Figure 1 Couper's illustration of the evolution of survey technology



Source: Couper, 2008, p.59

The social survey has adapted to new technologies in large part by necessity to survive in a changing environment. In a competitive research market, it is necessary to adapt, and indeed to show the ability to innovate by taking advantage of emerging technologies. For example, beyond an initial transitionary period, it was not realistic for survey practitioners to insist that respondents use a PC rather than a smartphone to complete an online survey. The adoption of new technologies also reflected the fact that they offered potential benefits, some of which were consciously acknowledged at the time of their adoption while others were serendipitous. One of the merits of technological development has been reduced costs. For example, the growth of telephone and online interviewing both had significant cost advantages over in-person interviewing and are significantly easier to deliver logistically. There has also been the promise, and in many cases the delivery, of improvements in data quality. Most obviously, the introduction of computer-assisted interviewing (whether CAPI, CATI or CASI) brought automated question routing and automated checks as interviewers entered responses. In addition, advances in survey programming facilitated a range of more sophisticated research techniques. For example in longitudinal research, dependent interviewing was used to reduce seam effects by reminding participants of past responses (Jäckle, 2009) and advanced programming supported visual displays of event history data which helped interviewers to elicit more accurate retrospective data than had been possible with question list surveys (Belli et al., 2001).

There have been many advances that draw on technologies developed outside of (academic) research. To take one example the development of easy mechanisms for recording and transferring sound files facilitated the augmentation of surveys with audio-CASI. This was

promoted as an approach to elicit more candid responses to sensitive questions, although arguably the advantages were overstated (Couper et al., 2009). Sound files were also used in other ways to improve data quality, for example by using pre-recorded word lists to deliver a cognitive function assessment consistently across all participants and across all waves of a longitudinal study (Steel et al., 2003). Similarly, technologies such as eye-gaze tracking technology which attracted considerable attention in advertising research, have been used as a primary research method to understand gambling behaviours (McGrath et al., 2018) and also as a methodological tool to support question design and testing (Romano and Chen, 2011) although they are likely to remain a niche methodology. More recently, new types of data were promoted and enhanced by developments during the COVID-19 pandemic, which has been a significant driver of methodological innovation and digitalisation, including, for example, the use of live video interviewing (West et al., 2021).

Using mobile devices as part of survey data collection has offered several advantages. For example, they have facilitated a number of research studies based on ecological momentary assessment, for example, to measure wellbeing (de Vries et al., 2021). Mobile apps can be used to facilitate diary keeping, for example recording purchases of foods to assess the impact of a public health campaign (Wrieden and Levy, 2016). By tapping into the inbuilt capability of the device itself, it has also been possible to extend the types of data that can be captured. For example, Kantar Worldpanel has made use of the ability of smartphones to scan barcodes (Jäckle et al., 2019b), while other studies have collected photographs to assess dietary intake (Sharp and Allman-Farinelli, 2014), or have captured digital trace data such as geolocation, accelerometer data, phone and text messaging logs and app use (Kreuter et al., 2020). Survey data collection approaches which use a mobile app offer several advantages: the survey does not need internet connection at the time of data collection, the app can access device capabilities such as GPS, pictures, videos, voice recording, barcode scanning, sound, and other sensors, and can send notifications and alarms (Callegaro, 2013). In addition, an app allows more control over how the survey will be displayed or interacted with. Technological developments have also led to significant changes in the way that surveys are managed, for example by improving the management of fieldwork through the use of call records and contact history data (West, 2011) and by improving the efficiency of coding and processing of data. Technological developments have underpinned the use of paradata for other purposes, for example verbal paradata to record interactions between interviewers and respondents, and automated information about the time taken completing a survey or individual questions (Couper and Kreuter 2011). These are examples of the rich data that is now available to methodologists and practitioners to understand more about the process of survey delivery.

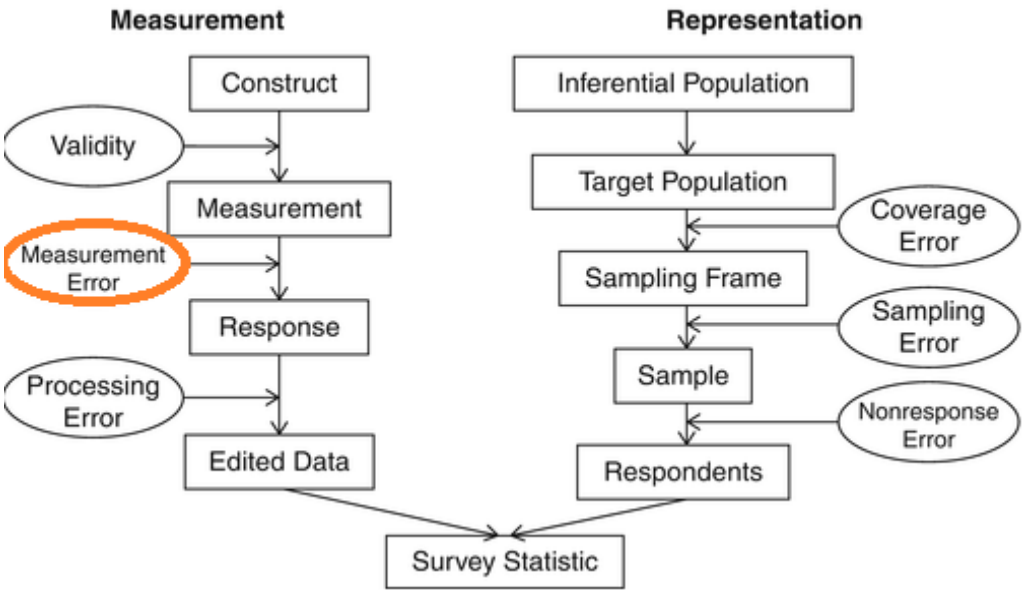
Future developments are likely to draw on new technologies, will respond to new areas of scientific enquiry where there is a desire to collect population level data and will probably respond to further shifts in the technological and social environment. During the COVID-19 pandemic, for example, lockdowns to prevent the spread of the disease led to fieldwork for many in-person surveys being put on hold or being rapidly transitioned to alternative telephone or online modes. In addition, public concern led to mass engagement with large-scale research studies, such as the Zoe app which involved daily reporting in a mobile app over a period of many months. Furthermore, the large-scale collection of biological samples to measure the prevalence of disease, for example in the ONS study, meant that millions of members of the public provided biomarkers for a national survey. All these trends may prove to have lasting impacts on the way that survey data collection and survey research is carried out in the future, and how it is perceived by the public. In this instance, the survey research community responded rapidly to a radical change in environment. In the past, this has not always been the case, and sometimes there have been tensions between rapid adoption and rejection of new technologies (Couper, 2008). This dynamism presents significant challenges to researchers, survey practitioners and methodologists. Key questions include when to consider the inclusion of allied technologies, when does a new technology reach a sufficient standard to justify its inclusion in a social survey, how to identify technologies that will have lasting value, and how to ensure that the devices are implemented effectively and can be properly evaluated.

1.4 Challenges of new technologies

However, while there are clearly many benefits to using new technologies in social surveys, it is also important to address the possibility that the introduction of technologies into the survey research process may have an impact on the quality of the data collected. The Total Survey Error framework, in which all elements of data production and collection are considered within a single schema is illustrated in Figure 2.

. This provides a classic approach to thinking about two major categories of error: errors of representation (in relation to coverage, sampling and nonresponse) and errors of measurement (particularly in relation to construct validity, measurement error and processing error) (Biemer, 2010, Groves and Lyberg, 2010).

Figure 2 Representation of Total Survey Error



Note: Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process (replicated in Groves and Lyberg, 2010, Figure 3)

Almost any technological change, whilst aiming to benefit survey data collection, can have negative effects on each of these error structures. This thesis focuses on errors of measurement associated with technology as highlighted (in orange) in Figure 2 above. It gives less attention to other errors, particularly errors of representation, though it is important to recognise the ways in which the use of technology in research may impact in these ways (Groves et al., 2011, Groves and Lyberg, 2010, Lessof and Sturgis, 2018).

This thesis provides three accounts of research activities that have been affected by technological change and examines measurement error or data quality. Two of the papers are related to different types of device effect, the first in surveys that collect biomeasures with different equipment and the second in online surveys that allow respondents to use different devices. The third is related to a relatively new technology platform, mobile app-based research. Each examines the process quality of the data collected. These papers are outlined in detail in the Section 1.5 below.

There are, of course, many other examples that illustrate the impact that technology may have on data quality in survey practice. For exampe, there have been several references to the measurement of gait speed which uses the simple technologies of a tape measure to demarcate the course and a stop-watch to time the respondent's walks. Here, the interviewer is relied upon to implement the chosen protocol consistently and may make errors, for example deciding when to switch the stop-watch on and off (Sustakoski et al., 2015). Consequently, researchers have shown an interest in using more advanced technologies such as a body worn monitor (Godfrey et al., 2015a) which may also extend the measures of physical capability, balance and gait that can

be collected and analysed (Godfrey et al., 2015b). However, implementing this approach requires additional costs and logistical effort and raises practical issue such as how to attach the monitor to the small of the respondents back, and how to ensure that interviewers are successful in initialising the monitor on every occasion.

A second example, is the assessment of cognitive function. Here, the use of sound files may help to standardise the delivery of a word recall measure. Another well-known element of cognitive assessment is the animal-naming assessment, in which a participant is asked to name as many animals as possible within one minute. Interviewers need to score this assessment in real time, not only recording the number of words mentioned, but also discounting repetitions and words that do not qualify as an animal name, such as "Mickey Mouse". This means that there is considerable scope for measurement error within the assessment – which may well vary by the characteristics of the interviewer, their hearing, manual dexterity and their own cognitive abilities. This calls for an increased use of technology to record the respondents attempt and to check that it has been coded correctly, although the act of recording may affect the interviewer's and respondents' behaviours.

There are many examples of methodological studies where researchers have considered the level of measurement error associated with the technology they use in their studies. It is important to acknowledge that the Total Survey Error framework does not completely represent all types of error. Indeed, Groves and Lyberg recognise that: "Any listing is bound to be incomplete, though, since new error structures emerge due to new technology, methodological innovations, or strategic design decisions such as using mixed-mode data collection. All error sources are not known, some defy expression, and some can become important because of the specific study aim, such as translation error in cross-national studies" (Groves and Lyberg, 2010, p854). In some areas, such as Big Data, attempts have been made to extend the TSE framework (Amaya et al., 2020, Link et al., 2014). Although some technologies may introduce data types with different error structures, as is the case with Big Data, for the purpose of this thesis, the traditional consideration of errors of representation and measurement are sufficient.

1.5 Focus of the thesis and overview of the three research papers

The aim of the thesis is to examine the impact of new or changing technologies on the quality of data collected through surveys. The thesis is made up of three original research studies (Chapters 2-4), which contribute new evidence to these discussions around the implications of technological change on data quality and measurement error. First, a short summary of each chapter is provided, setting out the respective aims and/or research questions, then the chapters are briefly compared, highlighting key differences.

Chapter 2 presents the findings from a methodological research project about the collection of a set of common physiological measures of health status – blood pressure, grip strength and lung function – which are commonly included in biosocial surveys and analysed by epidemiologists. **The key research question is whether the measurements collected differ depending on the make and model of specialist equipment which are used to collect them.**

The study was originally designed to support decision making by the Principal Investigators of the 1946 British birth cohort study, known as the National Survey of Health and Development (NSHD) at the MRC Lifelong Health & Ageing Unit, University College London (UCL), who needed to replace the equipment used for these measures in previous survey waves and required evidence to support future analyses of these measures so comparisons could be made across waves, or between studies which used different devices. Alongside the development and delivery of the 2015-16 NSHD survey fieldwork, a stand-alone randomised controlled trial of 118 people aged between 45-74 was conducted, using two sphygmomanometers (for blood pressure), four handgrip dynamometers (for grip strength) and two spirometers (for lung function). The primary analysis carried out in this chapter employs Bland and Altman plots (Bland and Altman, 1986, Bland and Altman, 1999), commonly used in the epidemiological literature. In addition, sensitivity analyses are conducted which go beyond the standard approaches, for example testing different definitions of each measure, and using multilevel regression modelling as an alternative way of estimating device effects.

The study was funded by CLOSER, a centre which provides research infrastructure to support British cohort surveys by providing data services and shared learning (O'Neill et al., 2019). Although this study focuses on specific equipment for three common measures of health status, the approach used has more generally applicability because many health surveys such as the Health Survey for England or NHANES, and many multidisciplinary studies such as the English Longitudinal Study of Ageing, the Health and Retirement Study (HRS), and *Understanding Society,* include physiological measures of this type, collected using technical equipment. While some measures use simple technologies that can be applied consistently across studies (see, for example Wang et al., 2015), many others, including these three, use complex medical equipment made by multiple manufacturers, with different models released over time. Differences in the technologies used by different studies, and changes in the technologies used in each study over time, introduce the risk of measurement error. Equipment comparison studies have been conducted before, but not for this combination of devices. This study represents one type of methodological research that is needed alongside the use of medical equipment within surveys.

Chapter 3 involves secondary analysis of a large-scale, online survey of young people aged 14-18 called the Science Education Tracker (SET) which is commissioned periodically by the Wellcome Trust to inform science education and strategy. **The key research question is whether data quality differs if respondents use a PC or a mobile device when answering an online survey.**

The context of this investigation is the growth of online surveys and the transition of responses to these surveys from PCs to the web browser of a mobile device. This has been accompanied by concerns that the different experience respondents may have using small devices which can be used on the go could be associated with differences in data quality. A strength of this study is that it is based on a large, random probability sample of young people who are digitally native. A quasi-experimental approach employing propensity score matching (specifically, inverse probability treatment weights or IPTW) is used to account for selection to device, making comparison of PC and mobile device responders possible. An additional aspect of the study is that, where possible, the matching process uses data that are exogenous to the device used to collect the survey data, such as linked administrative data. This study adds to a growing literature that explores device effects.

The aspect of technology that is investigated in **Chapter 4** is the development of a mobile app to deliver complex data collection tasks, such as diary keeping. **The key research question is whether a sample of survey respondents who agree to take part in an app-based spending diary engage fully – measured by compliance with different aspects of the study protocol – what factors are associated with adherence, and whether this is sustained over a one-month period.**

The study on which this analysis is based was funded by the ESRC and carried out at the University of Essex. A total of 2,432 respondents to the UK's *Understanding Society* Innovation Panel were invited to download a mobile app and record all their spending on goods and services for a month, by photographing receipts or reporting spending in the app. The analysis presented here is based on 268 adults who took part, drawing on 8308 observations and 3,454 photographed receipts.

The objective of the overall project was to understand whether an app-based spending study could be implemented successfully and, if the results were promising, to identify areas for further development. The app was evaluated from several perspectives. For example, a separate paper, Jäckle et al. (2019a), reported on initial response and representativity, while another paper considered outcome quality (Wenz et al., 2018). The purpose of the analysis presented in this thesis is to understand the quality of the data collected, by examining the extent to which the respondent engages with the process that is necessary to provide accurate data. This is

operationalised using the concept of adherence to protocol, borrowed from the medical literature (Couper, 2019), defined as how well respondents comply with four different aspects of the task, and how far they continue to do so over a one-month study period. The analysis uses multilevel regression models- linear, logistic and negative binomial to analyse these measures of adherence.

The mobile app study was intended to support economists in their search for better data collection tools to measure expenditure. In parallel, it was explicitly commissioned to support methodological investment in "transformative research". The findings from this study are relevant to those who are contemplating any research using mobile apps, and particularly those interested in using a diary methodology.

To summarise, the three papers in this thesis represent different aspects of data collection for social surveys and its use of technology. Table 1 below provides a brief comparison by theme, showing the differences in their funders, the research teams involved in the original studies, and that they are drawn from a range of social science disciplines – health, science education and economics. The comparison also highlights that the three chapters address different aspects of technological innovation, focusing on the quality of the data collected: whether the use of different models of equipment to collect biomeasures leads to differences in measurement (Chapter 2), whether data quality in an online survey varies depending on whether a PC or mobile device is used to respond (Chapter 3), and whether the expenditure data collected in a mobile app diary over a one month period can be assessed in terms of data quality (Chapter 4). The chapters also draw on different data types: biomeasures that could not be recorded without specific technologies (Chapter 2), a mix of survey data and survey process data (Chapter 3 and 4), as well as administrative and geographical data (Chapter 3) and data collected from photographs of receipts (Chapter 4). Similarly, the three studies use different approaches to investigate the effect of technological change. Chapter 2 is based on a stand-alone experimental design; Chapter 3 uses quasi-experimental methods to create balanced samples of young people who used PCs or mobile devices; Chapter 4 is exploratory and is based on analysis of a small-scale test of a mobile app study delivered as a follow up to a random probability longitudinal study. Finally, the three chapters also use very different statistical approaches, reflecting both the difference in the research questions addressed and their different disciplines. The concluding section of this thesis, Chapter 5, draws out key themes from the three papers.

Table 1 Summary information about the three papers (chapters 2-4)

Theme	**Chapter 2**	**Chapter 3**	**Chapter 4**
Funder	CLOSER	Wellcome Trust	ESRC
Research team	MRC Lifelong Health & Ageing Unit, UCL	Kantar Public	University of Essex

Table 1 cont.	Chapter 2	Chapter 3	Chapter 4
Discipline	Epidemiology	Science Education	Economics
Technological innovation	Comparison of the equipment used to collect three measures of health status: blood pressure, grip strength and lung function to establish whether medical devices measure differently	Investigation of an online survey of young people where participants use a PC or mobile device to respond, to establish whether the device used affects the quality of data provided	Exploration of whether participants adhere to a mobile app study to collect expenditure data by photographing receipts or providing summary information over one month
Purpose	To inform decision making about replacement equipment and to support analyses where devices are mixed	To inform decisions about future mobile app survey design and efforts to engage young people in surveys	To explore whether a mobile diary app can be used to measure expenditure over a one-month period
Data types	Biomeasures Survey responses	Survey responses, response behaviours, survey process data, geographical and administrative data	Survey responses, photographs and survey process data
Research methodology	Experimental: small-scale, stand-alone randomised controlled trial	Quasi-experimental: secondary analysis of Science Education Tracker (SET), based on large-scale cross-sectional data	Exploratory: bespoke sub-study completed by a sub-sample of the *Understanding Society* Innovation Panel
Sample definition	Sample of 118 adults, aged 45-74	Sample of 4081 young people, aged 14-18	Sample of 268 adults, aged 16+, who between them record 8308 spending events and take 3,454 photographs of receipts
Main statistical analysis methods	Primary analysis: Bland and Altman plots to test whether mean differences between devices vary by the magnitude of the measurement; paired t-tests to establish whether mean differences between devices effects are statistically significant. Secondary analysis: multilevel modelling to estimate device effects, taking account of clustering of multiple readings within each device, and controlling for additional covariates.	Logistic regression to compare a series of binary outcome measures, where the groups being compared are those who use a mobile device (for the primary analysis) and are matched to an equivalent sample who respond using a PC. The matching uses Inverse Probability of Treatment Weighting (IPTW).	Multilevel regression models (logistic, negative binomial and linear), which account for clustering effects in the data

Chapter 2 Comparison of different devices to measure blood pressure, lung function and grip strength: findings from a randomised repeated-measurements cross-over trial

2.1 Introduction

Surveys have provided a vehicle for researchers to collect a wide range of biomeasures and biomarkers. For example, In the UK this has been done in the Health Survey for England, the English Longitudinal Study of Ageing, the National Survey of Health and Development, Whitehall II and *Understanding Society*; and internationally, through NHANES, the Health and Retirement Study (HRS), the Survey of Health and Retirement in Europe (SHARE) and the Canadian Health Measures Survey (CHMS). These surveys include physiological measures such as blood pressure, lung and heart function; anthropometric measures such as height, weight, waist, and hip diameter; physical performance measures such as assessments of grip strength, gait speed and balance; and biological samples, most commonly from blood and saliva. These measures avoid the subjectivity of self-reports of health, and capture information that is not known by survey respondents. When these measures are collected from large, representative samples, alongside data about their characteristics, experiences, and behaviours, a wide range of biosocial investigations become possible, enabling researchers to compare populations and sub-groups, and to track changes in health and functioning over the life course.

These advances in research capability have, in part, resulted from the development of portable equipment and consistent protocols for a range of biomeasures, which have allowed survey nurses and interviewers to administer these assessments at scale, either in a home setting or in a research clinic. However, different makes and models of equipment made by different manufacturers have been adopted by different studies. Furthermore, almost inevitably, new models of equipment introduce improvements and older models need replacement because of damage or obsolescence. As a result, devices used to undertake these measures may differ between surveys and within surveys over time. If the differences between devices are random, this will increase the variance of the estimates obtained and so may reduce the precision of any findings. If, however, the measures provided by alternative devices differ *systematically*, with one device providing results which are, on average, higher or lower than another, this may bias the estimates and may lead to incorrect results and conclusions. Furthermore, the suspicion that

systematic differences exist between the measures provided by different devices, may discourage researchers from attempting comparisons across studies which use different devices, or within longitudinal studies where devices have been replaced over time. It is therefore important that the magnitude and direction of any systematic differences, or bias, be measured by conducting research which directly compares different devices.

This chapter illustrates this issue by examining three physiological measures where different devices have been used: to record blood pressure – both systolic (SBP) and diastolic (DBP), grip strength and lung function. These measures are commonly collected in cross-sectional and longitudinal biosocial surveys and are frequently used by studies in the CLOSER consortium. All three are non-invasive measures of physiological function that are practical for a survey nurse or interviewer to administer in a home or clinical setting using portable equipment. The advantages are that they avoid the subjectivity of self-reports of health, enable clinicians and researchers to track changes in health and functioning over the life course (Kuh et al., 2014) and are important biomarkers of healthy ageing (Lara et al., 2015). Their repeat assessment within longitudinal studies and their inclusion in many different surveys facilitates comparisons over time, across ages and cohorts (Dodds et al., 2014, Wills et al., 2011).

There have been a number of initiatives to encourage standardisation of the devices and protocols used (Gershon et al., 2013, Miller et al., 2005, Reuben et al., 2013, Standardization of methods, 1939) and to provide norms for these measures (NHANES, 1999, Scholes and Neave, 2017, Thomas et al., 2019). Nevertheless, different devices have been adopted by different surveys for a variety of practical reasons (Amnan et al., 1996, Goisis et al., 2014, Tolonen et al., 2015). The device used within a study may change over time as obsolete or outdated models are replaced with devices which may be more technologically advanced, and which improve or extend measurement, are less costly, more portable, or easier to use. Because medical devices of this kind are only subject to moderate regulation (Mohandes and Foley, 2010) such as CE marking (indicating European Conformity), the measures obtained from different makes and models of equipment may not be perfectly equivalent, with some devices possibly measuring, on average, higher or lower than others, thereby potentially introducing bias (Bridevaux et al., 2015, McFall et al., 2014, Mindell et al., 2011, Orfei et al., 2008, Wills et al., 2011). For example, some studies have shown differences between devices used to measure blood pressure (Bolling, 1994, Campbell and McKay, 1999, Skirton et al., 2011, Stang et al., 2006, Wan et al., 2010), grip strength (Guerra and Amaral, 2009, Kim and Shinkai, 2017, King, 2013, Svens and Lee, 2005) and lung function (Bridevaux et al., 2015, Gerbase et al., 2013, Hosie and Nimmo, 1988, Milanzi et al., 2019, Miller et al., 2005, Orfei et al., 2008).

These differences may have important implications for research findings. For example, it has been shown that a lack of adjustment for differences in devices used can have a marked influence on longitudinal trajectories of blood pressure (Wills et al., 2011). Similarly, artefactual findings attributable to a change in device have been seen in studies of lung function (Bridevaux et al., 2015, Orfei et al., 2008). Concerns about potential differences have led to study investigators discouraging potentially useful within-study and cross-survey analyses (McFall et al., 2014, Mindell et al., 2011).

However, while studies which compare different devices provide valuable evidence, they do not compare all the devices commonly used in cohort and longitudinal studies in the UK and other countries, and issues related to differences in devices are only occasionally discussed in a survey context.

2.1.1 Research questions

To address this gap, this chapter presents a stand-alone, randomised cross-over trial of 118 healthy adults aged 45-74, to estimate differences in measurements between devices used to assess:

- blood pressure measured using two sphygmomanometers,
- grip strength measured using four dynamometers, and
- lung function measured using two spirometers.

The purpose of the study is to identify any systematic differences in measurements when assessed using different devices.

It is important to acknowledge that this study focuses solely on device as a source of measurement error, although several other sources of measurement error may affect readings of blood pressure, grip strength and lung function (Amaral et al., 2012, Balogun et al., 1991, Bilo et al., 2017, Fess, 1981, Firrell and Crain, 1996, Handler, 2009, Incel et al., 2002, Jones et al., 2003, Miller et al., 2005, O'Driscoll et al., 1992, Roberts et al., 2011, Sousa-Santos and Amaral, 2017). The study was designed, developed, and implemented by the NSHD Principal Investigator team with researchers from Kantar Public and the thesis author. The funder of the study was CLOSER (O'Neill et al., 2019).

The remainder of the chapter is structured as follows: the data and methods are described in Section 2.2, the results are set out in Section 2.4, and the findings and implications are discussed in Section 2.5. A related paper demonstrates the potential importance of these findings for clinical research and practice (Cooper et al., 2021).

2.2 Data and methods

2.2.1 Study design and devices

A small, stand-alone cross-over trial was carried out in which a sample of individuals were randomly assigned to a sequence of measurements in a single visit following established guidelines (Moher et al., 2010, Schulz et al., 2010) using two sphygmomanometers to measure blood pressure, four handheld dynamometers to measure grip strength, and two spirometers to measure lung function. The makes and models of equipment used are shown in Table 2 and illustrated in Figure 3.

Table 2 Make and model of devices

Device type (and measurement)	**Device 1**	**Device 2**	**Device 3**	**Device 4**
Sphygmomanometer (blood pressure)	Omron 705-CP	Omron HEM-907	n/a	n/a
Hand-held dynamometer (grip strength)	Jamar Hydraulic Analog Hand Dynamometer	Jamar Plus+Digital Hand Dynamometer	Nottingham Electronic Handgrip Dynamometer	Smedley Spring-gauge Dynamometer
Spirometer (lung function)	Micro Plus by Micro Medical, a turbine spirometer	Easy on-PC by NDD, an ultrasonic flow-sensor spirometer	n/a	n/a

The two sphygmomanometers (Figure 3) are made by the same manufacturer (Omron), with the newer model (HEM-907) providing greater automation than its predecessor (705-CP).

Two of the dynamometers are electronic (shown on the right in Figure 3), one is spring gauge (shown far left in Figure 4), and the other is hydraulic (shown second left). One spirometer, the Micro Plus by Micro Medical (Figure 5), is an older, low-cost, hand-held device, while the other, the Easy on-PC by NDD, is a newer PC/tablet-based spirometer that includes specialist software which provides feedback and evaluates and records blows automatically,

Figure 3 Two sphygmomanometers: Omron 705-CP and Omron HEM-907

Note: Photograph of the two blood pressure devices, both shown with cuffs, used in the study

Figure 4 Smedley, Jamar Hydraulic, Jamar Plus+ and Nottingham Electronic dynamometers

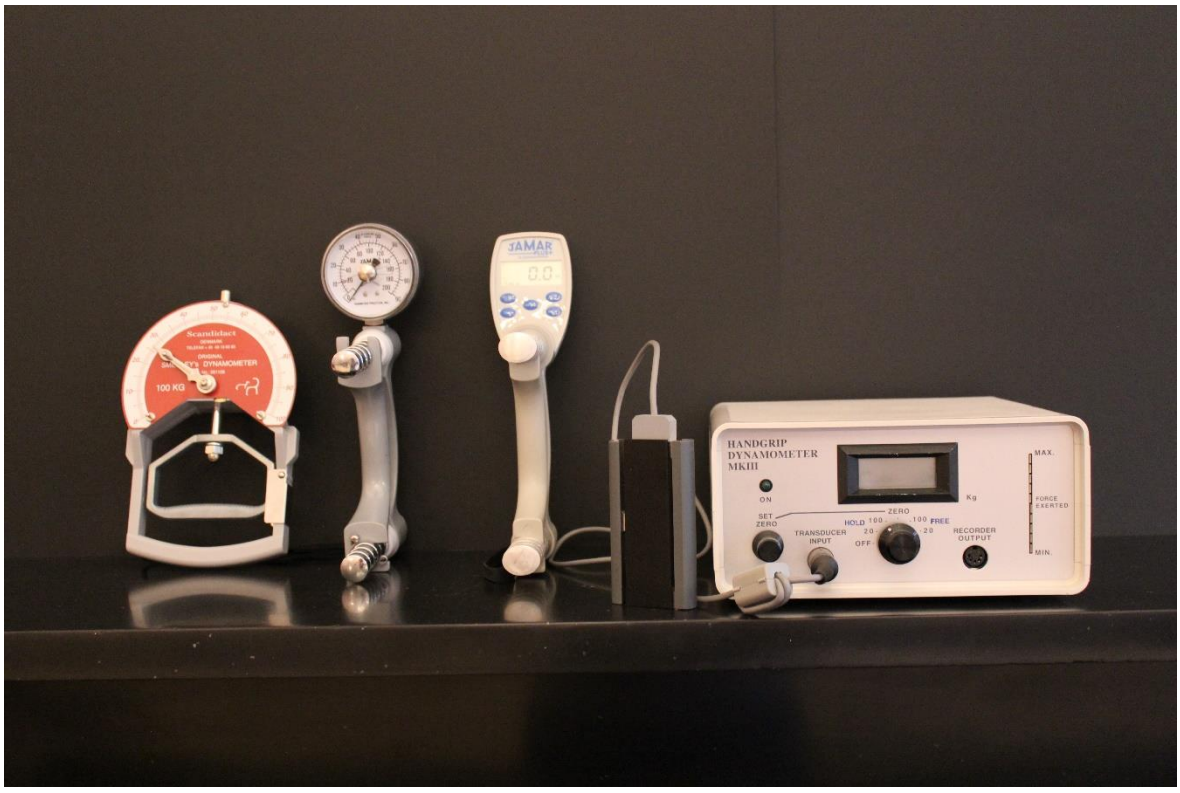
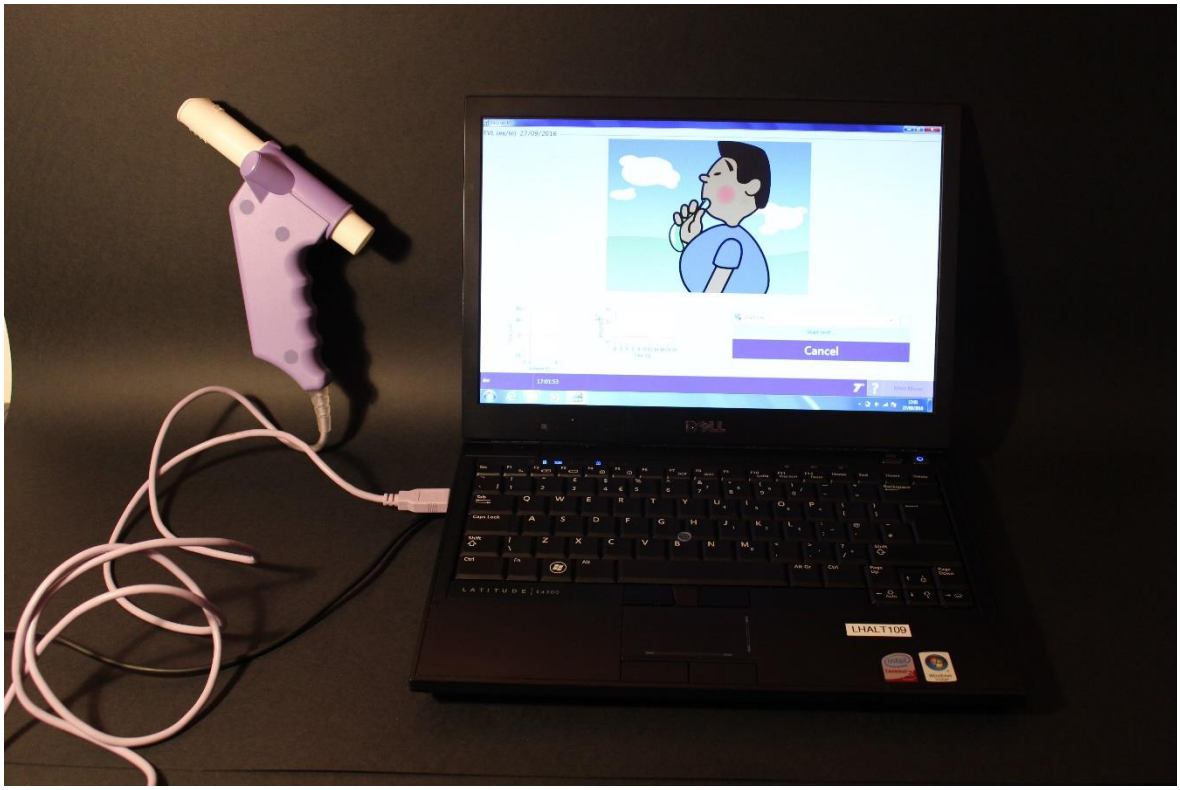


Figure 5 Micro Plus by Micro Medical, a turbine spirometer

Figure 6 Easy on-PC by NDD ultrasonic flow-sensor spirometer with on-screen feedback



2.2.2 Sample

The target sample, based on sample size calculations, was 120 men and women from the general population aged 45 to 74, comprising 20 men and women from each age group (45-54, 55-64 and

65-74). The principle underpinning the calculation of the required sample size required was that the sample should be large enough to ensure that the randomised cross-over trial would generate results which provide sufficient confidence in the estimation of differences in measurement between devices to achieve the study goals. The estimation of the sample size was based on the requirements for lung function, since this measure required most precision. It was calculated using an expected correlation between the two spirometers of between 0.8 and 0.9, following the method presented by Kaaks, Riboli and van Staveren (Kaaks et al., 1995, Minelli, 2014). This showed that a sample size of 100 subjects would be sufficient, which aligned with similar studies which rely on samples of between 30 and 220, with the large majority in the region of 100. A much larger sample, of perhaps 200, would have been needed to estimate a conversion (or correction) factor, for example using regression calibration. Here, the intention was simply to identify a reasonable estimate of the differences between devices.

Participants were recruited from a large database of members of the public who had participated in the TNS Omnibus survey, a market research study which uses a non-probability, multi-phase sampling design, random location quota sampling. The geography of Great Britain was stratified using 2011 Census small area statistics and the Postcode Address File to define sample points. Clusters of wards were selected within these sample points and within these clusters Census Output Areas were sampled. Participants were recruited within these areas, meeting a set of population quotas. Individuals who responded to the opening section of the interview were asked if they were willing to be re-contacted for future research purposes, and if they agreed they were added to a database which was available for sampling for follow-up studies. The TNS database was used to fill the study quotas of men and women between the ages of 45 and 74, who were living in London and the South-East.

An invitation letter and information sheet were posted out and this was followed-up by a telephone recruitment process. This included assessment of health-related exclusion criteria which were: a chest infection (such as influenza, pneumonia, bronchitis, severe cold) in the last 4 weeks; coughing up blood of unknown origin in the last 4 weeks; a heart attack or other heart complaint in the last 6 weeks; a stroke in the last 6 weeks; abdominal or chest surgery in the last 3 months; ever having been diagnosed with an aneurysm in chest, brain or stomach; a detached retina or eye surgery in the last 3 months; ear surgery in the last 3 months; a collapsed or punctured lung in the last 12 months; a blood clot in the lung in the last 3 months and; currently on medication for tuberculosis. Recruitment continued until the target number had been recruited within each age and sex group.

Eligible individuals were invited to attend a face-to-face assessment at the offices of the MRC Unit for Lifelong Health and Ageing in central London. Data collection took place between October

2015 and January 2016. Each was carried out within a one and a half hour time slot and the data was collected by one of seven researchers who were trained and tested in all relevant protocols. Participants and researchers were allocated to time slots according to their availability across the study period.

2.2.3 Assessment

Before the assessment began, the purpose of the study was explained, and all participants gave informed, written consent to participate. Ethical approval was given by UCL (Ethics Project Number: 6338/001) and, for analysis, by the University of Southampton (Ethics Project Number: 18498). The exclusion criteria were re-checked. In addition, before the grip strength assessment took place, participants were excluded if they had severely raised blood pressure (SBP ≥200mmHg or DBP ≥120mm Hg) or if either hand could not be assessed because of swelling or inflammation, severe pain or recent injury, or hand surgery in the last six months. Neither of these situations arose.

During the assessment, each participant was assessed in the sequence shown in Table 3. Blood pressure was measured consecutively on each device after a period of quiet rest. The remaining measures were ordered to ensure that there was sufficient time between the four grip strength measurements, and two spirometry measurements to avoid participants becoming fatigued. Height and weight were also measured, and a short self-completion questionnaire administered.

Table 3 Order of activities during assessment

1.	Introduction and consent module
2.	Three-minute rest period
3.	Blood pressure 1
4.	Two-minute rest period
5.	Blood pressure 2
6.	Height and weight
7.	Grip strength 1
8.	Lung function 1
9.	Grip strength 2
10.	10-minute break including paper self-completion
11.	Grip strength 3
12.	Lung function 2
13.	Grip strength 4
14.	Copy of assessment data and £50 gift voucher given to participant

For each physiological function, the order of device used in each assessment slot was determined before fieldwork began, using computer-generated random numbers within each age-sex strata (Moher et al., 2010, Schulz et al., 2010). Within each set of twenty sample members in a given age-sex stratum (45-54, 55-64, 65-74 M:F), individuals were randomly allocated, without replacement, to one of two possible orders of sphygmomanometer for blood pressure measurement (i.e. either Omron 705-CE then Omron HEM-907, or Omron HEM-907 then Omron 705-CE) in a 1:1 ratio, and to one of the two possible orders of spirometers for lung function measurement (i.e. Micro Plus then Easy on-PC, or Easy on-PC then Micro Plus) in a 1:1 ratio, and to one of 24 possible orders of the four dynamometers for grip strength (which can be called A, B, C and D for illustrative purposes) in random order with uniform distribution (i.e. ABCD, DCBA, ABDC, BACD and so on). A total of 120 data sheets were created providing the order of devices for each physiological function for each consecutive participant.

Participants received feedback on their results, advice to contact their GPs if their blood pressure was found to be elevated, and a £50 gift voucher which included costs for travel.

2.2.4 Measurement of blood pressure, grip strength and lung function

Standardised measurement protocols were implemented to control, to the extent possible, for potential sources of measurement error not related to device.

Blood pressure, for example, is affected by multiple factors (Jones et al., 2003) including the subject talking or actively listening, being exposed to cold, ingesting alcohol, having a distended bladder, recent smoking (Handler, 2009) and also to differences in measurement protocols such as arm position and cuff size (Bilo et al., 2017). Consequently, the study participant was provided with guidance before the assessment, and the assessment was conducted according to clear protocols. At the start of the assessment, they sat with legs uncrossed and their right arm resting comfortably, palm up, on a table, with the Omron positioned so that they could not see the display. The participant was asked to expose their right arm, making sure that rolled up sleeves did not restrict circulation and that any watches or bracelets had been removed. The sphygmomanometer cuff was positioned over the brachial artery with a large cuff made available where necessary. After three minutes of quiet rest, three readings were recorded using the first device, with a minute's rest between each reading. The device was then changed and, after a further two minutes of quiet rest, three readings were taken using the second device, again with a minute's rest between each reading. There was no talking until three readings on both devices had been completed. The same protocol was applied for both sphygmomanometers but the Omron HEM-907 provided greater automation, removing the need for the researcher to time the rest or record the result between each reading.

For grip strength, the values and precision of measurements have been shown to be influenced by a range of factors (Roberts et al., 2011, Sousa-Santos and Amaral, 2017) including whether allowance is made for hand size and hand-dominance (Incel et al., 2002), dynamometer handle shape (Amaral et al., 2012), position of the elbow (Balogun et al., 1991) and wrist during testing (O'Driscoll et al., 1992), setting of the dynamometer (Fess, 1981, Firrell and Crain, 1996), effort and encouragement, frequency of testing and time of day and training of the assessor (Fess, 1981, Roberts et al., 2011). In order to minimise these sources of measurement error, grip strength assessment was based on a published measurement protocol (Roberts et al, 2011). While seated in a chair with fixed arms, the participant was asked to place their forearm on the arm of the chair in the mid-prone position (the thumb facing up) with their wrist just over the end of the arm of the chair in a neutral but slightly extended position. Adjustments were made to each dynamometer to accommodate different hand sizes according to the make and model of the device. The dynamometer was held vertically and, on hearing the words “And Go”, the participant was encouraged, through strong verbal instructions, to squeeze as hard as possible for a few seconds until told to stop. For each device, two measurements were carried out in each hand in the sequence Left-Right-Left-Right. The value on the display was recorded to the nearest 0.1kg for the Jamar Plus+ Digital and Nottingham Electronic, to the nearest 0.5kg for the Smedley and to the nearest 1kg for the Jamar Hydraulic. The same protocol was applied to all dynamometers, including the Smedley, which is often assessed in a standing position. The researcher provided support where necessary.

For lung function, the accuracy of measurement relies primarily on optimal coaching: maximally deep breath, a rapid blast and appropriate encouragement as well as a full seal around the mouthpiece and correct body posture (Miller et al., 2005). Lung function measurements adhered to the American Thoracic Society/European Respiratory Society (ATS/ERS) lung function protocol (Miller et al., 2005). The procedure was explained and demonstrated, and the participant had a practice blow without completely emptying their lungs. The measurement was carried out with the participant standing unless they felt unable to do so. During measurement, maximum effort was encouraged verbally. In addition, the Easy on-PC was linked to a laptop with a cartoon representation of a child blowing up a balloon. This represents a real-time trace and as the participant is encouraged to exhale until the balloon pops, this helps ensure a maximal FVC is achieved. After each trial, the researcher recorded whether it satisfied the protocol, for example disqualifying blows if the participant did not form a tight seal around the mouthpiece or coughed during the procedure. In these instances, feedback was provided before the next attempt. Three valid measurements of lung function were obtained from each spirometer with participants having up to a total of five attempts to achieve these. The same protocol was applied for both

spirometers but with the Easy on-PC, the classification of blows and the feedback given was guided by the computer software.

Readings for blood pressure, grip strength and lung function using the Micro Medical spirometer were data entered twice, independently, and compared to ensure accuracy. Lung function readings taken using the Easy on-PC by NDD spirometer were downloaded directly from the laptop.

2.2.5 Other measures

Height was measured using a portable Marsden Leicester stadiometer and weight was measured using Tanita 352 scales according to standardised procedures, from which body mass index (BMI) was calculated as weight (in kg) divided by height (in m^2). Participants completed a two-page questionnaire shown in Figure 7. This gathered additional information on age, age at completion of full-time education, self-rated health, smoking history, medication use, and musculoskeletal, cardiovascular, and respiratory conditions which might influence performance on the functional tests.

These data were manually entered once with a sample of cases checked for quality purposes. In all cases, the dataset was successfully linked to a record of the randomised order of device administration and to field notes, including deviations from protocol such as the stated order. Pseudonymised datasets were used for analysis. The data set was anonymised, and a record of the linkage preserved by the NSHD Research Study Manager so that participants information could be used to link their data should that be required, for example if a participant requested information about their results.

Figure 7 Self-completion questionnaire completed during assessment

ID: ☐☐☐

MRC Unit for Lifelong Health and Ageing

UCL

CLOSER Calibration Study Questionnaire

This questionnaire provides us with information that we need in order to be able to interpret the results of your tests correctly. Please tick the appropriate box or give further details in the space provided. Please ask the nurse if you are unsure about anything.

All information you give us will be treated in the strictest confidence.

Age: _____ years Male/female

1. At what age did you finish your continuous full-time education at school or college?
   - Never went to school ☐
   - 14 or under ☐
   - 15 ☐
   - 16 ☐
   - 17 ☐
   - 18 ☐
   - 19 or over ☐

2. How is your health in general?
   - Excellent ☐
   - Very good ☐
   - Good ☐
   - Fair ☐
   - Poor ☐

3. Have you ever been told by a doctor that you have had:

a) a heart attack (myocardial infarction)?
   - No ☐
   - Yes ☐

b) angina?
   - No ☐
   - Yes ☐

c) other type of heart disease?
   - No ☐
   - Yes ☐

4. Have you ever been told by a doctor that you have hypertension or high blood pressure?
   - No ☐
   - Yes ☐

If **yes**, are you on any medication for hypertension/high blood pressure?
   - No ☐
   - Yes ☐

5. Do you have arthritis or other musculoskeletal conditions that affects your hands?
   - No ☐
   - Yes ☐

ID: ☐☐☐

6. Do you have difficulty because of long-term health problems holding something heavy like a full kettle or removing a stiff lid from a jar?
   - No ☐
   - Some difficulty ☐
   - A lot of difficulty ☐

7. Have you ever had:

a) eczema?
   - No ☐
   - Yes ☐

b) hayfever?
   - No ☐
   - Yes ☐

c) asthma?
   - No ☐
   - Yes ☐

d) chronic obstructive pulmonary disease (COPD), chronic bronchitis or emphysema?
   - No ☐
   - Yes ☐

e) Other respiratory problems
   - No ☐
   - Yes ☐

If **yes**, please say what these were _____

8. Do you currently take medication for any respiratory disease (such as asthma or chronic obstructive pulmonary disease)
   - No ☐
   - Yes ☐

If **yes**, what medication(s) do you take? Please specify _____

9. Do you currently smoke cigarettes?
   - No ☐
   - Yes ☐

If **yes**, how many cigarettes a day do you usually smoke? _____ cigarettes per day

At what age did you start smoking? _____ years

If **no**, please answer question 10

10. Have you ever smoked cigarettes?
   - No ☐
   - Yes ☐

If **yes**, how many cigarettes a day did you usually smoke? _____ cigarettes per day

At what age did you start smoking? _____ years

At what age did you stop smoking? _____ years

Thank you for completing the questionnaire

2.2.6 Primary outcome measures

Following standard practice in many epidemiological studies (Enarson et al., 2004, Miller et al., 2005, Powers et al., 2011, Roberts et al., 2011), the analysis was based on summary measures rather than analysis of all individual readings. For the primary analysis of blood pressure, the

mean of the second and third readings of SBP and DBP in mmHg were used. The first reading is commonly discarded on the basis that it tends to be inaccurate, though this approach is disputed (Salazar et al., 2015). For grip strength the maximum of the four readings in kilograms was used on the basis that this should be closest to the latent or underlying value. For lung function, following precise ATS/ERS criteria, the quality of readings were classified as quality A, cases where there were three or more readings and the highest two were within 100ml, as quality B where there were three or more readings and the highest were within 150ml, as quality C where there were two or more readings and the highest were within 200ml and as quality D where there were two readings but they were not within 200ml or there was only one reading. Any other cases were classified as E. For the primary analysis, the maximum FEV_1 and FVC from the highest quality readings (quality A or B) were used where the difference between the highest two readings was within 150 ml (Miller, 2005). Some participants were excluded from analyses due to missing readings (n=3 for blood pressure and n=12 for lung function). In addition, for lung function, just under a third (n=32 for FEV_1 and n=39 for FVC) of the remaining participants were excluded from the primary analysis because there were no readings of a sufficiently high quality.

2.3 Statistical methodology

2.3.1 Primary analysis

The composition of the achieved sample (Table 4) and the balance between groups was assessed for chance bias (Table 6). This involved a visual comparison of the sample characteristics by randomisation group, first for general measures (e.g., age, height, weight, BMI, age left education and self-reported health) and then for measures relevant to each physiological function (i.e., cardiovascular, musculoskeletal, and respiratory health). For blood pressure and lung function this involved comparing two randomisation groups – those who were assessed on one or other of the devices first. For grip strength, rather than compare 24 groups showing all possible combinations of device order, four groups were compared based on the dynamometer that was used first. The distributions are not expected to be perfectly balanced since the allocation is randomised, but any extreme imbalances would suggest that the randomisation process might have been incorrectly administered (Altman, 1985, NICE, 2013, Roberts and Torgerson, 1999).

As a next step, the reliability of each device was calculated. (Rabe-Hesketh and Skrondal, 2012). The measure of reliability ρ (rho) can be represented as follows:

$$\rho = \frac{Var\ (\zeta_j)}{Var\ (y_{ij})} = \frac{\psi}{\psi + \theta}$$

where

ρ (rho) is the between-subject variance (depicted as ψ) as a proportion of the total variance ($\psi + \theta$), that is, the sum of the between-subject variance (ψ) and the within-subject variance (θ), and where ζ_j (or zeta specific to each subject *j)* and y_{ij} is the response of unit *i* (here, a reading) in cluster *j* (here, an individual).

If the value of θ (which can also be seen as the level 1 residual) is zero then the reliability of the instrument will be 1, and as the value of θ increases, so the reliability of the instrument will be lower (Rabe-Hesketh and Skrondal, 2012). This aspect of the statistical methodology focuses on measurement variance within each device.

The statistical methods presented in the rest of this section focus on the bias that may result from systematic measurement differences between devices. First, to investigate order effects, that is to assess whether there may be confounding of practice effects with measurement effects, unpaired t-tests were used to compare the difference between the mean values of pairs of devices in one sequence (e.g., AB) compared to the opposite (e.g., BA). This shows whether there were larger learning effects when devices were ordered in one way than the other, which were therefore not compensated for by the random ordering of the devices. For grip strength, pairwise comparisons were made based on whether the measurement on one device was before or after another device, treating all possible alternatives as equivalent (e.g., comparing ABxx,AxBx,AxxB,xABx,xAxB,xxAB with Baxx,BxAx,BxxA,xBAx,xBxA,xxBA).

The difference in measurement was then calculated between pairs of devices and examined these visually using box plots. The mean, within-person difference between pairs of devices was then assessed using paired t-tests where the statistic value t can be calculated as the sum of the differences of each pair ($\sum d$), divided by the square root of n times the sum of the differences squared minus the sum of the squared differences, that is $n(\sum d^2) - (\sum d)^2$, divided by $n - 1$. This can be expressed as:

$$t = \frac{\sum d}{\sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n-1}}}$$

The assumption that the mean differences are normally distributed was checked using histograms, and then Bland and Altman plots were used to assess whether the mean differences observed varies depending on the magnitude of the measurements (Bland and Altman, 1986, Bland and Altman, 1999, Chhapola et al., 2015, Giavarina, 2015). Bland and Altman plots show the paired measurements on the y-axis and the average of the two instruments on the x-axis (Bland and Altman, 1986). As well as showing the mean difference in values between the two devices, Bland and Altman plots also show the 95% limits of agreement (LOA) (Chhapola et al., 2015). They

are commonly used for equipment comparison studies because they provide a clear visual illustration of the differences between devices – with bias indicated by deviation of the horizontal line from zero – and also show whether this remains constant as the magnitude of the measurements increase, which is indicated by whether there is a slope. Bland and Altman plots are used here rather than the more common Pearson correlation coefficient which is used to assess agreement between two measurements. However this is not a suitable measure since a high correlation coefficient does not necessarily show that there is good agreement between two measures which are being compared (such as measures taken using different devices); indeed data which have poor agreement can have a high correlation (Doğan, 2018).

2.3.2 Sensitivity analysis: using alternative outcome measures

To test the robustness of the main findings, a series of sensitivity analyses were performed, having

(i) excluded cases where fieldwork notes indicated that the device had been administered in the incorrect order (n=2 for blood pressure, n=5 for grip strength and n=1 for lung function).

(ii) removed extreme outliers (n=1 for blood pressure and n=2 for grip strength). These were initially identified visually using scatter plots. Cases which lay outside the upper and lower limits of a series of Box Plots (Schwertman et al., 2004) were then assessed for plausibility based on a detailed investigation of the data sheets which included researcher notes. Following review by the investigator team, a small number of cases which may have been a data collection or data recording error were treated as potential outliers; and

(iii) used alternative outcome definitions for each physiological function commonly used in analyses:

   a. for blood pressure this was the average of the three readings (Chobanian et al., 2003, Powers et al., 2011, Salazar et al., 2015) and using the second reading only (Hardy et al., 2004).
   b. For grip strength, this was the mean of the four readings (Massy-Westropp et al., 2011, Mathiowetz, 2002, Sousa-Santos and Amaral, 2017).
   c. For lung function, this was the highest of each, FEV_1 and FVC, from all available readings irrespective of whether they adhered to the ATS/ERS quality criteria. This involved retaining Easy on-PC readings of quality C to E which were excluded from the main analysis, only dropping the 12 cases where there were no valid readings.

2.3.3 Sensitivity analysis using an alternative multilevel modelling approach

Finally, multilevel modelling was carried out as an alternative statistical approach to estimate the difference between devices. This approach has several advantages in comparison with the approach taken for the primary analysis, which used Bland and Altman plots and comparison of means. The first is that it uses all readings rather than aggregated measures such as means and maximums, to account for variance between readings. Secondly, these models can include explanatory variables such as order of device which are observed but not accounted for in the preliminary analysis, as well as covariates such as age, sex and BMI which the preliminary analysis assumes are balanced due to randomisation. A third is that it makes it possible to explore the random structure of the models, though given constraints of time, this was only fully explored for one of the three measures in this study, blood pressure.

The multilevel modelling used the STATA 15.0 estimation command `mixed` where readings (level 1) were clustered within study participants (level 2) to control for the non-independence of measurements from the same person, with device included as a fixed effect. This can be represented statistically as a generalised linear random intercept model with a continuous dependent variable which is, in turn, the measure of blood pressure, grip strength or lung function. This can be represented by the formula:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

$$u_j \sim N(0, \sigma_u^2) \text{ and } e_{ij} \sim N(0, \sigma_e^2)$$

where i is the number of observations (in this case readings), the level 1 units,

j is the number of groups (in this case study participants), the level 2 units,

y_{ij} is the response for observation i (reading) in group j (study participant)

x_{ij} is an individual level covariate (In the baseline model this includes order of device, and in subsequent models this included a vector of covariates. Further information on the various models and covariates included is below.)

u_j is the level 2 residual, assumed to follow a normal distribution with variance σ_u^2 and

e_{ij} is the level 1 residual, assumed to follow a normal distribution with mean σ_e^2.

Careful model building was carried out. The baseline model, Model 1, includes only the variable device. Model 2 includes, in addition to device, covariates that account for the order in which the

devices were administered and the position of the reading in the sequence (1 to 3 for blood pressure, 1 or 2 for the dominant and non-dominant hands for grip strength, and 1 to 5 for lung function). Model 3 additionally includes covariates which reflect the characteristics of respondents in terms of age, sex and, for blood pressure only, BMI. The results of the estimated models (Model 1-3) are described in detail in Section 2.4.4 and Section 2.4.5.

To make full use of the potential benefits of multilevel modelling, the multilevel analysis of blood pressure was extended to include Model 4-6. For Model 4, a two-level model was built, with readings (level 1) nested within individuals (level 2). Unlike Model 1, this did not include device as a fixed effect, and it did not include any other covariates. Model 5 additionally included device as a random effect; this was a three-level model which accounted for clustering of readings (level 1), within device (level 2) and within individual (level 3). Again, this model did not include covariates. This made it possible to report on the partitioning of variance taking account of device. Finally, Model 6 additionally included researcher (the equivalent of survey or interviewer) as a random effect. This was a four-level model which accounted for clustering of readings (level 1), within device (level 2), within individual (level 3) and within researcher (level 4). Since exploration of the random structure was not the main focus of the analysis, this additional analysis is not reported for all measures.

Data cleaning and management were carried out using Excel, IBM-SPSS Version 22 and analysis was carried out using STATA 15.0.

2.4 Results

2.4.1 Analysis sample

During fieldwork 118 assessments were completed, with between 18 and 21 participants in each of the age-sex strata. Of the seven researchers, three carried out 20-30 assessments, two carried out 10-20 assessments and two carried out fewer than ten assessments.

Table 4 Achieved sample of individuals by age group and gender (n=118)

	Age group (in years)		
	45-54	55-64	65-74
Men	18	20	21
Women	20	19	20

Some participants were excluded from analyses due to missing readings (n=3 for blood pressure and n=12 for lung function). In addition, for lung function, just under a third (n=32 for FEV_1 and n=39 for FVC) of the remaining participants were excluded from the primary analysis because there were no readings of sufficiently high quality. For the other functions, there were only a

small number of outliers judged to be potentially implausible based on scatter plots (Appendix A), box plots (Appendix B) and case by case review using data sheets and notes (n=1 for blood pressure and n=2 for grip strength). These cases were nevertheless included in the primary analysis because there was no clear evidence that they were invalid. Moreover, in a small number of cases, devices had been administered in the incorrect order but were included in the primary analysis (n=2 for blood pressure, n=5 for grip strength and n=1 for lung function).

A visual assessment of the socio-demographic characteristics of the randomised groups suggests that they were reasonably well balanced in baseline characteristics and in key aspects of cardiovascular, musculoskeletal, and respiratory health (Table 6). The absence of any extreme imbalances provides reassurance that the random allocation was implemented successfully.

The reliability of all the devices included in the experiment was very good, ranging from 0.87 to 0.99 (Table 5).

Table 5 Reliability of the devices included in the experiment

Blood pressure	**SBP**	**DBP**
Omron 705-CE	0.90	0.89
Omron HEM-907	0.91	0.94
Grip strength	**Dominant hand**	**Non-dominant**
Nottingham	0.96	0.92
Jamar Plus+ Digital	0.95	0.93
Jamar Hydraulic	0.96	0.95
Smedley	0.92	0.87
Lung function	**FEV_1**	**FVC**
Micro Plus - A&B readings	0.97	0.98
Micro Plus - All readings	0.99	0.98
Easy on-PC - A&B readings	0.96	0.97
Easy on-PC - All readings	0.97	0.98

In experimental studies, where sample members participate in several treatments, differences may result from the order in which treatments are presented. Performance may improve because of practice or learning effects due to repetition of the task or may worsen due to fatigue effects because of boredom or tiredness. To avoid systematic order effects, the order of treatments is randomised. In this study, there was no evidence of order effects for blood pressure or for lung function (Table 7). For grip strength, there was some evidence of an order effect between the Nottingham Electronic and Smedley (difference= -3.08kg, 95% CI=-5.93, -0.23, p=0.03).

Table 6 Characteristics of the randomised group by order of device

	Blood pressure				Grip strength								Lung function			
	Omron 705-CP first (n=58)[1]		Omron HEM-907 first (n=60)		Jamar Hydraulic first (n=30)		Smedley first (n=28)		Nottingham Electronic first (n=30)		Jamar Plus+ Digital first (n=30)[1]		Micro Plus first (n=59)		Easy on-PC first (n=59)[1]	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age (years)	59.4	8.2	59.8	7.8	58.5	8.2	59.8	7.4	59.0	9.0	61.2	7.4	59.8	7.8	59.4	8.2
Weight (kg)	76.9	21.1	77.3	16.7	73.8	16.1	82.1	22.2	77.3	17.6	75.5	19.4	77.3	16.5	76.9	21.1
Height (cm)	168.5	9.0	167.6	8.9	166.5	8.2	170.2	9.3	165.9	9.9	169.6	7.9	168.2	9.6	168.5	9.0
BMI (kg/m^2)	27.5	4.6	27.4	4.9	26.5	4.7	28.2	6.0	27.8	4.6	27.3	3.6	27.2	4.5	27.5	4.6
	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
Men	29	50.0	30	50.0	14	46.7	15	53.6	14	46.7	16	53.3	30	50.9	29	50.0
Women	29	50.0	30	50.0	16	53.3	13	46.4	16	53.3	14	46.7	29	49.2	29	50.0
Age left full-time education	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
14	0	0.0	1	1.7	0	0.0	0	0.0	0	0.0	1	3.3	0	0.0	0	0.0
15	6	10.3	7	11.7	0	0.0	4	14.3	6	20.0	3	10.0	8	13.6	6	10.3
16	14	24.1	11	18.3	6	20.0	4	14.3	5	16.7	10	33.3	16	27.1	14	24.1
17	4	6.9	4	6.7	3	10.0	1	3.6	1	3.3	3	10.0	3	5.1	4	6.9
18	3	5.2	6	10.0	2	6.7	5	17.9	2	6.7	0	0.0	2	3.4	3	5.2
19 or over	31	53.5	31	51.7	19	63.3	14	50.0	16	53.3	13	43.3	30	50.9	31	53.5
Self-reported health is	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
…excellent	13	22.4	10	16.7	8	26.7	6	21.4	5	16.7	4	13.3	9	15.3	13	22.4
…very good	19	32.8	22	36.7	9	30.0	9	32.1	13	43.3	10	33.3	23	39.0	19	32.8
…good	20	34.5	22	36.7	7	23.3	12	42.9	11	36.7	12	40.0	20	33.9	20	34.5
…poor	6	10.3	5	8.3	6	20.0	1	3.6	1	3.3	3	10.0	7	11.9	6	10.3
…very poor	0	0.0	1	1.7	0	0.0	0	0.0	0	0.0	1	3.3	0	0.0	0	0.0
Cardiovascular health	Number	%	Number	%												
Doctor diagnosed condition[2]	4 (3)	6.9	9 (3)	15												
Diagnosed high blood pressure	18 (1)	31.0	19 (0)	31.7												
On medication for blood pressure	14 (0)	24.1	17 (0)	28.3												
Musculoskeletal health					Number	%	Number	%	Number	%	Number	%				
Dominant hand (right)					29	96.7	25	89.3	27	90.0	27	90.0				
Arthritis					6	20.0	5	17.9	4	13.3	5	16.7				
Some/lot of difficulty gripping					5	16.7	8	28.6	6	20.0	5	16.7				
Respiratory health													Number	%	Number	%
Ever had….Eczema													7 (-)	11.9	12 (-)	20.3
…Hay fever													18 (1)	30.5	22 (-)	37.3
…Asthma													7 (3)	11.9	6 (-)	10.2
…COPD, chronic bronc., emphysema													4 (2)	6.8	2 (-)	3.4
…Other respiratory problems													4 (3)	6.8	2 (1)	3.4
Taking medication for condition													4 (0)	6.8	2 (0)	3.4
Currently smokes cigarettes													13 (0)	22.0	8 (0)	13.6
Ever smoked cigarettes													21 (0)	35.6	27 (0)	45.8

Notes: Brackets by health conditions show number of missing cases because self-completion was incomplete. (1) One case had missing BMI (2) Includes heart attack, angina, and other heart conditions.

