*Journal of Survey Statistics and Methodology (2024) **00**, 1–30*

MEASURING EXPENDITURE WITH A MOBILE APP: DO PROBABILITY-BASED AND NONPROBABILITY PANELS DIFFER?

ANNETTE JÄCKLE *
CARINA CORNESSE
ALEXANDER WENZ
MICK P. COUPER

In this case study, we examine a novel aspect of data collected in a typical probability and a typical nonprobability panel: mobile app data. The data were collected in Great Britain in 2018, using the Innovation Panel of the UK Household Longitudinal Study and the Lightspeed online access panel. Respondents in each panel were invited to participate in a month-long study, reporting all their daily expenditures in the app. In line with most of the research on nonprobability and probability-based panel data, our results indicate differences in the data gathered from these data sources. For example, more female, middle-aged, and highly educated people with higher digital skills and a greater interest in their finances participated in the nonprobability app study. Our findings also show that resulting differences in the app spending data are difficult to eliminate by weighting. The only data quality aspect for which we do not find evidence of differences between the nonprobability and probability-based panel is behavior in using the spending app. This

Annette Jäckle is Professor at the Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK. Carina Cornesse is Doctor at the German Institute for Economic Research & University of Bremen, Mohrenstr. 58, 10117 Berlin, Germany. Alexander Wenz is Doctor at the Mannheim Centre for European Social Research, University of Mannheim, A5, 6, 68131 Mannheim, Germany. Mick P. Couper is Professor at the Institute for Social Research, University of Michigan, P.O. Box 1248, Ann Arbor, MI 48106, USA.

This research was funded by grants from the UK Economic and Social Research Council (ESRC) and the National Centre for Research Methods (NCRM) ES/N006534/1 and by ESRC funding for the Understanding Society survey (ES/N00812X/1). This work was also supported by the Collaborative Research Center SFB 884 “Political Economy of Reforms” (projects A8 and Z1), funded by the German Research Foundation (DFG). This study design and analysis was not preregistered.

*Address correspondence to Annette Jäckle, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, United Kingdom. E-mail: aejack@essex.ac.uk.

https://doi.org/10.1093/jssam/smae026

finding is contrary to the argument that nonprobability online panel participants try to maximize their monetary incentive at the expense of data quality. However, this finding is in line with some of the scarce existing literature on response behavior in surveys, which is inconclusive regarding the question of whether nonprobability online panel participants answer questions less conscientiously than probability-based panel respondents. Since the two panels in our case study differ in more aspects than the sample selection procedure, more research in different contexts is necessary to establish generalizability and causality.

KEY WORDS: Mobile app data; Nonprobability sample; Online panel; Probability sample; Spending diary.

Statement of Significance

Previous studies have shown that, while nonprobability panels are much cheaper to maintain than probability-based panels, estimates are often less accurate, even when taking account of differences in socio-demographic panel sample composition. In this paper, we examine expenditure data collected with a mobile app over a period of one month in a probability-based and a nonprobability panel. We find differences between the app study samples in who participates in the mobile app study and in the expenditure captured with the app, even after accounting for differences in panel sample composition, but no differences in how participants used the app.

1. INTRODUCTION

Many researchers use nonprobability online panels to collect large amounts of survey data relatively quickly at low cost. However, nonprobability online panels rely on volunteers who often participate in multiple panels (Hillygus et al. 2014) and who are mainly motivated by the monetary compensation (Keusch et al. 2014). They are commonly criticized for their poor performance in accurately representing the general population (MacInnis et al. 2018). However, probability-based panels face data quality issues too, and some have argued that ever-decreasing survey response rates make probability-based panels indistinguishable from nonprobability online panels in terms of data quality (Wang et al. 2015; Gelman et al. 2016). Nevertheless, previous studies have shown that probability-based panels produce more accurate estimates than nonprobability online panels (see Cornesse et al. 2020 for an overview).

So far, research on comparing results from probability and nonprobability online panels has focused on questionnaire-based survey data. In this paper,

we compare these panels on a new dimension: we examine what happens when panel members are asked to use a mobile app to record their spending every day for a month. We use data from a diary study on financial spending that was implemented in parallel in the *Understanding Society* Innovation Panel, a probability-based mixed-mode panel of households in Great Britain, and the Lightspeed UK nonprobability online access panel. The two panels differ in more characteristics than just the sample selection mechanism. This includes differences in their incentive schemes, survey mode design, and data collection frequency. Our investigation is therefore a "system comparison," testing for differences in data collected with the same tool in a rather typical probability panel and a typical nonprobability panel.

In both panels, participants were asked to install an app on their mobile device and use it to report all spending over a period of one month. We compare the app study data collected in the two panels to answer the following research questions:

RQ1: Do the app study participants in the probability-based panel have different characteristics than those in the nonprobability online panel?