Table 7 Assessment of order effects for all measures

Measure	Device (n)		Independent t-test			95% CI	
Device and order (A - B)	A	B	Diff	SE	p-value	Lower	Upper
SBP, mmHg (Mean of 2nd+3rd)							
Omron 705 - Omron 907	56	59	-2.32	1.37	0.109	-5.03	0.38
DBP, mmHg (Mean of 2nd+3rd)							
Omron 705 - Omron 907	56	59	-0.16	1.06	0.884	-2.26	1.95
Grip strength, kg (max of 4)							
Jamar Hydraulic - Smedley	59	59	-1.18	1.03	0.257	-3.23	0.87
Nottingham - Jamar Plus+	58	60	-1.92	1.16	0.099	-4.21	0.37
Jamar Plus+ - Jamar Hydraulic	60	58	-1.18	0.60	0.052	-2.37	0.01
Jamar Plus+ - Smedley	59	59	0.03	1.03	0.980	-2.02	2.07
Nottingham - Jamar Hydraulic	57	61	-1.78	1.18	0.133	-4.11	0.55
Nottingham – Smedley	60	58	-3.08	1.44	**0.034**	-5.93	-0.23
FEV_1, litres (ATS/ERS criteria)							
Easy on-PC - Micro Plus	39	35	-0.02	0.03	0.534	-0.08	0.04
FVC, litres (ATS/ERS criteria)							
Easy on-PC - Micro Plus	35	32	-0.06	0.06	0.256	-0.18	0.05

Diff=Difference; SE=standard error; CI=confidence interval

Before assessing average within-person differences between pairs of devices using paired t-tests, the difference in measurement between pairs of devices were visually assessed using box plots (Appendix B) and, using histograms of mean differences (Appendix C), it was established that the assumption that the mean differences are normally distributed was reasonable.

2.4.1 Blood pressure

The mean difference between the Omron HEM-907 and Omron 705-CP for SBP was 3.86 mmHg and for DBP was 1.35 mmHg, with the Omron HEM-907 measuring higher than the older Omron, as shown in Table 8.

Table 8 Comparison of means using paired t-tests; blood pressure

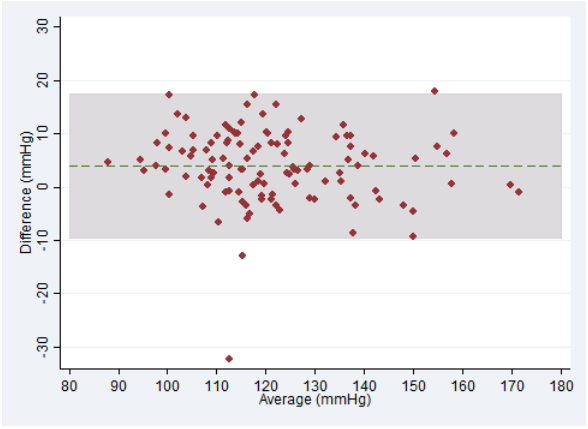
Blood pressure, mean of 2+3 (mmHg)	N	Paired t-test			95% CI		Limits of agreement	
		Mean diff	SE	p-value	Lower	Upper	Lower	Upper
SBP, Omron HEM-907 - Omron 705-CP	115	3.86	0.69	**<0.001**	2.50	5.22	-10.60	18.32
DBP, Omron HEM-907 - Omron 705-CP	115	1.35	0.53	**0.012**	0.30	2.39	-9.76	12.45

Diff=Difference; SE=standard error; CI=confidence interval
Comparison is of two sphygmomanometers, the Omron HEM-907 and Omron 705-CP

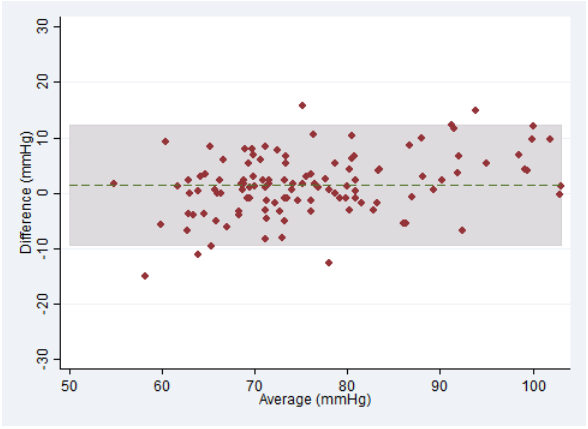
The Bland and Altman plots show that as the magnitude of the measurement increases, the mean difference between the two devices remains approximately constant (Figure 8). The limits of agreement were -10.60 and 18.32 mmHg for SBP and -9.76 and 12.45 mmHg for DBP (Table 8).

Figure 8 Bland and Altman plots: Blood pressure

(a) SBP Omron HEM-907: Omron 705-CP



(b) DBP Omron HEM-907: Omron 705-CP



2.4.2 Grip strength

Initial exploration of the four grip strength measures using ANOVA showed significant differences between devices. Pairwise analysis was then conducted. There was no evidence of differences when comparing two pairs of devices: first, the two electronic dynamometers, the Nottingham Electronic and Jamar Plus+ Digital (mean difference=0.29kg, 95% CI: -0.87, 1.44, p=0.623); and second, the hydraulic and spring-gauge dynamometers, the Jamar Hydraulic and Smedley (mean difference=0.23kg, 95% CI: -0.79, 1.26, p=0.654). However, on average there were differences of between 4 and 5kg between measurements from the four other combinations of devices comparing an electronic dynamometer with a hydraulic or spring-gauge dynamometer (Table 9).

Table 9 Comparison of means using paired t-tests; grip strength

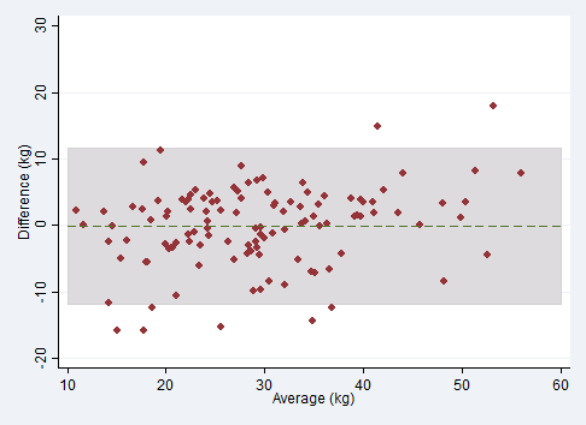
Grip, max of 4 readings (kg)	N	Paired t-test			95% CI		Limits of agreement	
		Mean Diff	SE	p-value	Lower	Upper	Lower	Upper
Jamar Hydraulic – Smedley	118	0.23	0.52	0.654	-0.79	1.26	-10.80	11.26
Nottingham Electronic- Jamar Plus+ Digital	118	0.29	0.58	0.623	-0.87	1.44	-12.11	12.69
Jamar Plus+ Digital - Jamar Hydraulic	118	4.45	0.30	**<0.001**	3.85	5.05	-2.02	10.92
Jamar Plus+ Digital – Smedley	118	4.68	0.52	**<0.001**	3.66	5.70	-6.28	15.65
Nottingham - Jamar Hydraulic	118	4.74	0.59	**<0.001**	3.57	5.91	-7.85	17.32
Nottingham – Smedley	118	4.97	0.73	**<0.001**	3.52	6.41	-10.56	20.50

Notes: Diff=Difference; SE=standard error; CI=confidence interval
Comparison is of four dynamometers; Jamar Hydraulic, Jamar Plus+, Nottingham Electronic, Smedley

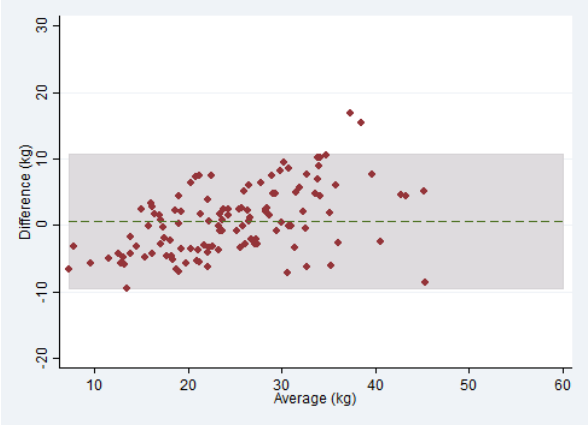
As shown in the table, the limits of agreement vary widely; for example, from between -2.02 and 10.12 kg for the pairing of the Jamar Plus+ and Jamar Hydraulic, to between -10.56 and 20.50 kg for the Nottingham Electronic and Smedley. The Bland and Altman plots (Figure 9) show that for three of the six pairings, the difference between the two devices appears approximately constant as the magnitude of the measurement increases. However, the three pairings which include the Smedley dynamometer show an increase in the difference between devices at higher magnitudes of mean grip strength (Figure 9: b, d, and f).

Figure 9 Bland and Altman plots: Grip strength

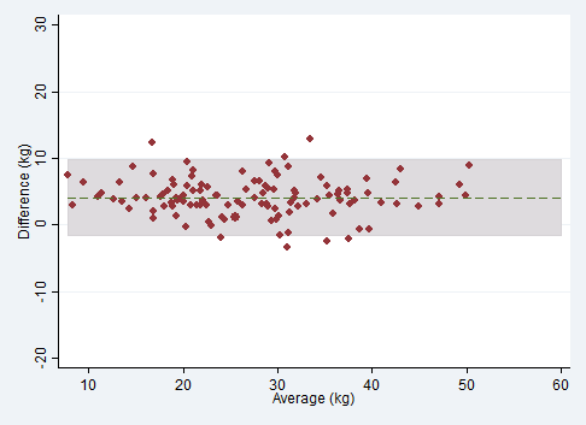
(a) Jamar Plus+ Digital: Nottingham



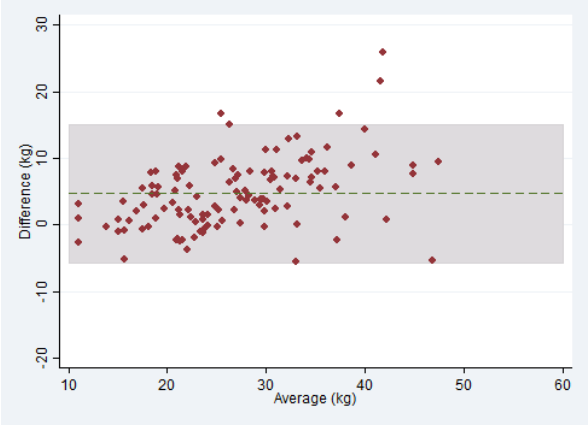
(b) Jamar Hydraulic: Smedley



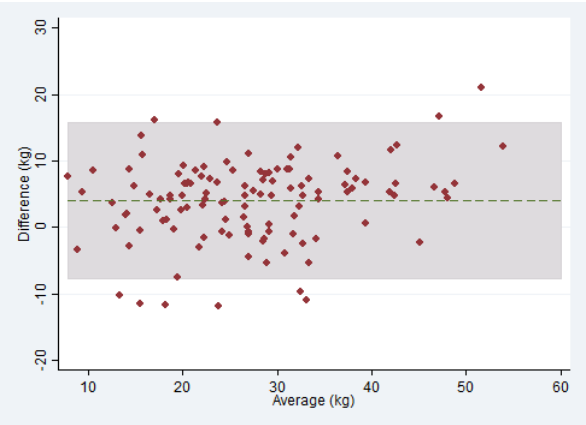
(c) Jamar Plus+ Digital: Jamar Hydraulic



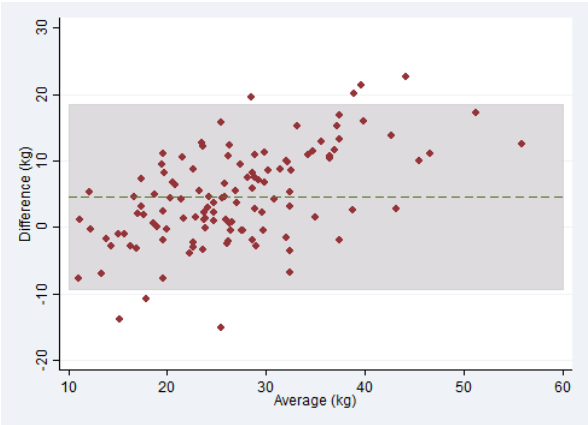
(d) Jamar Plus+ Digital: Smedley



(e) Nottingham: Jamar Hydraulic



(f) Nottingham: Smedley



2.4.3 Lung function

There was no evidence of a difference between devices in mean measures of FEV_1 (mean difference=0.00 litres, 95% CI: -0.03, 0.03, p=0.9)) but there were in FVC (mean difference=-0.47

litres, 95% CI:-0.53, -0.42, p<0.001) with the Easy on-PC measuring higher than the Micro Plus (Table 10).

Table 10 Comparison of means using paired t-tests; lung function

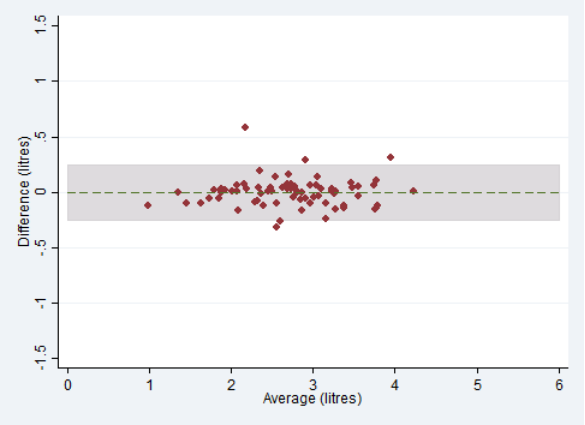
Lung function, maximum (litres), ATS/ERS criteria	N	Paired t-test			95% CI		Limits of agreement	
		Mean diff	SE	p-value	Lower	Upper	Lower	Upper
FEV_1, Micro Plus-Easy on-PC	74	0.00	0.01	0.880	-0.03	0.03	-0.25	0.25
FVC, Micro Plus- Easy on-PC	67	-0.47	0.03	**<0.001**	-0.53	-0.42	-0.92	-0.03

Notes: Diff=Difference; SE=standard error; CI=confidence interval
Comparison is of two spirometers to measure lung function; Micro Plus and Easy on-PC

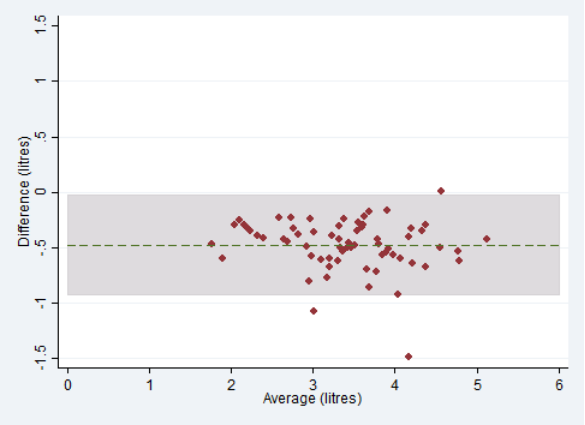
The Bland and Altman plots suggest that for FEV_1 as the average of each measure increases, the difference between the two devices remains approximately constant and close to zero (Figure 10a) with the limits of agreement between -0.25 and 0.25 litres. For FVC, the Bland and Altman plot also appears reasonably consistent as values of FVC increase (Figure 10b) but the limits of agreement are larger (-0.92 and -0.03).

Figure 10 Bland and Altman plots: Lung function

(a) Micro Plus: Easy on-PC, FEV_1



(b) Micro Plus: Easy on-PC, FVC



2.4.4 Results of the sensitivity analysis based on alternative measures

The following describes the results of the sensitivity analysis described in Section 2.3.2. There was evidence of a small number of additional order effects when wrongly ordered or outlying cases were removed and when alternative definitions were used, but the observed changes were small (Table 11). For example, for blood pressure, there was a small increase in the size of the observed order

effects for SBP which then appeared to be statistically significant at p<0.05, both when an outlier was removed (difference=-2.93, 95% CI=-5.36,-0.49, p=0.019) and when the mean of three readings was used (difference=-2.74, 95% CI=-5.25,-0.24, p=0.032). For grip strength, this is the case for the pairing of the Nottingham Electronic and Jamar Hydraulic when the mean of four readings was used (difference=-2.49, 95% CI=-4.64, 0.34, p=0.024) and for the Jamar Plus+ and Jamar Hydraulic pairing when the mean of the four readings was used (difference=-1.22, 95% CI=-2.25,-0.18, p=0.021) or when mis-ordered cases (difference=-1.24, 95% CI=-2.44,-0.03, p=0.044) or outliers were excluded (difference=-1.43, 95% CI=-2.52,-0.34, p=0.011).As stated earlier, these observed changes are small and do not materially impact on the results.

Table 11 Sensitivity analysis for order effects for all measures

	Device (n)		Independent t-test			95% CI	
	1st	2nd	Diff	SE	p-value	Lower	Upper
Blood pressure, mm Hg							
SBP: Omron 705 - Omron 907	705	907					
Primary, excluding mis-ordered	55	58	-1.91	1.36	0.162	-4.60	0.78
Primary, excluding outliers	56	58	-2.93	1.23	**0.019**	-5.36	-0.49
Mean of 3 readings	56	59	-2.74	1.27	**0.032**	-5.25	-0.24
Reading 2 only	56	59	-2.48	1.68	0.144	-5.82	0.86
DBP: Omron 705 - Omron 907	705	907					
Primary, excluding mis-ordered	55	58	0.27	1.02	0.791	-1.76	2.30
Primary, excluding outliers	56	58	-0.49	1.01	0.629	-2.50	1.52
Mean of 3 readings	56	59	-1.49	1.03	0.150	-3.53	0.55
Reading 2 only	56	59	-1.62	1.34	0.227	-4.27	1.02

Table 11 continued…	Device (n)		Independent t-test			95% CI	
	1st	2nd	Diff	SE	p-value	Lower	Upper
Grip strength, kg							
Jamar Hydraulic - Smedley	JH	SM					
Primary, excluding outliers	59	57	-0.59	0.96	0.539	-2.50	1.31
Mean of 4 readings	59	59	-1.14	0.94	0.229	-3.01	0.73
Nottingham - Jamar Plus+	NO	JP					
Primary, excluding mis-ordered	57	60	-1.86	1.16	0.113	-4.17	0.45
Mean of 4 readings	58	60	-2.15	1.09	0.051	-4.31	0.01
Jamar Plus+ - Jamar Hydraulic	JP	JH					
Primary, excluding mis-ordered	59	57	-1.24	0.61	**0.044**	-2.44	-0.03
Primary, excluding outliers	60	57	-1.43	0.55	**0.011**	-2.52	-0.34
Mean of 4 readings	60	58	-1.22	0.52	**0.021**	-2.25	-0.18
Jamar Plus+ - Smedley	JP	SM					
Primary, excluding outliers	59	59	0.03	1.03	0.980	-2.02	2.07
Mean of 4 readings	58	59	-0.38	0.96	0.692	-2.28	1.52
Nottingham - Jamar Hydraulic	NO	JH					
Primary, excluding mis-ordered	57	60	-2.09	1.14	0.069	-4.35	0.17
Primary, excluding outliers	57	60	-2.00	1.16	0.089	-4.30	0.31
Mean of 4 readings	57	61	-2.49	1.09	**0.024**	-4.64	-0.34
Nottingham – Smedley	NO	SM					
Primary, excluding mis-ordered	60	57	-3.00	1.45	**0.040**	-5.87	-0.14
Primary, excluding outliers	59	58	-3.45	1.40	**0.015**	-6.22	-0.68
Mean of 4 readings	60	58	-3.02	1.28	**0.020**	-5.55	-0.48
Lung function, litres							
FEV_1: Easy on-PC - Micro Plus	MM	Easy					
Primary, excluding mis-ordered	39	34	-0.02	0.02	0.528	-0.08	0.04
All cases, including C-E	54	52	-0.01	0.04	0.795	-0.08	0.06
FVC: Easy on-PC - Micro Plus	MM	Easy					
Primary, excluding mis-ordered	35	31	-0.06	0.06	0.287	-0.17	0.05
All cases, including C-E	54	52	-0.11	0.05	0.053	-0.22	0.00

Diff=Difference; SE=standard error; CI=confidence interval
Comparison is of two sphygmomanometers: Omron 705 (705) and Omron 907 (907); four dynamometers; Jamar Hydraulic (JH), Jamar Plus+ (JP), Nottingham Electronic (NO), Smedley (SM); two spirometers Easy on-PC (Easy) and Micro Medical Plus (MM).

Similarly, the estimates of differences between devices changed by small amounts when mis-ordered cases or outliers were excluded and when alternative summary measures were used, but the conclusions were unaltered (Table 12). For blood pressure, for example, the difference between sphygmomanometers increased fractionally in some instances. For grip strength, the mean differences between devices are broadly consistent whether the maximum or mean of the readings is used and whether wrongly ordered cases or outliers are excluded, but the magnitude of the

difference varies depending on the definition used. The key finding is that the estimates of difference all fall in the 4-5kg range, with minor differences as follows: for the Jamar Plus+ and Jamar Hydraulic the estimates of difference fall between devices range from 4.13-4.57kg; for the Jamar Plus+ and Smedley from 4.48-4.76kg; for the Nottingham and Jamar Hydraulic from 4.01-4.89kg; and for the Nottingham and Smedley from 4.64-4.92kg. The finding that there is no statistically significant difference between the Jamar Hydraulic and Smedley and between the Nottingham Electronic and Jamar Plus+ holds when outliers or mis-ordered cases are excluded, and when alternative definitions are used. The difference between the Jamar Hydraulic and Smedley based on the mean of four readings increases to 0.63kg but remains statistically non-significant (95% CI: -0.30,1.57, p=0.183). For lung function, the findings do not change when the single case administered in the wrong order was excluded. The estimates of differences between the devices only shift fractionally when lower quality readings were also included in the analysis.

Table 12 Sensitivity analysis for difference of means for all measures

		Paired t-test			95% CI		Limits of agreement	
	N	Mean diff	SE	p-value	Lower	Upper	Lower	Upper
Blood pressure, mm Hg								
SBP: Omron 907 - Omron 705								
Primary, excluding mis-ordered	113	3.85	0.69	**<0.001**	2.50	5.20	-10.35	18.06
Primary, excluding outliers	114	4.16	0.63	**<0.001**	2.92	5.40	-8.95	17.26
Mean of 3 readings	115	3.92	0.65	**<0.001**	2.65	5.19	-9.59	17.43
Reading 2 only	115	3.13	0.86	**<0.001**	1.45	4.81	-14.66	20.92
DBP: Omron 907 - Omron 705								
Primary, excluding mis-ordered	113	1.46	0.52	**0.005**	0.45	2.47	-9.16	12.08
Primary, excluding outliers	114	1.52	0.51	**0.003**	0.52	2.52	-9.05	12.08
Mean of 3 readings	115	1.49	0.52	**0.005**	0.47	2.51	-9.05	12.08
Reading 2 only	115	1.55	0.68	**<0.001**	0.22	2.87	-12.52	15.62

Diff=Difference; SE=standard error; CI=confidence interval
Comparison is of two sphygmomanometers, the Omron HEM-907 and Omron 705-CP

Table 12 Sensitivity analysis for difference of means for all measures (continued…)

		Paired t-test			95% CI		Limits of agreement	
	N	Mean diff	SE	p-value	Lower	Upper	Lower	Upper
Grip strength, kg								
Jamar Hydraulic - Smedley								
Primary, excluding outliers	116	-0.06	0.49	0.893	-1.02	0.89	-10.20	10.07
Mean of 4 readings	118	0.63	0.48	0.183	-0.30	1.57	-9.43	10.70
Nottingham - Jamar Plus+								
Primary, excluding mis-ordered	117	0.33	0.59	0.579	-0.83	1.49	-12.10	12.75
Mean of 4 readings	118	-0.12	0.56	0.832	-1.21	0.98	-11.86	11.62
Jamar Plus+ - Jamar Hydraulic								
Primary, excluding mis-ordered	116	4.44	0.31	**<0.001**	3.83	5.05	-2.08	10.95
Primary, excluding outliers	117	4.57	0.29	**<0.001**	4.01	5.13	-1.41	10.55
Mean of 4 readings	118	4.13	0.27	**<0.001**	3.60	4.65	-1.54	9.79
Jamar Plus+ - Smedley								
Primary, excluding outliers	117	4.48	0.48	**<0.001**	3.53	5.43	-5.65	14.61
Mean of 4 readings	118	4.76	0.49	**<0.001**	3.80	5.72	-5.60	15.12
Nottingham - Jamar Hydraulic								
Primary, excluding mis-ordered	117	4.89	0.58	**<0.001**	3.75	6.03	-7.32	17.10
Primary, excluding outliers	117	4.84	0.59	**<0.001**	3.68	6.00	-7.61	17.29
Mean of 4 readings	118	4.01	0.56	**<0.001**	2.91	5.10	-7.76	15.78
Nottingham – Smedley								
Primary, excluding mis-ordered	117	4.92	0.74	**<0.001**	3.46	6.37	-10.64	20.48
Primary, excluding outliers	117	4.79	0.72	**<0.001**	3.38	6.21	-10.35	19.94
Mean of 4 readings	118	4.64	0.66	**<0.001**	3.35	5.93	-9.22	18.50

Diff=Difference; SE=standard error; CI=confidence interval

Comparison is of four dynamometers; Jamar Hydraulic, Jamar Plus+, Nottingham Electronic, Smedley

Table 12 Sensitivity analysis for difference of means for all measures (continued…)

		Paired t-test			95% CI		Limits of agreement	
	N	Mean diff	SE	p-value	Lower	Upper	Lower	Upper
Lung function, litres								
FEV_1: Micro Plus - Easy on-PC								
Primary, excluding mis-ordered	73	0.00	0.02	0.875	-0.03	0.03	-0.26	0.25
All cases, including C-E	106	0.01	0.02	0.561	-0.03	0.05	-0.35	0.37
FVC: Micro Plus - Easy on-PC								
Primary, excluding mis-ordered	66	-0.47	0.03	**<0.001**	-0.53	-0.42	-0.92	-0.02
All cases, including C-E	106	-0.45	0.03	**<0.001**	-0.51	-0.40	-1.01	0.11

Diff=Difference; SE=standard error; CI=confidence interval
Comparison is of two spirometers to measure lung function; Micro Plus and Easy on-PC

2.4.5 Results of sensitivity analysis using multilevel modelling

The following sets out the results of the sensitivity analysis described in Section 2.3.3. The data were reanalysed using multilevel models, to make full use of the advantages of multilevel modelling approach, where separate readings of each physiological function were clustered within individuals, and covariates in the models were used to control for order or sequencing effects and sample composition. Three models were estimated for each physiological function, first with device only, then controlling for the order of readings and finally controlling for covariates. The random effects in all models were statistically significant. The results indicate that the estimates of differences between devices changed by only a small amount, although the standard errors around the estimates were reduced (Table 13 to Table 15). More specifically, looking at blood pressure in more detail (Table 13), and to facilitate comparison, two lightly shaded rows are included which first duplicate the main results based on the mean of the second and third readings (from

Table 8), and then duplicate the results based on the mean of three readings (from Table 12), which provides the best comparison for the multilevel analysis which also uses all three readings. Below these shaded rows, the estimated difference between devices using multilevel modelling can be seen. Model 1 provides nearly identical results but with slightly smaller standard errors for both SBP (28% smaller) and DBP (38% smaller) because of the much larger number of observations. Model 2 shows that controlling for measurement variables – specifically the order the devices were administered and the position of the reading within the sequence – fractionally reduces the estimate of difference, while Model 3 shows that controlling for key covariates (sex, BMI, and age) fractionally increases the estimate of difference. These refinements slightly improve the overall fit of the model based on AIC, but the estimates of difference are almost identical and fall well within the confidence intervals of the primary analysis.

Table 13 Sensitivity analysis using multilevel models: blood pressure

Omron 907 - Omron 705		Paired t-test			95% CI	
SBP, mm Hg	N	Mean	SE	p-value	Lower	Upper
Based on mean of 2nd+3rd	*115*	3.86	0.69	**<0.001**	2.50	5.22
Based on mean of 3 readings	*115*	*3.92*	*0.65*	**<0.001**	*2.65*	*5.19*
Multilevel models						
- Model 1: device only	689	3.93	0.46	**<0.001**	3.02	4.83
- Model 2: Order of readings	689	3.89	0.46	**<0.001**	3.00	4.79
- Model 3: M2 + demographics	683	3.91	0.46	**<0.001**	3.00	4.81
DBP, mm Hg	N	Mean diff	SE	p-value	Lower	Upper
Based on mean of 2nd+3rd	*115*	1.35	0.53	**0.012**	0.30	2.39
Based on mean of 3 readings	*115*	*1.49*	*0.52*	***0.005***	*0.47*	*2.51*
Multilevel models						
- Model 1: device only	689	1.53	0.32	**<0.001**	0.91	2.16
- Model 2: Order, dom, hand, seq	689	1.51	0.31	**<0.001**	0.90	2.13
- Model 3: M2 + demographics	683	1.56	0.31	**<0.001**	0.95	2.18

Notes: Model 2 included order of device, dominant/non-dominant hand, left/right and sequence of reading. Model 3 additionally included age, sex, and BMI. SE=standard error; CI=confidence interval

As explained in Section 2.3.3, the multilevel model for blood pressure was extended to explore the random structure of the data (Models 4-6). The random effects in Models 4 and 5 were significant and are described first. Based on the two-level model (Model 4), it can be seen that 86.2% of the observed variance is accounted for by variance between individuals, and the remaining 13.8% is residual (or within individual) variance. Examination of the three-level model (Model 5), which

additionally included device as a random effect, thereby accounting for clustering of readings (level 1), within device (level 2), within individual (level 3), showed that, for both SBP and DBP, 83.3% of variance was explained by variance between individuals, 7.3% was within device, and the remaining 9.4% was residual, or within-individual, variance. The results of this exploration of variance partitioning are presented in Appendix E. Finally, a four-level model (Model 6) was built. This additionally included researcher, thereby accounting for clustering of readings (level 1), within device (level 2), within individual (level 3), within researcher (level 4). However, the random effect for researcher was not significant and this was not pursued further.

For grip strength, the shaded rows in Table 14 replicate the main results based on the maximum of the four readings (from Table 9) and the mean of the four readings (from Table 12) to provide a direct comparison with the results from the multilevel modelling. Based on the multilevel modelling (using Model 3 which best fits the data based on AIC), the difference in measurement between the Nottingham Electronic and Jamar Plus+ changes very slightly, with a central estimate of difference of 0.20kg compared to 0.29kg based on the maximum reading and -0.12kg based on the mean reading but remains insignificant. However, the difference between the two dynamometers which are not electronic – the Jamar Hydraulic and Smedley – which increased in magnitude with the summary measure used was the mean of four readings – now, based on multilevel modelling, appears to be significant at $p<0.05$ (difference=0.64 kg, 95% CI=0.09, 1.20, p=0.023)[2].

The differences in measurements of the four combinations which mix an electronic with either a hydraulic or spring-gauge dynamometer remain statistically significant ($p<0.001$) with only small changes in the size of the average differences, and small reductions in standard errors. The estimated difference between devices for the Jamar Plus+ and Jamar Hydraulic varies between 4.13kg and 4.45kg. Regardless, all the values remain within the confidence interval from the multilevel modelling.

[2] Although the readings for the Smedley and the Jamar Hydraulic were rounded to the nearest 0.5kg and 1kg, a difference of 0.64kg is nevertheless meaningful because rounding increases the variance but does not introduce bias.

Table 14 Sensitivity analysis using multilevel models: grip strength

		Paired t-test			95% CI	
Jamar Hydraulic – Smedley	N	Mean diff (kg)	SE	p-value	Lower (kg)	Upper (kg)
Based on maximum	118	0.23	0.52	0.654	-0.79	1.26
Based on mean of 4 readings	118	0.63	0.47	0.183	-0.30	1.57
Multilevel models						
- Model 1: 4 separate readings	472	0.63	0.30	**0.037**	0.04	1.23
- Model 2: Order, seq, dom, hand	472	0.64	0.28	**0.023**	0.09	1.20
- Model 3: M2 and demographics	472	0.64	0.28	**0.023**	0.09	1.20
Nottingham - Jamar Plus+	N	Mean diff (kg)	SE	p-value	Lower (kg)	Upper (kg)
Based on maximum	118	0.29	0.58	0.623	-0.87	1.44
Based on mean of 4 readings	118	-0.12	0.55	0.832	-1.21	0.98
Multilevel models						
- Model 1: 4 separate readings	472	0.16	0.30	0.595	-0.43	0.75
- Model 2: Order, seq, dom, hand	472	0.20	0.28	0.484	-0.36	0.75
- Model 3: M2 and demographics	472	0.20	0.28	0.484	-0.36	0.75
Jamar Plus+ - Jamar Hydraulic	N	Mean diff (kg)	SE	p-value	Lower (kg)	Upper (kg)
Based on maximum	118	4.45	0.30	**<0.001**	3.85	5.05
Based on mean of 4 readings	118	4.13	0.27	**<0.001**	3.60	4.65
Multilevel models						
- Model 1: Separate readings	472	4.13	0.30	**<0.001**	3.53	4.72
- Model 2: Order, seq, dom, hand	472	4.15	0.28	**<0.001**	3.60	4.71
- Model 3: M2 and demographics	472	4.15	0.28	**<0.001**	3.60	4.71
Jamar Plus+ - Smedley	N	Mean diff (kg)	SE	p-value	Lower (kg)	Upper (kg)
Based on maximum	118	4.68	0.52	**<0.001**	3.66	5.70
Based on mean of 4 readings	118	4.76	0.49	**<0.001**	3.80	5.72
Multilevel models						
- Model 1: Separate readings	472	4.76	0.30	**<0.001**	4.17	5.35
- Model 2: Order, seq, dom, hand	472	4.80	0.28	**<0.001**	4.24	5.35
- Model 3: M2 and demographics	472	4.80	0.28	**<0.001**	4.24	5.35
Nottingham - Jamar Hydraulic	N	Mean diff (kg)	SE	p-value	Lower (kg)	Upper (kg)
Based on maximum	118	4.74	0.59	**<0.001**	3.57	5.91
Based on mean of 4 readings	118	4.01	0.55	**<0.001**	2.91	5.10
Multilevel models						
- Model 1: 4 separate readings	472	3.97	0.30	**<0.001**	3.37	4.56
- Model 2: Order, seq, dom, hand	472	3.95	0.28	**<0.001**	3.40	4.51
- Model 3: M2 and demographics	472	3.95	0.28	**<0.001**	3.40	4.51

Table 14 Sensitivity analysis using multilevel models: grip strength (continued…)

		Paired t-test			95% CI	
Nottingham – Smedley	N	Mean diff (kg)	SE	p-value	Lower (kg)	Upper (kg)
Based on maximum	118	4.97	0.73	**<0.001**	3.52	6.41
Based on mean of 4 readings	118	4.64	0.65	**<0.001**	3.35	5.93
Multilevel models						
- Model 1: device only	472	4.60	0.30	**<0.001**	4.01	5.19
- Model 2: Order, seq, dom, hand	472	4.60	0.28	**<0.001**	4.04	5.16
- Model 3: M2 and demographics	472	4.60	0.28	**<0.001**	4.04	5.16

Notes: Model 2 included order of device, whether hand was dominant, left or right and sequence of reading. Model 3 additionally included age and sex. SE=standard error; CI=confidence interval

Finally, for lung function as shown in Table 15, the use of multilevel modelling produces results very similar to the primary analyses (replicated from Table 10), even after controlling for a number of covariates. For FEV_1, there is no observed difference between measures for the two devices, supporting the finding reported from the primary analysis. For FVC, the difference between measures for the two devices remains statistically significant, and though slightly smaller, is not meaningfully different from the difference identified by the primary analysis (0.45 litres compared to 0.47 litres). However, the much larger number of observations results in a small reduction in the size of the standard errors.

Table 15 Sensitivity analysis using multilevel models: lung function

Micro Plus - Easy on-PC		Independent t-tests			95% CI	
FEV_1, litres	N	Mean diff	SE	p-value	Low	Upper
Primary analysis	74	-0.00	0.01	0.880	-0.03	0.03
Multilevel models						
- Model 1: all readings, device only	707	-0.01	0.01	0.297	-0.04	0.01
- Model 2: sequence	707	-0.01	0.01	0.252	-0.04	0.01
- Model 3: M2+age+sex +health	707	-0.01	0.01	0.249	-0.04	0.01
FVC, litres	N	Mean diff	SE	Sig	Low	Upper
Primary analysis	67	-0.47	0.03	**<0.001**	-0.53	-0.42
Multilevel models						
- Model 1: all readings, device only	705	-0.45	0.02	**<0.001**	-0.48	-0.42
- Model 2: sequence	705	-0.45	0.02	**<0.001**	-0.48	-0.42
- Model 3: M2+age+sex+health	705	-0.45	0.02	**<0.001**	-0.48	-0.42

Diff=difference; SE=standard error; CI=confidence interval
Notes: Model 2 includes sequence of readings. Model 3 additionally includes age, sex, and self-reported health.

2.5 Discussion

In a randomised cross-over study of 118 adults aged 45-74 years, evidence was found of differences in measurements of blood pressure, grip strength and lung function when assessed using different devices. For blood pressure, the newer Omron HEM-907 measured, on average, higher than the older Omron 705-CP (3.85 mm Hg for SBP and 1.35 mm Hg for DBP). For grip strength, the two electronic dynamometers were found to record measurements, on average, 4-5kg higher than either the hydraulic or the spring-gauge dynamometer, but there were only small differences when comparing the two electronic dynamometers or the hydraulic and spring-gauge dynamometers and these were not statistically significant. For lung function, the measures of FVC on the Easy on-PC by NDD were, on average, 0.47 litres higher than those for the Micro Medical, but there was no difference between measures of FEV_1. If different devices affected the level of variance observed in measurements, this would be relatively unproblematic, but the differences seen here are likely to be systematic and hence may indicate bias.

Only a few studies have compared combinations of these devices previously. For example, King compared the Jamar Hydraulic with the Jamar Plus+ dynamometer and, in contrast to our findings, reported that the electronic dynamometer had consistently lower readings than the hydraulic device (King, 2013). However, the study population was younger, with an average age of 32 years, and comprising only a convenience sample of 40 men and women. Another study reported a difference of 3.2kg when comparing the Smedley and Jamar Hydraulic dynamometers which contrasts with our finding of a measurement difference between these devices of an average of 0.23kg (Guerra and Amaral, 2009). However, this other study aimed to compare measurements in an older, smaller sample of 55 65–99-year-olds recruited from a retirement home and social day care centre. Another study (Kim and Shinkai, 2017), found that the Smedley dynamometer measured lower than the Jamar+ Digital, similar to our study, although in this other study there were other potentially important variations in measurement protocol – measures using the Smedley device were undertaken in a standing position and those using the Jamar device were undertaken seated.

A comparison of Micro Medical or other turbine spirometers with the Easy on-PC by NDD spirometer has not been identified. However, in a study of 35 volunteers, the Micro Medical turbine spirometer gave lower readings compared with the Vitalograph Micro pneumotachograph spirometer (Orfei et al., 2008), both for FEV_1 (0.24l) and FVC (0.34l). Another study of 49 volunteers found that a

pneumotachograph spirometer (Masterscreen) gave higher readings than the Easy on-PC by NDD (Milanzi et al., 2019), for FEV_1 (0.24l) and for FVC (0.37l). This is consistent with the finding in this study of similar FEV_1 between devices, although lower FVC in the turbine spirometer can be seen compared with the Easy on-PC by NDD spirometer.

An established way to determine what an 'important' difference is between measurements is the potential clinical significance of the differences between devices, with reference to published normative or predicted values of blood pressure, grip strength and lung function (Dodds et al., 2014, NHANES, 1999, Scholes and Neave, 2017). Based on analysis of age-related differences in mean blood pressure in the Health Survey for England 2016, the differences in SBP and DBP between devices that are observed are equivalent to an age difference of about five years. Similarly, using normative grip strength data (Dodds et al., 2014) it can be observed that a 4-5kg difference in median grip strength is equivalent to an age difference of approximately 5 years among men and approximately 10 years among women aged 65 years and above and is equivalent to even greater differences among younger adults. For lung function, based on the National Health and Nutrition Examination Survey (NHANES) III data (Thomas et al., 2019), predicted values for five-year age-groups (with male height of 175cm and female height of 160cm), show that a difference of 0.47l is equivalent to an age difference of around 15 years, between 45-75 years. These comparisons suggest that the differences observed between devices are likely to have important implications.

The focus of this research is on the implications the findings have for analyses which track change in the individual over time or make comparisons between groups across studies. However, systematic differences in measurements may also result in discrepancies in clinical diagnoses which use cut-points; for example, when identifying an individual as hypertensive or classifying them as sarcopenic. The study data on grip strength was used to demonstrate how this might affect the identification of low muscle strength (Cooper et al., 2021). The first page of this paper is reproduced in Appendix D.

A key strength of this study design is the use of the same standardised measurement protocols for all devices. This is important as for all three functional measures since the type of device used for assessment is only one of several factors which can affect measurements, and consistent application of protocols minimises variation which may result from these other factors. As described in Section 2.2.4, for blood pressure, these relate to the subject's correct preparation for the assessment, their compliance during it, their correct positioning and the proper application of the equipment (Bilo et al., 2017, Handler, 2009, Jones et al., 2003). For grip strength, these primarily relate to how the dynamometer is set up, how the participant is positioned, and how well the assessor is trained, particularly in providing effective encouragement to maximise performance (Amaral et al., 2012,

Balogun et al., 1991, Fess, 1981, Firrell and Crain, 1996, Incel et al., 2002, O'Driscoll et al., 1992, Roberts et al., 2011, Sousa-Santos and Amaral, 2017). For lung function, the other factors relate to the positioning of the participant and the coaching provided by the assessor (Miller et al., 2005).

There are other important differences between spirometers which may impact on results. For example, the Easy on-PC by NDD spirometer presents visualisation of the volume-time graph in real time. This means that the participant can be encouraged to blow until the curve has reached a plateau, that is, when the true FVC has been achieved. In the absence of this visual display the forced manoeuvre may be terminated prematurely, and the FVC underestimated. This is the most likely explanation for the substantially higher FVC values obtained using the Easy on-PC by NDD device than the Micro Medical device in this study.

These findings provide some reassurance about the consistency of measurement between specific device combinations (i.e., the Jamar Plus+ and Nottingham electronic; the Jamar Hydraulic and Smedley, and the Micro Medical and Easy on-PC by NDD for FEV_1) which may encourage further analyses that so far have been avoided. It may also provide useful evidence for investigator teams when a change of device is necessary. For other combinations, the results suggest that analysts may need to carry out sensitivity analyses or compute correction factors or device-specific reference equations when comparing values across time within a study or between studies that have been measured using different devices (Milanzi et al., 2019, Orfei et al., 2008). However, in the SAPALDIA study, using a group correction from a quasi-experimental study was found not to be adequate to correct for the change in spirometer, and an approach using spirometer-specific reference equations from longitudinal measurements to describe individualised corrections terms was preferred (Bridevaux et al.).

Maintaining consistency in the make and model of device used reduces the likelihood of measurement differences resulting from the device, but this is not always realistic given that equipment becomes obsolete. Where possible, there are benefits to standardising equipment across studies; for example, the National Institute of Health (NIH) Toolbox recommends and supports the use of the Jamar Plus+ for grip strength (Gershon et al., 2013). However, this is not always possible, and investigators consider multiple issues when making decisions. New technologies can improve measurement; for example, through automation (as is the case with the Omron 907), the transition from analogue to digital (as is the case with the transition from the Jamar hydraulic to Jamar Plus+ devices), or the introduction of visual encouragement and specific feedback (as provided by the Easy on-PC by NDD). Consequently, an important implication of this research is that investigators should include experiments to assess machine comparability when a new device is introduced into a study

or where comparison between studies using alternative equipment is planned. This could include randomised trials in a controlled environment similar to this study, ideally with the addition of randomisation of researchers to assessments to estimate research effects, as well as small scale in-field experiments providing a more realistic context. In the meantime, the research may provide survey researchers and practitioners with further evidence to support maintenance of existing equipment and reinforce interviewer and nurse training to ensure consistency in delivery of protocols.

The study has several strengths. The inclusion of different devices which are commonly used in a number of large-scale population-based studies in the UK and other countries means it has wide application. The sample size is sufficiently large to provide the statistical power necessary to detect small device effects. Participants were selected to include equal numbers of men and women across three age bands between the ages of 45 and 74 so the findings should hold for much of the population, although it is possible that results would have differed if younger or older age groups had been selected. The sample of volunteers was recruited from a large database of members of the general public. That said, the robustness of the findings does not depend on the nature of the sample, but the study's repeated measurements, cross-over design, which facilitates within-person measurement comparisons for each device, with randomisation of the sequence of measurements ensuring that the samples are balanced with respect to observed and unobserved characteristics. This means that differences in measurement can be confidently attributed to device.

The study also has a number of limitations. Some devices used in the CLOSER studies were not included in this experiment, such as the Vitalograph Escort used in *Understanding Society* in fieldwork in Scotland. The Smedley dynamometer is normally used with the study participant in a standing position; while this meant that protocols were applied consistently, this limits the applicability of the findings related to the Smedley.

In the primary analyses of lung function, a number of participants had to be excluded due to missing or low-quality readings, particularly on the NDD Easy on-PC, thus reducing the sample size and statistical power of these analyses. Although this is observed in studies which use experienced interviewers and survey nurses, this was most likely exacerbated by the approach taken here with a dedicated researcher team recruited and trained for this study alone. The results were not, however, affected by this as demonstrated by the sensitivity analysis using all available readings (regardless of quality). The sensitivity analyses considering outliers, incorrectly ordered tests and alternative coding of measures all showed that our results were robust. The same results are found whether

using standard analytical approaches based on summary outcome measures or using repeated measures on the same individuals within a variance components model.

The standard approach among epidemiologists is to use summary measures and a comparison of means using Bland and Altman plots. This is the primary analysis approach used in this paper. However, it is argued here that the approach of using all readings within a multilevel model is, perhaps, preferable. For blood pressure, the first reading is commonly discarded on the basis that it tends to be inaccurate, but this has been disputed (Salazar et al., 2015). For grip strength the maximum of the four readings in kilograms is selected on the basis that it should, by definition, be closest to the latent or underlying maximum value, but an alternative conception is that analysing multiple readings is more effective and accounts for variance. The same argument applies for lung function, though in addition, adhering to the ATS/ERS criteria means excluding almost one third of sample members for whom it was only possible to collect one reading, or where two readings were achieved but the difference between the readings was greater than 200ml. Although further training of the researchers may have reduced the number of missing cases, excluding lower quality cases almost inevitably reduces the sample and may have the effect of biasing the sample towards individuals with better physical or cognitive health.

With this in mind, the multilevel modelling offers an alternative statistical analysis approach and applied as a form of sensitivity analysis for this paper, has several advantages. Since the multilevel modelling takes account of all readings rather than using summary measures, it increases the number of readings available for analysis, reducing the size of the standard errors of the estimated differences. In the case of lung function, it very substantially increases the sample available for analysis when allied with a decision to reduce the quality standard applied to determining valid cases. Since using summary measures removes some of the variation in the dataset, multilevel modelling which uses all available readings for any given measure makes it possible to take account of the variability across measurements, since any reading includes an element of error (Rabe-Hesketh and Skrondal, 2012). In addition, with this statistical approach, the position of the reading in the sequence and the order of the devices (since order effects cannot be entirely ruled out) can be accounted for, whereas in the dominant approach used in the epidemiological literature, order effects are commented on in the narrative, but no adjustment is made for them in the comparison of means.

The random allocation of individuals to different groups should ensure these are reasonably well balanced in terms of demographic and other characteristics. For this study, balance was checked based on a visual assessment of a range of sample characteristics and more formal testing was not

carried out (NICE, 2013). In practice, an additional advantage of the multilevel modelling is that it takes account of any minor imbalances in the characteristics of the randomised samples and so minimises any remaining risk of chance bias (Altman, 1985, Roberts and Torgerson, 1999). In this case, included age, sex and body mass index in the final models did not have any profound effect on the results, as expected.

The study was prepared for an epidemiological journal and follows the primary approach used in that literature. Although multilevel models have been used in some cases, including by Bland and Altman (Rabe-Hesketh and Skrondal, 2012). Multilevel models were used here as part of sensitivity analysis to see if different approaches lead to the same or similar conclusions. The results show that both the Bland and Altman and comparison of means approach, and the multilevel modelling approach, indeed have led to the same conclusions. The multilevel modelling has many advantages, and the models were explored to their full power. For example, all readings were included, as was device (as a fixed effect), and relevant variables were included in the models such as the order in which devices were administered, the order reading were taken, and in the case of the dynamometer whether each reading was taken from the left or right hand, and hand dominance. However, the models only included a limited number of sample characteristics, and it could be argued that a greater number of characteristics should have been included to use multilevel modelling to its full power. Arguably, however, including these was unnecessary. Key demographic characteristics were included in the models to enable minor adjustments in case of any residual differences between the sample but, it is not appropriate to include a very large set of sample characteristics which should be approximately evenly distributed because the study had an experimental design, and randomisation should control for observed and unobserved characteristics of the sample. Crucially, the statistical modelling showed near identical results to the primary statistical methodology. In practice, for this analysis, the results of the multilevel modelling approach confirm the findings from the (standard) analytical approach using mean differences and Bland and Altman plots, generally more favoured by epidemiologists.

The small number of researchers who carried out the study were not experienced interviewers and survey nurses, and though the training that researchers received was thorough and involved an assessment, the standards reached may have been lower than would have been achieved using more experienced assessors. It is also possible that researchers varied in the extent to which they adhered to protocols to ensure that all assessments were consistently applied. Crucially, this is a within-person comparison study, and the same researcher assessed the same person on all machines. This means that while the individual measurements may have varied by researcher (due,

for example, to different levels of encouragement provided), it is not likely that they had a substantial impact on the differences in the measurements taken. Nevertheless, the issue of whether there is a researcher effect is still of interest. The study design did not account for researcher effects; researchers were not randomised to assessments and practical constraints meant that some researchers carried out many assessments while others carried out very few. During the analysis stage, resource and attention were focused elsewhere. Nevertheless, researcher effects were examined for blood pressure by inclusion of an additional random effect within the model (level 4). The random effect was not found significant in the multilevel model and resource and attention was focused elsewhere. Extending the multilevel modelling to account for possible researcher effects for grip strength and lung function is a potential area for future investigation. Perhaps most usefully this could be carried out within the context of a major field study with a large number of interviewers and survey nurses with varying degrees of experience, ideally with data that captures interviewer/nurse indicators such as age, sex and experience, both with respect to general surveys, biomeasures, or these assessments in particular.

Devices are often assessed relative to a gold standard, but in this paper, measurement error is considered within a survey context where readings from two or more devices may be used alongside each other. Future research could go further and assess device effects within a Total Survey Error framework which, alongside measurement error, would also consider coverage error, sampling error and non-response to the survey and the measurement itself (Groves et al., 2011).

Conclusions and implications for practice

In this randomised cross-over study measurement differences are shown between devices commonly used to assess blood pressure, grip strength and lung function which researchers should be aware of when carrying out comparative research between studies and within studies over time. There are four main conclusions that can be drawn from this study. Firstly, analysts may want to test the sensitivity of their findings to the device effects that were identified here, and perhaps compute correction factors or device-specific reference equations when estimating intra-individual changes in function over time using longitudinal studies that have switched device, or when comparing physiological measures within or across studies that use different devices. Second, investigators will want to consider these findings when selecting equipment to include in new studies or when changing equipment in longitudinal studies. More specifically, based on these examples, mixing electronic and other dynamometers may be more problematic than expected, while mixing measures of FEV_1 between the hand-held Micro Plus and the computer-based Easy on-PC may be less so. Third, introducing or changing equipment used within surveys will often require

methodological research to identify quality issues. Further trials are needed to replicate the comparison of these devices, to test the same devices with alternative protocols, and to test different device combinations, both in stand-alone studies of this kind and within larger observational surveys with greater variation in implementation. Finally, researchers carrying out these studies may want to consider the statistical approaches adopted here to assess device effects; in particular, the multilevel modelling approach that was used in the sensitivity analysis.

Chapter 3 Device effects: evidence from a large-scale mixed-device online survey of young people in England

3.1 Introduction

As outlined in Chapter 1, a key technological development affecting social research in recent years has been the emergence of mobile devices as a means of administering social surveys. Online surveys have increased substantially and a growing proportion of participants who complete them do so using the web browser on their smartphone or tablet rather than a PC. This shift has been respondent-driven and has followed the transition of many activities, such as shopping or banking, to mobile devices. Taking part in online surveys in this way is likely to be more convenient for participants, but the experience of responding may differ when questions are rendered on a small touchscreen, or if the context of using a mobile device means that survey response is subject to more interruptions or distractions or takes place in a less private environment. This device effect may have a negative effect on data quality and could affect survey estimates. Early studies carried out to investigate this effect identified a number of dimensions of data quality which may be affected by device, but the findings are somewhat mixed, and the question remains whether survey completion on a mobile device is associated with poorer data quality.

This chapter contributes to the literature on whether completing online surveys with a smartphone or tablet rather than a PC influences measurement quality. It takes a snapshot in time in 2016, several years after initial studies on this topic pointed to improvements in the design of online surveys in an attempt to make completion agnostic to the device used, and when respondents had become more familiar with mobile devices.

It is based on the Wellcome Trust Science Education Tracker (SET), a nationally representative online survey of over 4,000 pupils in school years 10 to 13 (aged 14-18) attending state-funded schools in England in 2016. The survey was optimised for completion on a smartphone and underwent thorough usability testing to ensure, as far as possible, that it was device agnostic. Respondents chose the device which they used to participate.

In addition to survey responses, the dataset includes information from three sources: geographical data about the characteristics of the area where each respondent lives, survey process data

including lapsed time between invitation and response, and for those who consented to data linkage, administrative records about the individual and their school. These additional data items are exogenous, that is, they are not affected by the process of responding itself or the device used.

The analysis of device effects applies quasi-experimental methods to control for possible selection effects. It uses data items which are, as far as possible, exogenous to the process of responding to the survey. Then, through a series of sensitivity analyses, it examines the effect of varying the matching specification employed, and the inclusion or exclusion of people who respond using medium or large tablets. Response behaviours are considered through 11 outcome variables, which capture (i) willingness to engage with the SET study beyond the act of responding to the survey itself, specifically by consenting to having their survey record data linked to administrative data, agreeing to receive findings from the study, and agreeing to be recontacted by the Wellcome Trust, (ii) satisficing behaviours and (iii) temporal aspects of participation such as completion time and whether there was a substantial interruption during survey completion. A discussion of a twelfth outcome variable, breakoff rates, is also included, but firm conclusions cannot be drawn because of limitations to the data that is available about these respondents.

Before describing the study and methodology in greater detail, this section sets out the research evidence about the growth of online surveys, the increase in use of mobile devices to respond to these surveys, the potential challenges this raises for data quality, and the evidence from earlier studies on device effects.

The substantial growth of online surveys is well documented (Mavletova, 2013, Mavletova et al., 2018, Tourangeau et al., 2018). Although response rates for online surveys are lower, and the representation of the general population is poorer when compared to face-to-face and telephone interviews (Couper, 2000), they nevertheless have been shown to have good measurement properties with reduced social desirability bias and less cognitive burden than telephone surveys (Tourangeau et al., 2013), and they allow for interactivity (Conrad et al., 2011).

The increase in the proportion of survey participants who choose to respond to online surveys using a smartphone or tablet rather than a PC is more recent but also remarkable (Clement et al., 2020, Couper et al., 2017, Gummer et al., 2019, Lugtig and Toepoel, 2016, Mavletova, 2013, Maslovskaya et al., 2019, Revilla, 2017, Revilla et al., 2016, Tourangeau et al., 2018). For example, Struminskaya et al. (2015) document the rapid growth in the use of mobile devices in three random probability internet panels. Indeed, in the UK, a mark of the phenomenal transition from PCs to mobile devices

is that, of the 89% of households who responded to Census 2021 online, almost two-thirds used either a smartphone (56.4%) or tablet (8.9%) (Office for National Statistics, 2021).

The shift from PCs to mobile devices was observed among young people before it became evident in the adult population (Maslovskaya et al., 2019). For example, although responding on a mobile device was discouraged in the 2016 data collection for the Longitudinal Study of Young People in England 2, 22.3% of 16–17-year-olds used a smartphone and 16.9% used a tablet to respond (Maslovskaya et al., 2019). By 2018, when Millennium Cohort Study members were asked to complete a short, online survey as part of the wider age-17 data collection, 71% chose to do so on a smartphone. This was at a time when Ofcom reported that 95% of 16-24-year-olds already owned a smartphone (Gilbert and Lindley, 2019), and one in eight were using only a smartphone to access the internet (Matthews et al., 2017).

The shift to responding to online surveys using mobile devices reflects several broad secular trends: increased mobile device ownership, higher specifications of smartphones with larger screen sizes which makes completion of a range of tasks easier, and the transition of many daily activities to mobile (Ofcom, 2021). For example, the latest data shows that 76% of adults in the UK use internet banking and 87% had shopped online within the last 12 months (Competition & Markets Authority, 2016, Office for National Statistics, 2020), while two-thirds (65.4%) of access to government information and services are provided through smartphones with a further 3.6% by tablet (GOV.UK, 2022).

Completing a survey on a mobile device may be convenient and offer some benefits (Fuchs, 2008), such as the potential to include harder-to-reach populations, or to incorporate new data types such as video, audio, or location (Poggio et al., 2015, Sugie, 2018, Toepoel and Lugtig, 2013). However, concerns have been expressed about the negative effects that a mobile device may have on data quality (Antoun, 2015, Antoun et al., 2017, Callegaro, 2013, Couper et al., 2017, Keusch and Yan, 2017, Schlosser and Mays, 2018, Tourangeau et al., 2017, Tourangeau et al., 2018), in particular by encouraging satisficing behaviours (Krosnick, 1991) where respondents with low motivation engage in suboptimal response strategies, including weak forms of satisficing behaviours (primacy, recency, and acquiescence) and strong forms (‘don’t know’ or no opinion, non-differentiation, random reporting and endorsing the status quo). The literature considers these issues alongside a range of associated measures of data quality, such as length of responses to open-ended questions, evidence of social desirability bias and poorer response accuracy, measures of non-response including breakoffs, and indicators of interview pace which may increase or decrease and may reflect satisficing behaviours (Roberts et al., 2019).

There are three main reasons why data quality may be negatively affected by the use of mobile devices, particularly smartphones. The first is that smaller screens provide less visibility for survey questions and response options (Couper et al., 2017, Mavletova et al., 2018), particularly for question types such as grid questions, which render particularly poorly. This may lead respondents to select visible options (Couper et al., 2004, Krebs and Höhne, 2020) or require additional scrolling (Tourangeau et al., 2013) which may, in turn, increase completion times (Couper and Peterson, 2017) and response burden (Krebs and Höhne, 2020). A second explanation is that the input capabilities of mobile devices differ substantially from a PC, and data entry on smartphones or tablets which use finger movements and tabs on touch screens may be both imprecise and effortful leading to increased errors and higher breakoffs (Antoun et al., 2017, Lugtig and Toepoel, 2016, Peytchev, 2009, Tourangeau et al., 2013), particularly if participants are unfamiliar with using mobile devices or lack dexterity or visual acuity (Olmsted-Hawala et al., 2021). Thirdly, there are differences between mobile devices (particularly smartphones) and PCs in the context of survey completion. Although a key opportunity provided by mobile devices is their portability, allowing them to be accessed at any time or place, this also increases the likelihood that other people will be present, reducing privacy. Furthermore, respondents are more likely to be on the move or in shared spaces, raising the possibility of distractions and interruptions (see for example, Mavletova, 2013, Sendelbah et al., 2016, Toninelli and Revilla, 2016). This exaggerates an existing problem faced by all self-administered surveys, that the researcher has little control of the environment in which the survey takes place (Clement et al., 2020). Indeed, the patterns of use do differ by device (Antoun et al., 2018, Couper et al., 2017, Deng et al., 2019, Wells et al., 2013). There may be other reasons why response behaviours may differ if, for example, participants associate a PC with more serious tasks than a mobile device.

Although this area of research is relatively recent, concerns about device effects have already prompted a number of studies (for example, see the reviews in Tourangeau et al., 2017, Couper et al., 2017, Clement et al., 2020, Keusch and Yan, 2017, Krebs and Höhne, 2020, Schlosser and Mays, 2018, Tourangeau et al., 2018). Some of these studies have shown that lower response or completion rates are associated with smartphones (Buskirk and Andrus, 2014, Couper et al., 2017, de Bruijne and Wijnant, 2013, Mavletova, 2013, Mavletova and Couper, 2013). Another fairly consistent finding is that completion times are slower on smartphones (Andreadis, 2015, Couper and Peterson, 2017, de Bruijne and Wijnant, 2013, Keusch and Yan, 2017, Maslovskaya et al., 2020, Struminskaya et al., 2015, Tourangeau et al., 2017), although occasionally no differences were found (Matthews et al., 2017) or PCs were found to be slower (Buskirk and Andrus, 2014). Slower completion times are cited as one explanation for the common observation that mobile responders

are more likely to break off before the survey reaches completion (see, for example Couper et al., 2017, de Bruijne and Wijnant, 2013, Buskirk and Andrus, 2014, Callegaro, 2010, Maslovskaya et al., 2020, Poggio et al., 2015, Stapleton, 2013). Some studies report that smartphone responders are more likely to skip items or show higher item non-response than PC responders (for example, Lugtig and Toepoel, 2016, Antoun et al., 2017, Keusch and Yan, 2017, Buskirk and Andrus, 2014, de Bruijne and Wijnant, 2013, Struminskaya et al., 2015); but others find no difference or report mixed evidence (Andreadis, 2015, Buskirk and Andrus, 2014, Couper et al., 2017, Maslovskaya et al., 2020).

Mixed results are also evident in other aspects of data quality. For example, although the expectation was that mobile respondents might select response options that were more visible without scrolling (Couper et al., 2004), this does not always appear to be the case (Tourangeau et al., 2017). Stronger primacy effects have been identified on mobile devices (Stapleton, 2013), but other studies did not observe a difference (Maslovskaya et al., 2020, Matthews et al., 2017, Mavletova, 2013, Toepoel and Lugtig, 2014, Wells et al., 2013). Some studies have found lower levels of agreement among smartphone responders (Tourangeau et al., 2017) but others find no differences (Matthews et al., 2017), while a more recent study has shown substantial differences in response behaviour between PCs and smartphones within each scale direction, with responses on smartphones more positive than on PCs (Krebs and Höhne, 2020). Some studies seem to find that non-differentiation or straightlining is higher on PCs (Keusch and Yan, 2017) but others report that it is higher on mobile devices (Barlas et al., 2015, McClain et al., 2012, Maslovskaya et al., 2020), or find mixed results (Matthews et al., 2017) or no difference (Antoun et al., 2017, Revilla and Couper, 2018, Tourangeau et al., 2017).

Similarly, while in some studies smartphone responders gave shorter answers to narrative open-ended questions (de Bruijne and Wijnant, 2013, Mavletova, 2013, Struminskaya et al., 2015), no differences were found in others (Lugtig and Toepoel, 2016, Toepoel and Lugtig, 2014). The findings in relation to half open or open questions were also inconsistent (de Bruijne and Wijnant, 2013, Peytchev, 2009, Wells et al., 2013). Some studies showed that smartphone responders were equally likely to disclose sensitive information (Maslovskaya et al., 2020, Matthews et al., 2017, Mavletova, 2013) but another study reported mixed results with significant differences in alcohol reporting but not in other sensitive attitudinal or behavioural differences (Mavletova and Couper, 2013).

A few other dimensions of data quality are considered occasionally in some studies (Roberts et al., 2019). For example, smartphone responders provided less accurate information about age and date of birth (Antoun, 2015), and errors have been shown to arise from a touchscreen with sliders and pickers (Buskirk and Andrus, 2014). There is some evidence that mobile users access online surveys

sooner after the invitation than those responding on a PC (Cunningham et al., 2013, Schlosser and Mays, 2018) and may be less likely to agree to a follow-up survey (Cunningham et al., 2013). No difference was found in the rate of consenting to data linkage (Maslovskaya et al., 2020, Matthews et al., 2017), but a small difference was found related to giving permission to capture GPS coordinates with smartphone responders being more positive (Toepoel and Lugtig, 2014).

These studies cover a period of about a decade, during which there have been significant changes in mobile device technologies and the survey software used to administer them. When respondents first elected to answer surveys on their mobile devices, the surveys were not prepared for this eventuality. Some surveys were configured to prevent respondents using a mobile device, while participants in other surveys were presented with messages that strongly encouraged completion on a PC (Maslovskaya et al., 2019). Efforts to adapt to the demands of the new technology included adjusting questionnaire design with shorter questions and response options, redesign of survey software layouts, and increased user-interface testing on multiple devices. The move towards device agnostic surveys seems to have resulted in fewer device effects (Antoun et al., 2017, Buskirk and Andrus, 2014), although some studies found few device effects even when the survey was *not* optimised for smartphones (Tourangeau et al., 2017, Tourangeau et al., 2018). Hence, research in this area continues to be needed to explore effects, as well as to research the process of adjusting layouts, for example how best to meet the needs of different respondent groups such as older respondents (Olmsted-Hawala et al., 2021).

In summary, while there is some evidence of greater satisficing among mobile respondents, the results tend to be mixed and the consensus seems to be that there are no major measurement error differences between those who complete a web survey on a mobile device or on a PC, especially where surveys are optimized for use on mobile devices, and, where differences do exist, these tend to be marginal (Antoun et al., 2017, Clement et al., 2020, Couper et al., 2017). Nevertheless, given mixed and sometimes contradictory results, there is agreement that further evidence is needed (Antoun et al., 2017, Couper et al., 2017).

One of the reasons that identifying device effects is difficult and remains contested is the fact that, unconstrained, different population groups choose to respond using different devices (Antoun, 2015, Clement et al., 2020, Maslovskaya et al., 2019, Revilla et al., 2016) which means that measurement and selection effects are confounded. Maslovskaya et al. provide evidence from six major UK surveys which have an online component and find that age, gender, employment status and household size all have a significant relationship with device, in line with findings internationally; and in these UK studies, they also observe associations with marital status, religion, the number of

children in the household, income, number of cars and frequency of internet use (Maslovskaya et al., 2019). People who respond using a mobile device tended to be younger, female, more likely to have a lower household income, and more likely to be renters and to be working (Maslovskaya et al, 2019). Furthermore, those who are only able to respond using a mobile device are particularly likely to be from more disadvantaged groups (Lugtig, 2020). Interestingly, the selection effects that are observed appear reasonably stable over time (Gummer et al., 2019). There is thus a risk of confounding between selection to device, and response behaviour, since the correlates for self-selection may also be related to measurement error (Lugtig and Toepoel, 2016); and any direct comparison of one group with the other will not provide meaningful findings. It might be anticipated that this would be particularly pronounced among young people. For example, data from the second cohort of the Longitudinal Study of Young People of England provides evidence that young people who are economically disadvantaged, and who have lower educational attainment, often lack access to a PC (Lessof et al., 2019, Lessof et al., 2016).

A mechanism is therefore needed to contend with selection to isolate measurement effects. In practice, the earliest studies of device effects were simple comparisons of responses from those who chose mobile devices for regular web surveys and took little account of selection (Couper, 2013c, McClain et al., 2012, Peterson, 2012), essentially describing the phenomenon of "unintentional mobile respondents" (Peterson, 2012). Following these early studies, a common approach has been to use experimental designs to isolate selection effects (Antoun et al., 2017, Couper, 2013c, Keusch and Yan, 2017, Mavletova, 2013, Mavletova and Couper, 2013, Mavletova and Couper, 2016, Toninelli and Revilla, 2016, Tourangeau et al., 2018) either within the administration of a smartphone survey or with randomised allocation to different devices; for example, using alternative formats for questions (Peytchev and Hill, 2010, Stapleton, 2013, Wells et al., 2013). These types of studies can be implemented quickly and at reasonable cost, making it possible to test a range of alternatives. Nevertheless, they have limitations (Clement et al., 2020). First, they often rely on either probability panels (such as GESIS, LISS, CentERpanel, Knowledge Panel, and MarketResponse) or non-probability panels (such as Netquest or the Russian Online Market Intelligence consumer access panel), so the respondents will already be experienced, and any response behaviours may be entrenched. They also sometimes use short, artificial question sets (Andreadis, 2015, Clement et al., 2020, Keusch and Yan, 2017, Toepoel and Lugtig, 2014, Schlosser and Mays, 2018). Several studies rely on small sample sizes (de Bruijne and Wijnant, 2013, Peytchev and Hill, 2010, Toepoel and Lugtig, 2014, Tourangeau et al., 2017). They may tell people which device to use or provide the device (Tourangeau et al., 2018) which may be unfamiliar and create an unnatural context which could exaggerate device effects (Clement et al., 2020, Lugtig and Toepoel, 2016), and in some

designs, respondents may not adhere to the allocated treatment (de Bruijne and Wijnant, 2013, Mavletova, 2013) although there have been attempts to address this by using a cross-over design (Mavletova and Couper, 2013). Finally, many of these studies set tablet responders aside and focus on comparing smartphone and PC responders (Lugtig and Toepoel, 2016).

Several alternative approaches have been used to analyse existing large-scale data sets for measurement effects while accounting for selection. This includes a variety of multivariate analyses (Maslovskaya et al., 2020), including longitudinal analysis of measurement error following a switch of device over two consecutive waves of a panel (Lugtig and Toepoel, 2016), a similar approach based on multiple waves and using multilevel models (Struminskaya et al., 2015), or using multivariate logistic regression and testing results using more than one cross-sectional sample (Clement et al., 2020). A final approach uses the quasi-experimental method of matching using propensity score analysis to make the profiles of PC and smartphone responders as similar as possible to each other (Matthews et al., 2017). This is also the approach used in this study (the methodology is explained in detail in Section 3.3.3).

3.1.1 Research questions

This research study is based on a large sample of young people for whom a wide array of data is available (including geographical, survey process and administrative data). It assesses whether the device used to complete an online survey affects data quality. The central question is addressed is:

1. Do the survey responses and behaviours of those who respond to the online survey using a smartphone or tablet (and who consent to administrative data linkage) differ from the responses of a matched sample who respond using a PC? And if they differ, how do they do so?

There are two supplementary research questions:

2. Are the results the same if only smartphone responders are compared with PC responders, or do they differ? And if they differ, how do they do so?
3. Are the results the same if the sample includes those who refuse data linkage, since they may be less compliant, and since some of the data that might help to match them to PC responders is unavailable, or do they differ? And if they differ, how do they do so?

The study that is used to answer these questions has some methodologically interesting features:

- While most studies of this kind focus on adults, this research adds to the examples that are based on young people who are digitally native (Couper and Peterson, 2017, Matthews et al., 2017).
- While much of the existing research is based on web panels with short questionnaires specifically designed for device experiments, this study is based on an authentic research study, with a nationally representative cross-sectional sample which is free of panel effects, and the survey is optimised for use on mobile devices so may be said to have ecological validity.
- Most importantly, the study uses propensity scores to balance the samples so that measurement effects are not confounded by selection effects; and it draws on a rich range of additional data types that are linked to the survey responses (geographical data, survey process data and administrative school records). This provides an opportunity to assess the use of exogenous confounder variables in the matching process.

The remainder of the chapter is structured as follows: Section 3.2 describes the SET study and the data used; Section 3.3 sets out the analytic method and presents some preliminary analysis necessary to understand the matching process; and Section 3.4 provides the main results, first focusing on the primary analysis and then providing a series of sensitivity analyses. Section 3.6 summarises and discusses the results, highlighting implications for survey research and practice and opportunities for further research.

3.2 Data

First, the survey and the analysis sample are described (Section 3.2.1), the outcome variables are defined (Section 3.2.2), and the covariates or potential confounders are examined (Section 3.2.3).

3.2.1 The Science Education Tracker and analysis sample

The 2016 Science Education Tracker (SET), known as “Pathways”, is a repeat cross-section online survey which provides evidence about the changing level of science engagement, education and career aspirations among young people aged between 14 and 18 (in school years 10 to 13) attending state schools in England (Hamlyn et al., 2017). The questionnaire content was developed using a series of nine focus groups, (see EdComs, 2016). The survey was funded by the Wellcome Trust.

SET is based on a random probability sample drawn from the 2014/15 National Pupil Database (NPD) and Individualised Learner Record (ILR). Invitations were issued to 8,124 students and fieldwork was

conducted by Kantar Public between June and August 2016, after the school exam period and before the start of the new academic year. Sample members were offered a £10 shopping voucher which they could download from an online portal and included Amazon, iTunes, River Island and Boots. The study had a comprehensive contact strategy including a prenotification letter sent to both parents and young people, a survey launch letter sent directly to young people if they were 16 or over and via their parent if they were under 16, and then up to four postal reminders in the sequence letter, postcard, letter, letter, again sent via parents if the young person was under 16. Reminder mailings were focused on groups with relatively low response rates. The achieved sample of 4,081 individuals represents a reported response rate of 50% (Hamlyn et al., 2017). Weights were calculated by Kantar Public to compensate for variations in response by different sample groups. The dataset is available through the UK Data Archive (Wellcome Trust, 2017). In addition, permission was given by the data owners for time stamp data and data about cases who did not complete the survey to be provided.

SET placed no restrictions on the internet-connected device that young people could use to respond to the survey (e.g., a desktop, laptop, netbook, smartphone, or tablet). Nine respondents used non-standard devices such as games consoles and are excluded from the analysis, reducing the sample from 4,081 to 4,072. Question content and layout were adapted for a small screen and usability testing and piloting was carried out to ensure the survey could be completed on all devices. Most students responded to the survey using a desktop, laptop, or netbook, which are referred to here as PC (63%, n=2572). A substantial minority (36.7%, n=1,500) used a mobile device. This was comprised of smartphone and small tablet responders (24.8%, n=1013), as well as medium (1.2%, n=50) and large (10.7%, n=437) tablet responders. Device type was recorded by the survey software and made available in the data set. The category of smartphones will have included some small tablets of similar screen size without telephone capabilities as these could not be distinguished based on the data available.

The following five types of data are available for analysis, providing a rich set of auxiliary variables for analysis: (a) survey responses provide substantive information about the respondents' characteristics and attitudes; (b) survey responses can be used to compute response behaviours such as counts of 'don't know' responses; and (c) survey process data provides information such as the nature of device used to respond to the survey, the date and time the survey was carried out, and time stamps at the start and end of each module revealing survey length (more detailed data about time taken was not collected). In addition, all survey records are linked to (d) geographical data, based on the pupil's home postcode, following a procedure which ensured the anonymity of respondents. Linkage to geographical data is based on the postal address used to issue survey

invitations. Permission is not required for this type of linkage, and it was conducted in a manner which ensured anonymity. In practice, 4 sample members were not successfully linked to the geographical data and are excluded from the analytic sample, further reducing the available sample from 4,072 to 4,068. Finally, (e) survey participants were asked for consent for their survey data to be linked to administrative records; for the 83.2% of respondents who agreed, data is available about the young person and their school from a combination of NPD and ILR[3] (referred to here and in the tables below as ADMIN). Several of these data types, such as the geographical and administrative data, were collected independently of the actions of the respondent or the type of device that they used, and so are exogenous to the mode of data collection.

Excluding those who used non-standard devices (n=9) or who could not be linked to geographical data (n=4), leaves a maximum sample size available for analysis of 4,068 individuals. There are two important considerations about how the analytical sample can be defined. This is important because it has a direct effect on the level of missingness in the analytical data set and impacts on which variables are available for analysis.

The first consideration is whether the analysis should include "all devices", i.e. PC, smartphone and tablet responders, maximising the size of the sample and minimising missingness, which may not be at random, or whether tablet responders should be dropped (n=487), allowing the analysis to focus on a "restricted devices" sample of PC and smartphone responders only, which may provide a sharper contrast in terms of screen size, features and context (Clement et al., 2020).

Including tablet responders in the analysis reduces the risk that the results are biased by dropping a substantial minority of respondents who may not be like PC and smartphone responders. In practice, the boundaries between devices are somewhat blurred (Lugtig and Toepoel, 2016) and the available evidence about tablets is limited so further research on the behaviour of tablet responders has been encouraged (Clement et al., 2020, Wells et al., 2013) . If tablet responders are to be included, a decision is then needed whether they should be combined with smartphones or PCs. It has been argued that tablets are more akin to PCs (Couper et al., 2017, Peterson et al., 2017, Wells et al., 2013), but smartphones and tablets do have important similarities, since both are mobile devices

[3] The SET was sampled from administrative data, but only limited information was provided to Kantar, sufficient to issue invitations to eligible individuals, alongside a unique anonymised identifier which made it possible to attach additional administrative variables to survey records for those who subsequently gave consent to linkage. This meant that the linkage process of administrative to survey data was successful whenever permission was given and was generally of a very high quality. However, as noted later, some individual records had missing data for some administrative variables.

with touch screens. Furthermore, as will be discussed later (see Section 3.3.1), the evidence from this study shows that the characteristics of smartphone and tablet responders are similar in many respects (Lugtig and Toepoel, 2016, Clement et al., 2020, Wells et al., 2013)

Secondly, the analysis can be conducted with a sample which includes both those who consent and who do not consent to data linkage; or the sample can be restricted to those who consent to data linkage only, for whom a rich additional source of administrative data is available. However, as is shown later in this chapter (Section 3.3.2), the 16.8% of the sample (n=694) who refused consent differ significantly on all observed characteristics and dropping them from the analysis clearly risks introducing bias since they cannot be treated as missing at random.

The sample can, therefore, be defined in alternate ways with different sample sizes and risks associated with missingness: first, including or excluding tablet users and, second, including or excluding respondents who did not consent to data linkage. Limiting the sample in either way increases the risk that the analysis will subject to missingness which cannot be assumed to be at random. However, including .

Table 16 below shows the different sample sizes available for analysis, taking account of both of these considerations – which devices are included and whether administrative data is available – simultaneously. As will be explained in more detail in Section 3.3.3, the primary analysis presented in this chapter is drawn from the 'all devices' sample, to include tablet responders who are similar to smartphone responders in most respects, but is limited to 'consenters' in order to have the widest array of variables possible (i.e., the primary analysis is based on a subsample of the 3,189 respondents shown in bold below).

To address the concerns about the possibility that cases are missing, but not at random, a series of sensitivity analyses are carried out to test the effect of defining the sample differently, based on the other cells shown in the table, with concomitant differences in the variables available for each respondent.

Table 16 Definition of primary analysis sample and samples for sensitivity analysis

Which devices are included ->	All devices sample (PC, Smartphone, and Tablet)		Restricted devices sample (PC and Smartphone only)	
Whether administrative -> data is available	Consenters and non-consenters (No ADMIN data)	Consenters only (ADMIN data)	Consenters and non-consenters (No ADMIN data)	Consenters only (ADMIN data)
Available sample	4,068	**3,189**	3,583	2,824

To this point, the focus has been on the main sources of missingness in the dataset due to non-consent to data linkage and, where tablet responders are excluded from the analysis, due to device chosen to respond to the survey. It should be noted that there is some additional missingness. This is either where administrative data is missing for some who consented to data linkage, or due to item non-response in the survey. This is summarised in an extension to this table in Section 3.3.3 (see Table 21). A visual representation of all sources of missingness in the data set are illustrated in the form of a flow diagram, showing where small numbers of sample members are lost for different reasons; this is provided for the 'all devices' sample in Figure 11 and for the 'restricted devices' sample of PC and smartphone responders in Figure 12, and the implications for analysis are considered further in Section 3.3.3.

Having described the data in some detail, the next sections describe the outcome measures and covariates (or potential confounders) derived for this analysis.

3.2.2 Outcome measures or indicators of data quality

In total, 12 outcome measures were identified, capturing different aspects of data quality, of which 11 data quality indicators form the focus of this analysis. These are listed in Table 17, and each is defined below. The twelfth outcome variable, breakoffs (Peytchev, 2009), can be seen as a measure of non-response, but here they are considered in terms of measurement error, as an indicator of data quality. This outcome is considered separately (for further details see Section 3.5) because most breakouts occurred too early in the survey for the individual to be included in the data set, so the variables needed for full analysis are not available. Furthermore, the occurrence of breakouts in this study is low, affecting just 1.2% of individuals who started the survey.

In Table 17, the left-hand column shows that all of the 11 main outcome variables, are, by definition, drawn from data collected at the time of the survey or 'treatment', that is, from the first three types of data described above, i.e. (a) direct responses to the survey questions, (b) computations from different types of survey response, and (c) survey process data. The right-hand side of the table

shows how they are grouped into three methodological categories: (i) engagement in broader aspects of the research, (ii) satisficing behaviours, and (iii) temporal aspects of participation. Each of the 11 outcomes are defined below and their frequency is shown in weighted percentages and unweighted numbers, based on the analytical sample size of 4,068 individuals.

Table 17 The three groups of outcome variables and their data source

Type of data	Outcomes (indicators of data quality)	Category
(a) Survey responses	1. Refuses data linkage 2. Refuses findings 3. Refuses recontact	(i) Engagement in broader aspects of the research
(b) Computations derived from survey responses	4. 'Don't know' responses 5. 'Don't know' responses to quiz 6. Straightlining 7. Agreement	(ii) Satisficing behaviours
(c) Survey process data	8. Speeding 9. Interruptions 10. Slow first module 11. Completion time (minutes)	(iii) Temporal aspects of participation

Note: Definitions of the variables and how they were derived are outlined below

Engagement in broader aspects of the research

Indicators of participant engagement with the research project, measured in terms of consent to supplementary requests, are rarely considered in the literature around device effects (Cunningham et al., 2013, Matthews et al., 2017) but are important. There are three measures of engagement in this study: agreement to data linkage which facilitates substantive and methodological analysis, agreement to receive findings which supports research impact, and consent to recontact, which is a precondition for inviting participation in future research.

1) Refuses consent to data linkage: During the survey respondents were told "The Department for Education holds information about your education. This includes the schools you've been to, the subjects and exams you've done, if you have a special educational need and if you have been eligible for free school meals." They were then asked, "Can we have your permission to link this information to your survey answers?" and are coded 0=consented or 1=refused. It should be noted that it was not possible to give partial consent – there is one administrative record for each respondent. Overall, 18.7% refused consent to data linkage (n=694).

2) Refuse to receive findings: Respondents were asked, "Would you like us to send you the findings of this survey when they are published next year?" and are coded 0=consented or 1=refused. Overall, 25.7% refused to receive results (n=989), though the response is missing in 4 cases.

Figure 11 Sample numbers and missingness for the 'all devices' sample

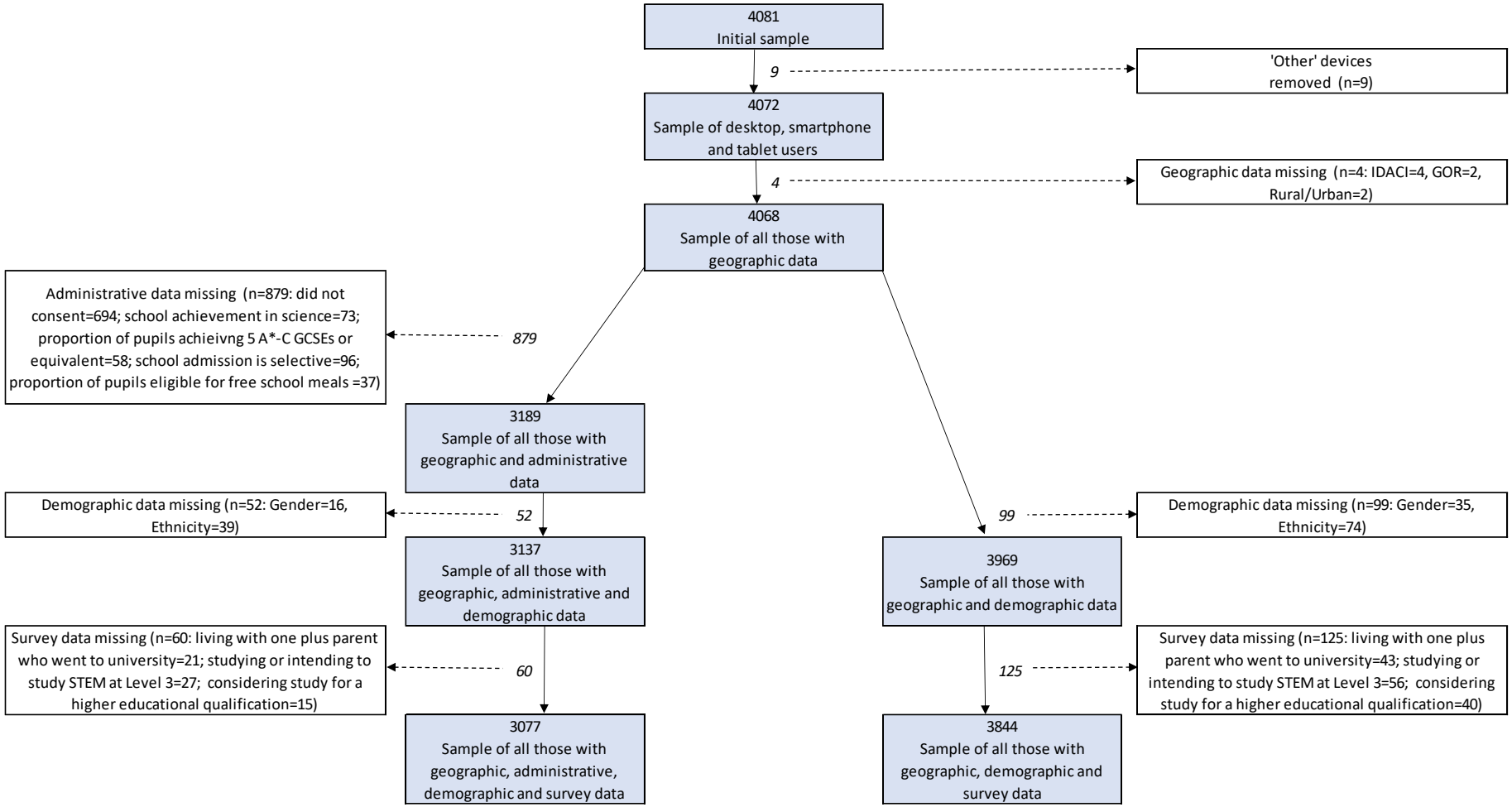
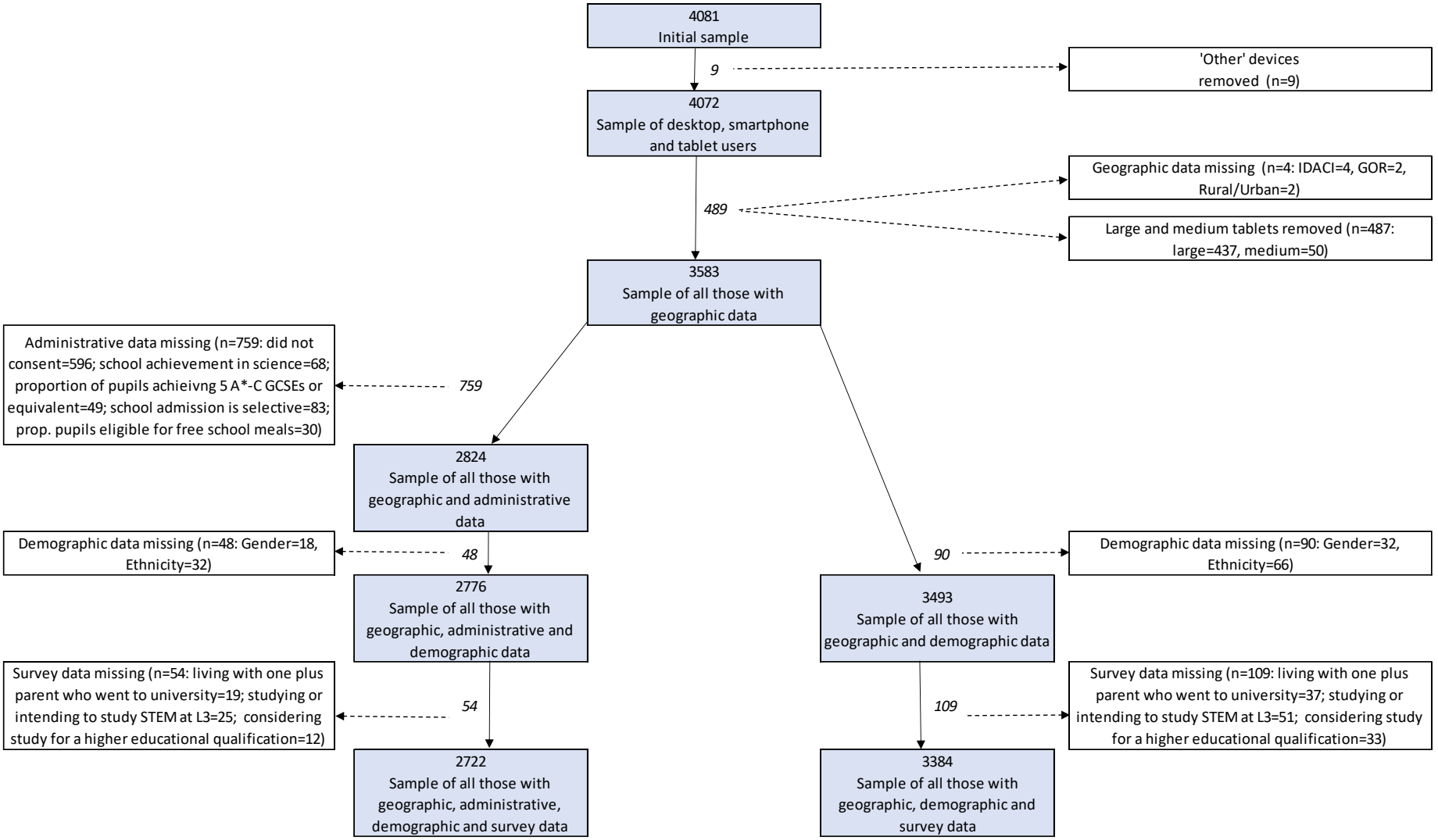


Figure 12 Sample numbers and missingness for the restricted devices sample (no tablets)



3) Refuses consent to recontact: Respondents also saw this text: "The Wellcome Trust may want to do follow up interviews with some people who have taken part in this study. In case they wanted to contact you, would it be alright for us to pass on your details to them – that is, your name, address, email address, telephone number and some of the answers you have given today?". Responses are coded 0=consented or 1=refused. Overall, 41.1% refused (n=1602); there are no cases where data is missing.

Satisficing behaviours

Item non-response reduces the sample available for analysis and, unless missing at random, may introduce bias and result in misleading estimates; two measures of item non-response are calculated using responses to survey questions. Two other measures of satisficing are also derived: nondifferentiation (or straightlining) and agreement.

4) 'Don't know' responses: The number of times a participant responded with "don't know" is counted across all survey questions, in both single-choice and multiple-choice questions, but excluding ten questions which constituted a science quiz, with those below the 90th percentile coded as 0=low item non-response and those at or above coded as 1=high. A high level of 'don't know' responses are observed in 11.5% of the sample (n=407); there are no missing cases. A measure which counted both 'don't know' responses and refuse responses was also considered, but the outcomes are so closely correlated (correlation of continuous variable 0.9244, tetrachoric correlation of binary variable 0.9785) that just one was selected.

5) 'Don't know' responses to science quiz: The number of 'don't know' responses to a ten-question science quiz were counted, with those with 0, 1 or 2 'don't know' responses coded as 0=low item non-response and those with 3 or more 'don't know' responses (where a shift was observed in the distribution) counted as 1=high. High levels of 'don't know' responses to the quiz questions are observed in 17.0% of the sample (n=647).

6*) Non-differentiation or straightlining:* All SET respondents were asked a single set of attitude questions. They were asked "How much do you agree or disagree with the following statements?" and were presented with the phrase "Careers that use science….", followed by seven statements: "are suitable for someone like me", "are difficult to get into", "require high grades", "are boring", "are more suited to men than women", "are open to anyone who has the ability regardless of their background" and "make a useful contribution to society". Respondents were offered a five-point Likert Scale ranging from strongly agree to strongly disagree (Likert, 1932)., where respondents fail to differentiate among their responses when answering items in a battery, is a recognised aspect of

satisficing, but there is no agreed standard technique for identifying this behaviour (AAPOR, 2010). Kim et. al., identify five distinct indices, each measuring a slightly different aspect of satisficing which are highly correlated (2019). The index chosen here – scale-point variation, based on the probability of differentiation (Linville et al., 1986) – measures the dispersion of ratings for each respondent by taking all rating scales into account and is better at detecting variation across rating scales than some other approaches. The probability of differentiation or P_d is defined as:

$$P_d = 1 - \sum P_i^2$$

where p_i is the proportion of the values rated at each scale point on a rating scale, and i indicates the number of scale points. If respondents use more scale points within a battery, the measure becomes larger, so that higher scores indicate less straightlining (Kim et al., 2019). This was implemented using the `respdiff` command in STATA 15.0 (Roßmann, 2017). Overall, 7.7% of respondents are identified as straightlining based on a threshold of 0.75 (n=311).

7) Agreement: The same set of attitude questions are used to identify the tendency towards agreement based on a threshold of 0.5, derived from the pattern of responses to successive questions. Overall, 5.7% of respondents (n=243) are identified as showing high levels of agreement.

Temporal aspects of participation

Behaviours such as very fast response times may indicate that a respondent is not fully engaged in the survey process, as may a very slow start to the survey or long interruptions during the survey. Slower response times may indicate that the survey is burdensome and could be associated with additional breakoffs. In the absence of more detailed information from survey process data, time stamps at the start and end of each module are used to derive the following four outcome measures.

8) Speeding: The average time a respondent took to answer each survey question is calculated by dividing the total time taken to answer the survey by the number of questions the respondent answered, derived from counts of survey responses. Participants who took less than 0.21 minutes per question are coded as 1=speeding, representing the quickest 10th percentile of respondents (n=406). All other respondents are coded as 0.

9) Interruptions: Respondents are coded as 1, having an interruption to their survey, if they took over 30 minutes to answer one or more modules. All other respondents are coded as 0. This threshold was chosen because it was close to the 95th percentile and corresponds with a shift in the distribution of long module durations. This amount of time, which represents one and a half times

the average duration of the full survey, is far in excess of the time that would reasonably be taken by a respondent to answer the questions contained in any module. Following this definition, 3.6% of respondents are identified as having an interruption to their survey (n=153).

10) Slow start: Respondents are identified as making a slow start to the survey and coded as 1 if they took 3 minutes or longer to complete the first module. All other respondents are coded as 0. The three-minute threshold corresponds with a shift in the distribution of start times. In total, 7.8% of young people who responded made a slow start (weighted percentage, unweighted n=307).

11) Overall completion time: The number of seconds between the time at the start of the first module and the time at the end of the last module of questions was calculated. Individual modules are right censored at the 99.5th percentile to remove a very small number of interviews with extreme lengths. The average time taken was 20.1 minutes (mean) or 19 minutes (median).

3.2.3 Covariates or potential confounder variables

Table 18 below lists the covariates included in the analysis. As discussed in the methodology (Section 3.3), these can be considered potential confounder variables. They are wide ranging and touch on the respondent's characteristics with respect to their socio-demographic and economic position, their local and school context, the possible salience of science and education in their lives, as well as a measure of their enthusiasm to participate in the survey. The left-hand column of the table shows that, in contrast to the outcome variables, most of the covariates are drawn from: (c) survey process data, (d) geographic data and (e) administrative data, that is, information collected independently of the survey or 'treatment', so avoiding the risk of endogeneity. However, some covariates are only available from (a) direct responses to survey questions, and so are at risk of endogeneity.

The right-hand column in Table 18 groups these variables into four sets. The first of these are variables available immediately prior to participation (PRE), which includes lapsed time from invitation to participation based on survey process data, and geographical variables based on where the sample member lives. The second of these is administrative data about the pupil and school available independently of the survey response, although conditional on agreement to link (ADMIN). The third set is demographic variables which were collected during the survey (DEM). Theoretically these variables could have been drawn from the NPD/ILR administrative data. Indeed, administrative data for year group and sex were used for sampling purposes; the population was divided by year group and sex was used as a stratifier, alongside science performance at school, overall performance at school, establishment type, region, and the Income Deprivation Affecting Children Index (IDACI).

However, administrative data for year group, sex and ethnicity were not made available for the purpose of analysis, even for respondents who gave consent to linkage. Instead, all respondents were asked these questions directly. An indirect result of this was that it was not possible to check the quality of the data linkage by examining common variables. The final set were also collected from the survey and capture interest in science and education and parental university experience (SUR). These four variable sets, PRE, ADMIN, DEM, and SUR, provide the building blocks for the matching methodology that is described in Section 3.3.

Table 18 Sets of covariates and their data sources

Data source	Covariates	Variable set
(c) Survey process data	Time from invite to response	Pre-participation (PRE)
(d) Publicly available geographical measures	Income Deprivation Affecting Children Index (IDACI) Government Office Region (GOR) Rural/Urban	
(e) Administrative data from the NPD and ILR	Special Educational Needs (SEN) status Free School Meal (FSM) status Highest science attainment School attainment 5A*-C GCSE/equivalent School admission policy Prop school Free School Meal (FSM) eligible	Administrative (ADMIN)
(a) Survey responses	Gender Year group Ethnicity	Demographic (DEM)
	Living with a parent who attended University Studies/intends to study L3 Maths/Science Considering higher educational qual	Survey (SUR)

Note: None of the covariates are drawn from source (b) computations derived from survey responses.

In the descriptions of the covariates below, the unweighted frequencies and weighted percentages are given based on the full analytical sample (n=4,068). More detailed information about the distribution of each variable is tabulated in Table 20**Error! Reference source not found.** (left-hand columns, heading highlighted in orange).

3.2.3.1 Collected immediately prior to participation (PRE)

This variable set includes a geographical indicator of socio-economic deprivation, two geographical markers, and a measure of propensity to participate. As explained earlier, 4 cases with missing values for the geographical data were dropped from the dataset during preliminary data cleaning.

Income Deprivation Affecting Children Index (IDACI): IDACI is one element of a wider Indicator of Multiple Deprivation (IMD). It is used in education research in the UK because it measures the proportion of all children aged 0 to 15 living in income deprived families in different local areas, and so is more specific than a more general measure of deprivation. IDACI is coded by quintile where 1=the most advantaged areas and 5= the least advantaged areas.

Government Office Region (GOR): This is coded by standard region 1=East Midlands, 2=East of England, 3=London, 4=North East, 5=North West, 6=South East, 7=South West and Wales, 8=West Midlands, 9=Yorkshire and Humber.

Rural/urban: This is an indicator of population density (coded 1=urban conurbation, 2=urban city and town, 3=rural).

Lapsed time to response: Survey process data is used to derive a measure of the respondent's engagement in the survey, or their propensity to participate. This is derived from the number of days between the invitation to participate and the date of online participation. This is coded as 1=responded between 0-13 days of the invitation, that is, before a reminder had been sent, 2=responded after the first reminder (14-28 days), 3=responded after the second reminder (29-59 days). Three quarters of the sample (76.4%) responded before the first reminder, with a further 15.4% responding after the first reminder and 8.2% responding after the second reminder. There is no missing data.

3.2.3.2 Administrative variables at person-level and school-level (ADMIN)

These variables provide information about the young person's individual status (i.e., whether they have special educational needs or are from an economically disadvantaged background), the likely salience of science in their lives (in terms of the individual's highest educational attainment in a science subject) and the educational support they are likely to derive from their school (measured by academic attainment at the school level, whether the school entry is selective, and whether the school faces the additional challenges of teaching an economically disadvantaged population). This data is unavailable for 694 respondents who did not consent to linkage.

Special Educational Needs (SEN): Administrative data provided information about whether the young person had been identified as requiring SEN support (coded as 0=no SEN, 1=some SEN provision). Overall, 13.8% of respondents have SEN status.

Free school meals (FSM): While IDACI provides a measure of socio-demographic deprivation in the local area, FSM provides a measure of the individual's current or past household deprivation.

Following common practice in education research in the UK, the measure used here is derived from two indicators, whether the individual was eligible for free school meals at that time (FSM), or in the previous six years (FSM6). In the policy environment, FSM status determines eligibility for the school to receive an additional 'Pupil Premium'. FSM status is coded as 0=not eligible for FSM now or in last 6 years, and 1= eligible for FSM now or in the last 6 years. Overall, 22.0% of respondents have FSM status.

Individual science attainment: An indicator of the individual's science performance at school is available from administrative data. For younger students in Years 10 and 11, this is based on Key Stage 2 (KS2) teacher assessed science level (level 3 or under, level 4 or level 5). These are assessments made at the end of junior school before the transition to senior school. For older students in Years 12 and 13, this is based on Key Stage 4 (KS4) science results. These are the examinations at the end of the first three years of senior school, that precede A levels or equivalents. The variable is coded as high (KS2=level 5 or KS4=two or more science GCSEs or equivalents at A*-B), medium (KS2=level 4 or KS4= two or more science GCSEs or equivalent at A*-C but not A*-B) or low (KS2=level 3 or under, KS4=not having two science GCSEs or equivalent at A*-C). This measure is not available for 767 sample members, 694 who did not consent to linkage and 73 who gave consent but for whom this administrative data is missing. Of those with valid data, 38.3% had the highest category of science attainment, 41.0% had medium level attainment and 20.7% had the lowest level of attainment.

School academic attainment: A measure of school attainment is based on the proportion of pupils at the individual's school achieving at least five GCSEs (or equivalent) at A*-C including English and Maths, with respondents in schools in the lowest quintile of school attainment coded as 1 and all others coded as 0. This measure is not available for 752 sample members, 694 who did not consent to linkage and 58 who gave consent but for whom this administrative data is missing.

School admissions policy: The admissions policy of the individual's school is coded as 0=comprehensive/modern, 1=selective. This measure is not available for 790 sample members, the 694 who did not consent to data linkage and an additional 96 who gave consent but for whom this administrative data is missing. Of those with valid data, 5.8% of respondents attended selective schools.

Eligibility for FSM in school: A measure of economic disadvantage in the student population is the proportion of pupils at the individual's school eligible for free school meals[Error! Bookmark not defined.], with respondents attending schools with the highest prevalence of FSM eligibility coded as 1 and all

others coded as 0. This data is not available for 731 sample members, 694 who did not consent to linkage and 37 who gave consent but for whom this administrative data is missing.

3.2.3.3 Demographic variables taken from survey data (DEM)

In addition to SEN which is derived from person-level administrative variables, three aspects of the individual's personal characteristics were included. In theory, these variables may be subject to endogeneity, but are not likely to have been substantially affected by the device used to report them.

Gender: Young people were asked to record their gender (coded as 0=male, 1=female). A total of 35 individuals did not respond to this question and are treated as missing. The sample was weighted to be representative of the school population.

School year group: Respondents were asked "Which academic year have you been in this past year?" or, if they were no longer in school or college, which year they would have been in if they were still studying. School year group is coded as Year 10=1, Year 11=2, Year 12=3 and Year 13=4. This acts as a proxy for age. The data was weighted with approximately equal proportions in each of the four year-groups. Neither age nor date of birth were collected directly to increase confidence in the privacy of responses, unless the respondent refused consent to linkage, in which case they were asked for age, qualifications and postcode.

Ethnic group: Respondents were asked to describe their ethnic group or background based on the ONS harmonised question. This is recoded into five summary categories, White=1, Mixed=2, Asian=3, Black=4, Other=5. This data is missing for 74 individuals.

Two other aspects of the respondent were considered but ultimately were not used in the analysis: household composition, derived from a set of questions about who the respondent lives with, and religion.

3.2.3.4 Survey variables (SUR)

The final three measures taken from survey responses capture interest in science and education, so may indicate the salience of the survey, and record whether the young person's parent had attended university, which can also be seen as a proxy for socio-economic position. Since these factors may have a bearing on sample members' interest in, and method of responding to the survey, they should be considered as potentially endogenous.

Studying maths or science at Level 3: Young people in Years 10 and 11 were asked whether, after year 11, they were planning to study for further qualifications at Level 3 (such as A levels or a National Vocational Qualification Level 3, in other words in academic qualifications taken at the end of senior school at approximately 18 years old. If so, they were asked how likely they were to study maths, biology, chemistry, physics, computer science or another science subject such as psychology, engineering, geology, or applied science. Students in Years 12 and 13 were asked equivalent questions to ascertain whether they were already studying those subjects. A combined variable, Level 3 science, is derived and coded 1=studying/plans to study science subjects at Level 3, 2=not studying/planning to study at Level 3, 3=undecided. This data is missing for 66 individuals. Of those with valid data, those studying or planning to study maths or science at Level 3 accounted for 43.9% of young people, those who did not accounted for a further 47.5% and 8.6% were undecided.

Higher education qualification: All students were asked whether they were thinking about going on to study for a higher education qualification in any subject with the response categories 1=Yes, a university degree, 2=Yes, another HE qualification (e.g., a Higher National Certificate, Higher National Diploma, Higher Education Diploma), 3=No, 4=Undecided. The 49 who responded 'prefer not to say' are coded as missing. Of those with valid data, 50.5% were considering a university degree, 6.5% a higher educational qualification, 14.3% were not intending to pursue either, and 29.1% were undecided.

Parents went to university: Respondents were asked whether one or more of the parents, foster parents, or parent's partners that they were living with (the questions were tailored to their household circumstances) had gone to university. Responses are coded as 1=one or more parents/parent figures attended university, 2=no parent/parent figure had attended university, 3=don't know. This data is missing for 57 young people. Less than a third were living with a parent figure who had attended university (31.6%) while the remainder did not or were uncertain.

As mentioned earlier, the main sources of missingness in the dataset are due to non-consent to data linkage and due to the analytical decision whether to exclude tablet responders from the analysis, but there is some additional missingness, both among those who consented to data linkage for whom some administrative data is missing, and due to item non-response in the survey (see Table 21 and the flow diagrams for 'all devices' in Figure 11 and for 'restricted devices' in Figure 12). The magnitude of missingness which results from these additional sources is relatively small. For example, based on the primary analysis sample (n=3,189) in Figure 11, just 52 individuals have missing data for demographic variables (gender=16, ethnicity=39), while 60 have missing data due to other item non-response to the survey (studying maths or science at Level 3=27, higher educational

qualification=15, parents went to university=21). Given the low numbers involved, the assumption made is that this level of additional missingness can be treated as missing at random.

3.3 Statistical methodology and analysis

This section first presents evidence that those who choose to respond on a mobile device differ from those who chose to respond using a PC, which necessitates the use of quasi-experimental methods to compare response behaviours using the 11 outcome codes described (Section 3.3.1), and controlling for covariates which potentially confound the relationship between device and outcome because they also influence selection (and hence are described as confounder variables). The method of analysis is described below – first for the primary analyses and then for a series of sensitivity analyses – using different definitions of the sample and different sets of covariates or potential confounder variables (Section 3.3.2). Finally, the effect of matching on rebalancing the samples is summarised (Section 3.3.4). This section focuses on the analytical method, but also includes preliminary analyses based on the SET data to understand the rationale for, and the effect of, matching. However, the substantive results of the research are not presented until Section 3.4.

3.3.1 Differences between PC and mobile device responders

Preliminary analysis shows that there are significant differences between survey responders who use a mobile device or PC for many characteristics, which supports the findings of other studies, although no significant differences were found for a few characteristics (such as school year group, ethnicity, SEN status and whether they are living in a rural or urban area). For example, in terms of demographic characteristics, while 49.0% of the sample are girls (as shown in Table 20, based on weighted data for the full sample where n=4,068), this differs markedly by device: 53.9% of mobile device responders are girls compared to just 46.0% of PC responders ($p<0.000$) (see the comparison in Table 19, left-hand columns).

The differences found are not just based on demographic variables, but multiple measures speak to a common theme, which is that young people who respond by mobile device are more disadvantaged than PC responders and are less likely to be engaged in education and science. A higher proportion of mobile device responders come from the most disadvantaged areas measured by IDACI (27.2% compared to 19.9% of PC responders). They are less likely to be living with at least one parent who went to university (23.1% compared to 36.9% of PC responders), are less likely themselves to be planning to study for a university degree (42.3% compared to 55.6%) and are less likely to be doing, or planning to do, Level 3 science (36.5% compared to 48.5% of PC responders).

For those for whom administrative data is available, objective measures support these findings: for example, it shows that young people who respond by mobile device are also more likely to have low past attainment in science (25.8% compared to 17.8%) and a higher proportion of them are eligible for Free School Meals (26.9% compared to 19.1%). Furthermore, their school context is also one of relative disadvantage: they are more likely to attend schools with high levels of students who are eligible for FSM (24.2% compared to 17.7%); they are more likely to attend schools which have lower attainment (23.6% compared to 18.2%); and they are less likely to attend selective schools (3.8% compared to 7.0%).

Returning to the comparison of mobile and PC responders, two conclusions can be drawn. First, there is clear evidence of differences in the characteristics of mobile device and PC responders, which confirms that a simple comparison of the response behaviours of those who chose to use different devices may suggest a measurement effect, which in practice is either wholly or partially confounded by selection. Secondly, these differences are not just demographic, but relate to the young person's socio-economic position and their experience of and attitude towards science and education. This increases the likelihood that the differences in propensity to respond using different devices will confound any comparison between devices. Indeed, a further indicator of differences in their propensity to participate is that young people who respond using a mobile device are more likely to have required a reminder before taking part (28.3%) than those who responded using a PC (20.6%). Therefore, a reasonable hypothesis is that matching solely based on the demographic characteristics of respondents is unlikely to fully account for selection in ways which are relevant to the theme of the Science Education Tracker – attitudes to science and education. In contrast, a similar comparison of smartphone responders and tablet responders (see Table 20, three final columns) shows that these two groups are *not* significantly different based on most characteristics, with the exception of IDACI, Free School Meals and whether the respondent's parents attended university. This provides some justification for the argument given earlier (in Section 3.2.1) that even though tablet and PC responders may be similar in several respects, in this study, it is reasonable to combine smartphone and tablet responders into a single group of mobile responders.

Table 19 Sample characteristics, and primary analysis sample before and after matching

		Full sample (n=4068)		**Primary analysis (all devices, consenters) PRE+DEM+ADMIN (n=3137)**			
		Weighted		Weighted		Matched	
		PC	Mob dev	PC	Mob dev	PC	Mob dev
PRE	**IDACI**	%	%	%	%	%	%
	1st (Advantaged)	20.9	14.6	22.2	16.1	17.6	17.7
	2nd	20.3	17.2	20.9	18.2	19.6	19.4
	3rd	18.9	19.2	18.9	19.4	19.9	19.9
	4th	20.1	21.8	19.3	21.2	20.3	20.0
	5th (Disadvantage)	19.9	27.2	18.8	25.1	22.6	23.0
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.999
	Region	%	%	%	%	%	%
	East Midlands	8.5	9.5	8.7	9.7	9.5	9.7
	East of England	11.8	11.8	11.8	12.3	12.6	12.7
	London	15.1	11.3	14.2	10.5	10.3	10.5
	North-East	4.8	5.4	5.0	5.4	5.1	5.1
	North-West	13.5	14.4	13.2	13.3	13.0	12.8
	South-East	17.3	15.8	17.7	16.2	16.4	16.4
	SW & Wales	10.5	7.9	11.2	9.0	9.3	9.4
	West Midlands	10.0	12.0	9.2	11.1	11.1	11.3
	Yorks. & Humber	8.5	11.8	9.1	12.5	12.7	12.1
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.006**		1.000
	Rural/Urban	%	%	%	%	%	%
	Urban conurbation	37.2	36.0	35.3	33.7	32.8	34.0
	Urban city/town	44.9	47.0	45.3	47.8	47.9	46.6
	Rural	17.9	16.9	19.4	18.5	19.3	19.4
	Total	100	100	100	100	100	100
	p-value		0.428		0.422		0.756
	Days from invite	%	%	%	%	%	%
	Pre reminder 0-13d	79.4	71.7	81.1	74.3	76.8	76.8
	Post reminder 1	14.0	17.6	13.6	16.8	15.4	15.6
	Post reminder 2	6.6	10.7	5.3	8.9	7.8	7.6
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.987
ADMIN	**SEN status**	%	%	%	%	%	%
	No SEN	86.6	85.4	87.4	87.3	89.5	88.8
	Some SEN	13.4	14.6	12.6	12.7	10.5	11.2
	Total	100	100	100	100	100	100
	p-value		0.377		0.951		0.596
	FSM status	%	%	%	%	%	%
	Not FSM	80.9	73.1	81.0	73.6	75.9	75.4
	FSM or FSM6	19.1	26.9	19.0	26.4	24.1	24.6
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.763
	KS2/KS4 science	%	%	%	%	%	%
	High	43.7	29.2	44.5	30.0	33.9	33.8
	Medium	38.6	45.1	38.6	45.8	46.9	46.3
	Low	17.8	25.8	16.9	24.2	19.2	19.9
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.909

		PC	Mob dev	PC	Mob dev	PC	Mob dev
ADMIN Continued	**School attainment**	%	%	%	%	%	%
	High	81.8	76.4	83.0	78.4	80.0	79.7
	Low	18.2	23.6	17.0	21.6	20.0	20.3
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.003**		0.863
	School selection	%	%	%	%	%	%
	Comp/Modern	93.0	96.2	92.9	96.2	95.5	95.5
	Selective	7.0	3.8	7.1	3.8	4.5	4.5
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.997
	School FSM	%	%	%	%	%	%
	Low	82.3	75.8	83.1	77.6	80.0	78.8
	High	17.7	24.2	16.9	22.4	20.0	21.2
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.437
DEM	**Gender**	%	%	%	%	%	%
	Male	54.0	46.1	51.3	44.0	41.1	41.1
	Female	46.0	53.9	48.7	56.0	58.9	58.9
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.995
	School year group	%	%	%	%	%	%
	Year 10	24.7	22.3	25.0	22.9	26.4	26.0
	Year 11	25.3	24.7	24.4	26.0	28.9	27.6
	Year 12	25.2	25.6	26.2	25.9	23.0	24.1
	Year 13	24.9	27.3	24.4	25.2	21.7	22.3
	Total	100	100	100	100	100	100
	p-value		0.227		0.535		0.824
	Ethnicity	%	%	%	%	%	%
	White	77.2	80.7	79.1	82.3	82.5	82.8
	Mixed	4.6	4.5	4.6	4.9	4.6	4.7
	Asian	12.3	10.0	10.8	8.2	8.1	7.8
	Black	4.8	3.7	4.6	3.5	3.6	3.6
	Other	1.1	1.1	0.9	1.1	1.2	1.1
	Total	100	100	100	100	100	100
	p-value		0.102		0.120		0.998
SUR	**Level 3 science**	%	%	%	%	%	%
	Doing or planning	48.5	36.5	49.9	38.8	45.5	41.2
	Not doing /plan	43.6	53.9	42.6	52.8	46.7	51.0
	Undecided	8.0	9.6	7.4	8.4	7.7	7.8
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		0.068
	Higher ed plan	%	%	%	%	%	%
	University degree	55.6	42.3	58.2	46.3	55.2	48.8
	Higher education	5.3	7.1	5.6	6.6	5.9	6.2
	Undecided	27.6	31.7	26.7	31.2	29.5	30.8
	No	11.5	18.9	9.4	15.9	9.5	14.3
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		**0.000**
	Parent to Uni	%	%	%	%	%	%
	1+ parent to Uni	36.9	23.1	37.9	25.7	34.0	27.0
	No parent to Uni	55.7	69.5	55.5	68.5	59.6	67.2
	Don't know	7.4	7.3	6.6	5.8	6.4	5.8
	Total	100	100	100	100	100	100
	p-value		**0.000**		**0.000**		**0.000**

Notes(1) The primary analysis sample includes responders using all devices who consented to data linkage. (2) See Table 20 for detailed information on missingness. (3) Totals may not add to 100 due to rounding.

3.3.2 Differences between consenters and non-consenters

Before setting out the matching methodology, it is important to extend the comparison of respondent characteristics, and, on this occasion, to examine the differences between those who do and do not consent to data linkage. This is important because of the potential bias introduced to the analysis if non-consenters are excluded, because they do not have the benefit of linked administrative data. This potential for bias is confirmed by looking at the differences in the characteristics between consenters and non-consenters for a few example variables, though detailed information for all covariates is provided in Table 20. For example, boys constitute 49.0% of consenters but they make up 59.5% of non-consenters. The oldest age group of students make up one-quarter of consenters (24.6%) but almost one third (31.2%) of non-consenters. Perhaps more importantly, non-consenters are likely to be less engaged in education; 33.8% of consenters say that one or both parents attended university compared to 21.9% of non-consenters, and 53.5% intend to go onto higher education compared to 36.9% of non-consenters. In fact, the distribution of all covariates differs significantly for data linkage consenters and non-consenters. It is not appropriate to assume that the consenting sample is unbiased, but this can be addressed using the study methodology.

3.3.3 Matching method, primary analysis, and sensitivity analysis

In the absence of an experimental design, quasi-experimental methods are needed to control for selection effects *ex post facto*. Theoretically, quasi-experimental methods would not have been necessary if SET respondents had been randomly allocated to device. However, in such a survey setting, random allocation is not normally possible. Methodological research was not the focus of the SET survey and there is no reason to expect that consideration would have been given to the idea of random allocation to device to identify measurement effects. Furthermore, many students would not have had access to both types of devices, and resource was not available to distribute devices, so a request of this kind would have significantly impacted the study.