RQ2: Are there differences between the panels in how participants use the app?

RQ3: Are there differences between the panels in expenditure estimates, i.e., the study's main outcome of interest?

RQ4: Do the differences between the panels in expenditure estimates remain after weighting?

2. CONCEPTUAL FRAMEWORK: DIFFERENCES BETWEEN PROBABILITY-BASED AND NON-PROBABILITY PANELS

There are two types of "panels:" (1) "traditional" panels, where participants are surveyed on a core set of topics repeatedly, typically at less frequent intervals, to provide longitudinal data and (2) "access panels," where participants are surveyed on a variety of different topics, often at frequent intervals. Both types of panels can be recruited using probability or nonprobability sampling methods, or a combination of both (Callegaro et al. 2014). Examples of traditional probability panels include the US Panel Study of Income Dynamics, the UK Household Longitudinal Study, and the German Socio-Economic Panel. Nonprobability access panels are typically online panels run by commercial companies, such as Lightspeed, YouGov, Toluna, and OnePoll. In addition, there are probability-based online panels that combine frequent data collection on different topics with a traditional panel element collecting longitudinal

data. Examples include the Understanding America Study and AmeriSpeak in the United States, the German Internet Panel, and the Dutch LISS panel.

The researcher's choice of the sample recruitment procedure has consequences for the resulting participant samples and can, thereby, influence the estimates calculated on the basis of the gathered data (Mercer et al. 2017). To recruit a probability-based panel, researchers draw a random sample of units (individuals or households) from a sampling frame, such as an address list or population register. Sampled units are then approached with a request to participate in the panel. Because the probability sampling process is usually time consuming and expensive, great effort is often expended to gain contact with the sampled units (e.g., multiple contact attempts) and establish cooperation with the request to participate in the panel (e.g., by offering multiple modes of data collection). Surveys are often conducted for a single sponsor (e.g., a university), compensated with monetary incentives, and take place at regular intervals (e.g., bi-monthly or annually).

To recruit a nonprobability online panel, researchers usually disseminate open invitations, whether via online advertisement, by sending invitation emails via newsgroups and mailing lists, or by placing a panel invitation question at the end of a pop-up survey (see Callegaro et al. 2014 for an overview of nonprobability online panel recruitment methods). Recently, it has also become popular to recruit nonprobability panels based on interactive features implemented in online media articles or via social media (see Zindel 2022 for an overview). The likelihood and frequency of exposure to the open invitations depends on whether people have a chance of being exposed (e.g., whether they have an Internet connection, whether and how often they visit any of the websites that display a particular ad, or whether they are enrolled in any mailing list over which an invitation is disseminated). The likelihood that a person who is exposed to an open invitation becomes aware of this invitation depends on whether the invitation catches their eye (e.g., whether a banner ad uses bright colors or an invitation email header contains a promise of attractive incentives). The likelihood that a person who is aware of an open invitation to volunteer then joins the nonprobability online panel depends on the person's time constraints, topic interest, motivation, Internet access constraints, and skills (e.g., digital literacy). Once a pool of people has volunteered to participate in a nonprobability online panel, researchers select panel members for a particular study. Sometimes this is done using quota sampling to achieve some balance with regard to a limited number of characteristics, such as age, gender, and geographic region. Nonprobability online panel members are typically exposed to a large number of survey requests (often several times a month or more frequently) for different sponsors or clients and are compensated with small monetary or in-kind incentives (e.g., points) for each completed survey. In addition, nonprobability panels often target participants based on their characteristics for population subgroup studies (e.g., on parents with young children) rather than conducting general population surveys.

In addition to differences in selection and recruitment methods, probability-based and nonprobability panels differ on many other dimensions, including how much effort is put into retaining sample members (i.e., reducing panel attrition), the degree of panel maintenance (e.g., removing fraudulent responses or dropping inattentive respondents), the frequency, type, and size of survey requests, and the range of topics covered. These design differences may all lead to differences in sample composition, motivation, and interest of sample persons to participate in specific tasks or activities.

Overall, we expect the differences in the probability-based and nonprobability panel recruitment and retention processes to affect who is selected into the respective panels, which, in turn, might lead to differences in app study data gathered on these panels.

3. EXPECTATIONS AND EVIDENCE FROM PREVIOUS EMPIRICAL STUDIES

A number of studies have examined whether it is sufficient to use samples from nonprobability online panels rather than investing in probability-based methods of survey data collection. For this purpose, researchers typically implement a questionnaire with identical survey questions, answer options, and fieldwork periods, among other design features kept identical, in at least one probability-based survey and at least one nonprobability online panel. While the nonprobability samples in this literature are usually online panels, the probability-based studies are compared to have varying designs: some are panel studies while others are cross-sectional studies, which may influence the results since the latter cannot be affected by attrition or conditioning biases.