The method used to do this is similar to the approach taken by Matthews et al., This simplifies the traditional matching process by using a single score based on propensity to be treated and matching using this score to make the profiles of PC and smartphone responders as similar as possible to each other (Matthews et al., 2017). With this approach, the survey is seen as the intervention or treatment. A propensity score is calculated which is the probability of being exposed to a treatment (i.e., use of a mobile device to respond to a survey), conditional on a set of observed baseline characteristics for any given individual, where 0 indicates that the event will not happen

Table 20 Characteristics of the different samples

	All devices (PC, Smartphone + tablet) n=4068		Compare consenters and refusers (n=4068)				Reduced devices (PC + smartphone) n=3583		Comparison of device users (n=1498)				
			Consenters only n=3374		Refusers only n=694		Sig		Smartphone (or small tablet) n=1013		Tablet (medium or large) n=485		Sig
PRE													
Income Deprivation Affecting Children Index (IDACI)	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
1st quintile – most disadvantaged	824	18.5	725	19.8	99	12.9		722	18.4	142	12.3	102	19.4	
2nd	826	19.1	701	19.5	125	17.3		729	19.2	180	16.8	97	18.2	
3rd	785	19.0	645	18.9	140	19.5		681	18.7	192	18.2	104	21.5	
4th	796	20.7	645	20.2	151	22.9		706	20.9	221	23	90	19.4	
5th quintile – least disadvantaged	837	22.6	658	21.5	179	27.5		745	22.8	278	29.8	92	21.4	
Total	**4068**	**100**	**3374**	**100**	**694**	**100**	**0.000**	**3583**	**100**	1013	100	485	100	0.000
Government Office Region (GOR)*	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	N	W %	N	W %	**Sig**
East Midlands	371	8.9	307	9	64	8.5		324	8.9	99	9.8	47	9.1	
East of England	491	11.8	411	12	80	11.2		435	11.9	128	12.1	56	11.2	
London	550	13.7	434	12.9	116	17.1		498	14.1	117	11.6	52	10.8	
North-East	196	5.0	167	5.2	29	4.0		178	5.1	60	6	18	4	
North-West	543	13.8	439	13.5	104	15.1		476	13.7	140	14.2	67	14.9	
South-East	692	16.8	593	17.3	99	14.5		605	16.7	150	15.1	87	17.3	
South-West and Wales	400	9.5	352	10.1	48	7.1		356	9.6	80	7.5	44	8.8	
West Midlands	434	10.8	340	9.9	94	14.3		379	10.7	125	12.3	55	11.4	
Yorkshire and Humberside	391	9.8	331	10.1	60	8.4		332	9.4	114	11.4	59	12.6	
Total	**4068**	**100**	**3374**	**100**	**694**	**100**	**0.000**	**3583**	**100**	1013	100	485	100	0.767

Table 20 continued…	All devices		Consenters		Refusers		Sig	Reduced dev		Smartphones		Tablets		Sig
Rural / Urban*	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Urban conurbation	1472	36.8	1170	35.2	302	36.8		1313	37.2	378	37.2	159	36.0	
Urban city and town	1840	45.7	1545	46.3	295	45.7		1616	45.5	467	47.0	224	47.0	
Rural (town and fringe/village)	756	17.5	659	18.5	97	17.5		654	17.3	168	15.8	102	16.9	
Total	**4068**	**100**	**3374**	**100**	**694**	**100**	**0.000**	**3583**	**100**	1013	100	485	100	0.154
Lapsed days (invitation to response)	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Before first reminder	3196	76.4	2693	78.2	503	68.9		2828	77.0	753	71.4	368	72.3	
After first reminder	581	15.4	467	14.7	114	18.3		499	15.0	161	17.2	82	18.5	
After second reminder	291	8.2	214	7.1	77	12.9		256	8.1	99	11.4	35	9.2	
Total	**4068**	**100**	**3374**	**100**	**694**	**100**	**0.000**	**3583**	**100**	1013	100	485	100	0.481
ADMIN VARIABLES – INDIVIDUAL														
Special Educational Needs status (SEN)	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
No SEN provision	2974	86.2						2635	86.2	710	85.1	339	86.1	
Some SEN provision	400	13.8						352	13.8	105	14.9	48	13.9	
Total	**3374**	**100**						**2987**	**100**	815	100	387	100	0.665
Missing	694							596		198		98		
Free school meals status (FSM)	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Not receiving FSM now or in past 6 years	2691	78.0						2379	77.8	588	70.1	312	79.7	
FSM now or in past 6 years	683	22.0						608	22.2	227	29.9	75	20.3	
Total	**3374**	**100**						**2987**	**100**	815	100	387	100	0.001
Missing	694							596		198		98		
Highest KS2/KS4 science achievement	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
High	1402	38.3						1253	38.7	242	26.5	149	35.1	
Medium	1345	41.0						1182	40.7	377	45.9	163	43.2	
Low	554	20.7						484	20.6	181	27.6	70	21.7	
Total	**3301**	**100**						**2919**	**100**	800	100	382	100	0.008
Missing	767							664		213		290		

Table 20 continued…	All devices		Consenters		Refusers		Sig	Reduced dev		Smartphones		Tablets		Sig
ADMIN VARIABLES - SCHOOL														
School attainment (Achieve 5 A*-C GCSE)	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Highest school attainment (4 top quintiles)	2698	79.8						2396	80.0	616	75.5	302	78.5	
Lowest school attainment (bottom qu'tile)	618	20.2						542	20.0	184	24.5	76	21.5	
Total	**3316**	**100**						**2938**	**100**	800	100	378	100	0.284
Missing	752							645		213		-326		
School admission policy	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Comprehensive/Modern/Non-selective	3060	94.2						2703	93.9	753	96.1	357	96.3	
Selective	218	5.8						201	6.1	36	3.9	17	3.7	
Total	**3278**	**100**						**2904**	**100**	789	100	374	100	0.854
Missing	790							679		224		-315		
Proportion of school pop eligible for FSM	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Lowest proportion of FSM (four quintiles)	2718	80.0						2410	80.0	611	74.2	308	79.5	
Highest proportion of FSM (most disadvantaged quintile)	619	20.0						547	20.0	196	25.8	72	20.5	
Total	**3337**	**100**						**2957**	**100**	807	100	380	100	0.058
Missing	731							626		206		105		
DEMOGRAPHIC VARIABLES (SURVEY)														
Sex	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Male	1923	51.0	1541	49.0	382	59.5		1702	51.2	415	44.8	221	49.0	
Female	2110	49.0	1813	51.0	297	40.5		1849	48.8	589	55.2	261	51.0	
Total	**4033**	**100**	**3354**	**100**	**679**	**100**	**0.000**	**3551**	**100**	1004	100	482	100	0.149
Missing	35		20		15			32		9		3		
Academic year	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Year 10	1108	23.8	944	24.7	164	19.8		963	23.4	241	20.3	145	26.8	
Year 11	1085	25.1	908	25.4	177	23.5		936	24.5	251	22.8	149	28.9	
Year 12	947	25.3	786	25.3	161	25.4		844	25.5	251	26.3	103	24.1	
Year 13	928	25.8	736	24.6	192	31.2		840	26.6	270	30.7	88	20.1	
Total	**4068**	**100**	**3374**	**100**	**694**	**100**	**0.002**	**3583**	**100**	1013	100	485	100	0.000

Table 20 continued…	All devices		Consenters		Refusers		Sig	Reduced dev		Smartphones		Tablets		Sig
Ethnic group	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
White	3156	78.5	2676	79.7	480	73.1		2759	77.9	787	79.6	397	82.9	
Mixed	179	4.6	156	4.8	23	3.5		166	4.8	52	5.4	13	2.6	
Asian	443	11.4	321	10.0	122	18.1		396	11.6	93	10.0	47	10.1	
Black	171	4.4	143	4.4	28	4.4		155	4.5	38	3.7	16	3.6	
Other	45	1.1	38	1.2	7	0.9		41	1.2	12	1.3	4	0.9	
Total	**3994**	**100**	**3334**	**100**	**660**	**100**	**0.000**	**3517**	**100**	982	100	477	100	0.180
Missing	74		40		34			66		31		8		
SUBSTANTIVE VARIABLES (SURVEY)														
If (will) study Level 3 Maths/Science	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Yes, does or plans to	1834	43.9	1582	45.7	252	35.9		1626	44.3	365	34.4	208	41.2	
Does not (currently) plan to	1833	47.5	1491	46.1	342	53.9		1598	47.1	535	55.6	235	50.4	
Undecided	335	8.6	268	8.2	67	10.2		299	8.6	93	10.1	36	8.5	
Total	**4002**	**100**	**3341**	**100**	**661**	**100**	**0.000**	**3523**	**100**	993	100	479	100	0.055
Missing	66		33		33			60		20		6		
Thinking about higher educational qual	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
University degree	2139	50.5	1877	53.5	262	36.9		1913	51.3	438	41.1	226	44.8	
Higher educational qualification	221	6.0	182	6.1	39	5.8		192	6.0	70	7.5	29	6.4	
Undecided	1162	29.1	941	28.4	221	32.4		1008	28.7	314	31.4	154	32.3	
No	497	14.3	353	12.0	144	24.9		429	14.0	177	20.1	68	16.5	
Total	**4019**	**100**	**3353**	**100**	**666**	**100**	**0.000**	**3542**	**100**	999	100	477	100	0.338
Missing	49		21		28			41		14		8		
If lives with parent who went to university	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**	**N**	**W %**	**N**	**W %**	**N**	**W %**	**Sig**
Lives with 1+ parents went to university	1349	31.6	1198	33.8	151	21.9		1204	32.1	221	20.8	145	28.3	
Does not live with parent who went to uni.	2378	61.0	1953	60.0	425	65.5		2078	60.5	701	71.7	300	64.8	
Does not know if parent went to university	284	7.4	197	6.2	87	12.6		252	7.4	76	7.5	32	6.9	
Total	**4011**	**100**	**3348**	**100**	**663**	**100**	**0.000**	**3534**	**100**	998	100	477	100	0.008
Missing	57		26		31			49		15		8		

and 1 indicates that it is certain to happen. This is a type of pseudo-randomisation which takes place after the events have occurred and data has been collected.

The treatment or exposure is normally binary (here, the dichotomy is whether the participant responds using a mobile device or a PC) and the propensity is therefore estimated using logistic regression. The purpose of the propensity score is to control for observed and measured confounders so that balance can be achieved between the exposed and unexposed groups, in other words the groups which receive different treatments.

An important consideration is which characteristics are potential confounders and should be included in the propensity score so that ignorability can be assumed (i.e., the matching process is sufficiently thorough that one can ignore how an individual was allocated to treatment or control). For this study it could be hypothesised that the potential confounders might include demographic characteristics, measures of socio-economic position, internet use, social media use and salience of the survey topic. In practice, the choice of confounders is often constrained by the available data. Ideally, these characteristics and behaviours should be observed independently of the treatment to avoid endogeneity. Endogeneity is a risk if the confounder variables are collected at the same time as the outcomes are measured; in particular, during the survey which is being evaluated. This argues for prioritising the use of independent variables which are collected ex-ante, such as information about the survey participant that is derived from the sample frame, or from linked data, rather than from the respondent itself.

Since propensity score matching, or indeed any quasi-experimental method, cannot account for unobserved or unmeasured confounders, it is important to identify and include as many potential confounders as possible. As a minimum, this tends to include baseline covariates that could confound the relationship between the treatment or exposure and the outcome, and often also includes covariates known to be associated with the outcome alone. The selection of variables to include in the model should be based on prior expectations about which covariates are likely to have an association with the outcome, drawing on the relevant literature. It is important that this set of covariates or confounders are not determined by assessing which of them are statistically significant within the logistic models used to estimate the propensity scores. No value is given to the identification of a parsimonious model.

While there are several approaches which have been used to perform propensity score analyses, including stratifying by the propensity score, propensity matching and multivariate adjustment, the one used here is inverse probability of treatment weighting (IPTW) (Chesnaye et al., 2021, Kibuchi,

2018, Stürmer et al., 2014). This is normally applied where there is a binary treatment (Austin and Stuart, 2015, Brown et al., 2020, Yoshida et al., 2017). Quite simply, weights are calculated for each individual which is the inverse of the probability of being exposed.

An advantage of IPTW is that multiple measures can be included in the specification, including continuous variables. Individuals in one group are effectively manipulated to make them compositionally equivalent to the other, based on the chosen set of potentially confounding variables, to deliver balanced groups akin to an experimental design. Once the sample has been balanced or adjusted for selection, any remaining effects are measurement effects; in this instance, they are device effects.

This method is applied to the 11 main outcome variables set out in Section 3.2.2 in the following way. First, based on weighted data, each outcome variable is compared by type of device (e.g., PC responders versus mobile device responders) and it is noted whether there is a statistically significant difference in the responding behaviours of these two groups. Any observed differences are likely to result from a combination of measurement and selection effects. A matching exercise is then carried out using Inverse Probability of Treatment Weighting (IPTW), an extension of propensity scores, to balance the samples, based on a specified set of covariates; the comparison of each outcome variable by type of device is then repeated. To the extent that the covariates specified in the matching exercise successfully capture selection, any remaining differences between devices can be attributed to a device effect. Where consenters and non-consenters are analysed, the weighted variable takes account of survey non-response. Additionally, where the analytical sample is limited to those who consented to data linkage, weighting also takes account of non-consent.

In the **primary analysis** the two data considerations set out in Section 3.2.1 were responded to, firs by using the 'all devices' sample, comparing PC with smartphone and tablet responders combined; and second, by selecting the sub-sample of consenters to include the administrative variables in the matching specification. It is hypothesised that the optimal specification of matching variables is a combination of the variable sets PRE, DEM, and ADMIN which provides a range of personal characteristics (gender, school year group, ethnicity, geographical location, and SEN status), measures of socio-economic disadvantage (FSM at the individual level, IDACI at the area level, and measures such as the proportion of pupils who are FSM eligible to provide school level context), topic salience (highest science attainment) and survey engagement (time from invite to response). This specification makes best use of the variables collected independently of the survey, but also includes demographic variables, while excluding the self-reported measures in the variable set SUR, which are most at risk of endogeneity. The analytical sample for the primary analysis is therefore all

device users who consented to data linkage, and thus for whom PRE, DEM and ADMIN variables are available (n=3,137). This sample is illustrated in Table 21 below, which extends Table 16, presented earlier, by summarising the way the sample is reduced when different variable sets are selected. The primary analysis appears on the left-hand side of the table – 'All devices sample' – and is labelled in red.

Table 21 Composition of the sample used in the primary and sensitivity analyses

Which devices are included ->	**All devices sample (PC, Smartphone, and Tablet)**		**Restricted devices sample (PC and Smartphone only)**	
Whether administrative -> data is available	Consenters and non-consenters (No ADMIN data)	Consenters only (ADMIN data)	Consenters and non-consenters (No ADMIN data)	Consenters only (ADMIN data)
Full sample Variable set PRE	4,068 Initial analytical sample	3,189 Sensitivity 1	3,583	2,824 Repeat sensitivity 1
Variable sets PRE+DEM	3,969 Sensitivity 2	3,137 Primary analysis	3,493 Repeat sensitivity 2	2,776 Repeat primary analysis
Variable sets PRE+DEM+SUR	3,844 Sensitivity 3	3,077	3,384 Repeat sensitivity 3	2,722

Note: This table extends the information provided in Table 16 on page 71

Three sensitivity analyses are then carried out to test the effect of varying the matching specification and sample definition. The **first sensitivity analysis** involves dropping the demographic variables (DEM) from the matching specification, to establish their importance, and given the concern that they are potentially endogenous. As before, the sample is based on 'all devices' and only includes consenters, but the specification of the matching variables is reduced to PRE+ADMIN. Since there is some missingness associated with the demographic variables which are no longer included in the specification, the size of the sample is slightly increased (n=3,189). This sample is illustrated in Table 21 as 'Sensitivity 1' in blue.

More radically, the **second sensitivity analysis** takes account of the potential for bias resulting from excluding non-consenters in the sample, as described above, and so reincludes them in the sample. This involves dropping the requirement for administrative variables (ADMIN) from the matching specification and matching with PRE+DEM variables alone. This increases the size of the sample and alters its composition, removing a potential source of bias (n=3969), shown in in Table 21 as 'Sensitivity 2' in blue. It represents the maximum possible sample with the greatest ecological validity, but at the same time relies on a weaker specification for matching.

The **third sensitivity analysis** also includes non-consenters in the sample but strengthens the matching specification by including additional survey variables (SUR). These are: living with at least one parent who attended university, which may act as an indicator of family support for education and of socio-economic position; whether studying or intending to study maths or science at Level 3; and whether considering a higher educational qualification. These two variables can be seen as indicators of whether the survey is likely to be salient to responders. The resulting specification is PRE+DEM+SUR with a slightly reduced sample size relative to the second sensitivity analysis because of missingness in the survey variables (n=3,844). It is shown in in Table 21 as 'Sensitivity 3' in blue.

Finally, **the primary analysis and three sensitivity analyses** are repeated to establish whether the findings are sensitive to the decision to include large and medium tablets within the definition of mobile devices. Each analysis is repeated, having removed tablet responders, so that the comparison is between PC responders and smartphone responders only, i.e., based on the reduced devices sample. The sample sizes for these repeated analyses are shown on the right-hand side of Table 21: for the primary analysis, n=2,776 (in green) and the first (n=2,824), second (n=3,493) and third (n=3,384) sensitivity analyses in purple. As mentioned earlier, Figures 11 and 12 provide a detailed illustration of the source of missingness in each of these samples.

3.3.4 The effect of matching and balance after matching

The purpose of matching is to adjust the sample so that it is balanced, and the effectiveness of this process can be investigated by comparing the distributions before and after matching for each specification. Indeed, one of the advantages of using propensity scores instead of multivariate models is that they allow for an assessment of the comparability of the treated and untreated groups following the analytical process (Stürmer et al., 2014). The right-hand columns in Table 19, provide a detailed example of this effect based on the sample selected for the primary analysis (n=3137): namely, all device users, who consented to data linkage, and thus for whom data is available for the variable sets PRE, DEM, and ADMIN. These columns show, for each covariate, a comparison of PC and mobile device responders, first based on weighted data[4] and then after matching.

As is expected, the matching process balances the samples of PC and mobile device responders considerably. For example, based on weighted data, 56% of mobile device responders are girls

[4] Because this is a subset of the full sample, the percentages are similar but not identical to those described in Section 3.3.1 and illustrated on the left-hand side of Table 21.

compared to 48.7% of PC responders, but after matching, 58.9% of both groups are girls. Similarly, based on weighted data, a higher proportion of young people who responded using a mobile device are eligible for Free School Meals (26.4%) compared to PC responders (19%) but following matching, the proportions are closer (24.6% and 24.1% respectively), and are no longer significantly different (p=0.763). Some characteristics of the sample are closer after matching but remain imbalanced. For example, based on weighted data, there is a 12.6 percentage point difference in the proportion who report that at least one parent had attended university. This reduces to a 7.6 percentage points after matching, but the difference remains highly statistically significant (p<0.000). The same is true of the young person's intention to study for a higher education qualification. The covariates where a significant difference remains in the characteristics of responders using different devices, are those which are not included in the matching specification.

While Table 19 above provided a detailed example of this effect based on the sample selected for the primary analysis, Table 22 below provides summary information for the 'all devices' sample, first for the primary analysis just described, but then also for the three sensitivity analyses. In this table, only the level of significance for each covariate is shown, indicating whether there is a significant difference in the distribution of each characteristic by device, using weighted data first and then matched data. The detailed information that underlies this summary table is provided in Appendix G. As expected, matching serves to reduce the difference between the characteristics in all cases, and the extent to which matching balances the sample depends on which variable sets are identified as potential confounders and are included in the matching specification.

As explained above, for the primary analysis, matching using the set of covariates PRE+ADMIN+DEM (first and second columns on the left-hand side of the table) results in a balanced sample in all respects, except for two of the survey variables: parental attendance at university and the young person's intention to study for a higher education qualification, neither of which were explicitly included in the matching process. In the first sensitivity analysis, where matching no longer includes the DEM variables (third and fourth columns), the differences between PC and mobile device responders are significant with respect to gender, ethnicity, and studying or intending to study maths or science at Level 3, which may reflect the impact that gender has on science uptake. For the second sensitivity analysis, increasing the sample to include non-consenters and matching with just PRE and DEM variables (fifth and sixth columns) leaves several additional significant differences remaining between PC and mobile device responders: the individual's FSM status, their highest science attainment, and aspects of their school (the admission policy and the proportion of school pupils eligible for FSM). Finally, in the third sensitivity analysis, defining the sample to include non-

Table 22 The effect of different matching specifications on the balance of the sample

			Primary		Sensitivity 1		Sensitivity 2		Sensitivity 3	
	Covariate set used in matching	->	PRE ADMIN DEM		PRE ADMIN		PRE DEM		PRE DEM SUR	
	Sample size	->	(n=3,137)		(n=3,189)		(n=3,969)		(n=3,844)	
	Covariates		Wtd	Mat	Wtd	Mat	Wtd	Mat	Wtd	Mat
PRE	IDACI		***		***		***		***	
	GOR		***		***		***		***	
	Rural/Urban									
	Time to respond		***		***		***		***	
ADMIN	SEN status									
	FSM		***		***		***	***	***	
	KS2/4 science		***		***		***	***	***	*
	School attainment		***		***		***		***	
	School admissions		***		***		***	*	***	
	School FSM eligibility		***		***		***	*	***	
DEM	Gender		***		***	***	***		***	
	Year group									
	Ethnicity					*				
SUR	Parent attend Uni		***	***	***	***	***	***	***	
	L3 maths/science		***		***	*	***	***	***	
	Higher education		***	***	***	***	***	***	***	

Notes: Wtd=weighted; Mat=matched; Analysis is based on all devices (PC, smartphone and tablet)
Key to results of significance tests: *** P ≤ 0.001, ** 0.001 < P ≤ 0.01, * 0.01 < P ≤ 0.05.

consenters and matching with PRE, DEM, and SUR variables (final two columns) reduces the difference between the samples in many respects, but the individual's highest science attainment remains significantly different. This demonstrates, *a priori*, that matching with different variable sets is likely to differentially account for selection effects when the outcome variables between devices are compared.

3.4 Results

3.4.1 Is there evidence of device effects after controlling for selection?

As explained above, the primary analysis is based on responders who used any device to respond (PC, smartphone, or tablet), for whom there is data for the variable sets PRE, DEM, and ADMIN (n=3,137), which is hypothesised is the optimal matching specification to control for respondents' propensity to participate in the survey using different devices. For the administrative variables to be available for analysis, the sample must be limited to those who consented to data linkage.

Therefore, it cannot be possible to present results for the first outcome, refusal to consent to data linkage, based on this specification. The results for the remaining 10 outcomes are presented in Table 23 below.

Table 23 Main analysis: all devices, consenters only, matching with PRE, ADMIN and DEM

All devices; consenters only n=3137	Before matching (weighted data) PC	Smart phone/ tablet	Diff	p	Post matching (PRE+ADMIN+DEM) PC	Smart phone/ tablet	Diff	P
(i) Engagement	%	%	ppt	p	%	%	ppt	p
1. Refuses data linkage	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
2. Refuses findings	18.7	22.4	-3.7	**0.015**	20.3	22.1	-1.8	0.278
3. Refuses recontact	33.4	32.7	0.7	0.820	34.6	31.5	3.1	0.105
(ii) Satisficing	%	%	ppt	p	%	%	ppt	p
4. 'Don't know' responses	6.3	9.8	-3.5	**0.000**	6.6	8.9	-2.3	**0.039**
5. Don't know to quiz	12.8	17.4	-4.6	**0.002**	15.2	16.5	-1.3	0.402
6. Straightlining	8.9	6.7	2.2	**0.041**	8.5	6.8	1.7	0.107
7. Agreement	6.1	5.3	0.8	0.305	5.9	5.6	0.3	0.729
(iii) Temporal	%	%	ppt	p	%	%	ppt	p
8. Speeding	9.3	8.4	0.9	0.500	8.8	8.4	0.4	0.696
9. Interruptions	2.6	5.2	-2.6	**0.000**	2.6	5.4	-2.8	**0.000**
10. Slow first module	7.0	7.9	-0.9	0.277	7.5	7.6	-0.1	0.881
	Mean (SD)	Mean (SD)	Diff	p	Mean (SD)	Mean (SD)	Diff	p
11. Completion time (minutes)	21.4 (10.3)	21.4 (10.1)	0.0	0.940	21.8 (11.7)	21.6 (8.7)	0.2	0.575

Notes: Grey shading: device effect <u>not</u> significant before or after matching
Yellow shading: device effect significant before and after matching
Blue shading: device effect significant before matching but not significant after matching.
Diff: percentage point difference in the outcome variable between device types;
Sig: statistical significance of that difference.

The left-hand-side of Table 23 above shows the comparison of PC and mobile device responders *before matching,* based simply on weighted data, first showing the percentage of PC responders who exhibited the behaviour defined by each outcome, then showing the percentage of smartphone and tablet responders, and finally by the difference shown in percentage points and the statistical significance of that difference. For five of the ten measures, shaded in grey, there are no observed device effects. Among these five outcomes, completion times are virtually identical, and mobile device responders show fractionally more compliant behaviours in three outcomes: agreement; speeding while PC responders show fractionally more compliant behaviours in the last outcome,

likelihood of a slow first module. However, all these differences are small, and none are statistically significant.

The right-hand-side of the table shows the results for each outcome *after matching*. As might be expected, as before, there are no observed device effects for these five outcomes. Although controlling for selection slightly reduces the observed measurement effects in some outcomes, for two, refusing recontact and completion times, matching results in a very slight increase in the observed difference between PC and mobile device responders. Nevertheless, all these differences remain insignificant.

For the remaining 5 outcome measures, the results *before matching* on the left-hand-side of the table suggest that these outcomes are subject to a device effect. In all but one, PC responders show more compliant response behaviours. The exception is straightlining, where 8.9% of PC responders exceed the threshold compared to 6.7% of mobile device responders, though the difference of 2.2 percentage points is small and only statistically significant at the 0.05 level (p=0.041). However, *after matching*, it is found that for 3 of these 5 outcome measures, controlling for selection reduces the difference such that no significant device effect is observed (shaded in light blue). This is the case for refusal to consent to receive findings, 'don't know' responses to the quiz questions, and straightlining . In some instances, the effect of matching is quite substantial but in others it is small. For example, in the case of item non-response measured by 'don't know' responses to the 10 quiz questions, controlling for selection through matching results in a substantial change; the difference before matching is 4.6 percentage points, with fewer PC responders than mobile device responders showing high item non-response to the quiz (12.8% compared to 17.4%, p=0.002), but after matching the difference is reduced to 1.3 percentage points and is not significant (p=0.402). In the case of straightlining, where the device effect is marginal before matching (p=0.041), controlling for selection reduces the difference further (from 2.2 to 1.7 percentage points), and, unsurprisingly, the device effect is no longer statistically significant (p=0.107). In this case, the shifting across the margin between significant and non-significant suggests there is no strong evidence of a device effect related to straightlining based on this specification.

After matching, **a clear device effect is observed for just two outcome measures** (shaded in yellow in Table 23. The first is 'interruptions', where one or more modules took over 30 minutes. In this case, matching has almost no effect on the difference between PC and mobile device responders. The second is item non-response, measured by the number of 'don't know' responses, where the measurement difference between devices reduces after matching, but remains significant.

In summary, in the primary analysis, after controlling for selection, a significant device effect is observed for just two outcome measures: interruptions and item non-response measured by 'don't know' responses. Even in these two cases, the magnitude of the device effect is quite small. In both instances, the results are relatively insensitive to the matching process so although there is a possibility that the inclusion of additional covariates in the matching specification might control for unobserved selection, these findings provide reasonably strong evidence that device is affecting measurement and warrants further consideration. These findings are shown in summary form in Table 24 below which presents, for each outcome, the difference in percentage points between PC and smartphone/tablet responders, and the statistical significance of this difference, both before and after the matching process. This table makes it possible to view the results of this primary analysis alongside the summary results for the three sensitivity analyses presented next.

3.4.2 Sensitivity of the findings to the specification of the matching process

Next, whether these findings are robust when three sensitivity analyses are conducted is considered. These also use the 'all devices' sample but each varies the specification of confounder variables included in the matching process, and consequently adjusts the sample used. Summary findings for these three sensitivity analyses are presented alongside the summary information for the primary analysis in Table 24 below.

Sensitivity 1. Dropping the demographic variables

The first sensitivity analysis is to test the effect of dropping the demographic variables (DEM) from the matching specification. Previous studies suggest that demographic characteristics are likely to be important in controlling for selection effects. However, the DEM variables are taken from the survey, so may be affected by endogeneity. Therefore, the matching process was repeated with the sample for whom PRE and ADMIN variables are available (n=3,189). In practice, **the results only change fractionally**; for example, in the differences observed for item non-response measured by 'don't know' responses to the quiz questions, where excluding demographic variables results in a slightly larger difference between device responders, but post matching, the difference between devices remains insignificant.

Sensitivity 2. Including non-consenters and dropping administrative variables

To this point, the specification for the matching has included administrative data, so the sample has been limited to those who consented to data linkage. As shown earlier, this introduces a potential bias since non-consenters have different characteristics to consenters, and they may also have

Table 24 Sensitivity to different specifications of matching variables (all devices)

Purpose ->	Primary analysis with optimal specification				Sensitivity 1: remove DEM				Sensitivity 2: include non-consenters with loss of ADMIN				Sensitivity 3: include non-consenters, addition of SUR			
Matching variables -> Available sample size ->	PRE+DEM+ADMIN n=3137				PRE+ADMIN n=3189				PRE+DEM n=3969				PRE+DEM+SUR n=3844			
	Difference PC - SP/Tablet				Difference PC - SP/Tablet				Difference PC - SP/Tablet				Difference PC - SP/Tablet			
	Before		After match		Before match		After match		Before match		After match		Before match		After match	
(i) Engagement	Diff ppt	Sig p	Diff ppt	Sig p	Diff ppt	Sig p	Diff ppt	Sig p	Diff ppt	Sig p	Diff ppt	Sig p	Diff ppt	Sig p	Diff ppt	Sig p
1. Refuses data linkage	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	-5.1	**0.000**	-3.7	**0.004**	-4.8	**0.001**	-2.1	0.119
2. Refuses findings	-3.7	**0.015**	-1.8	0.278	-3.7	**0.014**	-2.3	0.145	-4.9	**0.001**	-3.7	**0.011**	-4.4	**0.004**	-1.5	0.334
3. Refuses recontact	0.7	0.820	3.1	0.105	0.3	0.900	1.8	0.329	-1.7	0.315	-0.4	0.846	-1.7	0.325	1.6	0.363
(ii) Satisficing	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
9. Don't know response	-3.5	**0.000**	-2.3	**0.039**	-4.5	**0.000**	-2.9	**0.009**	-5.7	**0.000**	-4.5	**0.000**	-5.3	**0.000**	-3.0	**0.006**
11. Don't know to quiz	-4.6	**0.002**	-1.3	0.402	-4.5	**0.002**	-2.2	0.114	-5.1	**0.000**	-2.7	**0.032**	-4.7	**0.000**	-0.8	0.600
12. Straightlining	2.2	**0.041**	1.7	0.107	2.3	**0.024**	1.8	0.083	2.4	**0.007**	2.3	**0.014**	2.6	**0.005**	2.2	**0.019**
13. Agreement	0.8	0.305	0.3	0.729	0.7	0.432	0.6	0.501	0.9	0.235	1.0	0.223	0.8	0.283	0.2	0.725
(iii) Temporal	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
15. Speeding	0.9	0.500	0.4	0.696	0.6	0.589	0.4	0.721	1.0	0.354	0.8	0.396	1.0	0.350	0.7	0.532
16. Interruptions	-2.6	**0.000**	-2.8	**0.000**	-2.8	**0.000**	-2.9	**0.000**	-3.0	**0.000**	-3.2	**0.000**	-2.9	**0.000**	-3.1	**0.000**
17. Slow first module	-0.9	0.277	-0.1	0.881	-1.0	0.288	-0.5	0.696	-0.6	0.557	-0.1	0.942	-0.6	0.535	0.3	0.722
	Diff	p	Diff	p	Diff	p	Diff	p	Diff	p	Diff	p	Diff	p	Diff	p
18. Completion time	0.0	0.940	0.2	0.575	0.0	0.957	0.3	0.510	0.1	0.746	0.1	0.705	0.1	0.790	0.2	0.627

All devices: smartphone (SP) and tablet compared with PC

Diff: percentage point difference in the outcome variable between device types; Sig: statistical significance of the difference;

Grey shading: device effect not significant before or after matching; Yellow shading: device effect significant before and after matching; Blue shading: device effect significant before matching but not after matching

different response behaviours. As a result, the second sensitivity analysis includes respondents who did not consent to data linkage, increasing the size of the sample (n=3969), and reducing the matching specification to PRE+DEM variables only.

As found in the primary analysis, **no device effects are observed for five outcome variables**, either before or after matching (see grey rows in third and fourth sets of columns in Table 24). In contrast, while in the primary analysis just 2 of the remaining 5 outcomes showed a device effect *after matching*, with this weaker specification of the matching process, **a device effect is observed for all five other outcomes**, even after matching. In part, this may reflect the fact that even before matching, the larger sample which includes non-consenters is seen to have poorer response behaviours for both PC and mobile device users for several outcomes. For example, when respondents who refuse consent to data linkage are included, refusal to receive findings rises for PC responders (from 18.7% to 23.3%) and for mobile device responders (from 22.4% to 28.2%). Similarly, the proportion with high item non-response measured by 'don't know' responses increase for PC responders (from 6.3% to 8.7%) and for mobile device responders (from 9.8% to 14.4%). Furthermore, before matching, for three of the outcome measures (the two measures of item non-response and interruptions), the size of the difference between PC and mobile device responders also increased. For example, in the case of receiving findings, the difference between PC and mobile device responders widened by 1.2 percentage points. For these outcomes, matching resulted in a slightly reduced observed device effect or no change, but in no instances did matching using PRE+DEM reduce the difference sufficiently to render the device effect statistically insignificant. The change in composition of the sample is likely to have had some effect, since an increased proportion of reluctant participants may result in the detection of device effects if those who will not consent to data linkage behave differently with respect to device. However, a significant part of the explanation is that the matching specification is too weak, which is illustrated by the third sensitivity analysis, which follows.

However, before presenting the results of the third sensitivity analysis, there is one additional outcome variable to consider. It was not possible to analyse **consent to data linkage** in the primary analysis but dropping administrative variables from the matching specification and including non-consenters in the sample makes this possible. Table 24 above provides summary results for this analysis. The key finding is that, before matching, 21.2% of mobile device responders refuse data linkage compared to 16.1% of PC responders, a difference of 5.1 percentage points (p=0.000) suggesting a significant device effect. **After matching the difference reduces** to 3.7 percentage points **but remains significant** (p=0.004).

In summary, when the sample is increased to include survey participants who did not consent to data linkage, and the specification of the matching process is weakened by removing the administrative variables, selection effects cannot be controlled for to the extent observed in the primary analysis. **Six outcomes show a significant device effect**; consistent with earlier findings, **five show no device effect either before or after matching**. It seems likely that this is the result of the weaker specification of matching variables, although it may also be partially explained by the inclusion of non-consenters in the analytic sample, who have different characteristics and are more reluctant, so may behave differently with respect to device.

Sensitivity 3. Including non-consenters and including survey variables

The third sensitivity analysis builds on the second. Again, the sample includes responders who did not consent to data linkage; but here the specification for matching is strengthened by including additional survey variables (SUR). These variables are whether the student lives with at least one parent who attended university, which may act as an additional socio-economic indicator and a measure of likelihood of the family supporting education, and two variables related to survey salience. These are studying or intending to study maths or science at Level 3 and considering a higher educational qualification. These may be seen as proxies for the administrative variables; they have the advantage of being available for most members of the sample, but they are also subjective measures, and are collected at the same time as the outcome variables, introducing a risk of endogeneity. The resulting specification is PRE+DEM+SUR and has a slightly reduced sample size because of missingness in the survey variables (n=3844).

As before, **no evidence can be found of a device effect for the five outcomes measures reported earlier**, either before or after matching (shown in grey). The remaining results are close but not identical to the primary analysis. A significant **device effect for two outcome measures can still be observed: interruptions and item non-response measured by 'don't know' responses**. In addition, **straightlining shows a measurement effect** even after matching with PRE, DEM, and SUR variables. Matching with SUR variables does not therefore appear to be as effective at controlling for selection as matching with ADMIN.

For all the remaining outcomes – including refusal to consent to data linkage – **there is no evidence of a device effect** after matching where the specification includes the SUR variables, which effectively controls for selection. These are: refusal to receive findings, item non-response measured by 'don't know' responses to quiz questions, and refusal to consent to data linkage.

In summary, matching with covariates which include administrative variables appears to be most effecting in controlling for selection, and the variables also have the advantage that they are exogenous to the intervention. However, analysts cannot always rely on having appropriate covariates from external sources, and in this case these variables rely on consent to data linkage, which will inevitably result in some missingness, and this is unlikely to be at random. In the absence of administrative variables, matching with survey covariates controls for selection for many, but not all, outcome variables. This may partly reflect the different behaviours of non-consenters, but also suggests that survey variables are less successful than administrative variables at capturing the underlying differences in the characteristics of respondents who elect to respond by PC or by mobile device. Relying on a more basic set of variables (such as PRE and DEM) does not control effectively for selection effects, which suggests that matching should include measures that relate to study salience.

4. ***Restrict the comparison to PC and smartphone responders***

Finally, the analysis is repeated to establish whether the findings are sensitive to the exclusion of the 11.9% of pupils who responded using a medium (n=50, 1.2%) or large tablet (n=437, 10.7%). In fact, the characteristics of the reduced sample of PC and smartphone users (n=3585) are remarkably similar to the characteristics of all device users (PC, smartphone and tablet responders, n=4,068) as can be seen in Table 20. This may explain why the results of the sensitivity analysis show that **the results are generally insensitive to the exclusion of medium and large tablet responders**, as shown in Table 25 below.

However, there is one notable exception – survey completion time – and Table 26 below (on page 106) focuses on this, reprising the results from the primary analysis in the top row (drawn from Table 23 on page 98) with the results from the comparison of PC and smartphone responders on the second row. This shows that while in the primary analysis there is only a trivial difference in completion time between PC responders and mobile device users, there is a difference of about one minute when PC and smartphone responders are compared, which is statistically significant. This difference is observed based on weighted data and increases fractionally after matching.[5] It is important not to overstate the magnitude of this finding, but it does appear that **completion time is significantly faster for smartphone responders, even after controlling for selection effects**.

[5] When the matching is specified with SUR instead of ADMIN Variables, the direction and size of the difference is similar, but the statistical significance is even more marginal.

Table 25 Sensitivity to different specifications of matching variables (restricted devices)

Purpose ->	Primary analysis with optimal specification				Sensitivity: remove DEM				Sensitivity: include non-consenters with loss of ADMIN				Sensitivity: include non-consenters, addition of SUR			
Matching variables -> Available sample ->	PRE+DEM+ADMIN n=2776				PRE+ADMIN n=2824				PRE+DEM n=3493				PRE+DEM+SUR n=3384			
Difference PC -	Difference PC - SP				Difference PC - SP				Difference PC - SP				Difference PC - SP			
Pre/post matching ->	Before		After match		Before match		After match		Before match		After match		Before match		After match	
	Diff	Sig	Diff	Sig	Diff	Sig	Diff	Sig	Diff	Sig	Diff	Sig	Diff	Sig	Diff	Sig
(i) Engagement	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p
1. Refuses data linkage	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	-4.4	**0.000**	-3.3	**0.004**	-4.1	**0.001**	-1.4	0.119
2. Refuses findings	-3.5	**0.015**	-1.4	0.278	-3.6	**0.014**	-2.2	0.145	-5.1	**0.001**	-4.3	**0.011**	-4.2	**0.004**	-1.3	0.334
3. Refuses recontact	-0.9	0.820	2.9	0.105	-1.4	0.900	0.9	0.329	-2.5	0.315	-0.5	0.846	-2.5	0.325	1.6	0.363
(ii) Satisficing	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p
4. Don't know response	-4.5	**0.000**	-2.9	**0.039**	-5.7	**0.000**	-3.7	**0.009**	-6.7	**0.000**	-5.6	**0.000**	-5.9	**0.000**	-3.7	**0.006**
5. Don't know to quiz	-4.8	**0.002**	-0.4	0.402	-4.9	**0.002**	-2.1	0.114	-5.7	**0.000**	-3.3	**0.032**	-5.0	**0.000**	-0.4	0.600
6. Straightlining	2.4	**0.041**	2.2	0.107	2.6	**0.024**	2.1	0.083	2.8	**0.007**	2.9	**0.014**	3.0	**0.005**	2.8	**0.019**
7. Agreement	0.3	0.305	-0.1	0.729	0.2	0.432	0.3	0.501	0.6	0.235	0.9	0.223	0.4	0.283	-0.1	0.725
(iii) Temporal	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p	ppt	p
8. Speeding	-0.6	0.500	-0.8	0.696	-1.1	0.589	-1.1	0.721	-0.3	0.354	-0.4	0.396	-0.3	0.350	-0.6	0.532
9. Interruptions	-3.0	**0.000**	-3.4	**0.000**	-3.1	**0.000**	-3.2	**0.000**	-3.4	**0.000**	-3.9	**0.000**	-3.3	**0.000**	-3.7	**0.000**
10. Slow first module	-0.6	0.277	0.2	0.881	-0.8	0.288	0.0	0.696	0.0	0.557	0.4	0.942	0.1	0.535	1.0	0.722
	min	p	min	p	mins	p	mins	p	mins	p	mins	p	mins	p	mins	p
11. Completion time	0.9	**0.042**	1.2	**0.010**	0.9	**0.038**	1.2	**0.007**	0.8	0.056	0.9	**0.030**	0.8	0.061	0.93	**0.023**

Notes: The restricted sample includes PC and smartphone (SP) only i.e., medium, and large tablets users are excluded (small tablets are included in SP)
Diff ppt: percentage point difference in the outcome variable between PC and smartphone; Sig p: statistical significance of the difference
Grey shading: device effect not significant before or after matching
Yellow shading: device effect significant before and after matching
Blue shading: device effect significant before matching but not after matching

Table 26 Completion time and how it differs for the PC/tablet and smartphone sample

PRE+ADMIN+DEM	Before matching				Post matching			
Completion time (mins)	PC Mean (SD)	Device Mean (SD)	Diff	p	PC Mean (SD)	Device Mean (SD)	Diff	p
Consenters only PC+ v **SP/tab** (n=3137)	21.4 (10.3)	21.4 (10.1)	0.0	0.940	21.8 (11.7)	21.6 (8.7)	0.2	0.575
Consent and non-consent PC+ v **SP only** (n=2776)	21.4 (10.3)	20.5 (9.9)	0.9	**0.042**	21.8 (11.4)	20.6 (8.4)	1.2	**0.010**

One other minor difference relates to **straightlining**. In the primary analysis, there appeared to be a device effect of about 2.2 percentage points which is significant at the p<0.05 level, while after matching, no statistically significant device effect is observed. A similar pattern is observed when the matching specification is reduced to PRE+ADMIN variables. However, when the sample was increased to include non-consenters, there appeared to be a device effect even after matching. Now, with this final sensitivity analysis, when medium and large tablet responders are dropped and the primary analysis is repeated, there does not seem to be a significant device effect for straightlining either before or after matching, suggesting that tablet responders may be behaving differently to smartphone responders. In summary, **with respect to straightlining, the difference between devices is small and whether the difference is statistically significant fluctuates** depending on the specification of the sample and the matching process. There are at least three plausible explanations: that whether there is a device effect is marginal shown by the sensitivity to different specifications and cannot be relied upon; that the measure of straightlining is not of sufficient robustness to constitute a proper test; or that tablet responders genuinely behave differently to smartphone responders.

3.5 Discussion

3.5.1 Summary of the study

The aim of this study is to assess whether the device used to complete an online survey effects data quality. It provides an additional snapshot to the existing literature where the survey that is examined was fully optimised and benefited from extensive usability testing. It provides an additional example of a device effects study which is based on a large population of young people who are digitally native (Couper and Peterson, 2017, Matthews et al., 2017); uses an existing research study with a nationally representative cross-sectional sample, rather than a specially designed experiment based on an online panel; and uses IPTW to balance the samples, so that measurement effects are not confounded by selection effects (Matthews et al., 2017). A wide

range of data types are linked to the survey responses (specifically survey process data, geographical data, and administrative records from schools) which provides an opportunity to assess the use of exogenous confounder variables in the matching process. Because the size of the overall sample and the number of responders who used different devices were large, it was possible to test the sensitivity of the finding to various research decisions such as the inclusion or exclusion of tablet users and the specification of the matching process.

3.5.2 Summary of main results

In summary, the study identifies very few consistent device effects, and those that are observed are small. This provides reassurance for survey practitioners and analysts alike.

Significant device effects are identified for two of the 11 outcome variables investigated. The first is **'don't know' responses** (with high levels observed for 6.6% of PC responders and 8.9% of mobile device responders). This supports the findings of Mavletova and Couper (2014, 2015), Struminskaya et al. (2015) and Lugtig and Toepoel (2015). The second device effect is **interruptions** (observed for 2.6% of PC responders and 5.4% of mobile device responders) although given the very low levels of breakoffs observed, this does not seem to be a matter of concern for this study at least. These findings are consistent regardless of how the sample is defined and how the matching process is specified.

A third device effect is observed, but only when comparing PC responders with smartphone responders alone. Contrary to many examples in the literature, **completion times** for PC responders are found to be about a minute slower than smartphone responders. In past examples, surveys have been completed more slowly on smartphones (for example, Andreadis, 2015, Couper and Peterson, 2017, Gummer et al., 2019) or no difference in time taken has been found (Matthews et al., 2017). One possible explanation for these results being contrary to earlier research is that this survey is optimised well – in particular, it included very few grid questions which have been cited as leading to additional scrolling and slow response times on smartphones (Couper and Peterson, 2017), or it could reflect the fact that young respondents are adept at completing using their devices, though other studies of students have found contrary results (Couper and Peterson, 2017). Certainly, it was carried out considerably later than the cited examples cited which predominantly took place in 2013 and 2014 and the design would have been influenced by the learning from those studies (Couper and Peterson, 2017).

The results for **straightlining** are mixed; the primary analysis suggests no device effect, but when the sample is extended to include non-consenters, the difference between devices is increased and remains significant even after matching. It is possible that tablet responders genuinely

behave differently to smartphone responders; or that the device effect is so marginal that it is sensitive to different specifications; or that the measure of straightlining may not be sufficiently robust to constitute a proper test. Whatever the explanation, this finding does not contradict the existing literature.

For a further three outcomes (**refusing consent to data linkage**, **refusing findings** and **'don't know' responses to quiz questions**), what initially appeared to be a device effect before matching is no longer significant once selection had been taken into account, either with the primary specification (PRE+DEM+ADMIN) in the case of refusing findings and 'don't know' responses to quiz questions, or for all three examples, when including non-consenters, with the alternative specification (PRE+DEM+SUR). The absence of a device effect is consistent with earlier results (Maslovskaya et al., 2020, Matthews et al., 2017), but there do not appear to be direct comparators with the results for refusing consent to receiving findings or to responding to quiz questions.

For the remaining four data quality indicators investigated, no device effects were found, neither before nor after matching. This is the case for **consent to recontact**, **agreement**, **speeding**, and completing **the first module slowly**. Looking across the types of data quality issues considered in this study, no device effects were found in terms of **willingness to engage** in the broader aspects of the research study. The findings related to **satisficing** are mixed, with device effects observed for 'don't know' responses and, under certain conditions, for straightlining. As before, these findings do not represent any marked inconsistencies with earlier results.

Although the results show that two **temporal** outcomes are subject to small device effects – interruptions are more frequent among mobile responders and completion times are faster among smartphone responders: neither appear to have led to a significant proportion of breakoffs. Indeed, the overall **breakoff** rate in this study is low; indeed, lower than is found in earlier studies, perhaps because of the high incentive offered to young people, and the extensive development and usability testing carried out before the study began. Respondents who began the survey generally completed it successfully, regardless of the device they used. Although there is an indication that breakoffs are higher among mobile device responders, it is not possible to draw generalisable conclusions given the scarcity of the data and the absence of covariates for those cases with breakoffs.

3.5.3 Strengths and weaknesses of the study

There are several ways in which the SET survey provides a strong foundation to examine device effects which distinguishes it from other studies, though it also has limitations. The survey is based on a large-scale random probability sample drawn from an administrative database which covers all state schools, but it excludes approximately 7% of students who attend private schools. The £10 incentive payment, comprehensive contact strategy and thorough piloting and usability testing of the survey is likely to explain the 50% response rate which is high for an online survey. The most successful face-to-face surveys, such as the Crime Survey for England & Wales, achieve higher response rates, but this response rate compares well with studies such as the British Attitudes Survey and is much higher than many carefully developed online surveys which, even after multiple efforts and incentives, can have response rates as low or lower than 30% (Hamlyn et al., 2015, Kantar Public, 2021). That said, the most economically disadvantaged were less likely to participate, with response rates ranging from 56% among young people from the least disadvantaged areas[6] to 43% in the most disadvantaged, and an even more marked gradient by prior educational achievement. Nevertheless, the achieved sample is reasonably representative of the population of young people in England, and weights were calculated to compensate for variations in response by different sample groups (Hamlyn et al., 2017). Since the matching process included markers of deprivation, the comparison between responders using different devices should be robust, but the overall results may nevertheless be biased if the sample under-represents the most advantaged and disadvantaged young people, particularly if the middle of the distribution behaves differently with respect to device use.

Echoing the features set out by Clements (2020), SET has the benefit that it is cross-sectional, so responses are not subject to panel conditioning effects which may be the case with studies which rely on commercial or academic panels. It has a genuine research purpose, with a topic that is more or less relevant to the student population it addresses, and a reasonable survey length, so is likely to have prompted authentic survey respondent behaviour (Krebs and Höhne, 2020). Furthermore, respondents chose which device to respond with, so were not encouraged to respond in an artificial manner. These factors give the study a high level of ecological validity and distinguish it from some bespoke methodological studies in this field (Clements, 2020).

Perhaps most importantly, SET is based on a large group of young people who are digitally native and who are likely to be relatively homogeneous with respect to digital use. In addition, the large

[6] In this instance, deprivation is based on the highest quintile of the overall Index of Multiple Deprivation, rather than IDACI.

number of young people who chose to use a mobile device to respond, and the very substantial pool of potential controls, provides good support for a matching exercise.

The study implemented a quasi-experimental design to overcome selection effects using data collected independently of the survey, so avoiding the problem of endogeneity, in this instance geography-based indicators derived from home postcode, survey process data recording lapsed time between invitation and response and linked administrative data from school records. A minor limitation is that the three demographic variables (year group, sex, and ethnicity) could not be sourced from the administrative data and were instead collected during the survey, although these measures seem unlikely to be affected by device. While matching with administrative data sources offers benefits, this data source is not available for the 16.8% of respondents who refused consent to linkage. The sensitivity analysis addressed this by repeating the analysis with the full sample, including non-consenters, both with a reduced set of variables (PRE+DEM) and with alternatives (PRE+DEM+SUR), and generated similar results. The research also explored the effect of matching on more parsimonious and more extensive sets of potential confounders and the results show the benefit of matching with variables which capture a broader set of issues related to survey salience, rather than focusing solely on demographic variables, where necessary using proxy indicators collected during the survey.

The particular quasi-experimental approach used in this analysis, IPTW, has limitations. In particular, in its standard application, it only allows for a binary treatment which limits its use in real life scenarios (Austin and Stuart, 2015, Brown et al., 2020, Yoshida et al., 2017). In this study, an ideal scenario would have been to have tested three separate treatment groups: smartphone, tablet, and PC responders. Although it has been argued that tablets are more akin to PCs (Couper et al., 2017, Peterson et al., 2017, Wells et al., 2013), the review drawn on here did not identify any studies where these groups were actually combined and then compared with smartphones (Tourangeau et al., 2017) and smartphone and tablets can be considered similar in the sense that both are mobile devices. Consequently, in this study, smartphone and tablet responders were combined into one group and were compared with PC responders. Then sensitivity analysis was carried out with tablets removed from the sample which showed almost no differences, except that smartphone responders completed the survey faster than PC responders, while this finding was not significant for all mobile device responders. Sensitivity analyses were also carried out to assess the effect of including different sets of covariates which showed that using the combination of exogenous variables gathered from different sources was effective, but could be partly substituted by using survey variables, though these carry a risk of endogeneity. One of the assumptions that is necessary for treatment effects obtained using IPTW to be interpreted as

causal is that the treated and untreated are exchangeable, in other words that the risk of outcome would be the same had either group been exposed to the treatment. Full exchangeability can only be achieved if it is possible to identify and measure all potential confounders, but this is not realistic in observational research so only conditional exchangeability can ever be achieved (Chesnaye et al., 2021). Nevertheless, the objective for the researcher is to identify as many of the potential confounders as possible. In this study, careful consideration was given to the covariates included in the matching process. These included the students' demographic characteristics and socio-economic position, measures of their prior academic attainment, parental education, indicating the possible salience of science and education in their lives, their school environment and local geographic context, as well as a measure of their enthusiasm to participate in the survey. Two other aspects of the respondent were considered but were not used in the analysis; household composition, derived from a set of questions about who the respondent lives with, and religion, but these did not have a strong theoretical basis for inclusion and there is no evidence that they are associated with the type of device that individuals use to respond or to the study outcomes. Since the study was not designed to estimate causal relationships between device and response behaviours, some potential confounders were not included in the study design (Rosenbaum & Rubin 1983, Shadish et al 2012, Beal and Kupzyk 2013). For example, the study would ideally have included information about access to different internet-connected devices, and measures of confidence or hesitancy. It is possible that other confounders should have been included as well, though there are no other obvious omissions. Nevertheless, it is important to acknowledge the possibility that some important covariates are missing, and only conditional exchangeability can be said to have been achieved.

An additional minor limitation is that the assessment of balance in the samples was carried out using p-values. An alternative approach would also have been to checks on standardized mean differences, since there is a possibility that p-values could have been influenced slightly by sample size.

Turning to the outcome measures, some, such as the measures of engagement in survey related activities, are rarely covered in the literature, and are therefore a useful addition. However, some others are relatively weak, and may not test data quality adequately. For example, only one battery of attitude questions was asked of all respondents and could be used to derive measures of straightlining and agreement. There is also insufficient data to derive other outcomes such as bias towards agreeing, bias to the left of the scale, or tendency towards the centre. This may partly reflect the ecological validity discussed by Clements; methodological research has warned repeatedly against the excessive use of grids in online surveys, and they have also been identified as the cause of longer completion times on smartphones. As a result, a naturally occurring study

may lack sufficient grids to assess data quality satisfactorily. Some of the outcome variables are relatively weak in a different sense, which is that they point towards a lack of engagement by the respondent, but do not necessarily lead directly to poor data quality. This is the case, for example, for the slow start measure.

Ideally the analysis would have been able to draw on more extensive paradata, but this was not available. For example, this could have included: time taken to complete each question; measures of survey interruptions; whether respondents switched to other screens while carrying out the survey; and whether they used more than one device to complete the survey.

Careful consideration was given to the effect of missingness with respect to those who do not consent to data linkage and those who used a tablet to respond to the survey. The risk of bias due to missingness which his not at random as a result of missing administrative or survey data was shown to be small given the low numbers of missing cases involved. However, the assumption made that this level of additional missingness can be treated as missing at random could have been tested.

3.5.4 Implications for survey research and practice

The move towards completion of online surveys using mobile devices appears to be an irreversible trend. For example, while 24.8% of respondents used a smartphone to respond to SET in 2016, only three years later this figure had risen to 46% (Kantar, 2019). The public's preferences to respond on mobile devices have already been substantially accommodated by the adaptations to online survey introduced by practitioners, but efforts to improve the experience of survey completion on a small device to minimise any residual device effects should continue. The research also acts as a reminder that – certainly at the time this survey was carried out – mobile device responders remain different to PC responders and selection. In this study, the focus has been on controlling for these differences to compare responses and response behaviours, but it is also important to remember that several of the observed differences are not neutral; owning a PC is associated with socio-economic advantage, and the most disadvantaged sample members may continue to have limited access to the internet, often solely through a smartphone. The implication of this is that it is equally important to consider how smartphones can be used to include study participants who might otherwise be out of the focus of the research and to encourage them to respond. The evidence that response rates are lower for smartphones (Couper, 2000) should not overshadow the possibility that smartphones will encourage the inclusion of people who might otherwise not take part. It could be argued that the increasing

number of devices for data collection provide participants with greater availability and flexibility, which could slow the decline in response rates (Clement et al., 2020).

Efforts to disentangle selection and measurement effects are primarily concerned with isolating the measurement effect so that it can be addressed. An alternative perspective is to see both measurement and selection effects as areas of focus for ameliorative actions. Survey practitioners should continue to address the behaviours of respondents who are least engaged, particularly, but not only, among those who respond by smartphone. For example, they might experiment with approaches to encourage completion at home and without distractions, perhaps by increasing messaging about the importance of concentrating for the short period of the survey. It may be possible to target this messaging, given the technological capabilities associated with smartphones. For example, information about the device being used is automatically identified at the start of a survey; this could be used to send a tailored message acknowledging their choice of device and encouraging them to keep focused. More radically, sensors built into smartphones could be used to detect background noise, movement or switching between screens and trigger messaging to encourage increased concentration. Apps such as Waze take this approach, asking users to confirm that they are passengers, and not drivers, if motion is detected. An approach of this kind could only be implemented after careful consideration of technical and ethical issues and would need to be evaluated for acceptability and effectiveness. Alternatively, it may be sensible to target responders who are more likely to exhibit poor response behaviours and encourage them specifically, regardless of the device they use.

Regardless, while efforts continue to test for device effects, results of this research should encourage researchers who are considering matching as a quasi-experimental method, to seek covariates which are independent of the survey process, including administrative data, auxiliary data that is collected for sampling purposes, or survey process data. This can be effective, particularly where variables are included which are salient to the topic of interest – in this instance, data about the academic ability of the young people in question, and their potential interest in the topic. Without this data, some outcome measures appear to be subject to measurement effects that might otherwise have been controlled for. The study demonstrates that it is possible, to some extent, to substitute these ex-ante variables with measures collected at the time of the survey.

3.5.5 Further research

Although overall few consistent device effects are identified in this research, the persistent finding that survey respondents using mobile devices give higher rates of ‘don’t know’ responses suggests that further research is needed, either to ameliorate the effect or to understand

whether this might be the result of unobserved selection effects. The analysis presented here should be repeated using SET 2019, which was expanded to include younger school years. Other findings such as the faster completion times of smartphone users may warrant further examination using the SET 2016 dataset, to explore whether some a particular mechanism or interaction is at play, for example if smartphone users use different strategies to respond to quiz questions which may have an impact on time taken.

Further studies are also needed to assess device effects which may emerge as the proportion and composition of people who chose to respond to surveys using mobile devices may change (Gummer et al., 2019), or as survey designers incorporate new question designs or activities. Although only a small number of differences were found in the characteristics of smartphone and tablet responders, and only one significant difference (in response time) was found when tablet responders were dropped from the analytic sample, future research should consider either excluding tablet responders to avoid blurring any distinctions between smartphone and tablet use, or may wish to consider smartphone, tablet and PC responders as three distinct groups to provide more nuanced results.

This research has shown the benefits of including a wide range of covariates and drawing on external sources which are independent of the survey or treatment for variables to be used in matching process. Future research should explore both avenues further by collecting additional covariates related to topic salience, access to different devices, academic ability, and measures of confidence, as well as auxiliary data used for sampling or field management purposes, and linked data from geographic and administrative data where possible.

While this study has focused on whether there are observable device effects, an important theme has been the nature of the selection effects that have been observed, with young people who use mobile devices being, on average, less engaged and more disadvantaged. Further research is needed to improve data quality from responders who are more reluctant, and more likely to use smartphones.

In summary, after controlling for selection effects using matching, mobile device users have higher levels of item non-response particularly when measured by ‘don’t know’ responses and are more likely to have interruptions during survey completion. When the analysis is repeated by comparing PC responders with smartphone responders only, smartphone responders are also found to complete the survey more quickly. There is an indication that straightlining is higher for PC responders, but results are mixed. It seems reasonably likely that these measurement effects are the result of the interaction between the respondent and the device. Therefore, survey

practitioners should continue to test adaptations to questions, screen layout or instructions to reduce remaining measurement effects associated with mobile device use and should continue to test for these effects with additional case studies since the findings are not fully consistent. Since there may be some remaining selection effects which have not been fully accounted for, further research should capture additional measures of interest that could be incorporated in future quasi-experimental studies of this kind.

3.5.6 Breakoffs

Having presented the results for each of the 11 main outcome variables, the available evidence about breakoffs can be considered. These were not included in the deposited data and, by definition, cannot be analysed using the matching methodology given the absence of covariates. It was not possible to apply the methodology used for the 11 main outcome variables, and instead an alternative approach was necessary, comparing the proportion of respondents using each device who drop-out, using a chi-square test to examine the significance of any difference. The possibility that differences in the point at which respondents dropped out is associated with different devices was also assessed. There are 50 cases with breakoffs (or 1.2%[7]), which is a very low proportion compared to those observed in other studies of mobile device use, with few exceptions, which also reported very low breakoff proportions (Maslovskaya et al., 2020, Matthews et al., 2017). The main observation is that breakoffs were observed for a higher proportion of responders who used smartphones (1.8%, n=19) and tablets (2.4%, n=12) than PC responders (0.7%, n=19). Although these are noticeable differences and are statistically significant, it is not possible to say whether this is a meaningful device effect, given that it is not possible to control for selection since the necessary covariates are not available for matching. A closer inspection of the 50 cases with breakoffs suggests that responders using a PC may end their session slightly later and may be slightly more inclined to drop out at more complex questions than mobile device responders. However, the numbers here are too small for robust analysis, the differences may be the result of chance and, as before, may be explained by selection. Therefore, it is not possible to draw any firm conclusions about this outcome beyond commenting on its low prevalence.

[7] This percentage, and others in this section, are based on all those who started the survey using a PC, smartphone, or tablet, that is 4,072 cases included in the sample and 50 additional cases that broke off.

Chapter 4 Adherence to protocol in a mobile app study collecting photographs of shopping receipts

4.1 Introduction

The emergence of mobile devices, particularly smartphones, has been one of the key technological developments that has affected the way that surveys are delivered. Chapter 3 explored whether the choice to complete online surveys using the web browser on a mobile device, rather than a desktop, laptop, or notebook, affects the quality of participants' responses. This chapter shifts focus, to whether an app-based consumer spending diary, which includes captured images of receipts, can be used to collect high quality expenditure data. The analysis is based on the Spending Study, which was implemented through the *Understanding Society* Innovation Panel (IP9), a probability household panel in Great Britain (Jäckle et al., 2019a, Jäckle et al., 2021, Jäckle et al., 2019b, Read, 2019b, Wenz et al., 2019, Wenz et al., 2018). Quality is assessed using the concept of adherence, defined as how well respondents complied with four different aspects of the protocol for the diary. The paper then identifies characteristics associated with higher levels of adherence, and estimates changes in adherence over the one-month study period. The findings contribute to a broader discussion about whether mobile research apps can be used to deliver complex data collection tasks.

The study provides an early example of the ways that mobile research apps might be used to capture complex data and some of the challenges involved. Several strengths of the Spending Study data make this investigation possible: predictors of adherence are drawn from an earlier interview (IP9); the app technology provides paradata, making it possible to observe respondent interactions with the app; and the request to photograph receipts, which often provide date and time of purchase, creates the opportunity to study time lags between the spending event and its entry into the app. This research should be understood as an exploratory phase in a larger project, which would need further development and testing before being launched at scale within a major survey.

By way of context, the literature described in Sections 4.1.1 summarises the problems that economists have identified with the quality of expenditure data collected using survey recall methods and paper diaries. Section 4.1.2 then describes advances towards gathering expenditure data using a mobile app and identifies the different types of measurement error that an app would need to address to improve the quality of expenditure data that is collected. These issues are returned to in the discussion.

However, readers who prefer to focus on the research study itself will find that adherence is defined in Section 4.1.3, predictors of adherence in Section 4.1.4, and the specific research questions addressed in this paper are set out in Section 4.1.5.

4.1.1 Quality of expenditure data using recall and diary methods

Before focusing on the core concept of adherence, which is examined in this chapter, the context of this study is provided through a review of the literature about the importance, and difficulties, of gathering accurate expenditure data using existing recall questions and diary methods.

A wide range of macro and micro-economic questions rely on the availability of detailed information about household finances, and on consumer expenditure data in particular (Browning et al., 2014, Crossley and Winter, 2016, Deaton and Grosh, 2000) . However, economists have been constrained by the quality of these data, and specific puzzles remain unsolved. For example, survey data suggests the poorest households spend more than less poor households, but it is not known whether this counterintuitive finding reflects actual behaviour or is a result of measurement error (Bee et al., 2015, Brewer et al., 2013, Brzozowski and Crossley, 2011, Meyer and Sullivan, 2003).

The difficulties collecting consumer expenditure data using survey recall methods, whether with a complete set of expenditure questions in household budget surveys or shorter question sets, are well documented (Crossley and Winter, 2016). Expenditure is underestimated because memory declines with the length of the recall period (Sudman et al., 1996) and because quantities are hard to remember accurately (Gray, 1955), while it is overestimated if respondents telescope, reporting earlier purchases as if they fell within the study reference period (Neter and Waksberg, 1964, Rubin and Baddeley, 1989). Errors also arise when respondents have difficulty adding across different types of spending (Crossley and Winter, 2016). These issues are set against the backdrop of falling response rates to household budget surveys, which have been accompanied by a decline in the correspondence between survey-based estimates of household expenditure and aggregate expenditure derived from national accounts (Barrett et al., 2015, Crossley and Winter, 2016).

To address these shortcomings, national budget surveys, such as the US Consumer Expenditure Survey (CE) and the UK Living Costs and Food Survey (LCF), combine recall methods with diary approaches (Silberstein and Scott, 1991). Respondents are encouraged to report every item purchased each day over a given reference period – often two weeks – with encouragement provided by survey interviewers at key moments. Regular diary keeping is expected to reduce the number of purchases forgotten, and to improve the accuracy of information provided about each

item, as well as removing the cognitively difficult task of computing total spending.

However, in practice, participants do not comply fully with protocols, so the potential benefits are not realised (Crossley and Winter, 2016). For example, expenditure diaries do not eliminate recall errors because there are lapses in participants' diary-keeping (Silberstein and Scott, 1991). CE respondents are only asked to make diary entries on days they make purchases, so it is not a problem *per se* if there is no data entered on some days, but interviewers report having to complete records when they visit the household to collect diaries, suggesting that participants do not report spending contemporaneously (Collins et al., 2018, Silberstein and Scott, 1991). In addition, anecdotal evidence suggests the use of 'pocket-books', given to LCF participants to jot down details of expenditures to aid recall, is variable (Collins et al., 2018). Furthermore, while the expectation is that interviewer contact encourages good record keeping (Butcher and Eldridge, 1990), Collins reports that this varies and is unlikely to be effective, and evidence from double placement of diaries in CE indicates that the importance of interviewer contact between study-weeks in reducing error may be overstated (Johnson-Herring et al., 2009).

There is clear evidence from the CE (Silberstein and Scott, 1991, Stephens, 2003), the Canadian Food Expenditure Survey (Ahmed et al., 2006, Statistics Canada, 1996), the U.K. Family Expenditure Survey (FES) (Kemsley, 1961, Kemsley and Nicholson, 1960, Turner, 1961, Tanner, 1998) and the LCF (Ralph and Manclossi, 2016) that reports of spending decline with time, with rates of expenditure lower in the second of two weekly diaries, and within-week responses higher at the start of the week. Analysis of the CE in 1972-3 and 1987 shows the same pattern across survey years: a decline in reported expenditure by diary day, especially in the first week, and the mean expenditure of the second diary week 10% lower than the mean of the first week for food at home, food away from home and other expenses, and 20% lower for apparel (Pearl, 1979, Silberstein and Scott, 1991). Similarly, in the 2013 LCF diary data, the average number of items recorded in the diary decreased by 12% between week 1 and week 2 (Ralph and Manclossi, 2016), with the expenditure categories of 'takeaways brought home' and 'eating out' most affected: falling 23% and 16% respectively (Collins et al., 2018, p45, Ralph and Manclossi, 2016). This phenomenon is generally attributed to diary fatigue and has been observed over many years (e.g., Kemsley, 1961, McWhinney and Champion, 1974, Pearl, 1979, Sudman and Ferber, 1971, Turner, 1961).

The decline in reports of spending over the reference period may also partly reflect conditioning effects – behavioural responses to diary participation (Collins et al., 2018). However, conditioning is hard to measure (Silberstein and Scott, 1991), and the evidence is limited and mixed. Qualitative research with FES participants found little evidence that participants were adjusting their expenditure during the diary recording period (Ritchie and Thomas, 1992), but focus groups

with LCF and Expenditure and Food Survey interviewers suggests that respondents may, for instance, delay a big shop until the diary recording period has ended (Betts and Dickinson, 2015, Gatenby and Hunter, 2000), while an experiment based on the FES concluded that behavioural changes are not uniform or in one direction (Kemsley et al., 1980). Nevertheless, conditioning remains a potential source of error for expenditure diaries that is not encountered when recall methods are used.

Diaries are intended to provide high quality data about individual items purchased. However, respondents may provide a single cost for a list of items (combined entries) or provide insufficient detail (nonspecific entries). These data issues are referred to as non-specificity. Analysis of the 1987 CE survey found that a percentage of entries were either combined or nonspecific (7% for food at home), and accounted for a disproportionate level of expenditure: 26% for food at home, 37% for food away from home, and 11% for apparel (Silberstein and Scott, 1991). Th problem of non-specificity has persisted over time; the CE quality report shows a slight increase in the overall edit rate for reported expenditure where information is insufficient and requires imputation, rising from 9.4 to 10.8 per cent between 2010 and 2017 (Hubener et al., 2019). Although estimates of the extent of non-specificity in the LCF are not available, anecdotal evidence from LCF coders suggests it is a problem (Collins et al., 2018).

A particular phenomenon mentioned in the context of each of these error types is a day-one or peak reporting effect. For example, the largest decline in spending is observed between days 1 and 2, with expenditure reported on the first day of week 1 in CE, almost 50% greater than the overall estimate and 40% greater than the mean for week 1 (Pearl, 1979, Silberstein and Scott, 1991). Similarly, purchased items are more likely to be reported with the correct level of specificity on the first study day (Silberstein and Scott, 1991). Diary fatigue is the main explanation offered for this effect. Alternative explanations include: the novelty of diary-keeping results in an increase in purchases on that day, a form of conditioning; and telescoping, the tendency to report dates as falling within the reference period, or to round estimated dates forward in time (Sudman et al., 1996), a behaviour more generally associated with recall methods (Collins et al., 2018). However, the telescoping effect is lessened because the first week of reporting is bounded by the use of interviewers' examples from the respondents previous week of purchases, while the second week is bounded by the first-week diary pick-up (Tucker, 1992) and because telescoping would tend to occur when respondents reconstruct purchases after a delay, so later diary days would be more likely to be inflated than the first (Silberstein and Scott, 1991). Other explanations are also possible; for example, if participants 'practice' reporting using past purchases. Whatever the explanation, day-one effects are associated with diaries.

Diary methods are subject to three other forms of error which are common in recall methods. Because diary completion is burdensome, they tend to cover relatively short periods. The Diary of Consumer Payment Choice covers three days only (Angrisani et al., 2018, Edgar et al., 2013, Ralph and Manclossi, 2016). The shorter the reference period, the more likely it is that 'true' expenditure will be misrepresented because of infrequency problems, when large purchases happen to fall inside or outside the data collection period (Collins et al., 2018).

Diaries may also be more prone to specification error if diary keepers do not follow rules about reporting individual and household spending.

Although it is possible that social desirability effects would be higher in an interviewer-based survey (Poikolainen and Kärkkäinen, 1983), it is also possible that the expectation that an interviewer will check and collect an expenditure diary will influence spending behaviours.

4.1.2 The potential for improvements to quality using mobile devices

Given the concerns about collecting expenditure data using recall and paper diary methods described above, it is unsurprising that there is interest in whether data quality can be improved by using a digital diary. Indeed, the challenges of collecting accurate expenditure data in national budget surveys has been considered in methodological reviews (for example, see , Edgar et al., 2013, Ralph and Manclossi, 2016). Furthermore, serious consideration has been given to online and digital solutions, but, given the imperative for national budget surveys to maximise response rates and collect high quality data from all sectors of the population, National Statistical Organisations have not made more radical solutions, such as mobile apps, central to their strategies. Nevertheless, there is clear interest in learning from early tests of these approaches, given the rise in ownership of smartphones and tablets, and the increased interest in new forms of measurement which exploit the in-built capabilities of these technologies (Link et al., 2014, Lessof and Sturgis, 2018, Volkova et al., 2016). This chimes with a broader methodological interest in whether a mobile app, incorporating technological features of smartphones and tablets, can be used to address other complex measurement tasks which have real-life application.

Two kinds of mobile app-based expenditure diary have already been developed in the commercial sector (Jäckle et al., 2019b). The first supports personal budgeting, by inviting users to manually enter key information about every purchased item, exemplified by Dollarbird, Fudget, and Goodbudget (Foreman, 2022, Sharf, 2016), and by linking directly to bank accounts and financial

products through Open Banking arrangements (Competition & Markets Authority, 2016), exemplified by Yodlee, Yolt, Emma and Money Dashboard[8]. The second, exemplified by ReceiptPal and Receipt Hog[9], incentivise panel members to photograph receipts, so that market research companies can analyse consumer behaviour, including responses to special offers or advertising (Jäckle et al., 2019b).

Theoretically at least, expenditure diaries delivered through a mobile app may reduce some of the errors associated with paper diaries. In this section, many of the types of error identified in the literature on paper expenditure diaries above are highlighted (in bold) and discussed, as a means of exploring the potential that mobile data collection may have for error reduction. In the discussion at the end of this chapter, these concepts are revisited to assess whether there is any evidence that this potential could, with further development and testing, be met.

If delivering the expenditure diary using a mobile device means that spending is reported closer to the shopping event, then participants may be less likely to forget a purchase and may remember the details about it more accurately, thereby reducing **recall error**. There are two ways a mobile device could reduce the time lag before a purchase is recorded. First, it is beneficial by virtue of the general attributes of mobile device use which travel with the participant are familiar, unlike a paper diary or aide memoire, and are characterised as 'always on'. Indeed, advocates of mobile surveys promote the idea of 'in-the-moment' measurement where record-keeping occurs immediately after the relevant event and captures responses 'closer to the moment of experience' (Claxton, 2016). Secondly, and more specifically, an expenditure diary fielded through a mobile device can be designed to encourage daily reporting in specific ways; for example, building a diary-habit by requesting an entry each day, sending daily reminders through app notifications, and offering micro-incentives for daily use or for every purchase. In addition, a mobile spending diary which asks participants to photograph receipts using the in-built capability of mobile devices may combat recall error by making the physical receipt a trigger to make a diary entry, and by removing the need to memorise details about each purchase.

These same mechanisms may also mitigate **diary fatigue** by maintaining the motivation of the respondent. Although a mobile spending diary does not have the benefit of the interview

[8] There are many money management apps – limited information can be found from their commercial websites e.g. https://dollarbird.co/, http://fudget.com/ and https://goodbudget.com/ and at https://www.yolt.com/, https://emma-app.com/ and https://www.moneydashboard.com/. Some of these are regulated providers, see https://www.openbanking.org.uk/about-us/.

[9] Limited information can be found at https://support.receiptpalapp.com/hc/en-us/articles/360039259754-What-do-you-do-with-my-receipts- and https://receipthog.com/

motivating the respondent during the diary period, it may introduce accountability because the participant implicitly understands that their reporting behaviour can be tracked, and subsequently rewarded, which could substitute to some extent for the human encouragement normally provided through the interviewers who place diaries within the household, recontact them during the diary period, and provide support on collection. Incentives directed at rewarding completion each day, or over the full study period, may compensate for the burden of the diary activity. Furthermore, diary fatigue may be reduced if tasks involved in completing the mobile diary, such as entering information in a sequence of screens on an app, or photographing receipts, are less burdensome than equivalent tasks in standard expenditure diaries (Read, 2019b).

Moreover, if mobile devices are sufficiently successful at reducing diary fatigue, then it may also be possible to have longer diary periods. This would, by definition, reduce **infrequency problems**.

Photographed receipts provide raw data that can allow detailed expenditure data to be coded later. If an app-based study uses this approach to data collection, it is possible that this will reduce the number of occasions where respondents provide insufficient detail, or **non-specificity**, as detailed information should be available in a larger proportion of occasions (Wenz et al., 2018). However, using photographed receipts may also be associated with **missing data** (Jäckle et al., 2019b) if a receipt is not given to the customer after a shopping event, or is given but is mislaid, and there may be **item-missingness** if the receipt is incomplete, if some of the receipt is not photographed, or if it does not upload (Volkova et al., 2016).

It is time consuming and resource intensive to derive structured data from the plethora of differently organised receipts, whether this is done by manual data entry and coding, or by scraping data from images and using machine learning to automate data capture. Furthermore, **coding and processing errors** may occur when transferring information from the receipt, or when classifying information into spending categories, or if receipts have been photographed several times but duplicated data is not identified. Receipts also vary with respect to the auxiliary data they provide, such as time and location of the spending event, name of store, information about price reductions or multi-purchases, and whether a loyalty card was used. In summary, while some receipts provide detailed information about each purchased item, some even including item specific barcodes, on other occasions data may be limited or absent for a range of reasons (Jäckle et al., 2019b). As a result, the receipt may not be sufficient to capture all the information required for studies such as CE or LCF (Collins et al., 2018).

Another potential advantage of an app-based expenditure diary is that it may help to reduce **day-one effects** by making the correct start and end-date of the study clearer and creating a firmer boundary around the reference period, thereby reducing telescoping. Indeed, mobile apps could

be designed to reduce specific problems; for example, they might provide introductory screens which encourage participants to practice entering purchases before beginning the full study to reduce unexplained day-one effects. One commercial receipt-based app carries out near real-time checking of receipts, providing feedback to participants if they are not complying with the requirements; participants do not qualify for an incentive unless the purchases they enter are approved, so there is a direct mechanism, and an incentive, for within-study learning. This approach could also be used to discourage reporting of expenditure which precedes the start of the study.

An app-based expenditure diary clearly has potential advantages, but there are also reasons why this approach may be problematic. For example, apps of this kind are often used to support behaviour change, and it is plausible that using a mobile device might have a greater effect on behaviour, increasing **conditioning effects**, relative to paper diaries. More immediately, although the prevalence of smartphones and tablets continue to grow, ownership is not universal and unless suitable devices are loaned to participants, **coverage error** is likely (Fernee and Sonck, 2013, Sonck and Fernee, 2013). Moreover, ownership of a smartphone or tablet is not sufficient (Hargittai, 2001). Many apps are only developed for leading operating systems, so some device owners will be excluded; phones or tablets may not meet technical requirements for storage or memory; or participants may not have an adequate data plan and/or access to Wi-Fi. Although this is not unique to app studies, participants must be willing to participate in non-standard research activities (Revilla et al., 2019, Wenz et al., 2019) which may be burdensome (Bradburn, 1978, Read, 2019b). Even the initial stages of participation require multiple activities, such as completing a registration survey, downloading and installing an app, and deciding whether to accept notifications and beginning data collection. These may explain relatively low initial participation rates (Jäckle et al., 2019a). Unless subsequent data collection depends entirely on the passive flow of data from device to the research team, participants must then be willing to carry out a range of activities over a sustained study period, setting aside any residual concerns about privacy or data security using the app or sharing data. The request to complete additional tasks is not new to social surveys, which increasingly incorporate requests for physical or cognitive tests, biosamples and data linkages (Benzeval et al., 2016, O'Doherty et al., 2014, Sakshaug et al., 2012). The requirement to complete several types of activity is also observed in standard diary keeping. Nevertheless, there are distinct requirements where app-based diaries are concerned.

4.1.3 Conceptualising adherence

This review shows that there are many questions that need to be addressed to assess whether

new technologies in general, and app-based diaries in particular, will improve the measurement of expenditure, or other complex behaviours. There is a growing literature exploring these issues (Jäckle et al., 2021, Jäckle et al., 2019b, Keusch et al., 2020, Keusch et al., 2019).

Many paper to date focus on issues such as willingness to participate in app studies, coverage, burden, and initial response rates. For example, parallel papers analysing the *Understanding Society* Innovation Panel Spending Study have shown that initial participation in an app-based spending diary is low, though unbiased with respect to key variables (Jäckle et al., 2019a) and that subjective and objective burden are seemingly unrelated to each other but are associated with willingness to participate in similar tasks (Read, 2019b, Wenz et al., 2018). Other unrelated studies, have focused on coverage error (Keusch et al., 2020) and willingness to participate (Keusch et al., 2019).

A few studies go further and begin to examine whether participants carry out the set of related activities necessary to provide complete data over an extended period (Conrad et al., 2020, Sugie, 2018, Ting et al., 2017). Assessing this systematically is difficult because what constitutes full engagement is likely to vary, depending on the research purpose and study design, and is unlikely to be captured by a single measure. For example, where data about an ongoing behaviour is provided actively by participants, they would need to enter data on every occasion that the behaviour occurred throughout the study period, even though this activity might not take place every day or might take place multiple times on a given day. There are a number of other ways that capturing ongoing behaviours are complex.

To reflect the complexity of the set of tasks that may be associated with a mobile app study, and concept of '**adherence to protocol**' from the medical literature can be employed. This refers to a number of dimensions, such as taking the correct dose of a medicine, for the nominated number of times per day, with or without food as directed, and over the prescribed period (Couper, 2018, Jäckle et al., 2022, Nunes et al., 2009). This connotes active engagement rather than passive compliance (Osterberg and Blaschke, 2005, Tilson, 2004). Although there is no strict parallel with the medical term, drawing on the concept of adherence provides a useful framework for this chapter. Broadly, it can be defined in terms of initiating the activity according to protocol, following the specified regimen, and persisting for the full time-period. The 'protocol' is prescribed and may or may not be fully understood by the recipient or participant. In the next section, the concept is operationalised with respect to this study.

4.1.4 Conceptualising the predictors of adherence

Understanding what characteristics, behaviours or attitudes are associated with adherence to a

research app-based task may help to assess the likely success of future studies of this kind, identify population sub-groups who are less likely to adhere fully, and point to strategies which could improve data quality by overcoming or manipulating some of the barriers encountered. Adherence might be expected to vary by respondents' characteristics, behaviours or attitudes, and these can be conceptualised in three broad areas. First, adherence may vary depending on the **salience** of a research activity. Most obviously, it would be reasonable to expect that active people are be more likely to be compliant with exercise apps. In the case of an expenditure diary, salience might vary depending, for instance, on whether an individual felt in control of their spending, whether they felt responsible for household shopping, and/or whether they shopped frequently or rarely. Adherence might also be affected by an individual's more general **willingness to carry out research**, whether as a general interest in research, or in terms of the priority they might be willing or able to give to ultra-obligatory activities of this kind. It might also depend on the respondents' characteristics since certain attributes may be associated with different behaviours. For example, women may be more likely to shop, older participants may be less confident in using new technologies, and people with lower levels of education may be less interested in or less able to carry out complex research tasks. Therefore, **socio-demographic characteristics** are controlled for, to reduce confounding the effects of other predictors. For example, individuals who use their smartphone frequently and/or for multiple activities might simply have their phone or tablet to hand more often and might feel more confident or able to complete more complex data entry tasks, while those who had more concerns about data security or privacy might hesitate to share information on an ongoing basis. These, and basic demographic characteristics such as age, sex, and education, may help to explain adherence to an app-based research study, and may also help to understand persistence with these activities over time. A full list of potential predictors is listed in Section 4.2.4.

4.1.5 Research questions

In this paper, adherence to protocol in a mobile app study is examined using data collected from the Spending Study which formed part of the *Understanding Society* Innovation Panel, a probability household panel in Great Britain. Participants were asked to use the app every day for one month to record all purchases of goods and services made by any means, whether by cash, card, transfer, in-person or online; to record spending events by photographing shopping receipts or by entering summary information about the purchases made directly into the app; and to report days on which they did not spend any money. These data are used to examine the following questions:

1. To what extent do participants adhere to the Spending Study protocols?
2. Which participant characteristics and behaviours are associated with adherence?
3. Does adherence change over the course of the study month?

The remainder of the chapter is structured as follows: the research study and dataset that forms the basis of this investigation is described in Section 4.2, followed by the analytic and statistical method in 4.2.5, results in 4.4 and discussion in 4.5.

4.2 Data

The statistical models used to address these questions are set out in the methods section, 4.3. However, first, this section describes the *Understanding Society* Innovation Panel, which provides the foundation for this research (4.2.1), and the Spending Study itself (4.2.2). Four measures of adherence are then set out and explained, showing how they are operationalised (4.2.3), and the predictors or covariates employed in the analysis are described (4.2.4). Finally, a summary of the analytical sample is provided (4.2.5).

4.2.1 The *Understanding Society* Innovation Panel

The Innovation Panel is part of *Understanding Society*: the UK Household Longitudinal Study (University of Essex, 2018). The Innovation Panel was designed for methodological testing and experimentation, and is based on a stratified, clustered sample of households in England, Scotland, and Wales (Lynn, 2009). The original sample of 1,500 households were first interviewed in 2008 and followed annually, with refreshment samples of approximately 500 households added at waves 4 and 7. All household members aged 16 and older are eligible for a full, multi-topic interview, and followed if they move within the country. The study uses a sequential mixed mode design: at wave 5 a random third of the sample were issued to face-to-face interviewers; the remaining two-thirds were initially invited to participate online; and non-responders were followed up by face-to-face interviewers. In the final stages of fieldwork, any remaining sample members are followed up by telephone interviewers. This design, and allocation to modes, has been maintained in all waves (Institute for Social and Economic Research, 2020).

The Spending Study was implemented after fieldwork for wave 9 of the Innovation Panel (IP9) in May to September 2016. In preparation for the Spending Study, the IP9 interview included questions related to financial behaviour and mobile device usage. The household response rate for the wave 9 interviews was 84.7 percent, with 85.4 percent of eligible adults within participating households giving an interview (Institute for Social and Economic Research, 2020). Full documentation of the survey design and fieldwork is available at

https://www.understandingsociety.ac.uk/documentation/innovation-panel and the data are available from the UK Data Service at https://discover.ukdataservice.ac.uk/catalogue/?sn=6849.

4.2.2 The Spending Study

The Spending Study was carried out in collaboration with Kantar Worldpanel, who developed the app and implemented fieldwork between October 2016 and January 2017 (Jäckle et al., 2018a). All adult sample members in households where at least one person gave an interview in IP9 were invited to participate in the Spending Study, regardless of whether they had access to the internet, a suitable device, or had expressed willingness to participate in a study of this kind. Each sample member was sent an invitation letter by post and, if an email address was known, by email. Reminders were emailed twice a week for three weeks, and a final reminder letter was sent by post in the fourth week. Advance letters set out the incentives for participation: £2 or £6 for downloading the app (randomly allocated by household), £0.50 for each day they made an entry in the app and £10 as an end of study bonus if they had used the app every day.

Sample members were invited to complete a short registration survey which included questions about the participant's purchasing behaviours. At the end of the registration survey, they were asked to download the app, which was compatible with iOS and Android operating systems, and to report purchases of goods and services for a month. Using the app, participants could photograph and upload receipts, record summary information about purchases without a receipt, report a day without any spending, and access Frequently Asked Questions (see Figure 13 below).

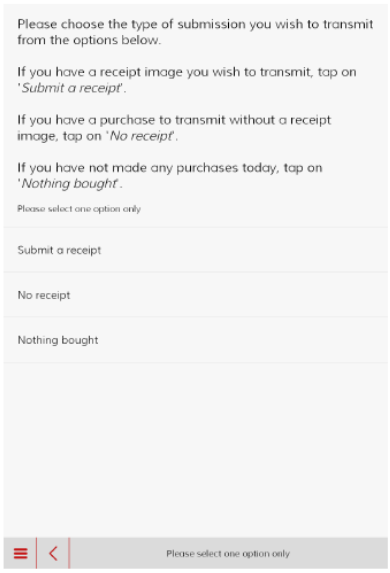
The app sent push notifications at 5pm each day to remind people to make an entry in the app, even if this was to record that they had not made any purchases on that day. Some of the key screens are presented above to give a sense of the app design, but full details can be found in the appendix to the User Guide (Jäckle et al., 2018b). At the end of each week in which participants made at least one app entry, they were reminded of their incentives earned and asked to complete a short online survey. At the end of the fieldwork period, participants were offered a further £3 incentive and asked to complete an online, end of project questionnaire, tailored to reflect their level of participation, to feedback on their experience. Non-respondents to the end of project questionnaire were sent a paper version by post, with a Freepost return envelope, but no incentive. Participants were sent gift vouchers by post. The maximum incentive participants could earn was either £30.50 or £34.50, depending on their experimental group (Jäckle et al., 2018a).

Of the 2,041 respondents in the IP9 interview, 16.5% completed the registration survey, 12.8% used the app at least once, and 10.2% used the app at least once in each of five consecutive

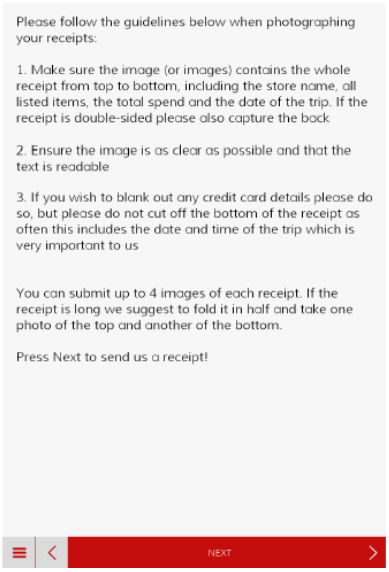
weeks (Jäckle et al., 2019a). It should be acknowledged that this low initial response must cast some doubt on the generalisability of the results, and the response was selective, though it was unbiased with respect to variables associated with expenditure (Jäckle et al., 2019a). The data is deposited at the UK Data Archive (University of Essex, 2018, 2021).

Figure 13 Key examples of the Spending Study app screens

a) Choosing a type of entry

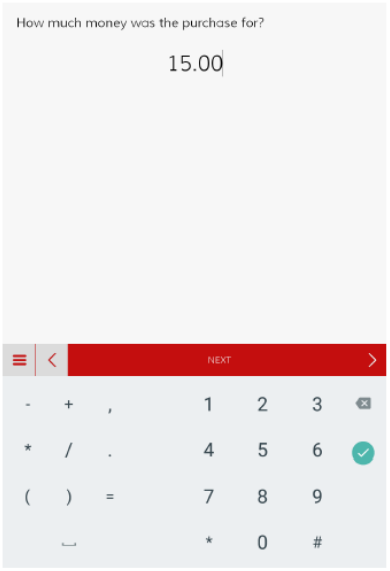


b) Steps to enter a receipt

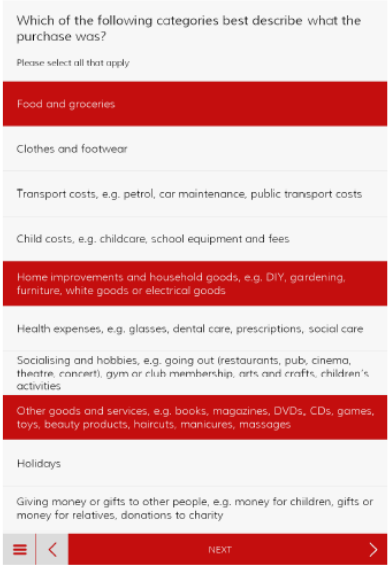


c) Process of entering receipt

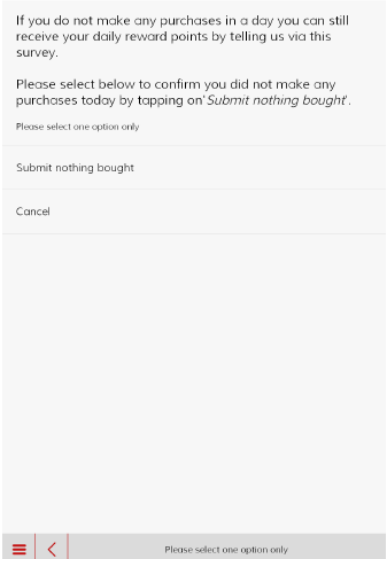
d) Direct entry of spending



e) Category of spending



f) Recording a 'no spend' day



4.2.3 Measures of adherence to the Spending Study protocols

Four aspects of adherence are operationalised, using the data recorded by the app (the type of app entry), paradata recorded by the app (the time and date of each app use), and information coded from the photographed receipts (the time and date of each spending event):

1. ***Whether the participant used the app, by day:*** This is coded as 1 if the participant used the app at least once on that day to photograph a receipt, enter spending events directly, or report no spending, and 0 if the participant did not use the app that day. There are thirty-one observations per participant, starting with the day on which they first used the app. This is a clear measure of adherence, because all participants should make at least one app entry each day.

2. ***Number of spending events reported, by day:*** This is coded as the number of times the participant used the app to report spending events that day, either by photographing a receipt or entering a spending event directly. There are thirty-one observations per participant, starting with the day on which the participant first used the app. Since participants do not necessarily make purchases every day and some will make more frequent purchases than others, this aspect of adherence is assessed by low and/or declining reports of spending events.

3. ***Whether spending events were reported by photographing a receipt or direct entry:*** This is coded as 1 if the spending event was reported by photographing a receipt, since this was the preferred approach to reporting a purchase, and 0 if it was entered directly. The number of observations for each participant equals the number of spending events reported. Participants were given a choice about how to report their spending, but an intrinsic aspect of the app was the facility to scan receipts, and the option to do so was consistently offered first, i.e., "If you have a receipt image you wish to transmit". Furthermore, from the research perspective, this was the preferred mode given the richness of the data from a receipt, compared to the summary information entered. Therefore, it represents a strong measure of adherence.

4. ***Lag between the time of the spending event, as shown on the receipt, and the timestamp on the photograph, showing when it was recorded:*** This is coded as the number of hours between the time stamp on the receipt and the time when the receipt was photographed. The number of observations for each participant equals the number of photographed receipts with non-missing timings data. The rationale for this measure of adherence is that events that are recorded close to the moment of experience are more likely to be remembered and reported accurately. This is particularly true for spending events entered directly into the app where there is no visual cue to make the app entry later, or visual reminder of the amount or categories of spending, as there is with a receipt. However, the time lag for spending events entered directly into the app cannot be measured, as the time of the event itself is unknown. The time lag for photographing

receipts is a valid measure and may reflect the respondent's promptness in recording spending events without a receipt.

4.2.4 Predictors of adherence to the Spending Study protocols

As discussed earlier, the predictors of adherence are conceptualised in terms of salience (expressed here in terms of financial control and purchasing behaviour), willingness to carry out research (expressed in terms of time constraints and past survey compliance), smartphone and tablet use (expressed in terms of data security concerns, frequency of use and number of different activities carried out on mobile devices) and sociodemographic variables as controls. During the development period for the spending study, careful consideration was given to the choice of covariates. To collect these, a short module of questions was added to IP9, ensuring that the covariates were not subject to endogeneity, and where necessary, questions were added to the spending study registration survey. The covariates associated with adherence were operationalised using data from IP9 and the Spending Study registration survey in the following ways.

Financial control: Participants who have an interest in monitoring their own financial situation might be more inclined to adhere to the protocols of an expenditure diary. For example, participants who keep a budget might be more likely to remember to make daily entries, or ask for receipts, or photograph them promptly. The opposite is also possible: participants with more chaotic spending behaviours may identify the Spending Study as an opportunity to reflect on their purchasing practice or exert some control. For this study, the measure of financial control is based on a question in IP9 and was coded as 1 if the answer to the question *"Now, thinking about different ways that people have of managing their finances, how, if at all, do you record your budget?"* was "I don't keep a budget", and coded as 0 otherwise.

***Purchasing behaviour*:** Individuals and households vary significantly in their shopping behaviours, which could affect how easy it is to recall all purchases and how burdensome it is to record them. For example, some may limit their purchases to a large, weekly shop while others may make multiple purchases daily, or combine larger shopping trips with smaller top-up shops. Some may primarily purchase products to consume in the home, while others may make frequent purchases of food and drink to be consumed 'on the go' or on leisure activities outside the home, which may be harder to recall. Whether the participant is responsible for the household shopping might also directly affect the amount of shopping they do.

For this study two measures of the respondent's purchasing behaviour, based on questions in the registration survey, are included: frequency of shopping, and whether the participant was the

main or joint shopper. For frequency of shopping, the response to the question "*How often do you spend money on goods and services?"* was coded into three categories: less than once a day, about once a day and several times a day. In response to the question *"Are you the person mainly responsible for buying goods and services in your household (excluding paying for rent, mortgage, and regular bills)?"* the responses "yes" and "jointly responsible with someone else" were coded as 1, and 'no' was coded as 0.

Time constraints: Even if a respondent has made the initial effort to participate in the Spending Study, those who are very busy may adhere to protocols less completely, or their engagement may decline over time. For example, busy participants may be less inclined to make an app entry every day, they may be less focused on collecting and retaining receipts, or they may delay recording purchases in the app. In practice, people with time constraints do not appear to be less willing to participate in mobile data collection activities (Wenz et al., 2019). However, there is evidence that they have lower response propensities to participate in surveys more generally (Abraham et al., 2006, Groves and Couper, 1998), which suggests that busyness may nevertheless be a factor in a study of this kind. This is operationalised for this study based on questions in IP9. Individuals were coded as 1 if the participant works for more than 40 hours (either employed or self-employed), has more than a one-hour, one-way commute to work, has young children under the age of five in the household, or has other caring responsibilities. Otherwise, they were coded as 0.

Past survey compliance: Participants who are more interested in research or identify strongly with a specific study may be more likely to adhere to the protocols, although it is uncertain whether compliance with standard survey activities is related to adherence to a non-standard activity of this kind. Nevertheless, past survey compliance was operationalised using a measure of item non-response at IP9. This was based on the proportion of eligible questions in the IP9 individual interview to which the participant answered 'don't know', refused or that were otherwise missing. This excludes ten questions about receipt of state welfare and pensions, which are repeated for each income source reported. Item non-response was coded as high if it is above the sample median value (4.1% of variables which the participant was eligible to answer were missing, 'don't know' or refused), and low if it is below. This classification, rather than the observed rate or a more complex measure of past survey compliance, was used for consistency with other analyses (Jäckle, 2019a).

Data security concerns: Participants might have concerns about the security of data transmitted with mobile devices, about providing spending data more generally, or about personal information contained in images of shopping receipts. Participants with such concerns might drop

out of the study more quickly, or reduce the intensity of their response, or may be inhibited from reporting sensitive or socially undesirable spending. Similarly, those who have concerns about using a mobile phone camera to provide images may be less likely to photograph and upload receipts and may instead enter that data directly into the phone. This concept was operationalised using the question in IP9 "*In general, how concerned would you be about the security of providing information in the following ways?"* which was followed by statements about different ways in which mobile devices might be used. The focus here is on the two statements most relevant to this study. When presented with the statement *"Download a survey app to complete an online questionnaire"*, responses were coded as 0 if the answer was not at all concerned, a little concerned or somewhat concerned, and coded as 1 if the answer was very concerned or extremely concerned. Responses for the statement *"Use the camera of your smartphone/tablet to take photos or scan barcodes"* were coded as 0 if the answer was not at all concerned and coded as 1 if it was a little concerned, somewhat concerned, very concerned or extremely concerned. The response scales for these two items were grouped differently. In practice, all study participants chose to download an app to take part in the Spending Study whatever the level of concern they expressed at the time of IP9, so the group of particular interest was identified as those who expressed strong or very strong concern yet took part, since any residual reluctance might influence their behaviour during the study. In contrast, interest lies in whether *any* level of concern about using the camera on the participant's mobile device predicted their subsequent behaviours.

Frequency of use and number of different activities carried out on mobile device: Participants who use their device frequently and for multiple purposes might be more willing and able to participate in mobile data collection tasks. Indeed, device familiarity has been shown to be associated with increased smartphone use to complete web questionnaires (Couper et al., 2017) and willingness to use mobile technologies (Wenz et al., 2019). However, it is also possible that proficient mobile device users may quickly learn to optimise their entries or satisfice, while less familiar users may be excited by the novelty factor or may take greater pains. This concept was operationalised based on questions in IP9 asked separately for smartphone and tablet. Responses were combined by giving the higher code priority. Frequency of device use is coded as 1 if the device is used daily, and 0 otherwise. Range of use is based on a question about which activities respondents carry out on their device, with the listed uses being browsing websites, email, taking photos, looking at content on social media websites/apps, posting content to social media websites/apps, making purchases, online banking, installing new apps, connecting to other electronic devices via Bluetooth, using GPS/location-aware apps, playing games and streaming videos or music. For the main analysis presented here, this was coded as low (0 or 1 activity), medium (2-8 activities) or high (9-12 activities), applied consistently for all subsequent analyses.

Socio-demographic characteristics: Participant characteristics may also explain variations in behaviour. For example, women may be more likely to shop, older participants may be less confident in using new technologies, and people with lower levels of education may be less interested in or able to carry out complex research tasks. Therefore, socio-demographic characteristics were controlled for, to reduce confounding the effects of other predictors. Three socio-demographic characteristics were controlled for, drawn from IP9: gender, age, and education. Gender was coded as 1 if female and 0 if male. Age was coded in six categories: 16-30, 31 to 60 in ten-year age bands, and 61 and over. Education was coded in three categories: degree; school or other higher qualification; and lower, no qualification or missing information.

4.2.5 Analysis sample

The analysis is based on two data sets, and a detailed explanation of how they were cleaned and prepared for analysis is provided in Appendix I. In summary, the first dataset is a long file which contains one record for each app entry. This app entries data set contains 9,386 records provided by 268 respondents (see Table 27 below). These records are comprised of 7,412 reported spending events (provided by 259 participants who reported at least one spending event) and 1,974 'no spend' days. In turn, the spending events are comprised of 2,820 direct entries and 4,592 photographed receipts, of which 3,454 had date and time information (provided by 236 participants who provided at least one photographed receipt with valid date and time information).

Table 27 Summary information about the analysis sample

	Based on activities during study period (days 1-31)	Number		
Participants	Used app at least once	268[1]		
	Reported at least one spending event	259		
	Photographed at least one receipt with valid date and time	236		
App entries	Photographed receipts with valid date and time	3,454		
	Photographed receipts without valid date and/or time	1,138		
	Total photographed receipts	*4,592*	4,592	
	Direct entries		2,820	
	Total reported spending events		*7,412*	7,412
	No-spend days			1,974
	Total app entries			*9,386*
Missing person-days (days 1-31)				2,522

Note 1: Two cases for whom there were no valid app entries were excluded, reducing the sample which formed the basis of response and bias in Jäckle et al., from 270 to 268 (2019a).

The long file of app entries was transformed to create a second dataset with 31 records for each of the 268 participants. In this 'study days' dataset, each record contains variables summarising app activity on 31 consecutive days of the study, with day 1 set to the first day the app was used i.e., the number of direct entries, the number of receipts photographed, whether the respondent recorded a 'no spend' day, and a flag showing if the respondent failed to make an entry that day. Taking all of this into account, Table 28 below shows, for each measure of adherence (row 1): the dependent variable used in the analysis (row 2); the data set used (row 3); the unit of analysis (row 4); the final number of observations available for analysis for each form of adherence (row 5); the different sample sizes (row 6); and, finally, the type of statistical analysis used (row 7) which is explained in the next section.

Table 28 Detailed information showing the genesis of the analytic dataset

1	Measure of adherence ->		Daily app uses	Number of spending events	Method of reporting spending events	Time lag, spending event to report
2	Dependent variable ->		Y=1 if receipt entered, direct entry or no spend 0 = if missing day	Y = count of number of spending events reported	Y=1 if photographed receipt 0=if direct entry	Y=hours between receipt time and photograph time
3	Data set used ->		Study days	Study days	App entries	App entries
4	Unit of analysis ->		One observation per day (days 2-31[1]) 268 x 30	One observation per day (days 1-31) 268 x 31	One observation per reported spending event	One observation per photo'd receipt
5	Observations in analysis ->		8,040	8,308	7,412	3,454
6	Individuals used in analysis ->		268	268	259	236
7	Type of multilevel regression analysis used ->		Logit	Negative binomial	Logit	OLS

Note 1: The analysis of daily app use is based on days 2-31 because, by definition, all participants made at least one valid app entry on their first study day, hence the number of records is 8,040 i.e., 268 x 30.

4.3 Statistical methodology

To examine the first research question, to what extent do participants adhered to the Spending Study protocols, the overall distribution of each of the four measures of adherence is described, showing how these vary between the participants. These are presented graphically using cumulative density functions.

The extent to which the four adherence measures are related to each other is then considered by examining their correlations. These are presented graphically using a scatter plot matrix. Kendall's tau-b, a nonparametric correlation estimator, is used to estimate the strength and significance of the associations between each combination. This statistical test is insensitive to outliers and has the advantage that it is possible to express the results in terms of the proportion of concordant or discordant pairs which is intuitive (Croux and Dehon, 2010). The formula for Kendall's tau a conveys the essential aspects of this test of association:

$$\tau = (C - D) \; / \; (C + D)$$

where *C* is the number of concordant pairs and *D* is the number of discordant pairs. The statistical significance can be calculated using a z-score (given the assumption of a normal distribution where the number of pairs are significantly higher than 10, which is the case here) where *n* is the number of pairs:

$$z = 3\tau * \sqrt{n(n-1) \; / \sqrt{2(2n+5)}}$$

In practice, Kendall's tau b follows the same principle but is slightly more complex than Kendall's tau a because the denominator takes account of ties. An alternative, Pearson correlation coefficient, was considered since this makes full use of the continuous data that is available and is generally well understood. In practice, the results were very similar using either statistical approach.

To address the second research question, and establish which participant characteristics, behaviours and attitudes are associated with more or less adherent behaviours for each of the four outcome variables, multivariate regression analysis is carried out with a vector of covariates. To account for the hierarchical nature of the data (in which observations were clustered within days and/or days within participants), all four models use mixed effects or multilevel regression. The exact type of model, and the analysis sample, varies depending on the nature of the measure of adherence, as follows:

4.3.1 Model 1: Daily app use

The probability of using the app (whether to report a spending event or a no-spend day), was modelled using multilevel logistic regression, with one observation per participant and study day,

using the STATA 15.0 estimation command `melogit`[10]. Model 1 accounts for the clustering of study days (level 1) within participants (level 2). By definition, all participants made at least one valid app entry on what was consequently defined as study day 1, resulting in complete collinearity. Therefore, this model is based on study days 2-31, with 8,040 observations for the 268 participants who used the app at least once. (Williams, 2012)

The statistical representation of this modelling approach requires, in the first instance, a reminder of the generalised linear random intercept model for a *continuous* dependent variable, which can be expressed as:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

$$u_j \sim N(0, \sigma_u^2) \text{ and } e_{ij} \sim N(0, \sigma_e^2)$$

where i is the number of observation (in this case study days), the level 1 units,

j is the number of groups (in this case individual study participants who used the app), the level 2 units,

y_{ij} is the response for individual i in group j

x_{ij} is an individual level covariate, although in practice, there is a vector of covariates

u_j is the effect of being in group j, otherwise known as the level 2 residual; this is the subject level random variable which distinguishes multilevel models from standard regression, and

e_{ij} is the measurement level random variable, that is, the more familiar residual or random error term (level 1).

In this case, the expected value of the dependent variable y_{ij} for a given $x_{ij} + u_j$ can be expressed as:

$$E(y_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$$

However, in the case where the dependent variable y_{ij} is a binary response as it is here (where app used = 1 and app not used = 0) for observation i (in this case days) in group j (in this case individuals), the expected value of the dependent variable y_{ij} can be expressed as:

$$E(y_{ij}) = \Pr(y_{ij} = 1) = \pi_{ij}$$

[10] An alternative approach was considered, based on the proportion of days an app entry was made, with one observation per participant.

Depending on the nature of the underlying distribution, the model is expressed with a link function, F^{-1} which can take several forms:

$$F^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$$

Here, where the link function F^{-1} is a logit, the model is represented as:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + u_j$$

$$u_j \sim N(0, \sigma_u^2)$$

In this case, β_0 is the log-odds that $y = 1$ when $x = 0$ and $u = 0$ and β_1 is the unit specific effect of x, in other words the effect on the log-odds of a 1-unit increase in x for observations (i.e., days) within the same group or u (i.e., individual). By extension, $exp(\beta_1)$ is an odds-ratio, comparing the odds for observations 1-unit apart (i.e., days) within the same group (i.e., individual).

The random element of this equation u_j, that is the level 2 residual, is the effect of being in group j on the log-odds that $y = 1$. In this study, the main purpose of modelling app use over study days within a multilevel model, where study days are clustered within individual, is to account for non-independence of study days, i.e., to control for differences in variance between individuals. This is because the primary research interest is how app use varies according to the characteristics and behaviours of individuals. Nevertheless, the random effects are also of interest, and were all significant.

An alternative statistical approach of modelling app use was considered. This involved calculating the proportion of study days an app entry was made, with one observation per participant and using the STATA command `fracreg logit`. However, there are a number of challenges fitting and interpreting the results of regression models where the dependent variable is a proportion. After careful consideration, `melogit` was used since this approach makes it possible to control for variation within the individual and means that a relatively consistent approach is used for the four models, each of which uses mixed effects models with either a second, or a second and a third level.

In this study, having carried out multilevel logistic regression by applying the command `melogit`, the results of the model are presented as Average Marginal Effects (AME), that is, the percentage point increase or decrease in the predicted probability of using the app on a given study day, associated with a one-unit change in the covariate (Williams, 2012). This is calculated using the STATA `margins, dydx()` command. Here, the average of the logistic probability density function

for all values of x are multiplied by the coefficient. This gives a number between 0 and 1 which represents the average change in probability when the value of x increases by one. The AME shows the change in the expected number of people using the app should one of the covariates change by a single unit, for example, looking at the difference between those who are the household main shopper and those who are not, holding all other covariates in the model constant. In this way, the AME isolates the expected change due to a single covariate and quantifies the impact of that covariate on the dependent variable, in this case the probability of using the app. The benefit of this approach is that rather than generating a result that will differ between individuals, the average marginal effect provides an average.

4.3.2 Model 2: Number of spending events

The number of spending events reported per day (including both photographs and direct entries) was modelled using multilevel negative binomial regression, a form of regression which is appropriate for predicting a count-based variable where the conditional mean is not equal to the conditional variance. This distinguishes it from the Poisson distribution, which is nested in the negative binomial model, but makes the assumption that variance is equal to the mean and is therefore sensitive to over-dispersion as is observed in the distribution of the dependent variable, where variance is greater than the mean. The analysis is based on one observation per participant and study day, using the STATA 15.0 estimation command `menbreg`. The modelling approach here accounts for the clustering of study days within participants (level 2). In this instance, all 31 study days are utilised (1-31) so the analysis is based on 8,308 observations for the 268 participants who used the app at least once.

The 2-level negative binomial regression model uses a link function F^{-1} which reflects the underlying distribution:

$$F^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$$

Here, the link function F^{-1} is $\log(y_{ij})$:

$$\log(y_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$$

$$u_j \sim N(0, \sigma_u^2)$$

A two-level negative binomial model can also be presented in the following way (as shown in the description of the `menbreg` command). For cluster $j, j = 1, \ldots, M$, the conditional distribution of $y_j = \left(y_{j1}, .., y_{j_{n_j}}\right)'$, given a set of cluster-level random effects u_j and the conditional overdispersion parameter α in a mean-overdispersion parameterisation, is:

$$f(y_j|u_{j,}\alpha) = \prod_{i=1}^{n_j}\left\{\frac{\Gamma(y_{ij}+r)}{\Gamma(y_{ij}+1)\Gamma(r)}p_{ij}^{r}(1-p_{ij})^{y_{ij}}\right\}$$

$$= \exp\left[\sum_{i=1}^{n_j}\{\log\Gamma(y_{ij}+r) - \log\Gamma(y_{ij}+1) - \log\Gamma(r) + c(y_{ij,}\alpha)\}\right]$$

Where $c(y_{ij},\alpha)$is defined as

$$-\frac{1}{\alpha}\log\{1+\exp(\eta_{ij}+\log\alpha)\} - y_{ij}\log\{1+\exp(-\eta_{ij}-\log\alpha)\}$$

and $r = {}^{1}/_{\alpha}$, $p_{ij} = {}^{1}/_{(1+\ \alpha\mu_{ij})}$ and $\eta_{ij} = x_{ij}\beta + \ z_{ij}u_j$

The estimation results are presented as the Incidence Rate Ratio (IRR); that is, the factor by which the predicted number of spending events increases or decreases, associated with a one-unit change in the covariate.

We should note, since the number of spending events is a count variable, this was initially analysed using a Poisson model, with the STATA command `mepoisson`, but this was tested and rejected due to over-dispersion (Dean and Lawless, 1989, Molla and Muniswamy, 2012). An alternative approach would have been to continue to use a Poisson model with adjustment for overdispersion (see for example the discussion in Molla and Muniswamy, 2012).

4.3.3 Model 3: Method of reporting spending

The probability of reporting a spending event by photographing a receipt rather than direct entry was modelled with a 3-level multilevel logistic regression, with one observation per recorded spending event using the STATA 15.0 estimation command `melogit`. This model (Model 3) accounts for an additional level; for the clustering of direct entries and photographed receipts within study days (level 2) and within participants (level 3). This can be represented statistically as follows, with a logit as the link function.

$$\log\left(\frac{\pi_{ijk}}{1-\pi_{ijk}}\right) = \ \beta_0 \ + \beta_1 x_{ijk} + v_{jk} + u_{jk}$$

$$v_k \sim N(0,\sigma_v^2), u_{jk} \sim N(0,\sigma_u^2).$$

This model 3 is based on 7,412 observations from the 259 participants who reported at least one

spending event. The estimation results are presented as Average Marginal Effects (AME); that is, the percentage point increase or decrease in the predicted probability of reporting a spending event by photographing a receipt, associated with a one-unit change in the covariate. As before, the purpose of using multilevel modelling is not to investigate the random effects associated with study day, or individual, *per se*. Rather, it is to correctly reflect the three-level structure of the data so as not to misattribute response variation to the level of interest which is the observation of spending events and to reveal how these vary according to respondent characteristics and behaviours, and also to allow correct estimation of effects given the clustering. Consequently, attention was given to establishing whether the multilevel models provided a better fit than a single level model and all three levels were significant, improving the model fit.

4.3.4 Model 4 Time lag between spending and reporting

The time lag between the spending event and when the receipt was photographed was modelled using a multilevel linear regression model, with one observation for each photographed receipt with full time and date information, using the STATA 15.0 estimation command `mixed`. Model 4 accounts for the clustering of photographed receipts within study days (level 2) and within participants (level 3). It is based on 3,454 observations recorded by the 236 participants who photographed at least one receipt. To account for the highly skewed distribution, the logarithm of time is modelled (in hours), and the results are presented as the percentage change in the time lag, which is associated with a one-unit change in the covariate.

$$y_{ijk} = \beta_0 + \beta_1 x_{ijk} + v_k + u_{jk} + e_{ijk}$$

$$v_k \sim N(0, \sigma_v^2),\ u_{jk} \sim N(0, \sigma_u^2) \text{ and } e_{ijk} \sim N(0, \sigma_e^2)$$

The two grouping variables, study days and participants, are treated as random effects. The respondent characteristics predicting adherence at the individual level (level 1) are treated as fixed effects. Likelihood ratio tests confirmed that accounting for clustering of observations within study day (for Models 3 and 4) and within participants (for all Models), by treating day and person identifiers as random effects, resulted in better fitted models than without random effects.

All four models include the same set of covariates (listed in section 4.3.5) to test which factors are associated with adherence and control for sociodemographic characteristics (age, gender, and qualifications). To assess the significance of the covariates in the four models, AIC was used to assess model fit.

4.3.5 Modelling the effect of time

To address the third research question "*does adherence change over the course of the study month*", each outcome measure was plotted across the 31 days of the study. Linear fits were estimated from the scatter plots using the `lfit` command in STATA 15.0, which calculates the predicted values from a linear regression of the measure of adherence on study day (starting at day 2). A test was then conducted to determine whether the trends observed are statistically significant. This was done by adding two indicators of time to the four regression models estimated to answer the second research question. This ensured that time was considered alongside participant characteristics, behaviours and attitudes which might be associated with more or less adherent behaviours. These are a binary indicator of whether the observation is from the first day of the study or a later day (where day one was coded as 1 and all other days were coded as 0), and a continuous variable identifying the study day (this ranged from 1 to31). Other possible specifications of time were tested but disregarded. These were including only the continuous study day identifier, quadratic or cubic terms, and splines. This was done by testing model fit using likelihood ratio tests to compare nested models and using AIC criteria to assess non-nested models with a consistent sample.

4.4 Results

Before presenting the main results, Table 29 provides descriptive statistics for each potential predictor of adherence to the Spending Study. For most indicators, fewer than 3 observations are missing in the original data. However, frequency of device use was missing for 15 participants, and concerns about security were missing for 17 participants. These include respondents who were not asked the corresponding questions in the IP9 interview because they had completed the interview in CATI and were therefore not routed into the self-completion module containing these questions; were CAPI respondents but declined to complete the self-completion section; had reported not using the internet; or were not using a mobile device to connect to the internet. Except where stated below with respect to education, missing observations for categorical variables are set to the modal category, and missing observations for continuous variables are set to zero. For all variables, the approach was consistent with that taken in the related study of coverage and participation rates (Jäckle et al., 2019a).

Table 29 Descriptive statistics for explanatory variables used to predict adherence

Concept	Covariate	Category	N	% or mean
Salience of finance topic	Budget keeping	Keeps a budget	137	51.1
		Does not keep a budget	131	48.9
Purchasing Behaviour	Frequency of shopping	Less than once a day	164	61.2
		About once a day	74	27.6
		Several times a day	30	11.2
	Role as shopper in household	Main or joint shopper	234	87.3
		Not main or joint shopper	34	12.7
Time Constraints	Time constraint	Is not time constrained	191	71.3
		Is time constrained	77	28.7
Past survey Compliance	IP9 item non-response rate	Low	178	66.4
		High	90	33.6
Data security concerns	Concern about survey app	Not, a little or somewhat	237	88.4
		Very or extremely	31	11.6
	Concern about using camera	Not at all	109	40.7
		Little, somewhat, very, extremely	159	59.3
Frequency and range of mobile device uses	Range of activities carried out on device	Low (none or 1)	20	7.5
		Medium (2-8 activities)	68	25.4
		High (9-12 activities)	180	67.2
	Frequency of use	Less often or never	29	10.8
		Everyday	239	89.2
Socio-demographic characteristics	Gender	Male	106	39.6
		Female	162	60.4
	Age	16-30	60	22.4
		31-40	61	22.8
		41-50	60	22.4
		51-60	48	17.9
		61+	39	14.6
	Qualifications	Degree	86	32.1
		School or other higher qual.	159	59.3
		Lower, none, missing	23	8.6

N = 268 participants

4.4.1 To what extent do participants adhere to the Spending Study protocols?

The first measure of adherence is based on daily app use. Figure 14a shows the cumulative density function for the number of study days on which each of the 268 participants used the app, either to report a spending event or a no spend day. As explained earlier, by definition, all participants made at least one valid app entry on their first study day, but 4.1% of participants did not use the app on any subsequent day. Most participants used the app quite intensively over the

study period: the mean number of app use days was 21.7, with a median of 24. About a third of participants (31.3%) used the app on at least 28 days, including 4.1% who used the app on every day of the study.

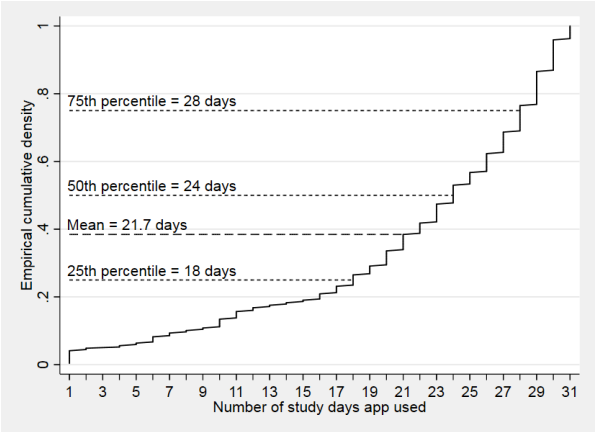
The second measure of adherence is based on the average number of spending events reported per day. The total number of spending events across the whole period ranged from zero to over eighty, with an outlying participant who recorded 131 spending events. Figure 14b shows the distribution of the mean number of spending events reported per day for the 268 participants. This shows that 3.4% of study participants did not record any spending events, while 4.1% recorded on average two or more spending events a day. The mean number of spending events per day was 0.89 with a median of 0.81.

The third measure of adherence is based on whether participants reported spending events by photographing receipts rather than entering them directly in the app. Figure 14c shows the distribution of the proportion of spending events reported with receipts, based on the 259 participants who reported at least one spending event. The graph shows that 5.0% of participants reported all spending events by directly entering them in the app, while 10.4% reported all spending events by photographing receipts. Most used a mix: on average, the proportion of spending events that participants reported by photographing a receipt was 0.61 with a median of 0.65.

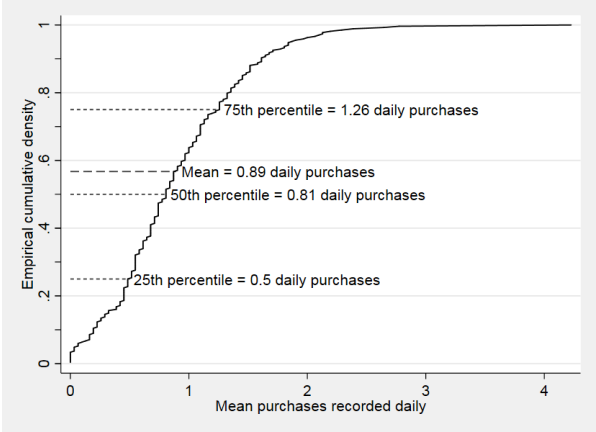
The fourth measure of adherence is based on the lapsed time between the moment a spending event takes place and the study participant photographing the receipt. Figure 14d shows the average time lag for each of the 236 participants who entered at least one valid receipt with full time and date information. The graph is truncated at 24 hours, which covers 94.9% of participants. The distribution is skewed with a mean time lag of 7.7 hours and a median of 4.7 hours. Just 2.5% of participants had an average time lag of less than an hour, 12.7% had an average time lag of less than two hours and a little under a quarter (23.7%) had an average time lag of less than 3 hours.

Figure 14 Variation in adherence between participants

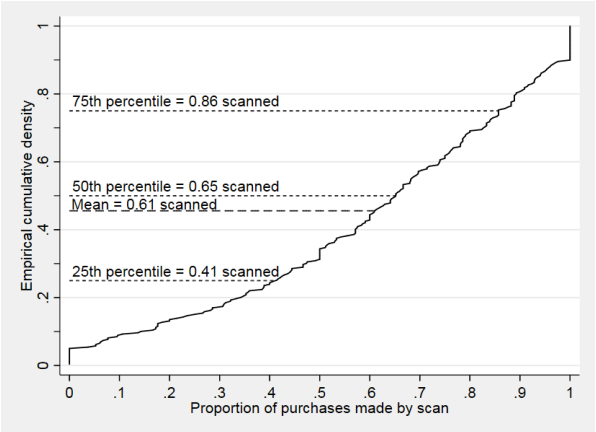
(a) Number of days participants used app, n=268



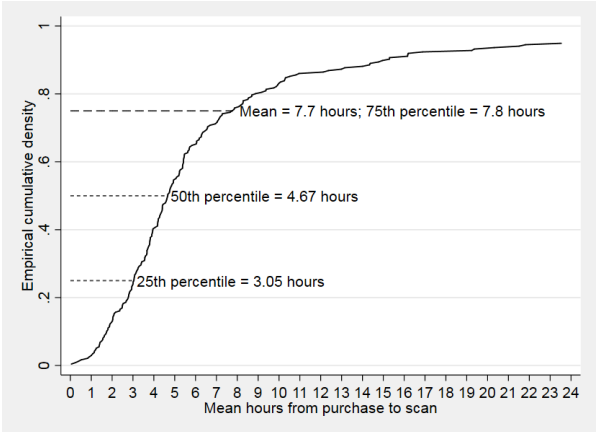
(b) Mean number of spending events reported daily, n=268



(c) Proportion of spending events participant recorded by photographing a receipt, n=259



(d) Mean time lag across all photographs made by participant, truncated at 24 hours, n=236



The scatter plots in Figure 15 examine whether participants who adhered to the Spending Study protocol in terms of one behaviour also adhered in terms of the others. For clarity, the graph excludes two outliers: one who, on average over the study period, reported more than four spending events per day and one where the average time from spending event to photographing the receipt was 118 hours. The correlation analysis was not sensitive to the inclusion or exclusion of these cases. The graph suggests a possible relationship between the number of days the app was used and the mean number of daily spending events (top row). At first consideration, this seems logical and even necessary, but some participants who shop rarely may record many 'no spend' days, while others who shop infrequently may report multiple spending events concentrated in only a few days. In practice, there is a positive correlation between the number of app use days and the mean number of spending events per day (Kendall's tau b = 0.404, $p<0.001$, n=268), which can be expressed as a positive correlation of 70.2% of possible pairs.