Most studies of this type focus on assessing how accurately probability-based and nonprobability surveys represent the intended target population. Some of the studies compare nonprobability online panels to probability-based face-to-face (Malhotra and Krosnick 2007; Loosveldt and Sonck 2008; Ansolabehere and Rivers 2013; Szolnoki and Hoffmann 2013; Dutwin and Buskirk 2017; Sturgis et al. 2018; Dassonneville et al. 2020) or telephone surveys (Szolnoki and Hoffmann 2013; Ansolabehere and Schaffner 2014; Gittelman et al. 2015; Pasek 2016; Dutwin and Buskirk 2017; Sohlberg et al. 2017; Legleye et al. 2018; Pennay et al. 2018). Other studies compare nonprobability online panels to probability-based online panels (Chan and Ambrose 2011; Steinmetz et al. 2014). A number of studies also try to disentangle potential mode and sampling effects by comparing samples from nonprobability online panels to probability-based offline surveys as well as probability-based online panels (Berrens et al. 2003; Chang and Krosnick 2009; Scherpenzeel and Bethlehem 2011; Yeager et al. 2011; Brüggen et al. 2016; Kennedy et al. 2016; MacInnis et al. 2018). Furthermore, some studies examine whether weighting adjustments improve the accuracy of

nonprobability online panels using a variety of weighting procedures, such as raking (Berrens et al. 2003; Chang and Krosnick 2009; Pasek 2016; Dutwin and Buskirk 2017; Sturgis et al. 2018), poststratification (Loosveldt and Sonck 2008; Yeager et al. 2011; Gittelman et al. 2015; MacInnis et al. 2018; Pennay et al. 2018), or propensity weighting (Berrens et al. 2003; Loosveldt and Sonck 2008; Steinmetz et al. 2014; Pasek 2016; Dutwin and Buskirk 2017; Sturgis et al. 2018). Little attention is usually paid to differences in study participation behavior between probability-based surveys and nonprobability online panels (for notable exceptions, see Chang and Krosnick 2009; Greszki et al. 2014, and Cornesse and Blom 2020).

While the existing literature compares probability-based surveys to nonprobability online panels, our study goes beyond survey data and instead focuses on an additional task that people are asked to do: installing and using a mobile app to report their spending over a month. In the following, we discuss relevant existing evidence from the survey literature and describe our expectations regarding the mobile app study.

RQ1: Do the app study participants in the probability-based panel have different characteristics than those in the nonprobability online panel?

Previous studies have focused on comparing the composition of probability-based surveys and nonprobability online panels by deriving aggregate indices, such as the absolute average bias (Kennedy et al. 2016), the largest absolute error (Yeager et al. 2011), or the root mean squared error (MacInnis et al. 2018). Nonprobability online panels have repeatedly been found to be more selective in that they deviate more from population benchmarks than probability-based surveys (Cornesse et al. 2020).

Given the focus on aggregate indices in the existing literature, it is difficult to identify common patterns at the variable level. However, it seems that in nonprobability online panels, there is often a stronger overrepresentation of people who are middle-aged (Malhotra and Krosnick 2007; Legleye et al. 2018; Dassonneville et al. 2020), female (Malhotra and Krosnick 2007; Legleye et al. 2018; Dassonneville et al. 2020), and highly educated (Malhotra and Krosnick 2007; Legleye et al. 2018; MacInnis et al. 2018; Dassonneville et al. 2020) compared to probability-based surveys. In our study, we, therefore, expect that app study participants from the nonprobability online panel are more likely to be middle-aged, female, and highly educated than those from the probability-based mixed-mode panel (H1.1).

Since nonprobability online panel recruitment usually relies on open invitations disseminated via the Internet and on volunteering rather than random selection, nonprobability online panels are also likely to systematically exclude or misrepresent some people beyond primary socio-demographic characteristics. For example, people without Internet access do not have a chance of being exposed to invitations and people who are frequently online have a particularly high chance of exposure to such invitations. In addition,

people who consider the costs of participation to be low (e.g., because they have high levels of digital skills) and the benefits of participation to be high (e.g., because the promised monetary or in-kind incentives appeal to them) are more likely to volunteer than people who consider the costs to be high and the benefits to be low. We, therefore, expect participants from the nonprobability panel to have higher levels of digital skills and greater interest in their finances than participants from the probability-based mixed-mode panel. Having higher levels of digital skills includes greater experience with and confidence in using digital devices (H1.2) as well as being more willing and less concerned to participate in additional smartphone-based tasks, such as taking photos or allowing GPS tracking (H1.3). Having greater interest in their finances includes being more involved in behaviors related to finances, such as budget keeping and bank balance checking (H1.4).

RQ2: Are there differences between the panels in how participants use the app?