Figure 15 Scatter plots of the four measures of adherence

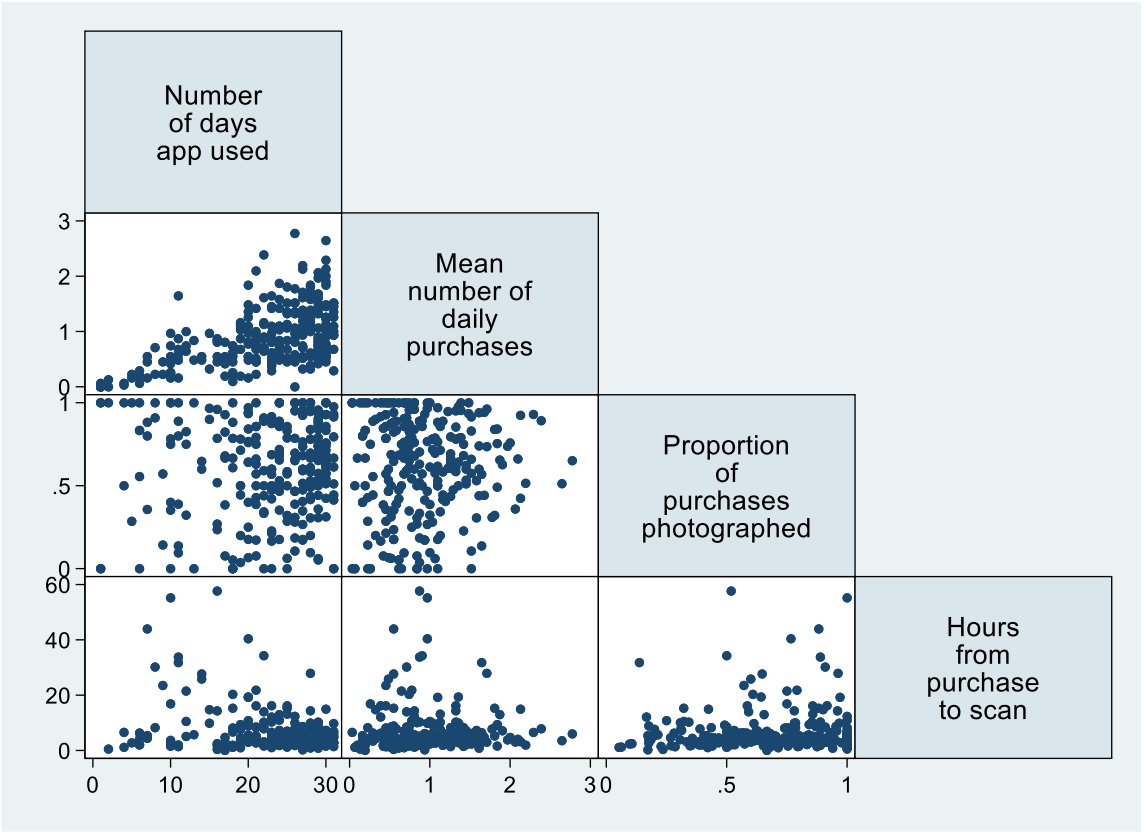


Figure 15 also suggests a possible association between the time lag from the moment the spending event takes place to the receipt being photographed and the other three outcomes (bottom row). In practice, only one of these relationships is statistically significant: there is a weak, negative relationship between the number of app use days and the average time lag between the moment a spending event takes place and the time the receipt is photographed (τ_b = -0.090, p=0.0451, n=236), which can be expressed as a negative correlation of 45.5% of possible pairs. This suggests that participants who use the app on more days tend to photograph receipts longer after the spending event. Overall, it seems the four measures of adherence to study protocol are largely independent of each other.

4.4.2 Which characteristics and behaviours are associated with adherence?

Results of the four final multilevel models used to examine the characteristics, behaviours and attitudes associated with more or less adherent behaviour are presented in Table 30 below. The bivariate relationships between each covariate and the four measures of adherence are reported in Appendix H. For each covariate, this shows each coefficient in relation to the reference category, its p-value, and also a joint test (here, a Wald test) showing whether the set of items which represent a categorical variable (such as age) are collectively significant within the model.

The discussion of results is focused on relationships which are statistically significant when controlling for all other covariates.

Although it was anticipated that participants with a stronger **interest in personal finance** might engage more fully with the study, whether the participant had reported keeping a budget was not associated with any of the four measures of adherence. Based on this result at least, adherence to the Spending Study does not seem to be associated with financial control.

It was also anticipated that differences in **purchasing behaviours** could affect the amount of reporting necessary to fully adhere to the study and so alter both the burden and salience of the task. However, **frequency of shopping** was not a statistically significant predictor of adherence. In contrast, participants who were not the **main or joint shopper** in the household appeared to be less engaged: they had an 11.7 percentage point lower predicted probability of using the app daily on a given day (AME=-0.117, SE=0.057, p=0.042) than those who were the main shopper or jointly responsible for household shopping. They were also 39.4% less likely to make a purchase (IRR=0.606, SE=0.096, p=0.002). These results seem to suggest that the Spending Study is less salient for those who are not responsible for household shopping.

Being a busy participant (those with long working hours, a long commute, or caring responsibilities) is associated with lower levels of adherence to the study protocols in two ways. In comparison with participants who were not classified as time constrained, they had a 9.1 percentage point lower predicted probability of using the app on a given day (AME=-0.091, SE=0.036, p=0.013) and were 31.9% slower to photograph their receipts after a spending event (0.319, SE=0.179, p=0.042).

Participants who had been **less compliant with the panel study in the past** were also associated with less adherent behaviours with respect to the Spending Study protocols. Compared to participants with low item non-response rates in the IP9 interview, those with **a high item non-response** rate had a 9.4 percentage point lower predicted probability of using the app on a given day (AME=-0.094, SE=0.033, p=0.005), and reported 22.8% fewer spending events (IRR=0.772, SE=0.073, p=0.007). However, past survey compliance was not associated with the probability of reporting spending events by photographing receipts, or with the delay in photographing receipts.

Participants who had reported being **very or extremely concerned about the security of providing data with a survey questionnaire app** (who nevertheless went on to take part in the Spending Study) appeared to show a weaker adherence to its protocols: they had a 12.4 percentage point lower predicted probability of using the app on a given day (AME=-0.124, SE=0.053, p=0.020), and reported 27.2% fewer spending events (IRR=0.728, SE=0.106, p=0.030) when compared to participants who were less concerned. Participants who had **expressed**

concerns about using the camera of their smartphone or tablet to take photos or scan barcodes for a survey were also associated with less adherent behaviours: they had a 9.7 percentage point lower predicted probability of using the app daily (AME=-0.097, SE=0.031, p=0.002), and reported 20.0% fewer spending events (IRR=0.800, SE=0.074, p=0.016) than participants who had expressed no concern. However, data security concerns were not associated with the method of entering spending events into the app, or the delay before doing so.

Frequency of mobile device use only helps to explain one of the four outcomes: participants who reported using their mobile device every day had a 16.7 percentage point higher predicted probability of photographing receipts rather than entering information directly in the app (AME=0.167, SE=0.074, p=0.025), compared with those who use their mobile device less frequently. Participants who used their device for a wide range of activities (reporting 9-12 activities) were 11 percentage points less likely to use the app on a given day (AME=-0.110, SE=0.036, p=0.002) than moderate users (reporting 2-8 activities). Alternative approaches to modelling intensity of device use were considered, using a continuous variable of the number of activities the participants carried out using their device (ranging from 0 to 12) where mean=8.8 and SD=3.4, and using a quadratic term. However, the best fitted model across the four outcome measures used a categorical measure of intensity of device use and provided a consistent approach.

The final set of predictor variables account for differences in the socio-demographic characteristics of participants. Although there are some associations in the bivariate relationships between socio-demographics and adherence outcomes, after controlling for other characteristics in the models there are no statistically significant associations between gender or education and the four measures of adherence, and just one significant association with age (based on a joint test, p=0.001). There, the suggestion is that those at the top of the age distribution are slower to photograph their receipts after a spending event, in comparison with younger age groups.

Table 30 Predictors of study protocol adherence

		Pr (used app)			Number of spending events			Pr (spending event entered by receipt)			Log (time of spending event - receipt being photographed)		
	Mean (SD)	0.701 (0.46)			0.893 (1.30)			0.620 (0.49)			7.43 hours (17.65)		
	Mixed effects model	**Logit**			**Negative binomial**			**Logit**			**OLS (log time)**		
Concept or covariate	Category	AME	SE	P-value	IRR	SE	P-value	AME	SE	P-value	exp(b)-1	SE* exp(b)	P-value
Financial control (Ref: does keep budget)	Does not keep budget	-0.037	0.031	0.220	0.921	0.082	0.357	-0.003	0.036	0.924	-0.090	0.103	0.405
Frequency of shopping (Ref: less than once a day)	About once a day	-0.017	0.035	0.634	1.138	0.118	0.213	-0.011	0.042	0.790	-0.163	0.109	0.174
	Several times a day	**-0.103**	**0.052**	**0.048**	1.155	0.165	0.312	0.004	0.058	0.940	-0.100	0.161	0.555
	Joint test, chi^2, p-value	*4.15*	*0.126*		*2.07*	*0.356*		*0.09*	*0.955*		*1.92*	*0.382*	
Role as shopper in HH (Ref: main shopper)	Not main shopper	**-0.117**	**0.057**	**0.042**	**0.606**	**0.096**	**0.002**	-0.003	0.066	0.969	-0.293	0.156	0.116
Time constraint (Ref: is not time constrained)	Is time constrained	**-0.091**	**0.036**	**0.013**	0.843	0.088	0.101	0.019	0.042	0.656	**0.319**	**0.179**	**0.042**
Past survey compliance (Ref: low IP9 item non-response)	High IP9 item non-response	**-0.094**	**0.033**	**0.005**	**0.772**	**0.073**	**0.007**	0.058	0.038	0.132	0.129	0.138	0.318
Concern about using survey app (Ref: not at all, little, somewhat concerned)	Very/extremely concerned	**-0.124**	**0.053**	**0.020**	**0.728**	**0.106**	**0.030**	0.002	0.061	0.974	0.412	0.281	0.083
Concern about using camera (Ref: not at all concerned)	Little/somewhat/v./extremely concerned	**-0.097**	**0.031**	**0.002**	**0.800**	**0.074**	**0.016**	0.000	0.037	0.992	0.043	0.120	0.713
Freq of mobile device use (Ref: less often or never)	Every day	0.047	0.064	0.463	0.941	0.171	0.739	**0.167**	**0.074**	**0.025**	-0.075	0.215	0.735
Number of activities done on device (Ref: 2-8 activities)	None or 1 activity	-0.156	0.080	0.050	**0.630**	**0.140**	**0.038**	0.003	0.088	0.976	0.243	0.288	0.399
	9-12 activities	**-0.110**	**0.036**	**0.002**	0.832	0.095	0.107	-0.068	0.045	0.130	-0.070	0.141	0.621
	Joint test, chi^2, p-value	*10.84*	*0.004*		*5.8*	*0.055*		*2.36*	*0.308*		*1.14*	*0.565*	
Gender (Ref: male)	Female	0.002	0.031	0.958	1.161	0.105	0.099	0.045	0.037	0.222	0.257	0.148	0.053
Age (Ref: 16-30)	31-40	-0.028	0.048	0.556	1.030	0.146	0.836	0.049	0.06	0.416	-0.218	0.145	0.186
	41-50	0.033	0.050	0.500	1.237	0.184	0.152	**0.125**	**0.062**	**0.043**	-0.146	0.166	0.415
	51-60	0.044	0.052	0.397	1.227	0.193	0.195	**0.168**	**0.064**	**0.009**	-0.160	0.170	0.390
	61+	-0.049	0.062	0.436	1.216	0.215	0.269	0.127	0.072	0.079	0.664	0.377	0.024
	Joint test, chi^2, p-value	*5.05*	*0.283*		*3.38*	*0.496*		*8.14*	*0.086*		*18.66*	*0.001*	
Qualifications (Ref Degree)	School/other higher qualification	0.025	0.034	0.467	0.897	0.087	0.262	-0.037	0.039	0.348	0.022	0.123	0.857
	Other, none, missing	0.066	0.059	0.262	0.829	0.151	0.302	0.089	0.068	0.193	-0.080	0.210	0.713
	Joint test, chi^2, p-value	*1.3*	*0.522*		*1.7*	*0.428*		*3.73*	*0.155*		*0.25*	*0.885*	
Goodness of fit (likelihood ratio test)	Wald, P>chi2	*64.93*	*0*		*76.98*	*0*		*29.22*	*0.046*		*41.42*	*0.001*	
Observations	N		8,040			8,308			7,412			3,454	
	Individuals		268			268			259			236	

Note: IRR=Incidence Rate Ratio (negative binomial regression); AME=Average Marginal Effect (logistic regression)

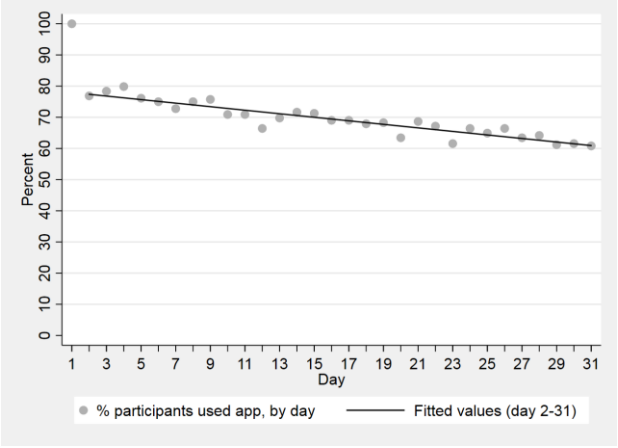
4.4.3 Does adherence change over the course of the study month?

Figure 16 plots each outcome measure across the 31 days of the study while Table 31 extends the regression models presented to answer the second research question, estimating the effects of time. Time is indicated by a binary indicator of first or subsequent study day and a continuous indicator of study day. The inclusion of the other covariates used in the models to address the second research question does not alter the estimated effects of time.

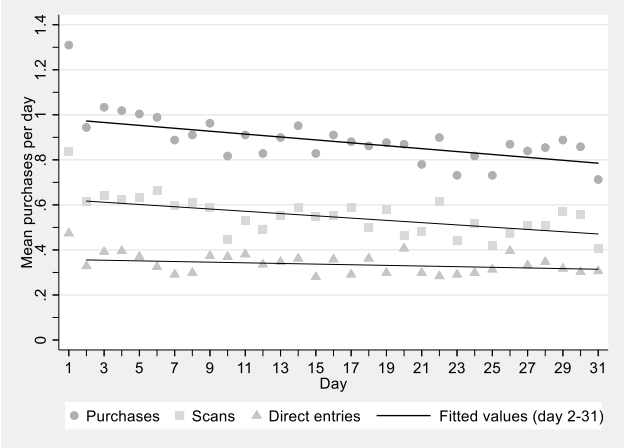
Figure 16a shows the percentage of participants who used the app at least once, by day. The scatter plot suggests a sharp drop in app use from 100% on day 1 to 76.9% on day 2. From day 2 onwards there is some fluctuation, but the trend is of a more gradual decline reaching 60.8% on day 31. The regression results in Table 31 confirms the monotonic decline, with a 0.6% lower predicted probability of using the app for each additional study day (SE=0.001, p<0.001).

Figure 16 Adherence to study protocol over time

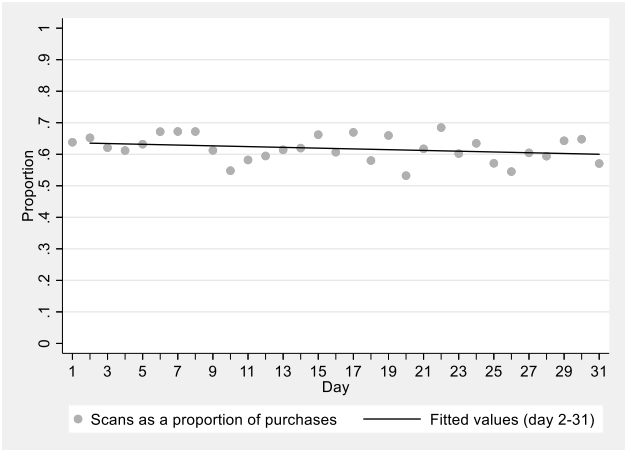
a) Percentage of participants who used the app, by day (n=268)



b) Mean spending events (photographs & direct entries) recorded by participants, by day (n=268)



c) Proportion of total number of spending events reported by photographing receipts, by day (n=259)



d) Average time lag between spending event and photograph, by day (n=236)

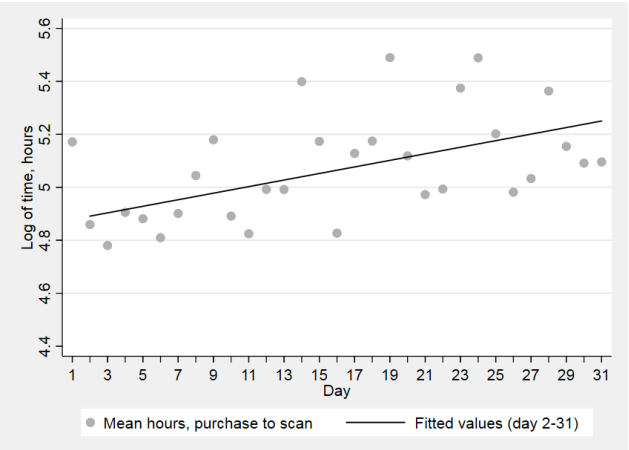


Figure 16b above shows changes in the average number of spending events by day. The top line includes both photographed and directly entered reports and shows an initial drop from a mean of 1.31 spending events on day 1 to a mean of 0.94 at day 2, followed by a more gradual decline to day 31 when the mean is 0.71. The best-fitted statistical model accounts for these two phases, showing that on day 1, there were 37.1% more spending events than on all subsequent days (IRR=1.371, SE=0.102, p<0.001), yet, for each additional day of the study, the rate of spending events falls by 0.9% (IRR=0.991, SE=0.002, p<0.001). Figure 16b also includes two supplementary lines, which show trends in the number of photographed receipts and directly entered spending events, which follow a similar pattern to that of total spending events, though direct entries seem to fall off more gradually.

Table 31 Change in adherence over time

	Pr (used app)		Number of spending events		Pr (spending event entered by receipt)		Log (time from spending event to receipt photographed)	
	Margin	SE	IRR	SE	Margin	SE	exp(b)-1	SE*exp (b)
First day	n/a	n/a	1.371***	0.102	-0.005 (ns)	0.028	0.489**	0.221
Study day	-0.006***	0.001	0.991***	0.002	-0.003***	0.001	0.011**	0.003
	Wald	P>chi2	Wald	P>chi2	Wald	P>chi2	Wald	P>chi2
	191.57	<0.001	142.96	<0.001	46.38	0.001	55.66	<0.001

Note 1: *** P ≤ 0.001, ** 0.001 < P ≤ 0.01, * 0.01 < P ≤ 0.05. The models include all covariates used in Table 30 as fixed effects. The indicator for 'first day' is not applicable in Model 1 because all 'day one' observations are excluded due to collinearity.

Figure 16c shows the proportion of total spending events reported by photographing receipts by study day. For this outcome there does not appear to be a day one effect. Instead, the proportion of entries made by photographing receipts is reasonably stable over time, with a gradual but continuing shift from photographs to direct entries. The estimates in Table 31 show a 0.3% drop in the predicted probability of recording a spending event by photographing a receipt on each additional study day (SE=0.001, p<0.000).

Most noticeable in Figure 16d, which is based on the log of time, is that the time lag between a spending event taking place and the study participant photographing a receipt fluctuates considerably across study days suggesting that the statistics in should be interpreted with caution. There is a noticeable 'day one' effect where the time lag between spending event and photograph drops substantially after the first day, suggesting an improvement in adherence once respondents get into the flow of the study, followed by a gradual worsening of adherence, with the time lag then increasing by 1.1%, on average with each additional day.

In summary, adherence to study protocol is associated with a decline over time for all four of the

outcome measures, and there is a day 1 effect observed for three of the measures.

4.5 Discussion

4.5.1 Interpretation of findings

One of the motivations for the Spending Study was that economists want to establish whether an app-based diary could be used to collect high quality expenditure data and compare favourably to paper expenditure diaries.

For this analysis, the focus was on process measures, conceptualised in terms of adherence. Since equivalent measures of quality are not available from paper expenditure diaries, it is not possible to benchmark against them. Nevertheless, it is possible to review the evidence about different types of error associated with reporting expenditure, set out in the introduction to this chapter (Section 4.1.2), and reflect on whether these appear to be ameliorated by applying new technologies.

In the absence of benchmark data, it is not possible to say whether **recall error** is lower when using the Spending Study app, but the findings are encouraging; one third of participants used the app on 28 or more days, with an average of entries made on 21.7 study days (median=24), rather than making entries retrospectively on only a few days, or at the end of the study, as has been observed with paper diaries (Collins et al., 2018, Silberstein and Scott, 1991). Furthermore, 94.9% of respondents photographed their receipts, on average, within 24 hours and, while further evidence is needed about the timing of direct entries, it seems likely that participants report their spending reasonably soon after the event, which should, in turn, reduce recall error. However, only a minority of respondents reported their spending events almost immediately after they occurred (2.5% with an average lapsed time of within an hour; median reporting lag 4.7 hours), dispelling the idea that mobile devices facilitate research “close to the moment of experience” or “in the moment” (Couper et al., 2017), which would have provided stronger support for a reduction in recall error.

Similarly, it is not possible to report definitively on whether the Spending Study reduced **diary fatigue** compared to paper expenditure diaries. The evidence reported in Section 4.1.1 draws on analyses from the CE, LCF and Family Expenditure Surveys of the UK and Canada and is based on changes in expenditure and number of items, rather than number of spending events or other metrics reported here. Nevertheless, this research provides evidence that the Spending Study has not eliminated diary fatigue; for all four outcome measures, a decline in adherence over the study

period can be seen. Indeed, the rate of decline is greater for photographing receipts than for direct entries, suggesting that photographing receipts is more burdensome and time consuming (Couper and Nicholls, 1998, Read, 2019b). Nevertheless, overall, the decline is gradual, with high levels of reporting even after four weeks.

This suggests that an app-based diary might be sustained over a month period, and if so, this would logically reduce the **infrequency problems** associated with the shorter reference periods used in paper expenditure diaries. However, it is not clear whether this finding would be replicated if a higher proportion of IP9 respondents had taken part in the Spending Study, since it is possible this would introduce a less enthusiastic set of participants who might behave differently. The findings might also differ if the sample had been freshly drawn, rather than having been selected from existing *Understanding Society* members where some panel conditioning effects might be expected.

The Spending Study shows a **day-one effect**, with a clear drop in adherence after the first day of participation in daily app use and number of spending events, as observed in traditional expenditure diaries, followed by a more gradual decline. In contrast, the time lag between the moment the spending event took place and app entry seems to *fall* after day one, suggesting a possible increase in adherence once a participant is 'up and running', though the signal here is weak.

As emphasised earlier, this study alone is not sufficient to draw firm conclusions on this issue. However, a benefit of using app technology is that times and dates are available for a sufficient number of shopping events to see when **telescoping** occurs (though only for photographed receipts for which date and time information are successfully captured) so it is possible to remove invalid cases, reducing the day-one effect that would otherwise have been observed. Furthermore, while Silberstein and Scott (1991) argued that telescoping in diaries would tend to occur late in the study period; in this study it is clear that telescoping mainly occurred on day one, and the following few days. In practice, even after spending events which fall outside the reference period have been removed using the additional data provided from the app technology, a day-one effect remains, which could be accounted for by a mix of increased recording of spending on day one because of novelty effects or practicing, and decreased recording of spending on subsequent days because of diary fatigue.

Finally, it is not possible to provide a clear comparison of **non-specificity** in the Spending Study with traditional diaries, that is the tendency of respondents to enter spending data with insufficient detail. Spending Study participants could choose to provide either a photograph of their receipt or a summary of their expenditure on each shopping occasion. In practice, about 61%

of spending events were recorded by photographing a receipt and where this was done, the data was often of high quality (Read, 2019b, Wenz et al., 2018), and has the advantage that it can be verified by the analyst and provides greater detail about expenditure. On all other occasions, however, summary information was entered directly into the app, simply providing amount spent (often combining items) and category of expenditure, so non-specificity remains an issue.

4.5.2 Strengths

In addition to its substantive aims, an interrelated motivation of the Spending Study was methodological; to explore whether a mobile app could be used to address a complex measurement task. This research provides an early, small-scale case study of how mobile device technology can be harnessed within a high-quality academic study. The study was methodologically innovative in several respects. Because the Spending Study was planned as a follow-up to IP9, a range of covariates believed to be associated with spending and reporting behaviours were collected *ex ante* through a bespoke questionnaire module. These variables, alongside standard socio-demographic variables from earlier waves of the *Understanding Society Innovation Panel*, made it possible to examine the antecedents of adherence, and therefore avoid endogeneity.

Furthermore, the use of mobile device technology meant that additional data such as the date and time of every app entry was captured automatically, and the date and time of spending events could be extracted from photographed receipts for the 46.4% of spending events that were reported in this way. Using these two sources of information, alone and in combination, it was possible to carry out analyses that are not possible using a traditional diary approach: spending events could be excluded if they fell outside the study reference period, patterns of app entries over time could be examined, and the time lag between spending event and app entry could be calculated. Although this data was not available for direct entries and lower quality photographs, where it was available, it provided useful substantive and auxiliary data. However, while photographed receipts may be preferred for several reasons, dropping the facility to enter summary information would have reduced the quality of the data with respect to total amount reported (Wenz et al., 2018), and may also have increased burden and discouraged continuing participation. The research also addresses a more general methodological question: how to evaluate a study where participants are asked to carry out a set of related activities over an extended time-period. This paper demonstrates that the concept of adherence, taken from the medical literature, can be implemented, and adds to the insights drawn from the initial response analysis (Jäckle et al., 2019a) and the comparison of expenditures (Wenz et al., 2018).

Finally, the research adds to the literature about the characteristics and behaviours associated with participation – and in this case adherence – to complex studies, which may help future research studies consider how to maximise engagement. . Our findings suggest that interest in research, the salience of the study, motivating people who are particularly busy, and addressing concerns about app security and data privacy are associated with compliance. Survey practitioners may want to consider, and further explore, these issues. Several of the behaviours most strongly associated with adherence relate to prior survey compliance, the respondents' device uses, and levels of concern about using mobile devices to carry out non-standard activities. Though the statistical models only identified a weak relationship, there was a possible indication that the participants most likely to adhere to a complex study of this kind may be moderate users who might be interested in the novelty of the study, rather than the most advanced users who may have been more likely to satisfice. This hypothesis warrants further research; if true, it suggests that increasing use of mobile devices will not necessarily be associated with increasing adherence to complex research activities as might have been expected.

4.5.3 Limitations

The Spending Study had strengths – it was a follow-up to a large-scale, random probability, longitudinal survey which provided a high-quality context from which to test a mobile app study. Furthermore, the number of observations recording spending was large: 8,040 daily records, 7,412 valid spending events and 3,454 receipts, providing an extensive analytic dataset.

However, it is also important to note its weaknesses: initial participation in the Spending Study was low, with just 16.5% of the invited sample completing the registration process, and fewer still downloading the app and reporting spending. Jäckle et. al. report that while response was selective, it was unbiased with respect to variables associated with expenditure (Jäckle et al., 2019a). Nevertheless, the low initial response must cast some doubt on the generalisability of the findings. Furthermore, some missingness was observed in the covariates used to predict adherence, and in some cases, such as qualifications, these were recoded following the approach taken in a related paper (Jäckle et al., 2019a), although simple imputation might have been more desirable.

In addition, while a number of covariates were included in the statistical models, and these were collected prior to the Spending Study so were not subject to endogeneity, further independent variables could have been included. Given that that the target population had already downloaded the app and made at least one app entry, variables recording access to mobile technologies (such as the different types of mobile device the respondent was able to use to connect to the internet, whether they had access to wi-fi at home and whether they had a data

plan) were not included in the analysis. However, these might have affected ongoing participation and could usefully have been added. Similarly, while some measures of confidence in device use were included, a self-reported measure of skill was not, and nor was a measure of financial position which may have acted as a proxy for socio-economic status or variations in spending (Jäckle et al., 2019a).

Furthermore, as part of the broader research project, the influence of device characteristics on data collection quality was considered in a separate paper (Read, 2019a). There, five device attributes were examined: whether the operating system was iOS or Android; whether the device was a smartphone or tablet; the device's Random-Access Memory (RAM); camera quality; and processor speed. This study found that three aspects of the device, the operating system (iOS or Android), whether a tablet or smartphone was used, and the device's RAM were associated with the duration of app use measured in seconds and whether the receipt was fully readable. It showed that none of these device characteristics were significant predictors of whether an app use was a photographed receipt, a manually entered purchase or a report of nothing bought (Read, 2019a). This suggests that device attributes did not affect one of the four outcome variables considered in this paper. On reflection, however, it would have been useful to have included device attributes, particularly the operating system and whether the device was a smartphone or tablet, within this paper.

Developing and implementing the study was time-consuming and costly, as was processing and analysing the dataset. Many further rounds of development and testing would be needed to optimise the design if it were to be implemented at scale as part of a major social survey. This needs to be kept in mind when weighing the costs and benefits of an innovative data collection activity of this kind. Whatever direction a future spending study might take (or a mobile-app study on any other topic), a strategy of repeated trials to test-and-learn is required.

4.5.4 Opportunities for further research

Following this research, a second trial of an app-based Spending Study was developed, but this focused on increasing initial participation by testing the effect of inviting people to participate during the *Understanding Society* interview and by offering a sequential app and web design (Jäckle et al., 2019a). Although this trial did not specifically offer opportunities to test ways of improving adherence, there are a range of additional trials that could be implemented to examine this further. For example, to increase daily app use, the effect of a larger daily or end-of-study incentive could be tested, or the effect of changing the messaging of daily reminders, or of providing a count of study days remaining. To increase the number of purchases reported, future

research could test the effect of offering an incentive for every spending event entered, rather than the first event each day. To increase the proportion of spending events reported by photographing a receipt, since this provides detailed, verifiable data, research could test the effect of offering higher incentives for photographed receipts rather than direct entries, providing more reassurance about data security. Further experiments could be used to test other aspects of the study design. During the study period, the process of photographing receipts might be encouraged by delaying the reminder notifications until later in the evening, since the number of direct reports rises immediately after this reminder is sent. Photographing receipts might also be encouraged when participants first join the study by including practice screens in the initial registration process. Practice screens might also provide an opportunity to signal that the one-month study period has now begun, anchoring the start of the reference period, and reducing the number of ineligible payments entered in the app, thereby reducing telescoping and day one effects.

Other design variations could be trialled to improve adherence, taking advantage of the app technology. Participants entering direct payments, or those who photograph a receipt with limited information, could be asked to enter the date and time each spending event took place. This would increase burden, but might encourage more prompt reporting, and would increase understanding of the time lag between spending and report. More radically, photographed receipts could be checked in real time and feedback could be given (as is done in one market research spending app), to ensure that receipts are valid and to train participants to check that date and time of the spending event has been captured.

In practice, as methods of payment continue to change, both technical and research effort will need to be made; for example, to ensure that payments made online or using mobile and wearable devices are reported well, and to capturing receipts which are sent digitally to the individual.

Even given significant improvements in the performance of a spending app, further research would still be required to allow direct comparison between app-based spending diary and paper expenditure diaries, focusing both on the estimates derived from the studies but also measures of process quality.

In the meantime, the search for other technological approaches to gathering accurate spending data should continue in a range of ways, such as seeking to access objective sources of data such as bank account information, to provide primary data about expenditure as well as to determine the quality or completeness of reports given in studies of this kind (Angrisani et al., 2017, Jäckle et al., 2019b).

Beyond the specific challenges of measuring spending, other app studies focusing on different types of measurement could be used to assess adherence, by, where possible, using the data generated by the technology itself for methodological purposes.

Chapter 5 Conclusion

In concluding, the findings are reviewed from the three papers, to examine specific aspects of the effect of technology on data quality and draw out broader implications about new technologies and their effect on social surveys. The differences between the studies are recapped, and then key findings for each are summarised (Section 5.1); the implications each study has for analysts, methodologists, and investigators and practitioners are then considered (Section 5.2). The discussion returns to the challenges presented by technology discussed in Chapter 1 and looks at how this applies to each of the papers (Section 5.3). The limitations of the three studies are identified, as are the limitations of the approach of examining specific technologies (Section 5.4); avenues for future research are then discussed (Section 5.5). The final paragraphs identify a few themes worth further consideration.

5.1 Key findings

This thesis presents three studies, each of which examines an aspect of the effect that technology has on the quality of the data collected. The three studies differ markedly, with respect to their funding, research teams, subject discipline, purpose, and the aspect of technology that they consider. The first study compares measures captured using different makes and models of equipment commonly included in biosocial surveys. The second considers the response behaviours of young people who took part in the Science Education Tracker (SET) survey, comparing those who responded using a PC with those who responded using a mobile device. The third study explores the quality of data collected in an app-based Spending Study based on four measures of adherence to protocol. Contrasting methodologies were used: a randomised repeated-measurements cross-over trial in Chapter 2, a quasi-experimental strategy using inverse probability treatment weights in Chapter 3, and a small-scale pilot within the context of a large, nationally representative survey in Chapter 4. The analytical methods used also differ; some of the methods follow the convention used in similar studies (such as Bland and Altman plots in Chapter 2) and some do not (such as multilevel modelling in Chapter 2 and matching methodologies in Chapter 3). Throughout, a wide range of additional types of data are included, that relate to that individual or their behaviours which are collected from a variety of sources.

Chapter 2 considers a set of biomeasures commonly collected through social surveys which make a valuable contribution to health research (Benzeval et al., 2016, Weir, 2018). These measures provide an example of how surveys can act as a vehicle for supplementary data collection in a representative population. By their nature, measures of physiological function (in this instance,

blood pressure, grip strength and lung function) frequently rely on specialised technical equipment which may vary from study to study and over time. This introduces the possibility of measurement error resulting from the device used. In a randomised cross-over study of 118 adults aged 45-74 years, there is evidence of differences in measurements when assessed using different devices. For blood pressure, the newer Omron HEM-907 measured higher, on average, than the older Omron 705-CP (3.85 mm Hg for SBP and 1.35 mm Hg for DBP). For grip strength, the two electronic dynamometers were found to record measurements, on average, 4-5kg higher than either the hydraulic or the spring-gauge dynamometer, but there were only small differences when comparing the two electronic dynamometers or the hydraulic and spring-gauge dynamometers, and these were not statistically significant. For lung function, the measures of FVC on the Easy on-PC by NDD were, on average, 0.47 litres higher than those for the Micro Medical, but there was no difference between measures of FEV_1.

The context for the research presented in Chapter 3 is the substantial growth in online surveys which seem likely to continue, and the rising proportion of respondents who complete these online surveys using the web browser of their mobile device. Although the communication technology involved in delivering an online survey to a PC, smartphone or tablet is essentially the same, there are differences in the screen size, mode of data entry, and how conducive the setting in which the survey is completed may be in terms of concentration and privacy. This introduces the possibility of measurement differences resulting from the device chosen to respond. In an initial analysis of 4,068 young people's responses to the Science Education Tracker survey, it appears that there are differences in the quality of responses which result from whether the young person chose to respond using a PC or using a mobile device. However, very few device effects are observed after controlling for selection effects – that is, when taking account of the clear differences in the characteristics of those who choose to use a mobile device rather than a PC – and the device effects that are observed are small. Respondents using a mobile device are more likely to provide a 'don't know' response and are more likely to have interruptions during survey completion. There is a small, somewhat inconsistent indication of greater straightlining among PC responders. Contrary to earlier studies, young people who respond using a smartphone have faster completion times than those who respond using a PC. It is not clear whether this has implications for data quality; breakoffs are low overall but higher for mobile device responders, although it is not possible to determine whether this is due to selection. This research represents an example of the way that a change in technologies used by the general population can affect the delivery of surveys which involve respondent self-completion.

The research study presented in Chapter 4 is an early example of app-based mobile research delivered in the context of a high quality, representative household panel survey, and has broader applicability in terms of the use of mobile research app technology. More specifically, the Spending Study presents findings from an app-based diary, designed to collect expenditure data over a one-month period from a sample of respondents to wave 9 of the *Understanding Society* Innovation Panel. The study considers the likely quality of data collected by the app by considering the engagement of participants, defined in terms of four measures of adherence to protocol, and the extent to which adherence is sustained over the duration of the project. The research identifies a reasonable level of engagement from the 268 individuals who agreed to participate. For example, the mean number of app use days in the one-month period was 21.7 and the mean number of spending events reported was 27.6. Almost all participants (96.6%) reported at least one spending event and of those, most (95%) used a combination of photographing receipts and making direct entries, or only photographed receipts, with 61% of all spending events reported by photographing receipts. Almost all of those (94.9%) who photographed one or more receipts which had relevant date information did so, on average, within 24 hours of the time of the spending event. Although adherence based on all four measures clearly declines across the study month, it remains reasonably high.

5.2 Implications and contributions

5.2.1 For analysts

Analysts should be aware of the differences in measures of blood pressure, grip strength and lung function when carrying out research studies which use different devices. They may want to test the sensitivity of their findings to these device effects, or compute correction factors or device-specific reference equations when estimating intra-individual changes in function over time using longitudinal studies that have switched device, or when comparing physiological measures within or across studies that use different devices. Although past advice has been to avoid analysis of lung function using mixed spirometers (McFall et al., 2014, Mindell et al., 2011), the evidence from this research suggests that measures of FEV_1 are consistent when measured with the Micro Medical and Easy-on PC. This may create the opportunity for new analyses of existing datasets. Further research would be needed to see whether the same is true of other device combinations.

The implications for analysts based on the SET study are slight since very limited device effects were identified. Analysts who intend to impute missing values may wish to take account of device used to respond, as well as the covariates associated with device selection. Since only a limited number of outcome codes were included in this analysis, analysts are encouraged to carry out

preliminary modelling to check device effects before fully specifying their analyses, perhaps particularly in terms of substantive responses which are not reported in Chapter 3.

The primary purpose of the Spending Study was developmental, so it is unlikely that the dataset would be used extensively for analysis of expenditures. Nevertheless, analysis carried out in parallel to this research by Wenz et al. adds to the evidence presented in this paper by focusing on outcome quality and comparing expenditure data collected in the app with benchmark data drawn from the LCF. The many differences in the methodologies of the Spending Study and LCF were overcome by a series of decisions: the analysis sampled two weeks of data from the Spending Study; focused on individual rather than household expenditure; aggregated data to allow comparison between average weekly expenditure in total and for specific categories; and used inverse probability weighting with a set of socio-demographic covariates to match the samples (age, gender, employment, income, house ownership, household size, number of children in the household, presence of computer and urban/rural indicator). Their conclusion was that the Spending Study provided a promising method for collecting high-level expenditure data, particularly when both photographed receipts and direct entries were combined; and were most promising for some categories of spending, for men, and for those with higher incomes (Wenz et al., 2018).

This provides further encouragement that a spending app could be used in this way. If it were, considerable care would need to be taken in generalising from the results, given the low initial participation rate. Furthermore, findings from this paper might encourage analysts to define eligibility of the sample more tightly than was the case here, where only two participants (for whom there were no valid app entries) were dropped. Additional criteria might be applied to identify a subset of participants who engaged sufficiently for their spending data to be considered robust; for example, by excluding respondents who did not report any expenditures, or by dropping the earliest and latest study days where the results might be influenced by day-one, learning or fatigue effects.

5.2.2 For methodologists

The equipment comparison study in Chapter 2 focuses on two sphygmomanometers, four dynamometers and two spirometers. The findings cannot be generalised to other device combinations or to other physiological measures. Furthermore, the cumulative evidence about these and other device combinations is slight and somewhat inconsistent. Additional equipment comparison studies are therefore needed to build a more robust body of evidence, both about the device combinations studied here, and to test other device combinations. As well as

conducting similar experiments in a controlled environment, comparison studies should be built into existing plans for biosocial surveys to understand the effect of technology in real-life environments. An approach of this kind would involve multiple interviewers or survey nurses, each with their own set of equipment, so additional arrangements would need to account for the specific interviewer and specific device. This type of research should be planned as part of the normal deployment of biosocial surveys so that robust evidence from multiple studies can be accumulated, both to support future equipment change and cross-study analyses; it should not be conducted only when the need for equipment changes is imminent, as was the case here. Finally, reflecting on the analytical approach taken in this comparison study, and the sensitivity analyses in particular, future analyses should consider using multilevel models which take account of all readings in place of the current norm of summary measures and Bland and Altman plots. A similar approach could be considered for equipment from other scientific disciplines which are incorporated in social surveys, if there are concerns about different makes of models of device which may be associated with increased measurement error.

When considering the implications for methodologists of the research presented in Chapter 3, related to device effects which might result from responding with a PC or mobile device, it is important to acknowledge that much of the literature uses experimental designs administered to known panel responders. One of the contributions that this study makes is that applies quasi-experimental methods to the SET study, using inverse probability treatment weights (Matthews et al., 2017), and so provides a useful example of an alternative approach based on a large, cross-sectional sample where responders answer an authentic questionnaire using the device of their choice (Clement et al., 2020). The chapter also demonstrates that matching with a broad set of variables – including measures related to the topic under investigation, such as parental interest in education and pupil attainment – strengthens the ability to control for selection effects. Furthermore, the study makes an original contribution by demonstrating the utility of matching with variables drawn from external sources, such as geographical databases, survey process data, and administrative variables which avoid the problem of endogeneity. This approach warrants consideration in future studies of this kind.

The Spending Study presented in Chapter 4 provides methodologists with an additional case study of an app-based diary. A key benefit of the study is that it took place within the *Understanding Society* Innovation Panel where priority is given to thorough methodological research. Furthermore, it was funded by the ESRC Transformative Research Scheme and the NCRM's Methodological Research Projects Scheme. Consequently, a literature review of new data sources and technologies used in measuring household finances was conducted, providing useful context for this research (Jäckle et al., 2021). In addition, careful attention was given to optimising the

study. For example, a specially designed module of covariates in IP9 was included, which took place prior to participants being invited to take part in the app study. This provides an excellent foundation to investigate response and bias (Jäckle et al., 2019a) and means that covariates used in the analysis presented in Chapter 4 are not subject to endogeneity.

While much was learned from the study, a key learning point is that multiple tests and revisions would be needed to develop the best possible app-based expenditure diary and study design. In evaluating and refining such an app, it is very important to reflect on the ultimate objectives of the project, given the almost inevitable need to trade-off competing goods. After the Spending Study was completed, additional funding made it possible to revise the study design and retest. In the interests of encouraging higher initial participation, the study team introduced a sequential design where a web equivalent was offered for those who were unwilling or unable to download an app. Creating a web equivalent of the app-based diary necessitated simplifying the data collection process and the facility to photograph receipts was dropped (Jäckle et al., 2022), so that spending could be recorded solely by direct entry. Offering the mobile web option increased initial participation but was not as successful as retaining participants as the app, with its inbuilt devices such as notifications to encourage continued use (Jäckle et al., 2022). If pursuing higher participation rates is the ultimate goal, it would be interesting to test a third approach, where text interviewing is used to collect daily reports each evening.

The research shows that it is not possible to rely solely on receipts to capture spending in a study of this kind; receipts are not always available, they may be lost, they are slightly more time consuming to provide (Read, 2019b) and some participants are reluctant, unwilling or unable to photograph them. This is confirmed by an analysis which shows that if the Spending Study had relied on receipts alone, and the facility to enter summary information had not been provided, estimates of total spending would have been lower and less accurate, particularly in some categories of expenditure (Wenz et al., 2018).

Arguably, though, simplifying the data collection process to pursue higher response rates means that the Spending Study no longer makes use of the full capability of mobile devices to capture photographs of receipts with their detailed listings of expenditure by item. In the view of some economists, this significantly reduces the value of the data collected (Griffith, 2018). This is a clear example of the tensions between the need to maximise participation, with the opportunities to extend measurement through innovative data collection, and the need for high quality data.

5.2.3 Implications for investigators and survey practitioners

Investigators with responsibility for decisions about future biosocial surveys will want to consider the findings in Chapter 2 when selecting equipment to include in new studies, or, by necessity, when changing equipment in longitudinal studies. Based on these specific examples, they will see that minor changes in equipment used to measure blood pressure may affect measurement. This is even more so with respect to spirometry, though measures of FEV_1 may be consistent across devices. In the case of grip strength, mixing electronic devices and other dynamometers seems particularly problematic. Most importantly, introducing or changing equipment used within surveys requires methodological research to identify quality issues and further trials are needed to replicate the comparison of these devices, to test the same devices with alternative protocols, and to test different device combinations, both in stand-alone studies of this kind and within larger observational surveys with greater variation in implementation. Indeed, given the mixed results from other ad hoc studies of this kind, it is arguable that experimental comparisons should be built into biosocial survey fieldwork on an ongoing basis to build more robust evidence that will support multiple studies, rather than relying on single experiments before key decisions.

There are also broader considerations. Given the historically low level of regulation of equipment of this kind, further engagement is needed with the medical equipment sector to support the use of equipment in research settings. Meanwhile, investigators should collaborate to standardise the equipment, protocols, training, and quality control mechanisms used for all such measures, across studies, wherever possible. This will reduce the development work needed to implement the measure within a study, will provide analytical support including norming, and may even reduce the costs of physical equipment if sharing arrangements are established between studies (Kapteyn et al., 2018). The NIH Toolbox (Gershon et al., 2013) provides an excellent framework of this kind, offering practical guidance for a range of measures of cognition, emotion and sensation, as well as motor measures. This includes a clear recommendation for standardisation of grip strength using the Jamar Plus+ (Reuben et al., 2013).

However, standardisation is challenging, not least because existing studies have already incorporated physiological measures using different equipment and so are at different starting points. Furthermore, when equipment needs replacement, investigators will want to select a device which optimises data quality, given constraints of supply and cost. As technology providers compete for market share, they innovate by offering additional measurements or benefits, such as removing the need for manual calibration, improving data quality by supporting compliance with protocol (for example, increasing the automation of the measurement process, reducing technician errors in recording and transferring data), and improving participant compliance (for

example, in the case of lung function measurement, by providing a visual cue to encourage individuals to exhale fully).

Therefore, technological change results in a tension between maintaining existing equipment to ensure consistent measurement and selecting new equipment which provides better measurement, making harmonisation elusive. This is further complicated if incorporating new technologies leads to more costly and/or less portable equipment, such as in the case of measuring lung function, where there is a divergence between studies which use the Easy on-PC and those which, particularly in developing countries, continue to use smaller, hand-held devices which are both portable and affordable. Indeed, cost constraints have led to some difficult design decisions, where, for example, lung function was measured during the *Understanding Society* Wave 2 and Wave 3 nurse visit with one spirometer in England and Wales and another in Scotland (McFall et al., 2014).

This issue is particularly evident in the study of biomeasures, but it also applies to the SET study and the Spending Study. In those cases, continued improvement of web surveys and mobile apps and the addition of new capabilities means that the tools of measurement are rarely stable. While this can create new opportunities, it also means that repeating the research using a consistent approach becomes difficult. In the case of medical equipment, principal investigators can at least store and repair their devices for many years. In the case of digital platforms which support web surveys and app diaries, the technology is changing at pace and is rarely under the control of the investigator. There are some exceptions: for example, where online software is specifically intended to support academic research (Wright, 2016), or where an app has been developed by a study team (Conrad et al., 2020). These initiatives provide a potential way forward but would require significant investment and evaluation.

The biomedical research discussed here provides another insight that has broader relevance for survey research. All biosocial surveys explicitly report the make and model of the devices they use so that these can be accounted for. In survey research, the extent to which research papers report on the characteristics of the survey instrument varies considerably. The device effects literature often reports simply on whether an online survey is ‘mobile optimised’ or not. Good examples do exist, such as where experiments have been conducted using slightly different question formats (for example, Mavletova, 2013, Mavletova et al., 2018). However, greater account needs to be given to the fact the body of research papers investigate slightly different technological features, which may account for the mixed results found. Further work on documenting the attributes of technologies used in survey research may be helpful.

In other respects, the findings of the SET study should be broadly encouraging to investigators and survey practitioners. The low level of device effects observed suggests that online surveys completed using a mobile device have broadly equivalent data quality. Efforts to improve the design of online surveys and the use of adaptive and responsive designs (Marcotte, 2011, Gustafson, 2016) to ensure they are device agnostic appear to have been reasonably successful and, in the case of this study, it seems likely that the very extensive questionnaire development process, which involved nine focus groups (EdComs, 2016), and the thorough usability testing carried out by Kantar Public, were at least in part responsible for the fact that the survey was completed successfully across different devices by 50% of the sample. Further work to reduce 'don't know' responses would nevertheless be valuable. Delivering surveys successfully on mobile devices is important for this demographic because a very high proportion of young people own or have access to a mobile device, while access to a PC is low, particularly among those who are economically disadvantaged. Indeed, practitioners should focus on encouraging response on a smartphone given the opportunity to encourage participation from more disadvantaged populations, which may reduce bias.

Findings from the Spending Study app in Chapter 4 can be read alongside preliminary findings by Wenz et al (2018) which show that the estimates of spending collected by the app are comparable with the LCF, although this analysis is based on benchmarking with a data source which is itself imperfect. Nevertheless, taken in combination, this evidence provides encouragement to investigators and survey practitioners that, with sufficient investment in development, an app-based spending study has the potential to collect high quality data. The intrinsic capabilities of app technology provide specific benefits such as the opportunity to photograph and transmit receipts and to issue daily notifications as reminders. The study design could be developed further, using the technological capabilities of the mobile device and app; for example, by more clearly delineating the start and end of the study reference period, or by counting down the number of remaining study days. There is clearly great potential to use survey process data to understand data quality and to evaluate proposed improvements in design. A key issue for investigators and practitioners to consider in this and other similar studies is whether to prioritise maximising participation or to focus on the potential for new measurement, given the likely trade-off between these two ambitions, as exemplified by decisions around the second Spending Study described earlier.

The equipment comparison study showed that it is challenging to gather consistent measures of phenomena that are well understood – in this case blood pressure, grip strength or lung function – when the technologies used for that measurement change. Measuring behaviours such as sleeping, eating, exercising, and spending are challenging in a different way, and spending is

perhaps even more so because it is a complex phenomenon. For example, it may be necessary to capture both individual and household expenditure, to account for spending on behalf of others, to allow for multiple methods of payment including joint credit cards, and so on. The task is made even more difficult because of the flux in spending behaviours and the technologies used to manage purchases in recent years. For example, there has been a reduction in the use of cash, an increase in online shopping, an increase in payments made with smartphones and smartwatches, a move away from the automatic printing of receipts and a growth in receipts being sent to people's email addresses. The emergence of open banking (Zachariadis and Ozcan, 2016) has increased individuals' use of apps for banking and to track spending, so this may create new possibilities for data collection for research purposes (Angrisani et al., 2017). However, the task is not an easy one, and innovations to capture spending using these new technologies are almost bound to need constant amendment and possibly re-conceptualisation. Arguably, since the underlying concept of spending is clear, the task should be to refocus on capturing key measurements using relatively traditional methods, and not be overwhelmed by the potential opportunities of new data types, at least until some form of stability is achieved.

5.3 Challenges raised by the three studies

In the introductory chapter to this thesis, some of the challenges that technology introduces to the research process were considered. Ethical issues were first considered, followed by respondent burden, and cost and logistics. The thesis then turned briefly to the Total Survey Error framework and errors of representation and measurement. Here, it returns to each of these topics in light of the findings in the three chapters.

5.3.1 Ethical issues

Each of the three studies reported in this thesis were subject to ethical review (as described in Chapter 1) and considered issues such as informed consent, privacy, and data security. The use of technology in these three studies does not introduce any major ethical issues but does highlight a few small issues worth consideration. With the equipment comparison study, changes in device used should probably be considered when making decisions about what information to feed back to respondents. Generally, following a survey which includes biomeasures, interviewers or survey nurses leave behind an information card which provides the participant with basic information such as anthropometric measures and blood pressure, which respondents often report informally as a benefit of participation, though the connection between participating and gaining medical information is discouraged by ethics boards. There are exceptions where feedback is not given,

such as when collecting genetic material, given the ethical complexities of revealing unsolicited information that may impact people's future lives; and feedback is not provided if the data is not considered reliable at the individual level (Lessof, 2009). The equipment comparison study reported in Chapter 2 suggests an additional criterion for holding back feedback where equipment is used and has changed, given the increased risk of measurement error which may mean that results in successive waves could be misleading.

There are two possible ethical issues related to the SET survey. The first is that all online surveys that are completed on a mobile device have a slightly higher likelihood of being carried out in a public setting, so may lead to concerns about the respondent having full privacy. However, it is not determined by the device alone, since in some instances responding on a smartphone may allow more privacy (for example, by retreating to a bedroom) than responding on a PC, which may be in a shared area. Looked at from a different perspective, making a survey such as SET easy to complete on a mobile device makes it more inclusive to young people who do not have access to a PC, who are more likely to be socially disadvantaged. Therefore, it can be argued that the use of mobile technologies improves the ethical position of the study, taken in the round.

The Spending Study, in contrast, may be seen to exclude groups of people who do not have a suitable smartphone or do not have the confidence or ability to use one to complete a mobile diary. Furthermore, it raises issues around data privacy, because it asks respondents to photograph receipts, which may contain sensitive information, including information about location and items purchased, and proper assurances would be needed that the data would not be disclosed to any authorities. This issue may be considered less acute if participants are asked to save their receipts for collection by the interviewer (Ransley et al., 2001, Timmins et al., 2013), although it could be argued that the anonymity of uploading data to an app avoids fear of judgement. Either way, some respondents had concerns about using mobile devices for research purposes, and they may need additional reassurances.

5.3.2 Respondent burden

Although some interviewers report that respondents like carrying out assessments which add interest to a long questionnaire, this is not always the case, and incorporating biomeasures in social surveys increases burden on participants. This might be the case particularly where equipment is used, either directly if, for example, grip strength causes discomfort for a participant with arthritis, or indirectly if an interviewer or survey nurse imposes on a household by unpacking large bags of equipment. These kinds of burden are justified with reference to the scientific value of the findings and these measures are only collected with informed consent.

The choice of device for the SET survey does not seem to have any obvious implications with respect to burden. Some studies have found that online surveys take longer to complete on a mobile device, leading directly to greater burden, but this is not found to be the case in this study. Indeed, smartphone responders had slightly shorter completion times than PC responders.

Read considers the level of burden experienced by participants in the Spending Study (Read, 2019b) and shows that within the app, some activities are more time consuming than others; for example, photographing receipts takes longer (on average 41 seconds) than reporting purchases without receipts (on average 30 seconds). However, the average time taken for all types of entry reduces as the study continues, and objective measures of time taken do not correlate well with subjective measures. Data was not collected on whether an app-based spending diary would be more or less burdensome than a paper equivalent, but both are certainly more burdensome than a set of recall questions during a face to face or online survey.

The salient issue here is that even if the activity is not considered burdensome by participants who are motivated to take part, incorporating technology in research studies does demand time and effort from the participant. Even where researchers claim that technologies reduce burden, such as by tracking energy use or exercise passively, this is rarely the case if an honest comparison is made with either foregoing the measurement overall or asking a series of recall questions in an existing survey.

5.3.3 Cost and logistics

As set out in Chapter 1, the cost of collecting biomeasures in social surveys is high. In addition to the cost of purchasing equipment, there are development and training costs, operational costs of maintaining, calibrating, and distributing equipment, and quality control costs. These costs are multiplied if the devices used change, and there are additional costs associated with harmonising data with studies (or past waves) which use different devices – this study and others of its kind can be seen as part of that cost. Earlier, the benefits and barriers to equipment sharing and standardisation, which would introduce savings, were discussed, but some barriers to achieving this were also identified. Regardless, the administration of these measures also takes up a considerable amount of survey time, often at relatively high survey nurse rates. For example, the NIH Toolbox report that grip strength takes 3 minutes to administer, which has direct costs and an opportunity cost in terms of simple questions and answers. The reason that these measurements continue, despite their relatively high costs, is because of the value of the (relatively) objective measures they provide.

In contrast, the SET study is based on an industry standard survey software that has already been developed to be device agnostic, making the running of the study relatively low-cost and low effort. However, considerable effort was put into developing these standards and, in the case of this study, into pre-testing and piloting the SET survey to ensure that it was implemented successfully (EdComs, 2016, Hamlyn et al., 2017). Similarly, the collection of survey process data and linkage to geographical datasets draws on programming that has been developed centrally, though this means there are some constraints in what is made available. The costs of linking to administrative data are rarely acknowledged but are very considerable, though these may fall as researchers gain experience in navigating the process of application and data receipt.

The Spending Study was also based on an existing technology platform which offered a range of capabilities, including taking and transmitting photographs. The design of the diary was also quite simple, with a limited number of screens. Despite this, a significant investment of time and effort was needed to agree and implement the Spending Study and embed it within the *Understanding Society* Innovation Panel. An app of this kind is only likely to be cost effective if it can be developed and used in a large project, or if it can be applied in multiple surveys. Furthermore, it would only be ready to be put into field after several additional rounds of development. The Spending Study was first mooted during an innovation workshop held by TNS BRMB in which a presentation was given by Kantar Worldpanel about the mobile app approach they were developing for their scanner-based in-home shopping study. Ultimately this technology was not appropriate for the Spending Study, but it is an example of a commercially funded development which is deployed at scale by a major market research organisation. In the social research world, these kinds of development are only possible given significant investment from government (for example, if the app were to support a national spending study), or as part of a major academic study such as *Understanding Society* with significant funding for innovation, as was the case here.

Having considered the issues of ethics, respondent burden, and cost and logistics, the paper now considers the themes covered by the Total Survey Error framework, which are representation and measurement.

As mentioned earlier, Groves and Lyberg account for the possibility that the TSE framework may need to be adapted to account for new phenomena: “Any listing is bound to be incomplete, though, since new error structures emerge due to new technology, methodological innovations, or strategic design decisions such as using mixed-mode data collection. All error sources are not known, some defy expression, and some can become important because of the specific study aim, such as translation error in cross-national studies.” (Groves and Lyberg, 2010, p854). Some adaptations to the framework have been mooted to address the challenges of Big Data (Amaya et al., 2020, Japec et al., 2015) and this may have some relevance to the collection of unstructured

data within the Spending Study mobile app. However, the examples of technology explored in this thesis sit broadly within the existing TSE framework, with technology providing one of many potential sources of measurement and representation error.

Indeed, the existing TSE framework provides a useful reminder that multiple sources of error need to be considered alongside each other. Chapter 2 shows that there are device effects when measuring physiological functions, but TSE emphasises that this is one of several sources of measurement error that may be at play. Errors such as failure to correctly implement protocols, or that result from poor preparation of the respondent also need to be considered. It also reminds us to consider other forms of error such as processing error, which is likely to be reduced by several of the newer devices, though this did not form a focus for the research presented here.

5.3.4 Errors of representation

In all three projects, the use of technology effects representation. As explained in the introductory chapter, in the case of collecting biomeasures, initial survey nonresponse will be compounded if assessments are carried out during a second visit from a survey nurse visit as some respondents will refuse or break their appointment. There will then be further loss of respondents who do not consent to a specific measurement or are deemed unable to complete it, if equipment is forgotten or fails, or if data is lost in transmission. While some of these types of missingness are random, others are informative, and individuals with worse physical or mental health are more likely to be excluded (Sakshaug et al., 2014). What is noticeable in this study is that while two measures – blood pressure and grip strength – have quite low levels of missingness, the third measure – lung function – has quite high missingness. Measuring lung function is quite physically demanding, and the quality standards set are high. In this trial, relatively inexperienced researchers delivered the spirometry assessment, and this may have increased missingness, which was even greater for the Easy-on PC (which imposes an assessment of whether each blow is valid) compared to the micro-medical (where the practitioner makes a subjective judgement).

The sensitivity analysis for lung function, which included measures of lower technical quality, reduced missingness, yet the comparison of the two spirometers yielded very similar results . From a Total Survey Error perspective, this seems a clear case where epidemiologists interested in population level analysis (rather than medical diagnosis) should seriously consider including these lower quality measurements to reduce missingness and increase representation. In practice, however, they may struggle to publish research that breaches the strict guidelines set out in journals focused on pulmonary health.

In contrast, in the SET survey, the opportunity to access the online survey through a smartphone or tablet may have *increased* participation of more disadvantaged students who are less likely to have access to a PC (Lessof et al., 2019, Lessof et al., 2016). Despite this, this group had poorer response rates and were under-represented (Hamlyn et al., 2017). Further research is needed to explore this issue, given evidence from experimental research which shows poorer response from mobile device users when participants are allocated randomly to device (Couper et al., 2017). Since raising participation of disadvantaged young people is likely to reduce bias, in addition to thinking about selection effects, this research should encourage further thought about the role mobile devices may play in making surveys more accessible to reluctant population sub-groups. At the same time, it may be necessary to think about how to encourage these young people to concentrate further while participating.

In the case of the Spending Study, the app was only available to people with access to a smartphone or tablet and who were willing and able to use their device for research purposes. Although ownership of mobile devices has increased, it is not universal, and relying on an app will inevitably have increased coverage error and non-response error. It is important to remember that evidence from the accompanying paper from the study, which looked at response and representation, showed how limited the uptake of the app study was (Jäckle et al., 2019a). Indeed, a fundamental concern is whether a research app can deliver a sufficiently high response rate without bias. This problem may be exacerbated if an app-based study requires that the respondent uses advanced features of their device such as the camera. In the case of the Spending Study, an attempt was made to reduce this risk by allowing participants to report expenditure without photographing a receipt. While strategies of this kind may reduce errors of representation, in this case it resulted in more complex data, with some structured and some unstructured elements.