Keusch et al. (2014) found that for 40 percent of the newly recruited members of a nonprobability online panel, the monetary incentives were an important motive for joining the panel. In addition, this monetary motivation was the strongest predictor of actual participation in panel survey waves as compared to other motives such as curiosity, entertainment, or novelty. Similarly, Sparrow (2006) found that 52 percent of newly recruited nonprobability online panel members stated that they participated in the panel because of the monetary incentive provided. Other reasons for participating, such as survey enjoyment and interest in the survey topic, were selected significantly less often. While the studies by Keusch et al. (2014) and Sparrow (2006) suggest that nonprobability online panel participants are mainly motivated by the promise of monetary incentives, the same does not seem to apply to probability-based panel participants. For example, in the LISS panel, only a minority (15.2 percent) of panel participants said that their most important motive for participating in the study was the financial rewards, while most participants said that their main reason for participating was either that they think it is important to contribute to science (16.4 percent), contribute to society (13.6 percent), or help the researchers (13.0 percent; numbers based on own calculations, data retrieved from www.lissdata.nl).

Research on the consequences of the incentive-oriented motivation of nonprobability online panel members for survey data quality is scarce and results are mixed. On the one hand, Cornesse and Blom (2020) found that straightlining in grid questions was significantly more likely in seven nonprobability online panels than in three probability-based online panels. On the other hand, Chang and Krosnick (2009) found the opposite when comparing a nonprobability online panel with a probability-based online panel. Similarly, Greszki et al. (2014) found that a nonprobability online panel performed worse in terms of survey response speed than a probability-based online panel, while

Chang and Krosnick (2009) found that a nonprobability online panel performed better in terms of random measurement error across multiple measures of the same construct than a probability-based telephone survey. Finally, Cornesse and Blom (2020) did not find any generalizable difference between nonprobability online panels and probability-based online panels with regard to survey item nonresponse or midpoint selection.

In line with the general finding in the literature that many nonprobability online panel members are motivated to volunteer to join the panel primarily for monetary reasons, we expect that app study participants from the nonprobability sample are more likely to apply strategies to maximize their monetary compensation (H2.1) and to minimize their effort (H2.2) than app study participants from the probability-based sample.

RQ3: Are there differences between the panels in expenditure estimates, i.e., the study's main outcome of interest?

Generally, many studies find that differences between probability-based and nonprobability panels matter for a diverse set of substantive outcomes, such as voting behavior (Malhotra and Krosnick 2007; Chang and Krosnick 2009; Sturgis et al. 2018), health behavior (Yeager et al. 2011), consumption behavior (Szolnoki and Hoffmann 2013), sexual behavior and attitudes (Erens et al. 2014; Legleye et al. 2018), and political attitudes (Malhotra and Krosnick 2007; Loosveldt and Sonck 2008).

In our study, we similarly expect differences in participant characteristics and study participation behavior to result in differences between samples in key outcomes. In particular, if women, middle-aged people, and highly educated people are more likely to participate in a nonprobability online panel than a probability mixed-mode panel (H1.1), this should result in differences in the distribution of expenditure across spending categories.

In addition, if app study participants from a nonprobability online panel maximize their incentive-to-effort-ratio by reporting fewer spending events than those from probability-based online panels (H2.1, H2.2), this should result in lower expenditure estimates. Compared to estimates from a probability-based panel, we therefore expect that estimates from the nonprobability panel suggest lower average expenditures, for example on eating out and other leisure activities (H3).

RQ4: Do the differences between the panels in expenditure estimates remain after weighting?

Most previous studies conclude that significant differences between probability and nonprobability samples remain after weighting (Schonlau et al. 2004; Duffy et al. 2005; Schonlau et al. 2007; Schonlau et al. 2009; Mercer et al. 2017; Smyk et al. 2021). Furthermore, some studies conclude that differences remain even after adding non-demographic characteristics to the weighting schemes (Lee 2006; Dutwin and Buskirk 2017; Mercer et al. 2017).

In our study we expect that if the probability and nonprobability samples differ in participant characteristics, study participation behavior, and substantive outcomes, it is unlikely that these differences all vanish after applying common weighting procedures based on socio-demographic characteristics. Including additional variables is likely to further reduce the differences between probability and nonprobability samples if the added variables are related to the nonprobability sample recruitment process and/or the study data of interest. For example, if nonprobability sample members use the Internet more frequently and with greater confidence than probability sample members, adding variables related to digital skills might help reduce differences in Internet usage behavior. We therefore expect that differences between probability and nonprobability samples will decrease when weighting for socio-demographic characteristics (H4.1), decrease further when adding financial behaviors (H4.2), and decrease further still when adding digital affinity measures related to the app use task (H4.3).

4. DATA

Spending Study 2 (University of Essex, Institute for Social and Economic Research 2022) was implemented in May to December 2018, using two different samples in Great Britain: a probability panel (the *Understanding Society* Innovation Panel) and a nonprobability panel (the Lightspeed UK online access panel). The samples are described below. This was a follow-up to an earlier study (Spending Study 1), carried out in 2016. The first study was only implemented in the probability-based Innovation Panel and, therefore, is not included in the present analyses. See Jäckle et al. (2018) for details of the first study.

Participants were asked to download a mobile app and use it for one month to report their spending. The same app was used in both samples and was compatible with iOS smartphones, Android smartphones, and tablets (see Supplementary Appendix 5 for screenshots of the app). The design and functionality of the app were based on findings from qualitative interviews with members of the general public about how the app could best support participants in reporting their daily expenditure (Suffield et al. 2018). Participants were asked to use the app to record all direct debits and standing orders that would automatically come out of their bank accounts during the month. In addition, they were asked to use the app every day to report all purchases, by selecting a purchase category and then entering the value of the purchase. On days on which they did not spend any money, they were asked to report "no purchases today" in the app. Sample members who did not use the app were invited to use a browser-based version instead. However, as take up of the browser-based version was very low in the Innovation Panel, this paper focuses on participants in both samples who used the app.

Participants were told that they could earn incentives for every day on which they used the app at least once (including to report that they had not made any purchases that day), a bonus if they used the app every day, and incentives conditional on completing the direct debit/standing order section and a debrief questionnaire at the end of the study. In the Innovation Panel, the daily incentive was £0.50, the bonus for completing the month was £10, the incentives for the direct debit section £1, and for the debrief questionnaire £3. In total, participants could earn up to £29.50. For the access panel, Lightspeed UK administered the incentives according to their standard rewards policy: panelists could earn a maximum of 500 points (equivalent to about £5) and could exchange their incentives for vouchers or charity donations. The wording of the invitations to the app study for both panels is included in Supplementary Appendix 1.

4.1 The Understanding Society Innovation Panel

The Innovation Panel is a stratified and clustered sample of households in Great Britain. It is part of *Understanding Society*: The UK Household Longitudinal Study and used for methodological testing and experimentation (University of Essex, Institute for Social and Economic Research 2021). The design of the Innovation Panel mirrors that of the main *Understanding Society* panel, with annual interviews of all household members aged 16+ years. The Innovation Panel wave 11 interview (IP11) fielded in May to September 2018 was used as the baseline survey for Spending Study 2: a random half of the respondents were invited to the Spending Study 2 within the IP11 questionnaire; the other half was sent an invitation by post, a couple of weeks after completing their IP11 interview. Although the within-interview invitation increased participation in Spending Study 2 compared to the postal invitation, there was no difference in the composition of the participant samples between the two treatment groups. The invitation treatment groups are, therefore, combined for the purposes of the analyses presented here.

IP11 was a mixed-mode survey. About two-thirds of sample households were allocated to web first: all adult household members received an invitation to complete their interview online. If they did not do so after several reminders, they were followed-up by a face-to-face interviewer. The remaining sample households were allocated to face-to-face interviewers, and nonrespondents were given the opportunity to complete the survey online in the final stages of fieldwork. Overall, 55 percent of respondents completed the survey with an interviewer and 45 percent completed it online. The Innovation Panel wave 11 household response rate was 73.2 percent, with 80.5 percent of eligible adults within those households completing individual interviews (AAPOR RR5, The American Association for Public Opinion Research 2023). For more details on IP11 fieldwork, see the Innovation Panel User

Guide at https://www.understandingsociety.ac.uk/documentation/innovation-panel/user-guide.

4.2 The Lightspeed UK Online Access Panel

Members of the Lightspeed UK online access panel were invited to complete a questionnaire that collected the same baseline information for Spending Study 2 as the Innovation Panel wave 11 questionnaire. This included socio-demographic characteristics, financial behaviors and position, mobile device access and usage, hypothetical willingness to do different types of tasks for a survey, and data security concerns. The fieldwork agency monitored quotas on age and gender for the baseline survey. At the end of this questionnaire, respondents were invited to participate in Spending Study 2 and to download the app. The Lightspeed implementation included a feedback experiment, whereby a random one-third of the sample were either told they would be able to see feedback about their spending within the app, were not told but able to view the feedback, or did not receive feedback. The feedback treatment had no effect on participation or reported spending, and therefore, the treatment groups are combined for the purposes of the analyses presented here.

4.3 Analysis Sample

For RQ1 (characteristics of app users from the two panels), we first compare the respondents to the two baseline surveys, that is, those invited to the app study. All other analyses are restricted to participants who used the app at least once to report a purchase, which is defined as providing a non-missing purchase amount in the "report daily purchases" section of the app. We use all entries participants made in the app within 31 days of first using the app.

In the Innovation Panel, 2,638 sample members gave a full interview and were eligible for Spending Study 2. Of these, 446 (16.9 percent, AAPOR RR6) used the app at least once to report a purchase, reporting a total of 12,579 purchases. In the access panel, 2,878 sample members completed the baseline survey and of these, 408 (14.2 percent, AAPOR RR6) used the app at least once, reporting a total of 11,517 purchases.