Some studies have attempted to overcome errors of representation which result from requiring participants to have access to and be willing to use a tablet or smartphone by offering devices to participants (Fernee and Sonck, 2013, Sonck and Fernee, 2013). The second Spending Study took a different approach and offered a web version of the spending diary for those who could not or would not download the app (Jäckle et al., 2022)...

5.3.5 Errors of measurement

The focus of this thesis has been on data quality resulting from technological change or new developments, and the findings have already been summarised and repeated. However, in the context of the Total Survey Error framework, it is worth drawing attention to other sources of measurement error that need consideration. In Chapter 2, the equipment comparison study,

other sources of error were mentioned, such as the adherence to protocol and the physical state of the participant (for example, whether they had smoked or eaten recently, or were well rested). In Chapter 3 substantial measurement error was not identified in the comparison between young people who completed the survey on a PC and those who completed it on a mobile device; however, this only gives a partial picture of the error that may exist in both accounts, and further research would be needed to understand the impact of the choice of an online survey, in comparison to the previous waves of the SET study, which were face-to-face. Finally, in the case of the Spending Study in Chapter 4, the concept of construct validity allows reflection on the way that the decision to use mobile app technology to collect expenditure had a formative impact on the nature of the data collected. For example, the study focused on individual rather than household spending because it was not practical to issue the app to multiple household members and collate the data; and it was necessary to restrict measures of spending to the day-to-day, which is likely to have focused minds on spending where receipts are given, rather than regular monthly payments or cash-in-hand payments. Furthermore, the spending app consideration also needs to be given to processing error. For example, in the Spending Study, receipts need to be scanned and coded accurately, and there may be data transcription or entry errors when collecting biomeasures.

5.4 Limitations

Each chapter of this thesis identifies some of the limitations in each of the three studies. For Chapter 2, the equipment comparison study, an important limitation is that only specific makes and models of equipment are considered, and more evidence is needed to build a robust evidence base about these and other combinations of devices. Future comparisons should seriously consider using multilevel modelling as the primary statistical approach which would allow a more systematic consideration of, for example, order effects and interviewer/researcher effects. More fundamentally, the study focuses on a single type of measurement error generated by device but does not consider other types of measurement error or errors of representation within a broader framework.

The SET data had several minor limitations. For example, the measures of quality were somewhat limited and additional covariates may have helped controlled further for selection. The SET study demonstrated the use of survey process data for both outcome and predictor variables; however, more extensive survey process data would have been valuable, including better measures of time taken, information about screen switching, and objective measures of distractions. A significant

constraint was that consent to link to administrative data was only given by 83% of the sample, limiting the substantive data for analysis, but also the opportunity to use these data for matching.

Perhaps the main limitations of the Spending Study are that, while the level of adherence is reasonably encouraging, objective data, such as information from participants' bank accounts, or equivalent data from studies using alternative methodologies such as paper expenditure diaries, are not available to enable a direct comparison of quality. Furthermore, the approach to conceptualising and measuring adherence is novel and worth further consideration, but the measures of quality could be developed in future research, and additional theoretical work on how best to think about engagement in complex survey tasks would be useful. It is also very important to remember that while the Spending Study app gathered a large volume of spending data, it did so from a relatively small sample. This provides interesting opportunities for research on consumption, but to be considered a suitable method for collecting spending data within a large-scale study such as *Understanding Society*, a much higher proportion of sample members would need to be persuaded to take part. Like SET, the Spending Study would have benefited from some additional covariates to support the analyses; for example, a multiple response item identifying which devices each respondent has access to, personality measures such as decisiveness, and objective measures of distractions such as sound and movement during the interview.

Looking across the studies, all three are limited in the sense that they cover very specific aspects of data quality. Furthermore, they each investigate technology at a specific moment in time and the specific elements of the research will inevitably become rapidly outdated. A broader perspective on the possible influences of technology on measurement might encourage a more strategic methodological programme that would more systematically consider issues of this kind and cumulate evidence across studies.

5.5 Future research

At the end of each chapter, opportunities for further research were identified. In the equipment comparison study, the importance of further comparisons of these and other device combinations was emphasised. This concluding discussion suggests that there would be value in further research drawing together different aspects of error associated with these measures within a TSE framework. For the SET study, it was suggested that continuing research is needed because mobile devices and app survey design will themselves continue to evolve. Furthermore, the proportion responding on a mobile device is likely to continue to rise and the composition of those who respond on a mobile device may change, which may be associated with a shift in

selection effects. More specifically, further research was suggested to better understand the remaining device effects – perhaps using an experimental approach – in order to test alternative modifications to questionnaire design to reduce 'don't know' responses on mobile devices. Further research may also be needed to understand whether the higher number of interruptions on mobile devices should be of concern. Despite limited evidence that device effects data quality, concerns are nevertheless expressed about the increased distractions that a respondent completing a survey on a mobile device may be experiencing. Research which uses the sensor capabilities of devices to capture sound and movement, and developments in survey process data to measure distracted behaviours such as screen switching, would open new ways of exploring these questions. Experiments to test ways of encouraging survey participants to focus as they respond would also be useful.

A series of small-scale investigations would be needed to understand how small adjustments to the Spending Study app could improve the quality of data provided. For example, the ability to send notifications every day at a fixed time seems likely to have increased daily participation, with a burst of entries immediately afterwards, but this could be investigated through further trials. Similarly, some day one effects could be observed and invalid spending reports removed by using survey process data. Small design changes could be used to reinforce the start and end of the reference period, which could then be evaluated. Another recommendation is to carry out various experiments with incentives to increase the capture of receipts. A comparison with paper diaries seems vital but this would need careful development. Paper diaries do not provide the rich source of survey process data that is available from mobile app studies. Consequently, a variety of methods including cognitive testing, observational approaches and qualitative research would be needed to understand how participants engage with diaries in both modes.

Perhaps the most consistent insight is that new technologies, and even minor shifts in existing technologies, require multiple methodological investigations or trials in order for them to be evaluated and developed. Confronted with the challenge of new technologies, National Statistical Organisations require sound evidence to deviate from gold standard methodologies, and survey organisations may shy away from the costs and risks of investing in uncertain research approaches. This may result in a "wait and see" approach when a more useful mantra is "test and learn". This is because the benefits as well as the problems associated with a technology need to be revealed in practice, and solutions identified. Serious consideration should be given to ways of encouraging many more small-scale studies supported by academia and government, and to ensure that methodologists can piggy-back existing survey fieldwork to build learning into existing projects.

In practice, opportunities for this kind of test-and-learn or agile methodological research are relatively rare. Perhaps the best example is provided by the many small, experimental studies of device effects in online surveys that were implemented by Couper, Mavletova and others (Antoun et al., 2017, Couper, 2013c, Keusch and Yan, 2017, Mavletova, 2013, Mavletova and Couper, 2013, Mavletova and Couper, 2016, Peytchev and Hill, 2010, Stapleton, 2013, Toninelli and Revilla, 2016, Tourangeau et al., 2018, Wells et al., 2013) which were criticised in Chapter 3 for being somewhat artificial (Clement et al., 2020). Some major studies have a framework for development projects, such as *Understanding Society*'s Innovation Panel, but even so, these studies tend to be annual, 'big bang' efforts rather than repeated, small-scale opportunities to keep revising and developing a new tool. An example of this is that a second version of the Spending Study was funded, but only a limited number of experiments were possible, and they focused (understandably) on increasing participation rates (Jäckle et al., 2022).

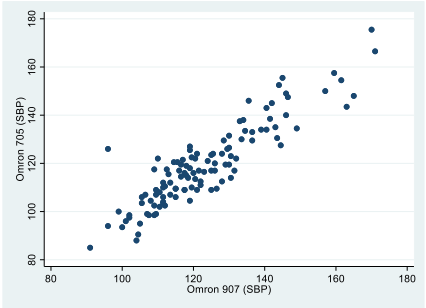
An additional issue is that the incentives to develop technologies, and then to review and refine them, are not well structured in academia or survey organisations. The imperative to publish in academia means that considerable time is spent analysing and reporting each experimental study and publishing results in key journals. While the publication process clearly builds shared knowledge, it may discourage a more agile approach to development – but this is problematic, because innovative technologies will never be implemented optimally on the first, second or even third occasion. In practice, multiple sequential experiments are often needed to refine a new research approach. A different – perhaps opposite – challenge is seen in research organisations which depend on winning competitive bids to secure work, where the large volume of projects creates many opportunities to test and develop new approaches and the need to compete encourages innovation, but places little emphasis on careful analysis, reporting and sharing of findings. Unsuccessful endeavours are more likely to be discarded than documented, even more so than is the case in academic research. Communication of successes tends to be focused on conference papers and marketing materials rather than carefully evidenced journal papers that would build shared knowledge. The optimal model lies somewhere in between.

Questions and responses are at the heart of the survey process and survey methodologists have developed rigorous methods to understand and strengthen question design, reflecting deeply on the interaction between the respondent and interviewer. This has included a consideration of the effect of changes in communication technologies and survey mode, and more recently changes in device. Far less attention has been given to the broader role technology has had in its role in the collection of supplementary data, whether through medical equipment, a mobile app which collects photographs, wearables, or sensors. The equipment used in surveys varies so widely, and comes from so many disciplines, that a unifying framework to consider all technologies seems

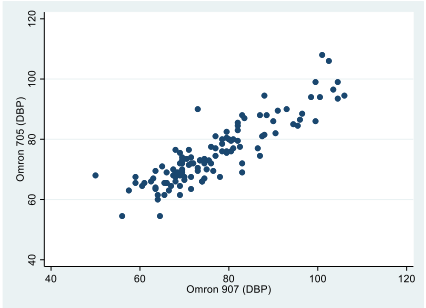
hard to achieve. However, a more coordinated and systematic consideration of the impact of technology could be of considerable value. A community of interest could bring together survey methodologists engaged in different aspects of technological innovation to consider cross-cutting themes, and to develop theory around the principles that should be considered. A network of this kind might scan and review emerging technologies in a timely way, mapping requirements for methodological investigation, and identifying opportunities to run early field tests and to share learning. Indeed, by including survey delivery agencies and National Statistical Offices, it may also be possible to identify increased opportunities for experiments and trials. This could lead to a more strategic effort to test and retest promising approaches, and to set aside approaches which are most likely to fail. Although there are no guarantees of success, it is possible that some technologies would be more robustly rejected, and others more effectively adopted. The discipline of survey methodology already has the appropriate tools and fundamental interest to test and develop survey questions; it could easily apply this same consistent focus to technological innovation.

Appendix A Scatter plots for all pairs of equipment

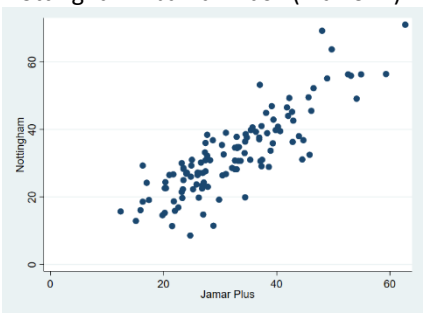
SBP: Omron 705 – 905 (Mean 2nd+3rd)



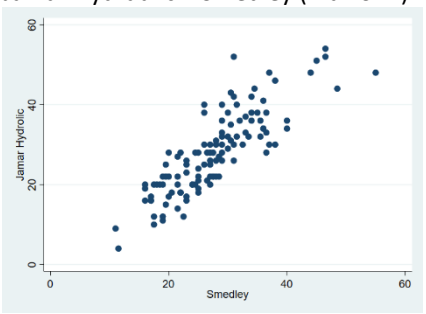
DBP: Omron 705 – 905 (Mean 2nd+3rd)



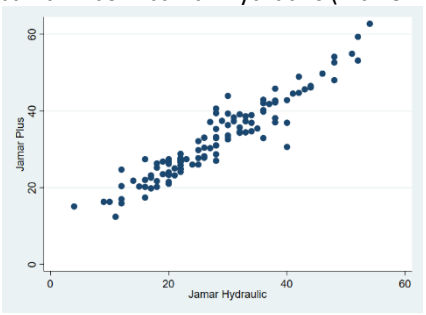
Nottingham – Jamar Plus+ (max of 4)



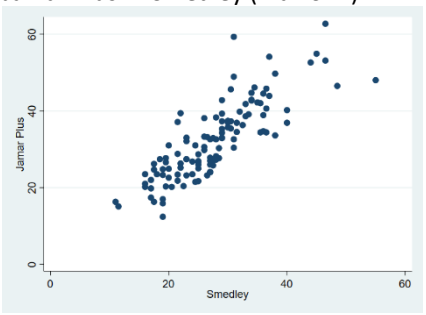
Jamar Hydraulic – Smedley (max of 4)



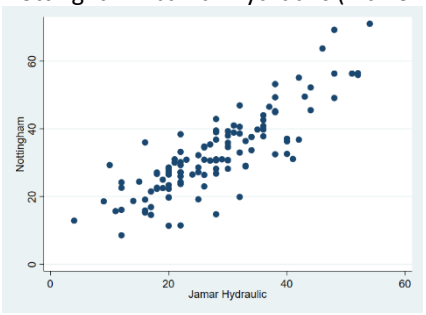
Jamar Plus+ - Jamar Hydraulic (max of 4)



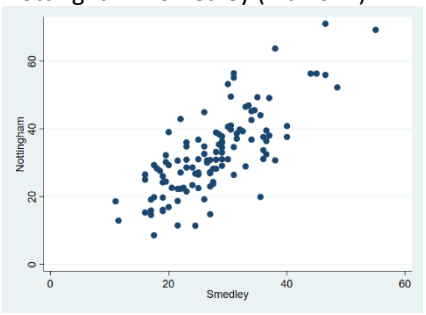
Jamar Plus+ - Smedley (max of 4)



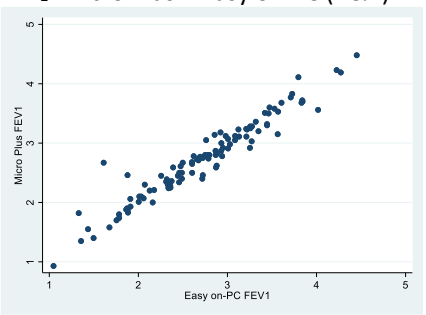
Nottingham – Jamar Hydraulic (max of 4)



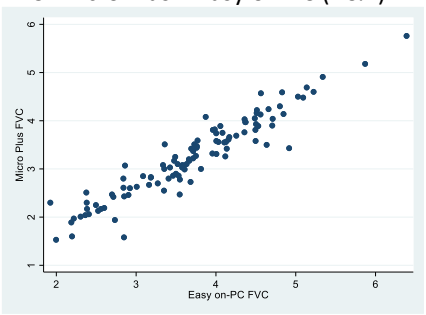
Nottingham - Smedley (max of 4)



FEV_1: Micro Plus – Easy on-PC (A&B)

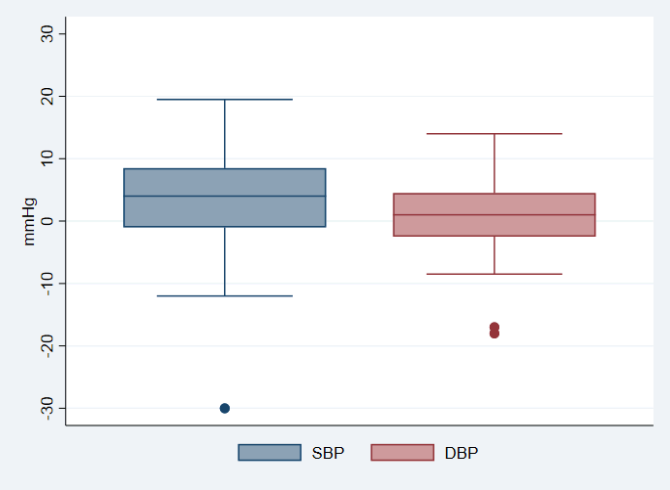


FVC: Micro Plus – Easy on-PC (A&B)



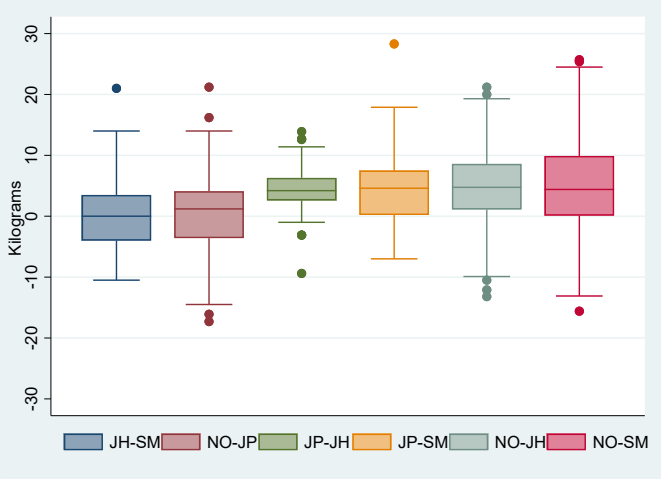
Appendix B Box plot of differences between devices

a) Box plot of differences between Omron HEM 907 and Omron 705-CP



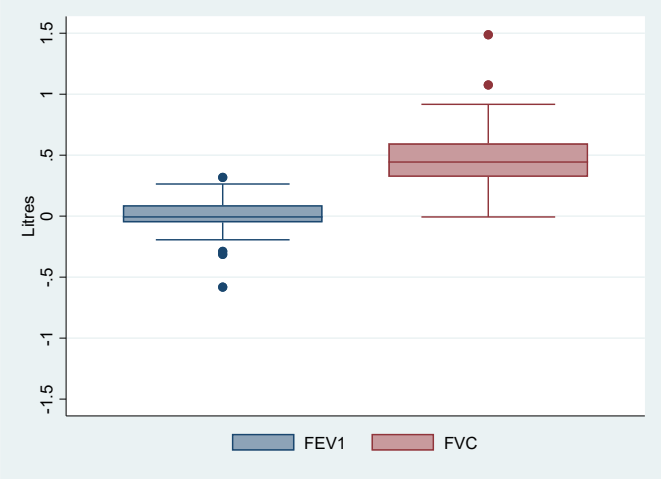
Notes: Mean of second and third readings

b) Box plot of differences between four dynamometers (max 4 readings)



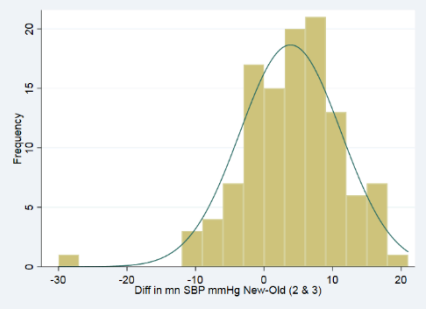
Notes: JH=Jamar Hydraulic, SM=Smedley, NO=Nottingham Electronic, JP=Jamar Plus+

c) Box plot of differences between Easy on-PC and Micro Plus (ATS/ERS criteria)

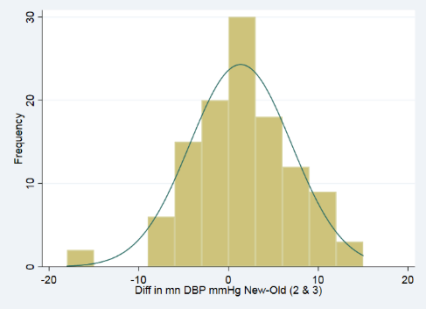


Appendix C Histograms of differences between devices

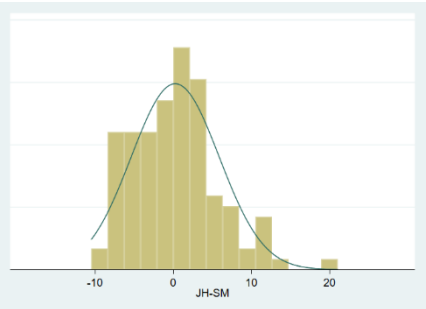
SBP: Omron 705 – 905 (Mean 2nd+3rd)



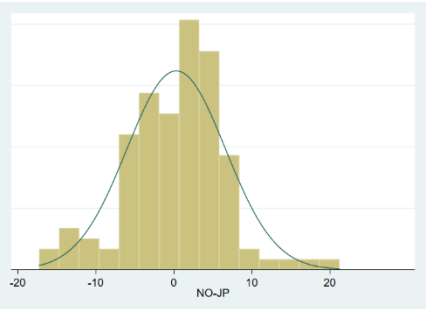
DBP: Omron 705 – 905 (Mean 2nd+3rd)



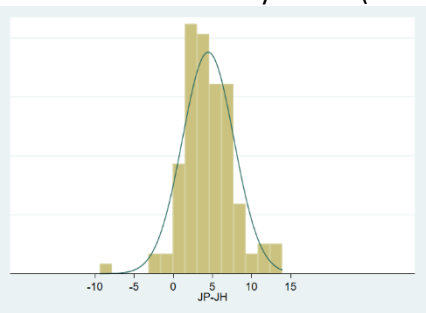
Nottingham – Jamar Plus+ (max of 4)



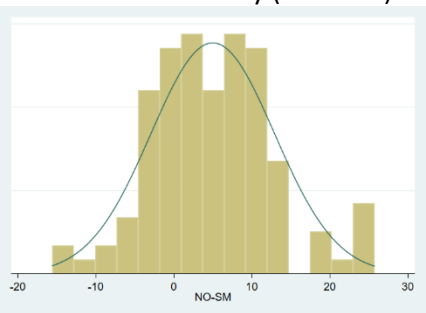
Jamar Hydraulic – Smedley (max of 4)



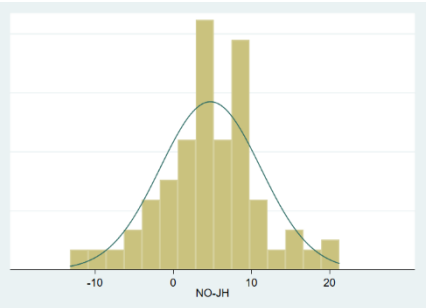
Jamar Plus+ - Jamar Hydraulic (max of 4)



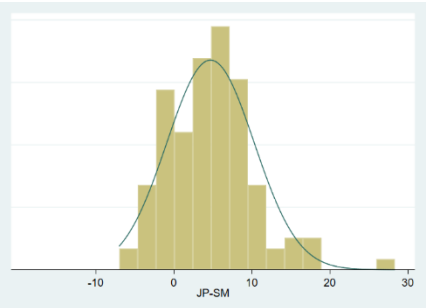
Jamar Plus+ - Smedley (max of 4)



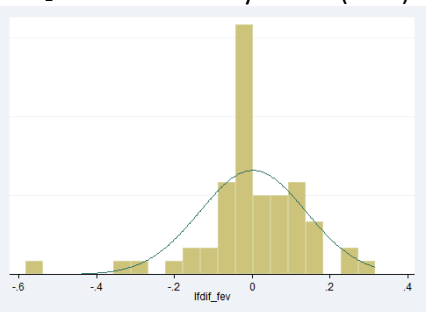
Nottingham – Jamar Hydraulic (max of 4)



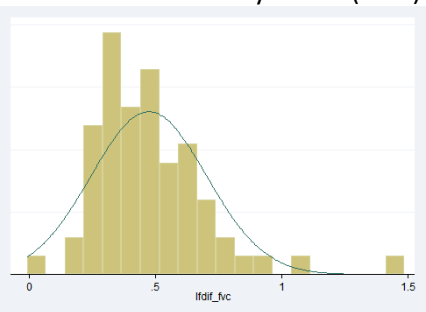
Nottingham – Smedley (max of 4)



FEV_1: Micro Plus-Easy on-PC (A&B)



FVC: Micro Plus – Easy on-PC (A&B)



Appendix D Multilevel models for blood pressure

Measure ->	SBP									DBP								
Model description ->	Device, allowing for clustering of readings within individual			Also controlling for order of device used and sequence of readings			Also controlling for any imbalance in sex, BMI and age			Device, allowing for clustering of readings within individual			Also controlling for order of device used and sequence of readings			Also controlling for any imbalance in sex, BMI and age		
Model number ->	1			2			3			1			2			3		
Variable	Coeff	SE	p value	Coeff	SE	p value	Coeff	SE	p value	Coeff	SE	p value	Coeff	SE	p value	Coeff	SE	p value
Intercept	119.6	1.54	0.000	121.0	1.62	0.000	77.2	12.01	0.000	75.4	1.02	0.000	76.6	1.07	0.000	64.5	7.68	0.000
Measurement at ind level																		
Device (Omron 705= ref)	3.9	0.46	0.000	3.9	0.46	0.000	3.9	0.46	0.000	1.5	0.32	0.000	1.5	0.31	0.000	1.6	0.31	0.000
Order of readings (1=ref)																		
2				-0.7	0.79	0.350	-0.7	0.80	0.378				-0.7	0.54	0.192	-0.7	0.54	0.214
3				-1.4	0.79	0.084	-1.3	0.80	0.094				-1.8	0.54	0.001	-1.7	0.54	0.001
4				-1.9	0.79	0.017	-1.8	0.80	0.021				-2.1	0.54	0.000	-2.1	0.54	0.000
5				-2.0	0.79	0.013	-2.0	0.80	0.013				-1.5	0.54	0.005	-1.4	0.54	0.008
6				-2.4	0.79	0.002	-2.3	0.80	0.004				-1.2	0.54	0.027	-1.1	0.54	0.038
Individual characteristics																		
Sex (Women=ref)							7.9	2.72	0.004							7.8	2.72	0.004
BMI							1.3	0.30	0.000							1.0	0.19	0.000
Age							0.1	0.17	0.698							-0.3	0.11	0.006
Random effects																		
Variance between indiv's	259.1	35.0		259.3	35.0		203.5	27.8		112.8	15.3		112.9	15.3		82.9	11.4	
Residual																		
Variance within individual	36.7	2.2		36.0	2.1		36.1	2.1		17.4	1.0		16.8	1.0		16.8	1.0	
n	689			689			683			689			689			683		

Appendix E Variance partitioning of blood pressure

DBP

	DBP (n=689 readings)					
	Variance between individuals			Variance between individuals and devices		
	Coeff	SE	p value	Coeff	SE	p value
Intercept	76.1	1.00	0.000	76.1	1.00	0.000
Random effects			Variance			Variance
Variance between individuals	112.7	15.26	86.2%	249.6	35.04	83.3%
Variance between devices	-	-	-	21.9	4.17	7.3%
Residual (within individual)	18.1	1.07	13.8%	28.2	1.86	9.4%

SBP

	SBP (n=689 readings)					
	Variance between individuals			Variance between individuals and devices		
	Coeff	SE	p value	Coeff	SE	p value
Intercept	121.6	1.52	0.000	121.6		0.000
Random effects			Variance			Variance
Variance between individuals	258.2	34.96	86.2%	249.6	35.04	83.3%
Variance between devices	-	-	-	21.9	4.17	7.3%
Residual (within individual)	41.4	2.44	13.8%	28.2	1.86	9.4%

Appendix F Publication related to Chapter 2

Accepted Article

Journal of Frailty, Sarcopenia and Falls JFSF

Short Communication

The impact of variation in the device used to measure grip strength on the identification of low muscle strength: Findings from a randomised cross-over study

Rachel Cooper[1]*, Carli Lessof[2]*, Andrew Wong[3], Rebecca Hardy[4]

[1]*Department of Sport and Exercise Sciences, Musculoskeletal Science and Sports Medicine Research Centre, Manchester Metropolitan University, Manchester, UK;*
[2]*National Centre for Research Methods, University of Southampton, Southampton, UK;*
[3]*MRC Unit for Lifelong Health and Ageing at UCL, London, UK;*
[4]*Cohort and Longitudinal Studies Enhancement Resources (CLOSER), UCL Social Research Institute, London, UK*
** equal contribution*

Abstract

Grip strength is commonly used to identify people with low muscle strength. It is unclear what impact the type of dynamometer used to measure grip strength has on the identification of low muscle strength so we aimed to assess this. Study participants were 118 men and women aged 45-74y from a randomised, repeated measurements cross-over study. Maximum grip strength was assessed using four hand-held dynamometers (Jamar Hydraulic; Jamar Plus+ Digital; Nottingham Electronic; Smedley) in a randomly allocated order. EWGSOP2 cut-points were applied to estimate prevalence of low muscle strength for each device. Agreement between devices was compared. Prevalence of low muscle strength varied by dynamometer ranging between 3% and 22% for men and, 3% and 15% for women. Of the 13 men identified as having low muscle strength by at least one of the four dynamometers, only 8% were identified by all four and 54% by just one. Of the 15 women classified as having low muscle strength by at least one of the four dynamometers, only 7% were identified by all four and 67% by only one. Variation in the measures of grip strength acquired by different hand-held dynamometers has potentially important implications when identifying low muscle strength.

Keywords: Cut-points, Hand-held dynamometer, Grip strength, Low muscle strength, Sarcopenia

There is increasing recognition of the important role of skeletal muscle for health and disease. This is exemplified by a growing awareness of the clinical importance of sarcopenia[1] - 'a progressive and generalised skeletal muscle disorder that involves the accelerated loss of muscle mass and function'[2] – which in the last 5 years has been assigned an ICD-10 code[3].

Despite progress, there remain well-documented challenges for clinical practice and research on sarcopenia[1,2]. One of the most important is the ongoing debate relating to how sarcopenia should be operationally defined[1]. Of a number of consensus definitions proposed over the last decade, the European Working Group on Sarcopenia in Older People's (EWGSOP) definition has gained considerable traction[4]. It was therefore noteworthy when an extended working group, EWGSOP-2, published a revised definition in 2019 to reflect updates to relevant evidence[5].

In working towards the aim of a true consensus definition and improved understanding of sarcopenia, each time a new definition is proposed it is important to compare this with

The authors have no conflict of interest.

Corresponding author: *Professor Rachel Cooper, Department of Sport and Exercise Sciences, Manchester Metropolitan University, All Saints Building, Manchester, M15 6BH, UK*

E-mail: *r.cooper@mmu.ac.uk*

Edited by: *George Lyritis*

Accepted *17 May 2021*

Appendix G The effect of matching on sample balance (all devices)

	PRE (n=4068) – Full sample				PRE DEM ADMIN n=3137 PRIMARY				PRE ADMIN (n=3189) SENSITIVITY 1				PRE DEM (n=3969) SENSITIVITY 2				PRE DEM SUR SENSITIVITY 3			
	Weighted		Matched		Weighted		Matched		Weighted		Matched		Weighted		Matched		Weighted		Matched	
	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T
IDACI	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
1st (Advantaged)	20.9	14.6	15.9	16.3	22.2	16.1	17.6	17.7	22.2	15.8	17.2	17.4	21.1	14.9	16.3	16.6	21.5	15.2	17.0	17.0
2nd	20.3	17.2	18.3	18.5	20.9	18.2	19.6	19.4	20.8	17.9	19.1	19.1	20.4	17.6	18.7	18.8	20.5	17.7	18.6	19.0
3rd	18.9	19.2	19.7	19.8	18.9	19.4	19.9	19.9	18.8	19.4	19.8	19.8	18.9	19.5	19.9	20.0	18.9	19.3	19.8	19.9
4th	20.1	21.8	20.8	20.8	19.3	21.2	20.3	20.0	19.4	21.7	20.8	20.4	19.8	21.6	20.6	20.5	19.4	21.5	20.4	20.4
5th (Disadvan.)	19.9	27.2	25.2	24.7	18.8	25.1	22.6	23.0	18.8	25.3	23.0	23.3	19.8	26.5	24.4	24.1	19.7	26.2	24.2	23.7
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		0.995		**0.000**		0.999		**0.000**		0.999		**0.000**		0.999		**0.000**		0.99
GOR	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
East Midlands	8.5	9.5	9.9	9.7	8.7	9.7	9.5	9.7	8.8	9.7	9.7	9.7	8.5	9.4	9.8	9.7	8.5	9.3	9.6	9.6
East of England	11.8	11.8	12.2	12.3	11.8	12.3	12.6	12.7	11.8	12.2	12.4	12.6	11.8	11.9	12.3	12.4	12.0	12.1	12.2	12.5
London	15.1	11.3	11.0	11.3	14.2	10.5	10.3	10.5	14.2	10.9	10.6	10.9	14.8	10.9	10.9	10.9	14.8	10.9	10.8	10.8
North-East	4.8	5.4	5.3	5.2	5.0	5.4	5.1	5.1	5.0	5.4	5.1	5.1	4.8	5.4	5.4	5.2	4.8	5.5	5.5	5.3
North-West	13.5	14.4	13.9	13.8	13.2	13.3	13.0	12.8	13.1	13.4	13.0	12.9	13.5	14.1	13.6	13.6	13.5	13.9	13.6	13.4
South-East	17.3	15.8	15.7	15.8	17.7	16.2	16.4	16.4	17.7	16.0	16.2	16.1	17.4	16.1	15.8	16.1	17.4	16.5	16.3	16.5
SW & Wales	10.5	7.9	8.0	8.3	11.2	9.0	9.3	9.4	11.1	8.8	9.0	9.2	10.7	8.0	8.1	8.4	10.7	8.2	8.2	8.5
West Midlands	10.0	12.0	12.0	12.0	9.2	11.1	11.1	11.3	9.4	11.1	11.2	11.4	9.8	12.1	12.0	12.0	9.7	11.7	11.6	11.7
Yorks. & Humber	8.5	11.8	11.9	11.5	9.1	12.5	12.7	12.1	9.0	12.6	12.7	12.2	8.7	11.9	12.0	11.7	8.6	12.0	12.2	11.7
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		1.000		**0.006**		1.000		**0.005**		1.000		**0.000**		1.000		**0.000**		1.00
Rural/Urban	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Urban	37.2	36.0	35.5	35.8	35.3	33.7	32.8	34.0	35.5	33.9	33.5	34.2	36.8	35.7	34.9	35.5	36.8	34.9	34.4	34.8
Urban city/town	44.9	47.0	46.4	46.1	45.3	47.8	47.9	46.6	45.3	47.7	47.4	46.5	45.0	47.1	46.6	46.1	44.9	47.6	47.2	46.6
Rural	17.9	16.9	18.2	18.0	19.4	18.5	19.3	19.4	19.2	18.4	19.1	19.3	18.2	17.2	18.5	18.3	18.2	17.5	18.3	18.6
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		0.428		0.971		0.422		0.756		0.447		0.891		0.447		0.930		0.290		0.93
Days from invite	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Pre reminder 0-	79.4	71.7	73.5	74.8	81.1	74.3	76.8	76.8	81.2	74.2	76.4	76.7	79.6	71.7	73.8	74.9	79.8	72.0	74.4	75.3
Post reminder 1	14.0	17.6	16.5	16.2	13.6	16.8	15.4	15.6	13.5	16.8	15.6	15.5	14.0	17.8	16.4	16.3	13.9	17.6	16.1	16.1
Post reminder 2	6.6	10.7	9.9	8.9	5.3	8.9	7.8	7.6	5.3	9.0	8.0	7.8	6.5	10.5	9.8	8.7	6.3	10.4	9.6	8.6
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		0.582		**0.000**		0.987		**0.000**		0.962		**0.000**		0.571		**0.000**		0.64

	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T
SEN status	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
No SEN	86.6	85.4	87.7	87.3	87.4	87.3	89.5	88.8	87.2	86.9	89.1	88.4	86.9	85.9	88.3	87.8	87.4	86.3	87.0	88.2
Some SEN	13.4	14.6	12.3	12.7	12.6	12.7	10.5	11.2	12.8	13.1	10.9	11.6	13.1	14.1	11.7	12.2	12.6	13.7	13.0	11.8
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		0.377		0.715		0.951		0.596		0.782		0.555		0.464		0.670		0.461		0.36
FSM status	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Not FSM	80.9	73.1	79.7	74.9	81.0	73.6	75.9	75.4	81.0	73.4	75.5	75.1	80.8	73.4	80.3	75.3	81.0	73.5	78.5	75.4
FSM or FSM6	19.1	26.9	20.3	25.1	19.0	26.4	24.1	24.6	19.0	26.6	24.5	24.9	19.2	26.6	19.7	24.7	19.0	26.5	21.5	24.6
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		**0.002**		**0.000**		0.763		**0.000**		0.822		**0.000**		**0.001**		**0.000**		0.05
KS2/KS4 science	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
High	43.7	29.2	44.7	33.1	44.5	30.0	33.9	33.8	44.5	29.7	33.7	33.5	43.7	29.4	44.6	33.4	43.9	29.9	39.6	33.9
Medium	38.6	45.1	39.7	45.7	38.6	45.8	46.9	46.3	38.7	45.7	46.6	46.3	38.5	45.3	39.9	45.8	38.4	45.1	42.1	45.5
Low	17.8	25.8	15.6	21.2	16.9	24.2	19.2	19.9	16.8	24.6	19.8	20.3	17.8	25.3	15.5	20.8	17.7	25.1	18.4	20.5
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		**0.000**		**0.000**		0.909		**0.000**		0.947		**0.000**		**0.000**		**0.000**		**0.01**
Schooattainment	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
High	81.8	76.4	79.8	77.9	83.0	78.4	80.0	79.7	83.1	78.2	79.6	79.5	81.8	76.7	80.3	78.2	81.9	77.2	78.8	78.7
Low	18.2	23.6	20.2	22.1	17.0	21.6	20.0	20.3	16.9	21.8	20.4	20.5	18.2	23.3	19.7	21.8	18.1	22.8	21.2	21.3
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		0.238		**0.003**		0.863		**0.001**		0.914		**0.001**		0.187		**0.003**		0.96
School selection	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Comp/Modern	93.0	96.2	93.1	95.4	92.9	96.2	95.5	95.5	92.9	96.3	95.7	95.6	93.1	96.1	93.1	95.4	93.0	96.1	94.2	95.3
Selective	7.0	3.8	6.9	4.6	7.1	3.8	4.5	4.5	7.1	3.7	4.3	4.4	6.9	3.9	6.9	4.6	7.0	3.9	5.8	4.7
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		**0.008**		**0.000**		0.997		**0.000**		0.918		**0.000**		**0.012**		**0.000**		0.18
School FSM	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Low	82.3	75.8	80.3	77.4	83.1	77.6	80.0	78.8	83.2	77.5	79.3	78.7	82.3	76.0	81.5	77.6	82.5	76.6	80.0	78.1
High	17.7	24.2	19.7	22.6	16.9	22.4	20.0	21.2	16.8	22.5	20.7	21.3	17.7	24.0	18.5	22.4	17.5	23.4	20.0	21.9
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		0.065		**0.000**		0.437		**0.000**		0.706		**0.000**		**0.012**		**0.000**		0.23
	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T	PC	SP/T

Gender	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Male	54.0	46.1	51.7	42.8	51.3	44.0	41.1	41.1	51.4	44.2	49.0	41.2	53.9	45.9	42.8	42.6	53.3	45.5	42.1	42.2
Female	46.0	53.9	48.3	57.2	48.7	56.0	58.9	58.9	48.6	55.8	51.0	58.8	46.1	54.1	57.2	57.4	46.7	54.5	57.9	57.8
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		**0.000**		**0.000**		0.995		**0.000**		**0.000**		**0.000**		0.868		**0.000**		0.93
School year	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Year 10	24.7	22.3	28.5	25.8	25.0	22.9	26.4	26.0	25.0	22.8	28.2	26	24.6	22.4	26.4	25.8	25.0	22.5	26.7	25.9
Year 11	25.3	24.7	26.5	26.7	24.4	26.0	28.9	27.6	24.6	25.8	26.3	27.4	25.0	24.9	27.2	26.9	24.9	25.0	27.5	27.0
Year 12	25.2	25.6	22.4	23.6	26.2	25.9	23.0	24.1	26.2	25.9	23.7	24.1	25.3	25.7	23.0	23.6	25.1	25.6	22.8	23.5
Year 13	24.9	27.3	22.6	23.9	24.4	25.2	21.7	22.3	24.2	25.5	21.7	22.6	25.1	27.0	23.4	23.6	24.9	26.8	23.0	23.5
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		0.227		0.289		0.535		0.824		0.532		0.620		0.377		0.960		0.333		0.91
Ethnicity	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
White	77.2	80.7	77.2	81.2	79.1	82.3	82.5	82.8	79.1	82.3	79.3	82.9	77.1	80.7	80.8	81.2	77.6	81	81.3	81.5
Mixed	4.6	4.5	4.5	4.5	4.6	4.9	4.6	4.7	4.5	4.9	4.0	4.7	4.6	4.5	4.4	4.4	4.6	4.5	4.5	4.4
Asian	12.3	10.0	12.4	9.6	10.8	8.2	8.1	7.8	10.8	8.2	10.9	7.8	12.3	10.0	9.9	9.6	12.0	9.9	9.4	9.4
Black	4.8	3.7	4.9	3.7	4.6	3.5	3.6	3.6	4.6	3.5	4.9	3.6	4.8	3.7	3.8	3.7	4.7	3.5	3.6	3.5
Other	1.1	1.1	1.1	1.1	0.9	1.1	1.2	1.1	0.9	1.1	0.9	1.1	1.1	1.1	1.1	1.1	1.1	1.2	1.2	1.1
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		0.102		**0.031**		0.120		0.998		0.113		**0.026**		0.102		0.998		0.121		0.99
Parent to Uni	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
1+ parent to Uni	36.9	23.1	35.3	24.8	37.9	25.7	34.0	27.0	38.2	25.6	34.5	26.9	36.7	23.3	35.0	25.0	37.1	23.6	25.4	25.3
No parent to Uni	55.7	69.5	57.3	67.9	55.5	68.5	59.6	67.2	55.2	68.7	58.6	67.3	56.0	69.5	58.5	67.8	55.9	69.3	67.9	67.6
Don't know	7.4	7.3	7.4	7.3	6.6	5.8	6.4	5.8	6.6	5.7	6.9	5.8	7.3	7.2	6.5	7.3	7.1	7.1	6.7	7.1
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		0.89
Level 3 science	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Doing or planning	48.5	36.5	47.8	38.9	49.9	38.8	45.5	41.2	49.9	38.7	45.7	41.1	48.6	36.7	47.6	39.1	48.7	37.1	38.8	39.5
Not doing /plan	43.6	53.9	43.6	52.3	42.6	52.8	46.7	51.0	42.5	52.7	46.0	50.9	43.4	54.2	44.1	52.6	43.4	54	52.3	52.3
Undecided	8.0	9.6	8.6	8.8	7.4	8.4	7.7	7.8	7.6	8.6	8.3	8.0	8.0	9.1	8.3	8.3	7.9	8.9	8.9	8.2
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		**0.000**		**0.000**		0.068		**0.000**		**0.036**		**0.000**		**0.000**		**0.000**		0.75
Higher ed plan	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
University degree	55.6	42.3	55.9	45	58.2	46.3	55.2	48.8	58.2	46.1	55.1	48.6	55.7	42.4	56.2	45.2	56.2	42.8	45.6	45.5
Higher education	5.3	7.1	5.0	6.7	5.6	6.6	5.9	6.2	5.5	6.6	5.8	6.2	5.4	7.1	5.1	6.6	5.3	6.9	5.9	6.4
Undecided	27.6	31.7	28.6	31.7	26.7	31.2	29.5	30.8	26.8	31.2	29.5	30.8	27.6	31.7	28.3	31.8	27.3	31.6	31.9	31.7
No	11.5	18.9	10.6	16.6	9.4	15.9	9.5	14.3	9.5	16	9.6	14.4	11.3	18.7	10.4	16.4	11.2	18.6	16.5	16.3
Total	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p-value		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		**0.000**		0.95

Appendix H Bivariate relationships: Adherence

Part One		Pr(used app)				Number of spending events			
		Logistic				Negative binomial			
Concept	**Category**	AME	P value	Wald test	Pr>Chi2	IRR	P value	Wald test	Pr>Chi2
Topic Salience	Keeps budget			1.9	0.164			3.0	0.082
	Does not	-0.045	0.164			0.845	0.082		
Frequency of shopping	Less than once a day			1.7	0.437			7.0	**0.030**
	About once a day	0.010	0.785			1.245	**0.046**		
	Several times a day	-0.063	0.255			1.393	**0.033**		
Role as Shopper in HH	Main or joint			7.9	**0.005**			26.2	**0.000**
	Not main/joint	-0.143	**0.007**			0.476	**0.000**		
Time constraint	Not time constrained			5.4	**0.020**			1.4	0.243
	Constrained	-0.084	**0.022**			0.882	0.243		
Past survey compliance	Low item non-resp			8.4	**0.004**			7.7	**0.006**
	High item non-resp	-0.100	**0.004**			0.753	**0.006**		
Concern about using app	Not/little/somewhat			4.7	**0.031**			5.6	**0.018**
	Very/extremely	-0.113	**0.038**			0.694	**0.018**		
Concern about using camera	Not at all			11.0	**0.001**			12.6	**0.000**
	Little/extremely	-0.110	**0.000**			0.708	**0.000**		
Freq of mobile device use	Less than daily			2.6	0.109			0.8	0.386
	Everyday	0.086	0.122			1.148	0.386		
Number of activities on device	None or 1	-0.183	**0.010**			0.550	**0.003**		
	2-8 activities (ref)			9.7	**0.008**			9.5	**0.009**
	9-12 activities	-0.088	**0.010**			0.799	**0.043**		
Gender	Male			0.0	0.972			2.2	0.136
	Female	0.001	0.972			1.160	0.136		
Age	16-30			7.5	0.111			12.3	**0.016**
	31-40	0.001	0.986			1.267	0.095		
	41-50	0.066	0.166			1.532	**0.003**		
	51-60	0.114	**0.019**			1.530	**0.005**		
	61+	0.038	0.487			1.446	**0.021**		
Qualifications	Degree			0.3	0.843			6.5	**0.039**
	School/higher quals	0.019	0.587			0.833	0.082		
	Other, none, missing	0.000	0.997			0.641	**0.018**		
Observations	N	8,040				8,038			
	Individuals	268				268			

AME=Average Marginal Effect; Ref=reference category; Wald test shows chi-square value then associated p-value.

Part Two		Pr(spending event entered by receipt)				Log(time from spending event to receipt photographed)			
		Logistic				Linear			
Concept	**Category**	AME	P value	Wald test	Pr>Chi2	1-exp(b)	P value	Wald test	Pr>Chi2
Topic Salience	Keeps budget (ref)			0.2	0.659			0.0	0.882
	Does not	0.016	0.659			0.017	0.881		
Frequency of shopping	Less than 1/day (ref)			0.5	0.782			2.8	0.251
	About once a day	-0.022	0.600			0.195	0.101		
	Several times a day	0.020	0.735			0.114	0.522		
Role as Shopper in HH	Main or joint (ref)			1.8	0.184			3.1	0.081
	Not main/joint	-0.079	0.190			0.300	0.081		
Time constraint	Not time con (ref)			0.2	0.658			0.2	0.654
	Time constrained	-0.018	0.659			-0.060	0.654		
Past survey compliance	Low item non-r (ref)			3.1	0.077			1.0	0.327
	High item non-response	0.068	0.073			-0.132	0.327		
Concern about using app	Not/little/somewhat			0.6	0.426			4.1	**0.044**
	Very/extremely	0.047	0.417			-0.490	**0.044**		
Concern about using camera	Not at all (ref)			0.0	0.967			1.5	0.215
	Little/extremely	-0.002	0.967			-0.156	0.215		
Freq of mobile device use	Less than daily (ref)			1.2	0.280			3.4	0.066
	Everyday	0.067	0.286			0.310	0.066		
Number of activities on device	None or 1	-0.087	0.079			0.177	0.262		
	2-8 activities (ref)			6.9	**0.032**			4.9	0.088
	9-12 activities	-0.106	**0.039**			-0.226	0.130		
Gender	Male (ref)			0.4	0.556			2.4	0.119
	Female	0.022	0.557			-0.207	0.119		
Age	16-30 (ref)			14.1	**0.007**			20.0	**0.001**
	31-40	0.071	0.197			0.078	0.643		
	41-50	0.136	**0.012**			0.052	0.757		
	51-60	0.189	**0.001**			0.046	0.791		
	61+	0.157	**0.009**			-0.912	**0.001**		
Qualifications	Degree (ref)			5.9	0.054			0.3	0.871
	School/higher quals	-0.051	0.184			0.001	0.996		
	Other, none, missing	0.099	0.128			-0.121	0.622		
Observations	N	7,412				3,454			
	Individuals	259				236			

AME=Average Marginal Effect; Ref=reference category; Wald test shows chi-square value then associated p-value.

Appendix I Sample definition for Spending Study

This appendix sets out a detailed description of how the two datasets analysed in Chapter 4 were cleaned and how they are comprised. The key information from this table is summarised in Table 27 on page 134 in the body of this thesis, but this very detailed information is necessary to permit replication of the analysis carried out here, and so is included as an appendix.

The first, 'app entry' dataset contains a minimum of one record for each study participant and a maximum of 134 (mean 44.0, SD=18.8). This was cleaned to remove ineligible cases based on four criteria (rows i-iv). First, although the study period was intended to be one month, the app was not designed to block entries at the end of that period so 2,008 app entries that were made more than 31 days after the participant's first valid app entry were dropped (row i). Secondly, since the analysis in this chapter relies on having basic information about the nature of each app entry (whether the participant photographed a receipt, directly entered a spending event, or recorded a no-spend day), 14 entries where information about 'activity type' was missing were also dropped (row ii). Part of the analysis uses information about the time lag between the moment a spending event occurs (based on data extracted from the photographed receipt) and the receipt being photographed (based on paradata from the app), so thirdly, 50 entries where the date on the photograph of the receipt was *after* the app entry were dropped (row iii[11]) and fourthly, 49 entries where the date on the receipt was before the date the individual made their first app entry, that is before the start of the study reference period (row iv[12])[13].

Having removed identifiable, invalid cases, the cleaned data set of app entries contains 9,386 records (row v). This includes 4,592 photographed receipts (row A plus row B) and 2,820 direct entries of spending events (row C), amounting to a total of 7,412 reported spending events which are either photographed or directly entered purchases comprising one observation per reported spending event (row A plus B plus C). These were entered by the 259 participants who recorded at least one spending event. Also, there are 1,974 app entries which record a 'no spend' day (row D).

[11] The 65 instances where this was observed were checked for data entry errors and 15 were corrected. These anomalous cases may have been the result of errors in the time recorded from the app (for example if participants were overseas) or errors on the printed receipts.

[12] Originally there were 66 cases of this kind but 17 were corrected following a review of the photographed receipts.

[13] It is important to note that these two final categories of ineligible app entries were only discoverable for photographed receipts where the time and date of purchase could be identified. It was not possible to verify the majority of app entries; the 1,138 photographed receipts with no date or time (row B), the 2,820 direct entries (row C) or the 1,974 'no spend' days (row D) where there is no independent data to allow verification of this or any other kind.

Table 32 Detailed information showing the genesis of the analytic dataset

		Full data set	Measure of adherence			
			Daily app use	Number of spending events	Method of reporting spending events	Time lag, spending event to report
Excluded observations						
i	Number of invalid app entries dropped as they were made 32+ days after first entry	2,008				
ii	Number of invalid app entries dropped as they had a missing activity type	14				
iii	Number of invalid receipts with date & time, receipt date after photograph date (15 corrected)	50				
iv	Number of invalid receipts with date & time, receipt date is before first app entry (17 corrected)	49				
Types of app entry						
A	App entry – valid receipt with date & time	3,454	*	*	*	*
B	App entry – photographed receipt, no date & time [14]	1,138	*	*	*	-
C	App entry – direct entry of spend	2,820	*	*	*	-
D	App entry – record of no spend day	1,974	*	* (counted as '0')	-	-
Number of app entries						
v	Number of app entries included in dataset/analysis	9,386	9,386 A+B+C+D	9,386 A+B+C+D	7,412 A+B+C	3,454 A only
vi	Once data is transformed to one observation per day per participant, calc. missing days		2,506	2,522 (counted as '0')	-	-
vii	Unit of analysis		One obs per day (days 2-31[15]) 268 x 30	One obs per day (days 1-31) 268 x 31	One obs. Per reported spending event	One obs. Per photo'd receipt

[14] Comprised of 942 missing date and time + 19 missing date and time once outliers checked + 50 missing time only + 127 missing date only.

[15] As explained later in this chapter, the analysis of daily app use is based on days 2-31 because, by definition, all participants made at least one valid app entry on their first study day, hence the number of records is 8,040 i.e., 268 x 30.

viii	Observations in analysis		8,040	8,308	7,412	3,454
ix	Individuals used in analysis		268	268	259	236

Of the 4,592 photographed receipts (rows A and B), 3,454 entries (recorded by 236 participants, shown in row A) had the date and time information needed to derive the time lag between the time a spending event occurs and photographing the receipt. In the remaining 1,138 receipts (row B) either the date, or the time, or both were missing, for example because the image only captured part of the receipt.

In the process of cleaning the data, the sample for this analysis was reduced from the 270 participants reported in the analysis of the study's response and bias to 268 participants (Jäckle, Burton, Couper et al., 2019). In the two cases that were dropped, the participant had only made a single entry in the app, and on both occasions the entry had been invalid; once because the activity type was missing, and once because the spending event that was reported had been made before the start of the participant's study reference period.

The final long file of app entries was transformed to create a second dataset with 31 records for each of the 268 participants. In this 'study days' dataset, each record contains variables summarising app activity on 31 consecutive days of the study, with day 1 set to the first day the app was used. Each record contains the number of receipts photographed that day (ranging from 0 to 13, mean=0.55, SD=1.08) and the number of direct entries made (ranging from 0 to 8, mean=0.34, SD=0.72). In addition, a flag was generated for days where the participant recorded no-spend (removing instances where this was entered on the same day as a report of a spending event or reducing the number of 'no spend' days to 1 where these were recorded several times on a single day)[16]. Finally, where no app entry was made on a given day, a flag was generated to connote a missing value (row vi). Taking all of this into account, row viii shows the final number of observations available for analysis for each form of adherence based on different units of analysis (row vii) and consequently different sample sizes (row ix). Ultimately (row viii), the 'study days' dataset forms the basis of the analysis of app use (where n= 8,040 i.e. 268 x 30 days; day 1 is excluded since by definition all participants made at least one valid app entry on their first study day), and of the number of spending events reported each day (where n=8,308 i.e. 268 x 31), while the 'app entry' dataset forms the basis of analysis of the method used for entering spending events (where n=7,412 for photographed receipts or direct entry) and the lapsed time between a

[16] In 104 instances spread over 63 participants (i.e., approximately 24% of the total of 268 participants) 'no spend' was entered more than once in a day. In 38 cases this was one extra time (approximately 14% of sample), in 19 cases this was two extra times (approximately 7% of sample), in 4 cases this was three extra times (approximately 1%), and there was 1 person who entered this 7 times and 1 person who entered it 9 times on a day.

spending event taking place and entering it (where n=3,454 based on photographed receipts with time and date information).

References

1939. STANDARDIZATION OF METHODS OF MEASURING THE ARTERIAL BLOOD PRESSURE: A JOINT REPORT OF THE COMMITTEES APPOINTED BY THE CARDIAC SOCIETY OF GREAT BRITAIN AND IRELAND AND THE AMERICAN HEART ASSOCIATION. *British heart journal,* 1**,** 261-267.

AAPOR 2010. Research synthesis: AAPOR report on online panels. . *Public Opinion Quarterly,* 74**,** 711-781.

ABRAHAM, K. G., MAITLAND, A. & BIANCHI, S. M. 2006. Nonresponse in the American Time Use Survey: Who Is Missing from the Data and How Much Does It Matter? *Public Opinion Quarterly,* 70**,** 676-703.

AHMED, N., BRZOZOWSKI, M. & CROSSLEY, T. F. 2006. Measurement errors in recall food consumption data. *IFS Working Paper,* W06/21.

AL BAGHAL, T., B, A., AUSPURG, K., BLAKE, M., BOOKER, C., CROSSLEY, T., D'ARDENNE, J., FAIRBROTHER, M., IACOVOU, M., JÄCKLE, A., KAMINSKA, O., LYNN, P., NICOLETTI, C., OLDFIELD, Z., PUDNEY, S., SCHNETTLER, S., UHRIG, S. C. N. & J, W. 2014. Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments. *Understanding Society Working Paper Series.* Colchester, Essex: University of Essex.

ALTMAN, D. G. 1985. Comparability of randomised groups. *Journal of the Royal Statistical Society: Series D (The Statistician),* 34**,** 125-136.

AMARAL, J. F., MANCINI, M. & NOVO JÚNIOR, J. M. 2012. Comparison of three hand dynamometers in relation to the accuracy and precision of the measurements. *Brazilian Journal of Physical Therapy,* 16**,** 216-224.

AMAYA, A., BIEMER, P. P. & KINYON, D. 2020. Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology,* 8**,** 89-119.

AMNAN, S. L., COOPER, R., ROTIMI, C., MCGEE, D., OSOTIMEHIN, B., KUDIRI, S., KINGUE, S., MURUQ, W., FRUSER, H., FORRESTER, T. & RUINFORD, W. 1996. <Ataman_19961-s2.0-0895435696001114-main.pdf>. *Journal of Clinical Epidemiology,* 49**,** 869-877.

ANDREADIS, I. 2015. Web Surveys Optimized for Smartphones: Are there Differences Between Computer and Smartphone Users? *2015,* 9.

ANGRISANI, M., FOSTER, K. & HITCZENKO, M. 2018. The 2015 and 2016 Diaries of Consumer Payment Choice: Technical Appendix. *Research Data Reports.* Federal Reserve Bank of Boston.

ANGRISANI, M., KAPTEYN, A. & SAMEK, S. 2017. Real Time Measurement of Household Electronic Financial Transactions in a Population Representative Panel. *The 7th Conference of the European Survey Research Association* Lisbon, Portugal.

ANTOUN, C. 2015. Who Are the Internet Users, Mobile Internet Users, and Mobile-Mostly Internet Users?

Demographic Differences across Internet-Use Subgroups in the U.S. *In:* TONINELLI, D., PINTER, R. & DE PEDRAZA, P. (eds.) *Mobile Research Methods.* Ubiquity Press.

ANTOUN, C., COUPER, M. P. & CONRAD, F. G. 2017. Effects of Mobile versus PC Web on Survey Response Quality: A Crossover Experiment in a Probability Web Panel. *Public Opinion Quarterly,* 81**,** 280-306.

ANTOUN, C., KATZ, J., ARGUETA, J. & WANG, L. 2018. Design Heuristics for Effective Smartphone Questionnaires. *Social Science Computer Review,* 36**,** 557-574.

ARMSTRONG, D. 2019. The social life of data points: Antecedents of digital technologies. *Social studies of science,* 49**,** 102-117.

AUSTIN, P. C. & STUART, E. A. 2015. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine,* 34**,** 3661-3679.

BALOGUN, J. A., AKOMOLAFE, C. T. & AMUSA, L. O. 1991. Grip strength: effects of testing posture and elbow position. *Archives of physical medicine and rehabilitation,* 72**,** 280-283.

BARLAS, F. M., THOMAS, R. K. & GRAHAM, P. 2015. Purposefully Mobile: Experimentally Assessing Device Effects in an Online Survey. *AAPOR 2015.* Hollywood, Florida.

BARRETT, G., LEVELL, P. & MILLIGAN, K. 2015. A Comparison of Micro and Macro Expenditure Measures across Countries Using Differing Survey Methods. *In:* CARROLL, C. D., CROSSLEY, T. F. & SABELHAUS, J. (eds.) *Improving the Measurement of Consumer Expenditures.* Chicago: University of Chicago Press.

BEE, C., MEYER, B. & SULLIVAN, J. 2015. The Validity of Consumption Data: Are the Consumer Expenditure Interview and Diary Surveys Informative? *In:* CARROLL, C. D., CROSSLEY, T. F. & SABELHAUS, J. (eds.) *Improving the Measurement of Consumer Expenditures.* Chicago: University of Chicago Press.

BELLI, R. F., SHAY, W. L. & STAFFORD, F. P. 2001. Event History Calendars and Question List Surveys: A Direct Comparison of Interviewing Methods*. *Public Opinion Quarterly,* 65**,** 45-74.

BENZEVAL, M., KUMARI, M. & JONES, A. M. 2016. How Do Biomarkers and Genetics Contribute to Understanding Society? *Health Economics,* 25**,** 1219-1222.

BETTS, P. & DICKINSON, E. 2015. Data collection methodology review for work stream 1 – under-reporting of expenditure. *National Statistics Quality Review of the Living Costs and Food Survey.* London.

BIEMER, P. P. 2010. Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly,* 74**,** 817-848.

BIEMER, P. P. 2018. Big Data Can't Replace Surveys, But They Can Work Together. Available: https://www.rti.org/insights/big-data-can%E2%80%99t-replace-surveys-they-can-work-together.

BILO, G., SALA, O., PEREGO, C., FAINI, A., GAO, L., GŁUSZEWSKA, A., OCHOA, J. E., PELLEGRINI, D., LONATI, L. M. & PARATI, G. 2017. Impact of cuff positioning on blood pressure measurement accuracy: may a specially designed cuff make a difference? *Hypertension research : official journal of the Japanese Society of Hypertension,* 40**,** 573-580.

BLAND, J. M. & ALTMAN, D. G. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet***,** 307-310.

BLAND, J. M. & ALTMAN, D. G. 1999. Measuring agreement in method comparison studies *Statistical Methods in Medical Research,* 8.