4.4 Respondent Characteristics

The respondent characteristics used to examine differences in sample composition are derived from the baseline questionnaires for both samples (the Innovation Panel wave 11 questionnaire can be accessed at https://www.understandingsociety.ac.uk/documentation/innovation-panel/questionnaires. The access panel baseline questionnaire is documented in Jäckle et al. 2019). The characteristics include standard socio-demographic variables,

financial behaviors, and measures of digital affinity. For most variables, the questions were asked in the same way in both samples. Among app users, the rate of missing items was between 0 percent and 1.8 percent in the Innovation Panel data and between 0 percent and 0.2 percent in the access panel data. Missing observations were set to the modal answer categories for the sample. As a robustness check, we replicated the analyses of app users in tables 1, 2, 3, 4, and 5 using complete cases only, dropping 5.6 percent of the Innovation Panel sample and 0.7 percent of the access panel sample. The patterns of differences between the samples are comparable. Across all tests, there are two that reach significance at the $p < .05$ level when the data with imputed values are used, but not in the complete case analyses; there are five tests that reach significance when the complete cases are used. As the estimates using imputed values are slightly more conservative, we present those here.

The socio-demographic characteristics included are gender (male, female), age (16–35, 36–55, 56+), highest educational qualification (degree, A/AS level [∼13 years of schooling], GCSE/CE level [∼11 years of schooling], no formal qualification), whether the respondent is in work (employed or self-employed in the prior week), whether living as a couple (yes, no), and the number of children aged under 16 living in the household (0, 1, 2+).

The financial behaviors include whether the respondent keeps a budget (yes, no), how often they check their bank balance (most days, at least once a week, less frequently), and whether they check their balance using an app on a mobile device (yes, no). The wording of these questions is documented in Supplementary Appendix 1.

The measures of digital affinity include whether the participant uses the following devices to connect to the Internet: PC or laptop (yes, no), smartphone (yes, no), tablet (yes, no); how frequently they use their smartphone (every day, less frequently); the number of different types of activities they do on their phone (summed count of 13 different activities); and self-rated smartphone skills (five-point scale coded as beginner, intermediate, advanced). The measures also include the average willingness to do different types of activities with their mobile device for a survey (four-point scales asking about willingness to do each of eight activities), and the average concerns about the security of providing data in these ways (five-point scales). In order to generate categorical indicators of willingness and data security concerns, the mean scores are rounded to the nearest integers and labelled according to the original response categories. Except for the questions about how the respondent connects to the Internet, the questions were routed on using a smartphone to connect to the Internet. The wording of these questions is documented in Supplementary Appendix 1. For respondents who did not use a smartphone (six in the access panel sample and twenty-seven in the Innovation Panel sample), the frequency of smartphone usage and number of

activities are set to zero. The smartphone skills, willingness, and data security concerns are set to the lowest skill/willingness categories and to the highest concern category.

4.5 App Usage Behaviors

The indicators of app usage behavior are derived from the app paradata in the same way for both samples. The indicators are coded into categories and summarize the following for each participant:

- The number of times the participant clicked on the landing page of the app (coded into terciles).
- The number of days on which the participant used the app at least once to report a direct debit or standing order, to report a purchase, or to report a "no spend day" (coded into terciles).
- The number of direct debits or standing orders reported (coded as 0, below the median, above the median).
- The total number of daily purchases reported (coded into terciles).
- The average number of daily purchases per day on which the participant used the app (coded as <1, 1–2, >2).
- The total number of small purchases reported, defined as costing less than £3 (coded as 0, below the median, above the median).

4.6 Measures of Spending

We examine three aspects of the spending reported by participants in the app: the total value of direct debits and standing orders, the value of daily purchases by category, and the total value of direct debits. The sixteen categories to record daily purchases (based on work by d'Ardenne and Blake 2012) were food and groceries; eating and drinking out; clothes and footwear; transport and car; child costs; home improvements and household goods; health expenses; socializing and hobbies; books, magazines, films and music; games and toys; haircuts, manicures and massages; holidays; gifts and donations; rent (not direct debit/standing order); bills (not direct debit/standing order); and other purchases or payments.

5. METHODS

We test for differences between the two samples using χ^2 tests, which account for the clustered and stratified sample design of the Innovation Panel (see Supplementary Appendix 6 for the PRICSSA checklist). In these analyses, we first compare the two samples to answer RQ1–RQ3. For RQ1 we, in

addition, estimate propensity models of the probability that respondents who completed the baseline surveys participated in the app study. We calculate Average Marginal Effects from separate logit models for the two panels, regressing whether the respondent used the app on the respondent characteristics used to examine RQ1. Then, we examine whether differences in the measurement of spending remain after controlling for sample composition differences and/or app usage behavior using propensity weights (RQ4). Since some respondents of the probability sample survey were invited to the app study during a face-to-face interview instead of a web survey, we repeated all analyses using only the probability sample app participants who were invited during the web survey as a sensitivity analysis to try to account for the fact that the probability-based panel has a mixed-mode design, whereas the nonprobability panel only uses the online mode of survey data collection (see Supplementary Appendix 3 tables 2–6). Because the results with and without the app study participants invited during the face-to-face interview are broadly comparable, we only show the results for the full samples in the results section of this paper.

To examine the effectiveness of adjusting for differences in sample composition, we use propensity score weighting to match the sample composition of the nonprobability sample to that of the probability sample. We follow the approach of McCaffrey et al. (2004). To compute the weights, we combine the two samples into one dataset and estimate probit models of the probability that a participant is in the probability sample rather than the nonprobability sample. Based on these models, we compute the predicted probability for each participant i of being in the probability sample, $p(x_i)$. The propensity score weights are then computed as $w_i = p(x_i)/[1 - p(x_i)]$ for the nonprobability sample and $w_i = p(x_i)/p(x_i) = 1$ for the probability sample. The denominators of the weights adjust for the differences in characteristics between the two samples, while the numerator weights the pooled sample to match the characteristics of the probability sample: Respondents in the nonprobability sample who have characteristics that are not typical in the probability sample will have a $p(x_i)$ close to zero and therefore a weight close to zero. We estimate three separate probit models and use these to calculate three sets of weights (Supplementary Appendix 2 table 1): weight 1 adjusts only for socio-demographic characteristics, weight 2 in addition includes financial behaviors, and weight 3 in addition includes digital affinity. The sequential addition of variables, starting with socio-demographic characteristics, follows standard approaches used in comparisons of data from probability and nonprobability samples (e.g., Kennedy et al. 2016; Kocar and Baffour 2023). The area under the curve (AUC) statistics reported in Supplementary Appendix 2 table 1 indicate that the fit of the weighting model improves with the addition of each set of covariates: the AUC for weighting model 1 is 0.6357, for model 2 it is 0.7429, and for model 3 it is 0.8071.

6. RESULTS

In the following, we present our results of comparing the app study samples from the probability-based mixed-mode panel and the nonprobability online panel following our research questions RQ1–RQ4. We first focus on the unweighted estimates in tables 1, 2, and 3 to address RQ1–RQ3. To address RQ1, tables 1, 2, and 3 contain comparisons of the baseline survey respondents (i.e., everyone invited to the app study) in addition to comparisons of the app users. Table 6 provides an overview of the research questions, hypotheses, and whether the hypotheses are supported by the results.

RQ1: Do the app study participants in the probability-based panel have different characteristics than those in the nonprobability online panel?

Table 1 shows the socio-demographic characteristics that the participants from the two panels reported in the baseline questionnaire. Focusing first on the app users (last four columns), we find that the nonprobability panel participants are more likely to be female, middle-aged, highly educated, and living in a household with at least one child than the probability-based panel participants, supporting H1.1. The only examined socio-demographic variables on which we did not find any significant differences between app study participants are employment status and partnership status. The average absolute difference in socio-demographic characteristics between the panels is 5.4 percentage points. When comparing the baseline survey respondents, these differences are all already present, and in fact larger, with an average absolute difference of 7.8. The exceptions are age (with more respondents in the youngest group in the nonprobability panel) and more people in work in the nonprobability panel. Supplementary Appendix 4 documents propensity models for the two panels, using the baseline survey respondents to predict the probability of participating in the app study. In the nonprobability panel, women and those in the middle age group were more likely to use the app (+3.3 percentage points, $p < .05$ and +5.9 percentage points compared to the youngest age group, $p < .001$); in the probability panel, those in households with two or more children were less likely to participate in the app study compared to those with no children (–4.5 percentage points, $p < .05$). There are also significant differences in the reported financial behaviors of the app study participants (table 2 last four columns). The nonprobability panel participants are much more likely to keep a budget (75.7 percent versus 41.9 percent), check their bank balance on most days, and check their bank balance using a mobile app than the probability-based panel participants, supporting H1.4. The average absolute difference in financial behaviors between the panels is 13.7 percentage points. Comparing the baseline survey respondents, these differences are again already there and larger, with an average absolute difference of 15.2. In the propensity model (Supplementary Appendix 4), those who keep a budget and those who check their balance infrequently were less likely to use

Table 1. Differences in Socio-Demographic Characteristics Between the Probability and Nonprobability Samples