BLOM, A. G. & KORBMACHER, J. 2018. Linking Survey Data to Administrative Records in a Comparative Survey Context. *In:* VANNETTE, D. L. & KROSNICK, J. A. (eds.) *The Palgrave Handbook of Survey Research.* Palgrave Macmillan.

BOGART, L., BENIGER, J. R., BRODY, R. A., CRESPI, I., DAVIS, J. A., DAVISON, W. P., DICHTER, E., LANG, G. E., LANG, K., MENDELSON, H., ROGERS, E. M., DEARING, J. W., BALL-ROKEACH, S. J., ROKEACH, M., ROPER, B. W., SHEATSLEY, P. B., SHLAPENTOKH, V., WORCESTER, R. M., YANKELOVICH, D. & STOEZEL, J. 1987. The Future Study of Public Opinion: A Symposium. *The Public Opinion Quarterly,* 51**,** S173-S191.

BOLLING, K. 1994. *The Dinamap 8100 calibration study : a survey carried out by Social Survey Division of OPCS on behalf of the Department of Health*, London : H.M.S.O., 1994.

BOYLE, J. 2020. Declining survey response rates are a problem — here's why. Available from: https://www.icf.com/insights/health/declining-survey-response-rate-problem [Accessed 5 June 2022 2022].

BRADBURN, N. M. 1978. Respondent burden. *In:* REEDER, L. G. (ed.) *Health survey research methods.* Washington DC: US Government Printing Office.

BREWER, M., ETHERIDGE, B. & O'DEA, C. 2013. Why are households that report the lowest incomes so well-off? *University of Essex Department of Economics Discussion Paper Series* [Online].

BRICKA, S., ZMUD, J., WOLF, J. & FREEDMAN, J. 2009. Household Travel Surveys with GPS:An Experiment. *Transportation Research Record,* 2105**,** 51-56.

BRIDEVAUX, P. O., DUPUIS-LOZERON, E., SCHINDLER, C., KEIDEL, D., GERBASE, M. W., PROBST-HENSCH, N. M., BETTSCHART, R., BURDET, L., PONS, M., ROTHE, T., TURK, A., STOLZ, D., TSCHOPP, J. M., KUENZLI, N. & ROCHAT, T. 2015. Spirometer Replacement and Serial Lung Function Measurements in Population Studies: Results From the SAPALDIA Study. *Am J Epidemiol,* 181**,** 752-61.

BROWN, D. W., DESANTIS, S. M., GREENE, T. J., MAROUFY, V., YASEEN, A., WU, H., WILLIAMS, G. & SWARTZ, M. D. 2020. A novel approach for propensity score matching and stratification for multiple treatments: Application to an electronic health record–derived study. *Statistics in Medicine,* 39**,** 2308-2323.

BROWNING, M., CROSSLEY, T. F. & WINTER, J. 2014. The Measurement of Household Consumption Expenditures. *Annual Review of Economics,* 6**,** 475-501.

BRZOZOWSKI, M. & CROSSLEY, T. F. 2011. Viewpoint:Measuring the well-being of the poor with income or consumption: a Canadian perspective. *Canadian Journal of Economics,* 44**,** 88-106.

BULMER, M. 2001. Social Survey, History of. *In:* SMELSER, N. J. & BALTES, P. B. (eds.) *International Encyclopedia of the Social & Behavioral Sciences.* Oxford: Pergamon.

BULMER, M., BALES, K. & SKLAR, K. K. 1991. *The social survey in historical perspective, 1880-1940*, Cambridge : Cambridge University Press, 1991.

BUSKIRK, T. D. & ANDRUS, C. H. 2014. Making Mobile Browser Surveys Smarter:Results from a Randomized Experiment Comparing Online Surveys Completed via Computer or Smartphone. *Field Methods,* 26**,** 322-342.

BUTCHER, R. & ELDRIDGE, J. 1990. The Use of Diaries in Data Collection. *Journal of the Royal Statistical Society. Series D (The Statistician),* 39**,** 25-41.

CALDERWOOD, L. & LESSOF, C. 2009. Enhancing Longitudinal Surveys by Linking to Administrative Data. *In:* LYNN, P. (ed.) *Methodology of Longitudinal Surveys.*

CALLEGARO, M. 2010. Do You Know Which Device Your Respondent Has Used to Take Your Online Survey? *Survey Practice,* 3**,** 1-12.

CALLEGARO, M. 2013. From Mixed-mode to Multiple Devices: Web Surveys, Smartphone Surveys and Apps: Has the Respondent gone ahead of us in Answering Surveys? *International Journal of Market Research,* 55**,** 317-320.

CAMPBELL-HALL, V., CLEGG, S., V, D. G. & BOLLING, K. 2010. British Crime Survey – Interpersonal violence question development.

CAMPBELL, N. R. C. & MCKAY, D. W. 1999. Accurate blood pressure measurement. *Why does it matter?,* 161**,** 277-278.

CHAPLIN GRAY, J., GATENBY, R. & SIMMONDS, N. 2009. Millennium Cohort Study Sweep 3 Technical Report.

CHATZITHEOCHARI, S., FISHER, K., GILBERT, E., CALDERWOOD, L., HUSKINSON, T., CLEARY, A. & GERSHUNY, J. 2018. Using New Technologies for Time Diary Data Collection: Instrument Design and Data Quality Findings from a Mixed-Mode Pilot Survey. *Social Indicators Research,* 137**,** 379-390.

CHESNAYE, N. C., STEL, V. S., TRIPEPI, G., DEKKER, F. W., FU, E. L., ZOCCALI, C. & JAGER, K. J. 2021. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal,* 15**,** 14-20.

CHHAPOLA, V., KANWAL, S. K. & BRAR, R. 2015. Reporting standards for Bland–Altman agreement analysis in laboratory research: a cross-sectional survey of current practice. *Annals of Clinical Biochemistry: An international journal of biochemistry and laboratory medicine,* 52**,** 382-386.

CHOBANIAN, A. V., BAKRIS, G. L., BLACK, H. R., CUSHMAN, W. C., GREEN, L. A., IZZO JR, J. L., JONES, D. W., MATERSON, B. J., OPARIL, S. & WRIGHT JR, J. T. 2003. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *hypertension,* 42**,** 1206-1252.

CLAXTON, B. 2016. The case for mobile research: closer to the experience (part 1). Available from: https://www.peopleforresearch.co.uk/blog/2016/09/why-choose-nativeye-mobile-research/# last accessed 05/02/2020 [Accessed 30th September 2016.

CLEMENT, S. L., SEVERIN-NIELSEN, M. K. & SHAMSHIRI-PETERSEN, D. 2020. Device effects on survey response quality. A comparison of smartphone, tablet and PC responses on a cross sectional probability sample. *Survey Methods: Insights from the Field.* .

COBIAC, L. J. & SCARBOROUGH, P. 2021. Modelling future trajectories of obesity and body mass index in England. *PLOS ONE,* 16**,** e0252072.

COLLINS, D., COMANARU, R., CROSSLEY, T. F., LEPPS, H., PIGGOT, H. & PILEY, S. 2018. Living Costs and Food Survey: Review of the expenditure diaries. Unpublished manuscript.

COMPETITION & MARKETS AUTHORITY 2016. Retail banking market investigation: Final report.

CONNELLY, R. & PLATT, L. 2014. Cohort Profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology,* 43**,** 1719-1725.

CONRAD, F. G., SCHOBER, M. F., ANTOUN, C., HUPP, A. L. & YAN, H. Interviewing by Textng: Costs, Efficiency and Data Quality. AAPOR, 2014 Anaheim, California.

CONRAD, F. G., SCHOBER, M. F., ANTOUN, C., HUPP, A. L. & YAN, H. Y. 2017. Text Interviews on Mobile Devices. *Total Survey Error in Practice.*

CONRAD, F. G., TOURANGEAU, R., COUPER, M. & ZHANG, C. 2011. Interactive interventions in Web surveys can increase response accuracy. *2011 AAPOR Conference.* Pheonix, Arizona.

CONRAD, J., WIESE, M., ANDONE, I., KOCH, S., MARKOWETZ, A., ALEXY, U. & NOTHLINGS, U. 2020. Development and feasibility testing of the smartphone-based dietary record app NutriDiary (beta version). *Proceedings of the Nutrition Society,* 79**,** E84-E84.

CONVERSE, J. M. 2017. Survey Research in the United States: Roots and Emergence 1890-1960. Routledge.

COOPER, R., LESSOF, C., WONG, A. & HARDY, R. 2021. The impact of variation in the device used to measure grip strength on the identification of low muscle strength: Findings from a randomised cross-over study. *J Frailty Sarcopenia Falls,* 6**,** 225-230.

COUPER, M., ANTOUN, C. & MAVLETOVA, A. 2017. Mobile Web Surveys: A Total Survey Error Perspective. *In:* BIEMER, P., ECKMAN, S., EDWARDS, B., DE LEEUW, E., KREUTER, F., LYBERG, L., TUCKER, C. & WEST, B. (eds.) *Total Survey Error in Practice.* New York: Wiley.

COUPER, M. P. 2000. Review: Web Surveys: A Review of Issues and Approaches. *The Public Opinion Quarterly,* 64**,** 464-494.

COUPER, M. P. 2008. Technology and the Survey Interview/Questionnaire. *In:* CONRAD, F. G. & SCHOBER, M. F. (eds.) *Envisioning the Survey Interview of the Future.* Chichester: John Wiley & Sons.

COUPER, M. P. 2013a. Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods,* 7**,** 145-156.

COUPER, M. P. 2013b. Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *5th conference of the European Survey Research Association.* Lubliana, Slovenia.

COUPER, M. P. 2013c. Surveys on Mobile Devices: Opportunities and Challenges; Web Surveys for the General Population: How, Why and When? (Training Materials). London: Survey Research Center and Joint Program in Survey Methodology, University of Michigan.

COUPER, M. P. 2018. *RE: Proposal to use the concept of adherence.* Type to TEAM, S. S. R.

COUPER, M. P. & NICHOLLS, W. L. 1998. *The History and Development of Computer Assisted Survey Information Collection Methods in Collection,* New York, John Wiley and Sons.

COUPER, M. P. & PETERSON, G. J. 2017. Why Do Web Surveys Take Longer on Smartphones? *Social Science Computer Review,* 35**,** 357-377.

COUPER, M. P., TOURANGEAU, R., CONRAD, F. G. & CRAWFORD, S. D. 2004. What they see is what we get: response options for web surveys. *Soc. Sci. Comput. Rev.,* 22**,** 111–127.

COUPER, M. P., TOURANGEAU, R. & MARVIN, T. 2009. Taking the Audio Out of Audio-CASI. *Public Opinion Quarterly,* 73**,** 281-303.

CRIMMINS, E. M., KIM, J. K., LANGA, K. M. & WEIR, D. R. 2011. Assessment of cognition using surveys and neuropsychological assessment: the Health and Retirement Study and the

Aging, Demographics, and Memory Study. *The journals of gerontology. Series B, Psychological sciences and social sciences,* 66 Suppl 1**,** i162-i171.

CROSSLEY, T. F. & WINTER, J. K. 2016. Asking Households About Expenditures: What Have We Learned? *In:* CARROLL, C., CROSSLEY, T. F. & SABELHAUS, J. (eds.) *Improving the Measurement of Consumer Expenditures. Studies in Income and Wealth.* Chicago: University of Chicago Press.

CROUX, C. & DEHON, C. 2010. Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications,* 19**,** 497-515.

CUNNINGHAM, J. A., NEIGHBORS, C., BERTHOLET, N. & HENDERSHOT, C. S. 2013. Use of mobile devices to answer online surveys: implications for research. *BMC Research Notes,* 6**,** 258.

DE BRUIJNE, M. & WIJNANT, A. 2013. Comparing Survey Results Obtained via Mobile Devices and Computers:An Experiment With a Mobile Web Survey on a Heterogeneous Group of Mobile Devices Versus a Computer-Assisted Web Survey. *Social Science Computer Review,* 31**,** 482-504.

DE VAUS, D. 2013. *Surveys in social research*, Routledge.

DE VRIES, L. P., BASELMANS, B. M. L. & BARTELS, M. 2021. Smartphone-Based Ecological Momentary Assessment of Well-Being: A Systematic Review and Recommendations for Future Studies. *J Happiness Stud,* 22**,** 2361-2408.

DEAN, C. & LAWLESS, J. F. 1989. Tests for Detecting Overdispersion in Poisson Regression Models. *Journal of the American Statistical Association,* 84**,** 467-472.

DEATON, A. & GROSH, M. 2000. Consumption. *In:* M, G. & GLEWWE, P. (eds.) *Designing Household Survey Questionnaires for Developing Countries: Lessons from Ten Years of Living Standards Measurement Survey Experience.* Washington, DC: The World Bank.

DENG, T., KANTHAWALA, S., MENG, J., PENG, W., KONONOVA, A., HAO, Q., ZHANG, Q. & DAVID, P. 2019. Measuring smartphone usage and task switching with log tracking and self-reports. *Mobile Media & Communication,* 7**,** 23 - 3.

DEPARTMENT FOR WORK AND PENSIONS 2022. Family Resources Survey: background information and methodology. *In:* BRANDON-BRAVO, A., BRITTON, A., CAMERON, C., CONNOLLY, A., HASSAN, M., JONES, H., LOMAS, S., MCCAUGHEY, C., OWEN, J. & WARHURST, C. (eds.). London: Department for Work and Pensions,.

DODDS, R. M., SYDDALL, H. E., COOPER, R., BENZEVAL, M., DEARY, I. J., DENNISON, E. M., DER, G., GALE, C. R., INSKIP, H. M., JAGGER, C., KIRKWOOD, T. B., LAWLOR, D. A., ROBINSON, S. M., STARR, J. M., STEPTOE, A., TILLING, K., KUH, D., COOPER, C. & SAYER, A. A. 2014. Grip strength across the life course: normative data from twelve British studies. *PLoS One,* 9**,** e113637.

DOĞAN, N. Ö. 2018. Bland-Altman analysis: A paradigm to understand correlation and agreement. *Turkish Journal of Emergency Medicine,* 18**,** 139-141.

EDCOMS 2016. Wellcome Trust: SET Development 2016.

EDGAR, J., NELSON, D., PASZKIEWICZ, L. & SAFIR, A. 2013. The Gemini Project to Redesign the Consumer Expenditure Survey: Redesign Proposal. Gemini Design Team, Bureau of Labor Statistics.

ELLIOTT, J. & SHEPHERD, P. 2006. Cohort profile: 1970 British birth cohort (BCS70). *International journal of epidemiology,* 35**,** 836-843.

ENARSON, D. A., KENNEDY, S. M. & MILLER, D. L. 2004. Measurement in epidemiology. *Int J Tuberc Lung Dis,* 8**,** 1269-73.

ERENS, B., PHELPS, A., CLIFTON, S., MERCER, C. H., TANTON, C., HUSSEY, D., SONNENBERG, P., MACDOWALL, W., FIELD, N., DATTA, J., MITCHELL, K., COPAS, A. J., WELLINGS, K. & JOHNSON, A. M. 2014. Methodology of the third British National Survey of Sexual Attitudes and Lifestyles (Natsal-3). *Sexually Transmitted Infections,* 90**,** 84-89.

FERNEE, H. & SONCK, N. 2013. Is everyone able to use a smartphone in survey research. *Survey practice,* 6.

FESS, E. 1981. Clinical assessment recommendations. *American society of hand therapists***,** 6-8.

FIRRELL, J. C. & CRAIN, G. M. 1996. Which setting of the dynamometer provides maximal grip strength? *The Journal of hand surgery,* 21**,** 397-401.

FOREMAN, D. 2022. *Best Budgeting Apps Of April 2022* [Online]. Forbes. Available: https://www.forbes.com/advisor/banking/best-budgeting-apps/ [Accessed 22 April 2022].

FORMANEK, T., KAGSTROM, A., WINKLER, P. & CERMAKOVA, P. 2019. Differences in cognitive performance and cognitive decline across European regions: a population-based prospective cohort study. *European Psychiatry,* 58**,** 80-86.

FUCHS, M. 2008. Mobile Web Surveys: A Preliminary Discussion of Methodological Implications. *In:* CONRAD, F. G. & SCHOBER, M. F. (eds.) *Envisioning the Survey Interview of the Future.*

GATENBY, R. & HUNTER, P. 2000. Expenditure and Food Survey 2000 Pilot: Final Report. London.

GERBASE, M. W., DUPUIS-LOZERON, E., SCHINDLER, C., KEIDEL, D., BRIDEVAUX, P. O., KRIEMLER, S., PROBST-HENSCH, N. M., ROCHAT, T. & KUNZLI, N. 2013. Agreement between spirometers: a challenge in the follow-up of patients and populations? *Respiration,* 85**,** 505-14.

GERSHON, R. C., WAGSTER, M. V., HENDRIE, H. C., FOX, N. A., COOK, K. F. & NOWINSKI, C. J. 2013. NIH toolbox for assessment of neurological and behavioral function. *Neurology,* 80**,** S2-6.

GIAVARINA, D. 2015. Understanding Bland Altman analysis. *Biochem Med (Zagreb),* 25**,** 141-51.

GIBSON, R., FIELDHOUSE, E., GREEN, J. & SOUTHERN, R. 2016. iBES Aggregate Twitter Daily Campaign File v1.0: Explanatory Notes.

GILBERT, E. & LINDLEY, L.-J. 2019. Designing a device agnostic online survey for 17 year olds: Experiences of the Millennium Cohort Study. *European Social Research Association.* Faculty of Economics and Business at the University of Zagreb, Croatia.

GODFREY, A., DEL DIN, S., BARRY, G., MATHERS, J. C. & ROCHESTER, L. 2015a. Instrumenting gait with an accelerometer: a system and algorithm examination. *Med Eng Phys,* 37**,** 400-7.

GODFREY, A., LARA, J., DEL DIN, S., HICKEY, A., MUNRO, C. A., WIUFF, C., CHOWDHURY, S. A., MATHERS, J. C. & ROCHESTER, L. 2015b. iCap: Instrumented assessment of physical capability. *Maturitas,* 82**,** 116-22.

GOISIS, A., BROWN, M., KUMARI, M. & SULLIVAN, A. 2014. Overview of bio measures in longitudinal and life course research. *CLOSER Resource Report.* London: Institute of Education, University of London.

GOV.UK. 2022. How people access GOV.UK: Breakdown of desktop , mobile and tablet usage on GOV.UK over time. *The National Archives* [Online]. Available: https://webarchive.nationalarchives.gov.uk/ukgwa/20210315092414/https:/www.gov.uk/performance/site-activity/device-type [Accessed 27 January 2022].

GRAY, P. G. 1955. The Memory Factor in Social Surveys. *Journal of the American Statistical Association,* 50**,** 344-363.

GRIFFITH, R. 11 June 2018 2018. *RE: Comments made at a meeting held between ISER, IFS and Kantar Worldpanel to discuss the spending study and plans for a nutrition study.*

GROVES, R. M. & COUPER, M. P. 1998. When Interviewers Meet Householders: The Nature of Initial Interactions. *Nonresponse in Household Interview Surveys.*

GROVES, R. M., FOWLER JR, F. J., COUPER, M. P., LEPKOWSKI, J. M., SINGER, E. & TOURANGEAU, R. 2011. *Survey methodology*, John Wiley & Sons.

GROVES, R. M. & LYBERG, L. 2010. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly,* 74**,** 849-879.

GUERRA, R. S. & AMARAL, T. F. 2009. Comparison of hand dynamometers in elderly people. *The Journal of Nutrition, Health & Aging: Clinical Trials and Aging,* 13.

GUMMER, T., QUOß, F. & ROßMANN, J. 2019. Does Increasing Mobile Device Coverage Reduce Heterogeneity in Completing Web Surveys on Smartphones? *Social Science Computer Review,* 37**,** 371-384.

GUSTAFSON, A. 2016. *Adaptive Web Design: Crafting Rich Experiences With Progressive Enhancement,* United States of America, Pearson.

HAMLYN, B., FITZPATRICK, A. & WILLIAMS, J. 2015. Investigating the viability of moving from a face-to-face to an online/postal mode: evidence from a series of methodological studies 2012–2015. Cabinet Office.

HAMLYN, R., MATTHEWS, P. & SHANAHAN, M. 2017. Young people's views on science education: Science Education Tracker Research Report.

HAND, C. 2020. Biology and being green: The effect of prenatal testosterone exposure on pro-environmental consumption behaviour. *Journal of Business Research,* 120**,** 619-626.

HANDLER, J. 2009. The importance of accurate blood pressure measurement. *The Permanente Journal,* 13**,** 51.

HARDY, R., WADSWORTH, M. E., LANGENBERG, C. & KUH, D. 2004. Birthweight, childhood growth, and blood pressure at 43 years in a British birth cohort. *International Journal of Epidemiology,* 33**,** 121-129.

HARGITTAI, E. 2001. Second-level digital divide: Mapping differences in people's online skills. *First Monday,* 7.

HIRANI, V., TABASSUM, F., ARESU, M. & MINDELL, J. 2010. Development of New Demi-Span Equations from a Nationally Representative Sample of Adults to Estimate Maximal Adult Height. *The Journal of Nutrition,* 140**,** 1475-1480.

HOSIE, H. E. & NIMMO, W. S. 1988. Measurement of FEV1 and FVC. Comparison of a pocket spirometer with the Vitalograph. *Anaesthesia,* 43**,** 233-8.

HUBENER, E., KNAPPENBERGER, C., LEE, Y. & TAN, L. 2019. The 2018 CE Data Quality Profile. *Consumer Expenditure Surveys Program Report Series.* US Bureau of Labor Statistics.

INCEL, N. A., CECELI, E., DURUKAN, P. B., ERDEM, H. R. & YORGANCIOGLU, Z. R. 2002. Grip strength: effect of hand dominance. *Singapore medical journal,* 43**,** 234-237.

INSTITUTE FOR SOCIAL AND ECONOMIC RESEARCH 2020. Understanding Society – The UK Household Longitudinal Study, Innovation Panel, Waves 1-12, User Manual. Colchester: University of Essex.

JÄCKLE, A. 2009. Dependent interviewing: A framework and application to current research. *In:* LYNN, P. (ed.) *Methodology of Longitudinal Surveys.*

JÄCKLE, A., BURTON, B., COUPER, M. P. & LESSOF, C. 2019a. Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. *Survey Research Methods,* 13.

JÄCKLE, A., BURTON, B., WENZ, A. & READ, B. 2018a. *Understanding Society* The UK Household Longitudinal Study: spending study 1 User Guide. Colchester: University of Essex.

JÄCKLE, A., BURTON, B., WENZ, A. & READ, B. 2018b. *Understanding Society* The UK Household Longitudinal Study: spending study 1 User Guide - Appendix C: App Screenshots. Colchester: University of Essex.

JÄCKLE, A., COUPER, M. P., GAIA, A. & LESSOF, C. 2021. Improving Survey Measurement of Household Finances: A Review of New Data Sources and Technologies. *Advances in Longitudinal Survey Methodology.*

JÄCKLE, A., GAIA, A., LESSOF, C. & COUPER, M. P. 2019b. A review of new technologies and data sources for measuring household finances: implications for total survey error. Understanding Society at the Institute for Social and Economic Research.

JÄCKLE, A., WENZ, A., BURTON, J. & COUPER, M. P. 2022. Increasing Participation in a Mobile App Study: The Effects of a Sequential Mixed-Mode Design and In-Interview Invitation. *Journal of Survey Statistics and Methodology*.

JAPEC, L., KREUTER, F., BERG, M., BIEMER, P., DECKER, P., LAMPE, C., LANE, J., O'NEIL, C. & USHER, A. 2015. Big Data in Survey Research: AAPOR Task Force Report. *Public Opinion Quarterly,* 79**,** 839-880.

JOHNSON-HERRING, S. A., KING, S. L., TAN, L. & OLSON, T. 2009. The Effects of Double Placements in the Consumer Expenditure Diary Survey.

JONES, D. W., APPEL, L. J., SHEPS, S. G., ROCCELLA, E. J. & LENFANT, C. 2003. Measuring blood pressure accurately: new and persistent challenges. *JAMA,* 289**,** 1027-30.

JONES, K. H., HEYS, S., TINGAY, K. S., JACKSON, P. & DIBBEN, C. 2019. The Good, the Bad, the Clunky: Addressing Challenges in Using Administrative Data for Research. *International Journal of Population Data Science,* 4.

JOWELL, R., ROBERTS, C., FITZGERALD, R. & EVA, G. 2007. *Measuring attitudes cross-nationally: Lessons from the European Social Survey*, Sage.

KAAKS, R., RIBOLI, E. & VAN STAVEREN, W. 1995. Sample Size Requirements for Calibration Studies of Dietary Intake Measurements in Prospective Cohort Investigations. *American Journal of Epidemiology,* 142**,** 557-565.

KANTAR PUBLIC 2021. Community Life Surve Technical Report 2020/21. London: Kantar Public.

KAPTEYN, A., BANKS, J., HAMER, M., SMITH, J. P., STEPTOE, A., VAN SOEST, A., KOSTER, A. & HTAY WAH, S. 2018. What they say and what they do: comparing physical activity across the USA, England and the Netherlands. *Journal of epidemiology and community health,* 72**,** 471-476.

KEMSLEY, W. F. F. 1961. The Household Expenditure Enquiry of the Ministry of Labour. Variability in the 1953-54 enquiry. *Applied statistics,* 10**,** 117-135.

KEMSLEY, W. F. F. & NICHOLSON, J. L. 1960. Some Experiments in Methods of Conducting Family Expenditure Surveys. *Journal of the Royal Statistical Society: Series A (General),* 123**,** 307-328.

KEMSLEY, W. F. F., REDPATH, R. U., HOLMES, M. & DIVISION, G. B. O. O. P. C. S. S. S. 1980. *Family expenditure survey handbook : sampling, fieldwork, coding procedures and related methodological experiments : an account of the operations carried out by Social Survey Division of OPCS on behalf of the Department of Employment* London, H.M.S.O.

KEUSCH, F., BAHR, S., HAAS, G. C., KREUTER, F. & TRAPPMANN, M. 2020. Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey. *Sociological Methods & Research,* 0**,** 0049124120914924.

KEUSCH, F., STRUMINSKAYA, B., ANTOUN, C., COUPER, M. P. & KREUTER, F. 2019. Willingness to Participate in Passive Mobile Data Collection. *Public Opinion Quarterly,* 83**,** 210-235.

KEUSCH, F. & YAN, T. 2017. Web Versus Mobile Web:An Experimental Study of Device Effects and Self-Selection Effects. *Social Science Computer Review,* 35**,** 751-769.

KIBUCHI, E. M. 2018. *An Investigation of methods for Improving Survey Quality.* PhD, University of Southampton.

KIM, A., GUILLORY, J., BRADFIELD, B., RUDDLE, P., HSIEH, Y. P. & MURPHY, J. Information Exposure and Sharing Behavior of e-Cigarette Users: Do Survey Responses Correlate with Actual Tweeting Behavior? AAPOR 71st Annual Conference: Reshaping the Research Landscape: Public Opinion and Data Science, 12-15 May 2016 2016 Austen, Texas.

KIM, M. & SHINKAI, S. 2017. Prevalence of muscle weakness based on different diagnostic criteria in community-dwelling older adults: A comparison of grip strength dynamometers. *Geriatr Gerontol Int,* 17**,** 2089-2095.

KIM, Y., DYKEMA, J., STEVENSON, J., BLACK, P. & MOBERG, D. P. 2019. Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys. *Social Science Computer Review,* 37**,** 214-233.

KING, L., TAYLOR, H., YOUNG, L., OATES, C. & HALE, S. 2019. NTS Travel Diary Discovery Report. London: Department for Transport,.

KING, T. I. 2013. Interinstrument reliability of the Jamar electronic dynamometer and pinch gauge compared with the Jamar hydraulic dynamometer and B&L Engineering mechanical pinch gauge. *Am J Occup Ther,* 67**,** 480-3.

KREBS, D. & HÖHNE, J. K. 2020. Exploring Scale Direction Effects and Response Behavior across PC and Smartphone Surveys. *Journal of Survey Statistics and Methodology,* 9**,** 477-495.

KREUTER, F., HAAS, G.-C., KEUSCH, F., BÄHR, S. & TRAPPMANN, M. 2020. Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges Around Privacy and Informed Consent. *Social Science Computer Review,* 38**,** 533-549.

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

KUH, D., KARUNANANTHAN, S., BERGMAN, H. & COOPER, R. 2014. A life-course approach to healthy ageing: maintaining physical capability. *Proc Nutr Soc,* 73**,** 237-48.

LAMBERT, D. 1993. Measures of Disclosure Risk and Harm. *Journal of Official Statistics,* 9**,** 313-331.

LANGA, K. M., RYAN, L. H., MCCAMMON, R. J., JONES, R. N., MANLY, J. J., LEVINE, D. A., SONNEGA, A., FARRON, M. & WEIR, D. R. 2020. The Health and Retirement Study Harmonized Cognitive Assessment Protocol Project: Study Design and Methods. *Neuroepidemiology,* 54**,** 64-74.

LARA, J., COOPER, R., NISSAN, J., GINTY, A. T., KHAW, K. T., DEARY, I. J., LORD, J. M., KUH, D. & MATHERS, J. C. 2015. A proposed panel of biomarkers of healthy ageing. *BMC Med,* 13**,** 222.

LESSOF, C. 2009. Ethical Issues in Longitudinal Surveys. *In:* LYNN, P. (ed.) *Methodology of Longitudinal Surveys.*

LESSOF, C., ROSS, A. & BRIND, R. 2019. Multiple disadvantage and KS4 attainment: evidence from LSYPE2. *Department for Education Research Report.*

LESSOF, C., ROSS, A., BRIND, R., BELL, E. & NEWTON, S. 2016. Longitudinal Study of Young People in England cohort 2: health and wellbeing at wave 2. *Department for Education Research Report.*

LESSOF, C. & STURGIS, P. 2018. New Kinds of Survey Measurements. *In:* VANNETTE, D. L. & KROSNICK, J. A. (eds.) *The Palgrave Handbook of Survey Research.* Cham: Springer International Publishing.

LEVENSON, J. S. & GRADY, M. D. 2016. The influence of childhood trauma on sexual violence and sexual deviance in adulthood. *Traumatology,* 22**,** 94-103.

LIKERT, R. 1932. A technique for the measurement of attitudes. *Archives of Psychology,* 22**,** 5-55.

LINK, M. W., MURPHY, J., SCHOBER, M. F., BUSKIRK, T. D., HUNTER CHILDS, J. & LANGER TESFAYE, C. 2014. Mobile Technologies for Conducting, Augmenting and Potentially Replacing Surveys. *Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research.* Deerfield IL: American Association for Public Opinion Research.

LINVILLE, P. W., SALOVEY, P. & FISCHER, G. W. 1986. Stereotyping and perceived distributions of social characteristics: An application to ingroup-outgroup perception. *Prejudice, discrimination, and racism.* San Diego, CA, US: Academic Press.

LUGTIG, P. Jan 31, 2020 2020. Mobile-only web survey respondents. Available from: https://www.peterlugtig.com/post/mobile-only-web-survey-respondents/ [Accessed 22 May 2022 2022].

LUGTIG, P. & TOEPOEL, V. 2016. The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey:Effects on Survey Measurement Error. *Social Science Computer Review,* 34**,** 78-94.

LYNN, P. 2009. Sample design for Understanding Society. *Understanding Society Working Paper.* Colchester: University of Essex.

MARCOTTE, E. 2011. *Responsive Web Design*, A Book Apart.

MASLOVSKAYA, O., DURRANT, G. B., SMITH, P. W. F., HANSON, T. & VILLAR, A. 2019. What are the Characteristics of Respondents using Different Devices in Mixed-device Online Surveys? Evidence from Six UK Surveys. *International Statistical Review,* 87**,** 326-346.

MASLOVSKAYA, O., SMITH, P. & DURRANT, G. B. 2020. Do respondents using smartphones produce lower quality data? Evidence from the UK Understanding Society mixed-device survey. *National Centre for Research Methods Working Paper* [Online], 1/20.

MASSY-WESTROPP, N. M., GILL, T. K., TAYLOR, A. W., BOHANNON, R. W. & HILL, C. L. 2011. Hand Grip Strength: age and gender stratified normative data in a population-based study. *BMC Res Notes,* 4**,** 127.

MATHIOWETZ, V. 2002. Comparison of Rolyan and Jamar dynamometers for measuring grip strength. *Occupational Therapy International,* 9**,** 201-9.

MATTHEWS, P., BELL, E. & WENZ, A. 2017. Surveying young people in the smartphone age. *Social Research Practice,* 5**,** 2-11.

MAVLETOVA, A. 2013. Data quality in PC and mobile web surveys. *Social Science Computer Review,* 31**,** 725-743.

MAVLETOVA, A. & COUPER, M. P. 2013. Sensitive Topics in PC Web and Mobile Web Surveys: Is There a Difference? *Survey Research Methods,* 7.

MAVLETOVA, A. & COUPER, M. P. 2016. Grouping of Items in Mobile Web Questionnaires. *Field Methods,* 28**,** 170-193.

MAVLETOVA, A., COUPER, M. P. & LEBEDEV, D. 2018. Grid and Item-by-Item Formats in PC and Mobile Web Surveys. *Social Science Computer Review,* 36**,** 647-668.

MCCLAIN, C., CRAWFORD, S. D. & DUGAN, J. P. 2012. Use of Mobile Devices to Access Computer-optimized Web Instruments: Implications for Respondent Behavior and Data Quality. *American Association for Public Opinion Research (AAPOR) 67th Annual Conference.* Orlando, Florida.

MCFALL, S., PETERSEN, J., KAMINSKA, O. & LYNN, P. 2014. Understanding Society The UK Household Longitudinal Study: Waves 2 and 3 Nurse Health Assessment, 2010 – 2012, Guide to Nurse Health Assessment. Colchester, Essex: ISER, University of Essex.

MCGRATH, D. S., MEITNER, A. & SEARS, C. R. 2018. The specificity of attentional biases by type of gambling: An eye-tracking study. *PLOS ONE,* 13**,** e0190614.

MCWHINNEY, I. & CHAMPION, H. 1974. The Canadian experience with recall and diary methods in consumer expenditure surveys. *In:* BERG, S. V. (ed.) *Annals of Economic and Social Measurement.* Cambridge, MA: National Bureau of Economic Research.

MELZER, D., GARDENER, E., LANG, I., MCWILLIAMS, B. & GURALNIK, J. 2006. Measured physical performance. *In:* BANKS, J., BREEZE, E., LESSOF, C. & NAZROO, J. (eds.) *Retirement, health*

and relationships of the older population in England: The 2004 English Longitudinal Study of Ageing. Institute for Fiscal Studies.

MEYER, B. D. & SULLIVAN, J. X. 2003. Measuring the Well-Being of the Poor Using Income and Consumption. *Journal of Human Resources,* 38**,** 1180-1220.

MILANZI, E. B., KOPPELMAN, G. H., OLDENWENING, M., AUGUSTIJN, S., AALDERS-DE RUIJTER, B., FARENHORST, M., VONK, J. M., TEWIS, M., BRUNEKREEF, B. & GEHRING, U. 2019. Considerations in the use of different spirometers in epidemiological studies. *Environ Health,* 18**,** 39.

MILLER, M. R., HANKINSON, J., BRUSASCO, V., BURGOS, F., CASABURI, R., COATES, A., CRAPO, R., ENRIGHT, P., VAN DER GRINTEN, C. P., GUSTAFSSON, P., JENSEN, R., JOHNSON, D. C., MACINTYRE, N., MCKAY, R., NAVAJAS, D., PEDERSEN, O. F., PELLEGRINO, R., VIEGI, G., WANGER, J. & FORCE, A. E. T. 2005. Standardisation of spirometry. *European Respiratory Journal,* 26**,** 319-38.

MINDELL, J., BIDDULPH, J. P., HIRANI, V., STAMATAKIS, E., CRAIG, R., NUNN, S. & SHELTON, N. 2012. Cohort Profile: The Health Survey for England. *International Journal of Epidemiology,* 41**,** 1585-1593.

MINDELL, J., CHAUDHURY, M., ARESU, M. & JARVIS, D. 2011. Lung function in adults. *In:* CRAIG, R. & HIRANI, V. (eds.) *Health Survey for England 2010.* The Health and Social Care Information Centre.

MINELLI, C. 9 and 10 March 2014 2014. *RE: Deliberation about necessary sample size for the equipment comparison study.* Type to COOPER, R.

MOHANDES, A. & FOLEY, K. A. 2010. Medical Devices: Adapting to the Comparative Effectiveness Landscape. *Biotechnology Healthcare*.

MOHER, D., HOPEWELL, S., SCHULZ, K. F., MONTORI, V., GOTZSCHE, P. C., DEVEREAUX, P. J., ELBOURNE, D., EGGER, M. & ALTMAN, D. G. 2010. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ,* 340**,** c869.

MOLLA, D. T. & MUNISWAMY, B. 2012. Power of Tests for Overdispersion Parameter in Negative Binomial Regression Model. *IOSRJM,* 1**,** 29-36.

NETER, J. & WAKSBERG, J. 1964. A Study of Response Errors in Expenditures Data from Household Interviews. *Journal of the American Statistical Association,* 59**,** 18-55.

NHANES 1999. NHANES Spirometry Normative Values. *In:* VITALOGRAPH (ed.).

NICE 2013. The social care guidance manual: Appendix D Methodology checklist: cohort studies.

NUNES, V., NEILSON, J., O'FLYNN, N., CALVERT, N., KUNTZE, S., SMITHSON, H., BENSON, J., BLAIR, J., BOWSER, A., CLYNE, W., CROME, P., HADDAD, P., HEMINGWAY, S., HORNE, R., JOHNSON, S., KELLY, S., PACKHAM, B., PATEL, M. & STEEL, J. 2009. Clinical Guidelines and Evidence Review for Medicines Adherence: involving patients in decisions about prescribed medicines and supporting adherence. London: National Collaborating Centre for Primary Care and Royal College of General Practitioners.

O'DOHERTY, K., JASZCZAK, A., HOFFMANN, J. N., YOU, H. M., KERN, D. W., PAGEL, K., MCPHILLIPS, J., SCHUMM, L. P., DALE, W., HUANG, E. S. & MCCLINTOCK, M. K. 2014. Survey field methods for expanded biospecimen and biomeasure collection in NSHAP Wave 2. *J Gerontol B Psychol Sci Soc Sci,* 69 Suppl 2**,** S27-37.

O'DRISCOLL, S. W., HORII, E., NESS, R., CAHALAN, T. D., RICHARDS, R. R. & AN, K.-N. 1992. The relationship between wrist position, grasp size, and grip strength. *The Journal of hand surgery,* 17**,** 169-177.

O'NEILL, D., BENZEVAL, M., BOYD, A., CALDERWOOD, L., COOPER, C., CORTI, L., DENNISON, E., FITZSIMONS, E., GOODMAN, A. & HARDY, R. 2019. Data resource profile: cohort and longitudinal studies enhancement resources (CLOSER). *International journal of epidemiology,* 48**,** 675-676i.

OFCOM 2021. Online Nation: 2021 Report. *In:* OFCOM (ed.).

OFFICE FOR NATIONAL STATISTICS 2020. Internet access – households and individuals, Great Britain: 2020. *Statistical bulletin.*

OFFICE FOR NATIONAL STATISTICS 2021. Delivering the Census 2021 digital service: How the technical aspects of the Census 2021 digital service were built, for interest of

digital professionals across government. Office for National Statistics.

OFSTEDAL, M. B., BOUND, J. & KIM, M. H. 2013. Physical Performance, Self-Rated Health and Mortality among Older Adults in the US and England: Preliminary Draft. *2013 meeting of the Population Association of America.* New Orleans, LA.

OLMSTED-HAWALA, E., NICHOLS, E. & WANG, L. Numeric Keypads or Character Keyboards for Numeric Entries on Surveys and Forms: Surprising Results from Older Adults Using Mobile Devices. 2021 Cham. Springer International Publishing, 213-227.

ORFEI, L., STRACHAN, D. P., RUDNICKA, A. R. & WADSWORTH, M. E. 2008. Early influences on adult lung function in two national British cohorts. *Arch Dis Child,* 93**,** 570-4.

OSTERBERG, L. & BLASCHKE, T. 2005. Adherence to medication. *N Engl J Med,* 353**,** 487-97.

PAQUETTE, R., TANTON, C., BURNS, F., PRAH, P., SHAHMANESH, M., FIELD, N., MACDOWALL, W., GRAVNINGEN, K., SONNENBERG, P. & MERCER, C. H. 2017. Illicit drug use and its association with key sexual risk behaviours and outcomes: Findings from Britain's third National Survey of Sexual Attitudes and Lifestyles (Natsal-3). *PLOS ONE,* 12**,** e0177922.

PARK, A., BRYSON, C., CLERY, E., CURTICE, J. & PHILLIPS, M. 2013. *British Social Attitudes: the 30th Report,* London, NatCen Social Research.

PEARL, R. B. 1979. Reevaluation of the 1972-73 US Consumer Expenditure Surveys. *US Bureau of the Census Technical Papers.* Washington DC: US Government Printing Office.

PETERSON, G. 2012. What can we learn from unintentional mobile respondents. *Council of American Survey Research Organizations (CASRO) Journal,* 2012-13.

PETERSON, G., GRIFFIN, J., LAFRANCE, J. & LI, J. 2017. Smartphone Participation in Web Surveys. Choosing between the Potential for Coverage, Nonresponse, and Measurement Error. *In:* P, B. P., DE LEEUW, E. D., ECKMAN, S., EDWARDS, B., KREUTER, F., LYBERG, L. E., TUCKER, C. N. & T, W. B. (eds.) *Total Survey Error in Practice.* New York: Wiley.

PEYTCHEV, A. 2009. Survey Breakoff. *Public Opinion Quarterly,* 73**,** 74-97.

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

PIERCE, M. B., ZANINOTTO, P., STEEL, N. & MINDELL, J. 2009. Undiagnosed diabetes—data from the English longitudinal study of ageing. *Diabetic Medicine,* 26**,** 679-685.

POGGIO, T., BOŠNJAK, M. & WEYANDT, K. W. 2015. Survey Participation via Mobile Devices in a Probability-based Online-Panel: Prevalence, Determinants, and Implications for Nonresponse. *Survey practice,* 8**,** 2849.

POIKOLAINEN, K. & KÄRKKÄINEN, P. 1983. Diary gives more accurate information about alcohol consumption than questionnaire. *Drug Alcohol Depend,* 11**,** 209-16.

POWER, C. & ELLIOTT, J. 2006. Cohort profile: 1958 British birth cohort (national child development study). *International journal of epidemiology,* 35**,** 34-41.

POWERS, B. J., OLSEN, M. K., SMITH, V. A., WOOLSON, R. F., BOSWORTH, H. B. & ODDONE, E. Z. 2011. Measuring Blood Pressure for Decision Making and Quality Reporting: Where and How Many Measures? *Annals of Internal Medicine,* 154**,** 781-788.

PUDNEY, S., HANCOCK, R. & SUTHERLAND, H. 2006. Simulating the Reform of Means-tested Benefits with Endogenous Take-up and Claim Costs*. *Oxford Bulletin of Economics and Statistics,* 68**,** 135-166.

RABE-HESKETH, S. & SKRONDAL, A. 2012. *Multilevel and longitudinal modelling using STATA*, STATA Press.

RALPH, J. & MANCLOSSI, S. 2016. Living Costs and Food Survey. *National Statistics Quality Review.* London: Office for National Statistics.

RANSLEY, J. K., DONNELLY, J. K., KHARA, T. N., BOTHAM, H., ARNOT, H., GREENWOOD, D. C. & CADE, J. E. 2001. The use of supermarket till receipts to determine the fat and energy intake in a UK population. *Public Health Nutrition,* 4**,** 1279-1286.

READ, B. 2019a. The influence of device characteristics on data collection using a Mobile App. *Understanding Society Working Paper Series.* Understanding Society at the Institute for Social and Economic Research.

READ, B. 2019b. Respondent burden in a Mobile App: evidence from a shopping receipt scanning study. *Survey Research Methods,* 13**,** 45-71.

REUBEN, D. B., MAGASI, S., MCCREATH, H. E., BOHANNON, R. W., WANG, Y. C., BUBELA, D. J., RYMER, W. Z., BEAUMONT, J., RINE, R. M., LAI, J. S. & GERSHON, R. C. 2013. Motor assessment using the NIH Toolbox. *Neurology,* 80**,** S65-75.

REVILLA, M. 2017. Are There Differences Depending on the Device Used to Complete a Web Survey (PC or Smartphone) for Order-by-click Questions? *Field Methods,* 29**,** 266-280.

REVILLA, M. & COUPER, M. P. 2018. Testing different rank order question layouts for PC and smartphone respondents. *International Journal of Social Research Methodology,* 21**,** 695-712.

REVILLA, M., COUPER, M. P. & OCHOA, C. 2019. Willingness of Online Panelists to Perform Additional Tasks. *2019,* 13.

REVILLA, M., TONINELLI, D., OCHOA, C. & LOEWE, G. 2016. Do online access panels need to adapt surveys for mobile devices? *Internet Res.,* 26**,** 1209-1227.

RITCHIE, J. & THOMAS 1992. Referred to in Collins, D (2018) Living Costs and Food Survey: Review of the expenditure diaries. Unpublished manuscript.

ROBERTS, C., GILBERT, E., ALLUM, N. & EISNER, L. 2019. Research Synthesis: Satisficing in Surveys: A Systematic Review of the Literature. *Public Opinion Quarterly,* 83**,** 598-626.

ROBERTS, C. & TORGERSON, D. J. 1999. Understanding controlled trials: baseline imbalance in randomised controlled trials. *Bmj,* 319**,** 185.

ROBERTS, H. C., DENISON, H. J., MARTIN, H. J., PATEL, H. P., SYDDALL, H., COOPER, C. & SAYER, A. A. 2011. A review of the measurement of grip strength in clinical and epidemiological studies: towards a standardised approach. *Age and Ageing,* 40**,** 423-9.

ROFIQUE, J., HUMPHREY, A. & KILLPACK, C. 2012. National Travel Survey 2011 GPS Pilot, a technical report on the pilot survey management and data collection. London: Department for Transport.

ROMANO, J. C. & CHEN, J. M. 2011. A Usability and Eye-Tracking Evaluation of Four Versions of the Online National Survey of College Graduates (NSCG): Iteration 2. *Census Working Papers.* United States Census Bureau.

ROßMANN, J. 2017. RESPDIFF: Stata module for generating response differentiation indices *Statistical Software Components S458315.* Boston College Department of Economics.

RUBIN, D. C. & BADDELEY, A. D. 1989. Telescoping is not time compression: A model of the dating of autobiographical events. *Memory & Cognition,* 17**,** 653-661.

SAKSHAUG, J. W., COUPER, M. P., OFSTEDAL, M. B. & WEIR, D. R. 2012. Linking Survey and Administrative Records:Mechanisms of Consent. *Sociological Methods & Research,* 41**,** 535-569.

SAKSHAUG, J. W., OFSTEDAL, M. B., GUYER, H. & BEEBE, T. J. 2014. The Collection of Biospecimens in Health Surveys. *Health Survey Methods.*

SALAZAR, M. R., ESPECHE, W. G., AIZPURUA, M., SISNIEGUEZ, C. E., SISNIEGUEZ, B. C., DULBECCO, C. A., MARCH, C. E., STAVILE, R. N., FERRARI, E. H., CORREA, M., MACIEL, P. M., BALBIN, E. & CARBAJAL, H. A. 2015. Should the first blood pressure reading be discarded? *J Hum Hypertens,* 29**,** 373-8.

SCHLOSSER, S. & MAYS, A. 2018. Mobile and Dirty:Does Using Mobile Devices Affect the Data Quality and the Response Process of Online Surveys? *Social Science Computer Review,* 36**,** 212-230.

SCHOBER, M. F. & CONRAD, F. G. 2008. Survey Interviews with New Communication Technologies: Synthesis and Future Opportunities. *In:* CONRAD, F. G. & SCHOBER, M. F. (eds.) *Envisioning the Survey Interview of the Future.*

SCHOLES, S. & NEAVE, A. 2017. Health Survey for England 2016: Physical activity in adults. *Leeds: Health and Social Care Information Centre*.

SCHULZ, K. F., ALTMAN, D. G., MOHER, D. & GROUP, C. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med,* 7**,** e1000251.

SCHWERTMAN, N. C., OWENS, M. A. & ADNAN, R. 2004. A simple more general boxplot method for identifying outliers. *Computational Statistics & Data Analysis,* 47**,** 165-174.

SENDELBAH, A., VEHOVAR, V., SLAVEC, A. & PETROVČIČ, A. 2016. Investigating respondent multitasking in web surveys using paradata. *Computers in Human Behavior,* 55**,** 777-787.

SHARF, S. 2016. *12 Free Apps To Track Your Spending And How To Pick The Best One For You* [Online]. Forbes. Available: https://www.forbes.com/sites/samanthasharf/2016/03/02/12-free-apps-to-track-your-spending-and-how-to-pick-the-best-one-for-you/ [Accessed 4 February 2020].

SHARP, D. B. & ALLMAN-FARINELLI, M. 2014. Feasibility and validity of mobile phones to assess dietary intake. *Nutrition,* 30**,** 1257-1266.

SHELBY, M. H. 1931. The Social Survey The Idea Defined and its Development Traced (reprinted in 1970). Russell Sage Foundation.

SILBERSTEIN, A. R. & SCOTT, S. 1991. Expenditure Diary Surveys and Their Associated Errors. *Measurement Errors in Surveys.*

SKIRTON, H., CHAMBERLAIN, W., LAWSON, C., RYAN, H. & YOUNG, E. 2011. A systematic review of variability and reliability of manual and automated blood pressure readings. *J Clin Nurs,* 20**,** 602-14.

SONCK, N. & FERNEE, H. Using smartphones in survey research: a multifunctional tool. 2013.

SONNEGA, A., FAUL, J. D., OFSTEDAL, M. B., LANGA, K. M., PHILLIPS, J. W. & WEIR, D. R. 2014. Cohort Profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology,* 43**,** 576-585.

SOUSA-SANTOS, A. R. & AMARAL, T. F. 2017. Differences in handgrip strength protocols to identify sarcopenia and frailty - a systematic review. *BMC Geriatr,* 17**,** 238.

STANG, A., MOEBUS, S., MOHLENKAMP, S., DRAGANO, N., SCHMERMUND, A., BECK, E. M., SIEGRIST, J., ERBEL, R., JOCKEL, K. H. & HEINZ NIXDORF RECALL STUDY INVESTIGATIVE, G. 2006. Algorithms for converting random-zero to automated oscillometric blood pressure values, and vice versa. *Am J Epidemiol,* 164**,** 85-94.

STAPLETON, C. E. 2013. The Smart(Phone) Way to Collect Survey Data. *Survey Practice,* 6**,** 1-7.

STATISTICS CANADA 1996. Food Expenditure Survey, Public-use Microdata Files. *In:* INCOME STATISTICS DIVISION, S. C. (ed.). Ottawa.

STEEL, N., HUPPERT, F., MCWILLIAMS, B. & MELZER, D. 2003. Physical and cognitive function. *In:* MARMOT, M., BANKS, J., BLUNDELL, R., LESSOF, C. & NAZROO, J. (eds.) *Health, wealth and lifestyles of the older population in England: The 2002 English Longitudinal Study of Ageing.* London: The Institute for Fiscal Studies.

STEPHENS, M., JR. 2003. "3rd of tha Month": Do Social Security Recipients Smooth Consumption Between Checks? *American Economic Review,* 93**,** 406-422.

STEPTOE, A., BREEZE, E., BANKS, J. & NAZROO, J. 2012. Cohort Profile: The English Longitudinal Study of Ageing. *International Journal of Epidemiology,* 42**,** 1640-1648.

STRUMINSKAYA, B., WEYANDT, K. & BOSNJAK, M. 2015. The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel. *2015,* 9.

STURGIS, P., KUHA, J., BAKER, N., CALLEGARO, M., FISHER, S., GREEN, J., JENNINGS, W., LAUDERDALE, B. E. & SMITH, P. 2018. An assessment of the causes of the errors in the 2015 UK general election opinion polls. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 181**,** 757-781.

STÜRMER, T., WYSS, R., GLYNN, R. J. & BROOKHART, M. A. 2014. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *J Intern Med,* 275**,** 570-80.

SUDMAN, S., BRADBURN, N. M. & SCHWARZ, N. 1996. *Thinking about answers: The application of cognitive processes to survey methodology,* San Francisco, CA, US, Jossey-Bass.

SUDMAN, S. & FERBER, R. 1971. Experiments in Obtaining Consumer Expenditures by Diary Methods. *Journal of the American Statistical Association,* 66**,** 725-735.

SUGIE, N. F. 2018. Utilizing Smartphones to Study Disadvantaged and Hard-to-Reach Groups. *Sociological Methods & Research,* 47**,** 458-491.

SUSTAKOSKI, A., PERERA, S., VANSWEARINGEN, J. M., STUDENSKI, S. A. & BRACH, J. S. 2015. The impact of testing protocol on recorded gait speed. *Gait Posture,* 41**,** 329-31.

SVENS, B. & LEE, H. 2005. Intra- and inter-instrument reliability of Grip-Strength Measurements: GripTrack™ and Jamar® hand dynamometers. *The British Journal of Hand Therapy,* 10**,** 47-55.

TANNER, S. 1998. How Much Do Consumers Spend? Comparing the FES and National Accounts. *In:* BANKS, J. & JOHNSON, P. E. (eds.) *How reliable is the family expenditure survey? Trends in incomes and expenditures over time.* London: Institute for Fiscal Studies.

THOMAS, E. T., GUPPY, M., STRAUS, S. E., BELL, K. J. & GLASZIOU, P. 2019. Rate of normal lung function decline in ageing adults: a systematic review of prospective cohort studies. *BMJ open,* 9**,** e028150.

TILSON, H. H. 2004. Adherence or Compliance? Changes in Terminology. *Annals of Pharmacotherapy,* 38**,** 161-162.

TIMMINS, K. A., MORRIS, M. A., HULME, C., EDWARDS, K. L., CLARKE, G. P. & CADE, J. E. 2013. Comparability of methods assigning monetary costs to diets: derivation from household till receipts versus cost database estimation using 4-day food diaries. *European Journal of Clinical Nutrition,* 67**,** 1072-1076.

TING, Y., MACHADO, J., HELLER, A., BONILLA, E., MAITLAND, A. & KIRLIN, J. 2017. The Feasibility of Using Smartphones to Record Food Purchase and Acquisition. *AAPOR 72th Annual Conference.* New Orleans, Louisiana.

TOEPOEL, V. & LUGTIG, P. Mobile Pevices a Way to Recruit Hard-to-reach Groups? Results from a Pilot Study Comparing Desk Top and Mobile dDvice Surveys. 5th Conference of the European Survey Research Association, Ljubljana, Slovenia, 2013.

TOEPOEL, V. & LUGTIG, P. 2014. What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users. *Social Science Computer Review,* 32**,** 544-560.

TOLONEN, H., KOPONEN, P., NASKA, A., MANNISTO, S., BRODA, G., PALOSAARI, T., KUULASMAA, K. & PROJECT, E. P. 2015. Challenges in standardization of blood pressure measurement at the population level. *BMC Med Res Methodol,* 15**,** 33.

TONINELLI, D. & REVILLA, M. Smartphones vs PCs: Does the device affect the web survey experience and the measurement error for sensitive topics?-A replication of the Mavletova & Couper's 2013 experiment. Survey Research Methods, 2016. 153-169.

TOURANGEAU, R., CONRAD, F. G. & COUPER, M. P. 2013. *The Science of Web Surveys,* New York, Oxford University Press.

TOURANGEAU, R., MAITLAND, A., RIVERO, G., SUN, H., WILLIAMS, D. & YAN, T. 2017. Web Surveys by Smartphone and Tablets Effects on Survey Responses. *Public Opinion Quarterly,* 81**,** 896-929.

TOURANGEAU, R., SUN, H., YAN, T., MAITLAND, A., RIVERO, G. & WILLIAMS, D. 2018. Web Surveys by Smartphones and Tablets:Effects on Data Quality. *Social Science Computer Review,* 36**,** 542-556.

TUCKER, C. 1992. The estimation of instrument effects on data quality in the consumer expenditure diary survey. *Journal of Official Statistics,* 8**,** 41-61.

TURNER, R. 1961. Inter-Week Variations in Expenditure Recorded During a Two-Week Survey of Family Expenditure. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 10**,** 136-146.

UNIVERSITY OF ESSEX 2018. Institute for Social and Economic Research: Understanding Society: Spending Study 1, 2016-17. [data collection]. UK Data Service.

UNIVERSITY OF ESSEX 2021. Institute for Social and Economic Research: Understanding Society: Innovation Panel, Waves 1-13, 2008-2020. [data collection] 11th Edition. DOI: http://doi.org/10.5255/UKDA-SN-6849-14: UK Data Service.

VOLKOVA, E., LI, N., DUNFORD, E., EYLES, H., CRINO, M., MICHIE, J. & NI MHURCHU, C. 2016. "Smart" RCTs: Development of a Smartphone App for Fully Automated Nutrition-Labeling Intervention Trials. *JMIR Mhealth Uhealth,* 4**,** e23.

WADSWORTH, M., KUH, D., RICHARDS, M. & HARDY, R. 2006. Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int J Epidemiol,* 35**,** 49-54.

WAN, Y., HENEGHAN, C., STEVENS, R., MCMANUS, R. J., WARD, A., PERERA, R., THOMPSON, M., TARASSENKO, L. & MANT, D. 2010. Determining which automatic digital blood pressure device performs adequately: a systematic review. *J Hum Hypertens,* 24**,** 431-8.

WANG, Y.-C., BOHANNON, R. W., KAPELLUSCH, J., GARG, A. & GERSHON, R. C. 2015. Dexterity as measured with the 9-Hole Peg Test (9-HPT) across the age span. *Journal of Hand Therapy,* 28**,** 53-60.

WEIR, D. 2018. Biomarkers in Representative Population Surveys. *In:* VANNETTE, D. L. & KROSNICK, J. A. (eds.) *The Palgrave Handbook of Survey Research.* Cham, Switzerland: Springer International Publishing.

WELLCOME TRUST 2017. Wellcome Science Education Tracker, 2016. [data collection]. UK Data Service.

WELLS, T., BAILEY, J. T. & LINK, M. W. 2013. Filling the Void: Gaining a Better Understanding of Tablet-Based Surveys. *Survey Practice,* 6**,** 1-9.

WENZ, A., JÄCKLE, A. & COUPER, M. P. 2019. Willingness to use mobile technologies for data collection in a probability household panel. *Survey research methods,* 13**,** 1-22.

WENZ, A., JACKLE, A., CROSSLEY, T., COUPER, M. P., BURTON, J. & WINTER, J. 2018. Quality of Expenditure Data Collected with a Receipt Scanning App in a Probability Household Panel. *BigSurv18.* Barcelona, Spain.

WEST, B. T. 2011. Paradata in Survey Research. *Survey Practice,* 4.

WEST, B. T., ONG, A. R., CONRAD, F. G., SCHOBER, M. F., LARSEN, K. M. & HUPP, A. L. 2021. Interviewer Effects in Live Video and Prerecorded Video Interviewing. *Journal of Survey Statistics and Methodology,* 10**,** 317-336.

WILLIAMS, R. 2012. Using the margins command to estimate and interpret adjusted predictions and marginal effects. *Stata Journal,* 12**,** 308-331.

WILLS, A. K., LAWLOR, D. A., MATTHEWS, F. E., SAYER, A. A., BAKRA, E., BEN-SHLOMO, Y., BENZEVAL, M., BRUNNER, E., COOPER, R., KIVIMAKI, M., KUH, D., MUNIZ-TERRERA, G. & HARDY, R. 2011. Life course trajectories of systolic blood pressure using longitudinal data from eight UK cohorts. *PLoS Med,* 8**,** e1000440.

WOODHALL, S. C., SOLDAN, K., SONNENBERG, P., MERCER, C. H., CLIFTON, S., SAUNDERS, P., DA SILVA, F., ALEXANDER, S., WELLINGS, K., TANTON, C., FIELD, N., COPAS, A. J., ISON, C. A. & JOHNSON, A. M. 2016. Is chlamydia screening and testing in Britain reaching young adults at risk of infection? Findings from the third National Survey of Sexual Attitudes and Lifestyles (Natsal-3). *Sexually Transmitted Infections,* 92**,** 218-227.

WRIEDEN, W. L. & LEVY, L. B. 2016. 'Change4Life Smart Swaps': quasi-experimental evaluation of a natural experiment. *Public health nutrition,* 19**,** 2388-2392.

WRIGHT, A. 2016. REDCap: A tool for the electronic capture of research data. *Journal of Electronic Resources in Medical Libraries,* 13**,** 197-201.

YOSHIDA, K., HERNANDEZ-DIAZ, S., SOLOMON, D. H., JACKSON, J. W., GAGNE, J. J., GLYNN, R. J. & FRANKLIN, J. M. 2017. Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology,* 28**,** 387-395.

ZACHARIADIS, M. & OZCAN, P. 2016. The API Economy and Digital Transformation in Financial Services: The Case of Open Banking. *SSRN Electronic Journal*.

ZHANG, L., GUO, L., WU, H., GONG, X., LV, J. & YANG, Y. 2019. Role of physical performance measures for identifying functional disability among Chinese older adults: Data from the China Health and Retirement Longitudinal Study. *PLoS One,* 14**,** e0215693.