	Baseline survey respondents				App users			
	Probability sample	Nonprobability sample	Delta	*p*	Probability sample	Nonprobability sample	Delta	*p*
Female	55.1	69.4	14.3	<.001	57.8	72.5	14.7	<.001
Age								
16–35	22.7	34.3	11.6		35.9	35.5	–0.3	
36–55	33.3	38.6	5.3		41.9	48.5	6.6	
56+	44.0	27.1	–16.9	<.001	22.2	15.9	–6.3	.023
Education								
Degree	39.3	41.9	2.6		45.3	45.6	0.3	
A/AS level (∼13 years of schooling)	13.4	25.7	12.3		17.0	28.4	11.4	
GCSE/CSE level (∼11 years of schooling)	30.0	27.6	–2.4		29.8	22.8	–7.0	
No formal qualification	17.3	4.8	–12.5	<.001	7.8	3.2	–4.7	<.001
In work	55.2	66.1	10.9	<.001	72.4	73.8	1.4	.485
In couple	61.5	59.3	–2.2	.099	61.9	60.5	–1.3	.587
No. kids in household under 16								
None	74.5	69.5	–5.0		67.9	60.0	–7.9	
One	12.7	15.0	2.4		15.9	18.6	2.7	
Two or more	12.9	15.5	2.6	.003	16.1	21.3	5.2	.040
Number of observations	2,638	2,878			446	408		
Average absolute difference			7.8				5.4	

NOTES.—$p = p$-values from χ^2 tests for differences in distributions between the two samples.

Table 2. Differences in Financial Behaviors Between the Probability and Nonprobability Samples

	Baseline survey respondents				App users			
	Probability sample	Nonprobability sample	Delta	*p*	Probability sample	Nonprobability sample	Delta	*p*
Keeps a budget	38.9	79.4	40.5	<.001	41.9	75.7	33.8	<.001
Checks bank balance								
Most days	20.7	32.8	12.1		31.8	44.9	13.0	
At least once a week	40.2	40.5	0.3		41.9	40.2	–1.7	
Less frequently	39.1	26.7	–12.4	<.001	26.2	15.0	–11.3	<.001
Checks balance using mobile app	23.5	33.9	10.4	<.001	45.7	54.2	8.4	.001
Number of observations	2,638	2,878			446	408		
Average absolute difference			15.2				13.6	

NOTES.—$p = p$-values from χ^2 tests for differences in distributions between the two samples.

Table 3. Differences in Digital Affinity Between the Probability and Nonprobability Samples

	Baseline survey respondents				App users			
	Probability sample	Nonprobability sample	Delta	*p*	Probability sample	Nonprobability sample	Delta	*p*
Uses desktop/laptop for internet	78.0	92.3	14.3	<.001	84.5	96.3	11.8	<.001
Uses smartphone for internet	77.5	84.5	7.1	<.001	95.5	98.5	3.0	<.001
Uses tablet for internet	62.8	65.2	2.4	.018	72.2	70.8	–1.4	.540
Uses smartphone every day	60.4	68.2	7.7	<.001	82.3	86.5	4.2	.030
No. activities on smartphone								
0–9	65.5	60.0	–5.6		42.6	40.4	–2.2	
10–11	20.6	21.5	0.9		31.6	31.9	0.2	
12	13.8	18.6	4.7	<.001	25.8	27.7	1.9	.628
Smartphone skills								
Beginner	37.2	23.9	–13.4		10.8	5.9	–4.9	
Intermediate	21.0	25.5	4.4		20.2	25.2	5.1	
Advanced	41.7	50.7	8.9	<.001	69.1	68.9	–0.2	.002
Willingness tasks on smartphone								
Very	5.8	14.7	8.9		17.0	35.5	18.5	
Somewhat	23.0	28.0	5.0		39.9	48.0	8.1	
A little/not at all	71.2	57.3	–13.9	<.001	43.0	16.4	–26.6	<.001
Data security concerns								
Not at all	6.9	9.8	2.9		18.2	30.4	12.2	
A little	26.8	26.6	–0.3		45.3	41.9	–3.4	
Somewhat/extremely	66.3	63.6	–2.6	<.001	36.5	27.7	–8.9	<.001
Number of observations	2,638	2,878			446	408		
Average absolute bias			6.4				7.0	

NOTES.—$p = p$-values from χ^2 tests for differences in distributions between the two samples.

the app in the non-probability panel. In both panels, respondents who check their bank balance using a mobile app were more likely to use the app.

Table 3 (last four columns) shows the digital affinity characteristics that the app study participants from the two sources reported in the baseline questionnaire. We find that the nonprobability panel participants are more likely to use desktop computers, laptops, and smartphones to connect to the Internet, use a smartphone every day, and have intermediate smartphone skills. However, there are no significant differences in the use of tablets to connect to the Internet and the number of activities that participants do on their smartphones. Overall, this finding partly supports H1.2. Furthermore, we find that participants in the nonprobability panel are more likely to be willing to do additional tasks on their smartphones, and be unconcerned about the security of mobile data collection than the probability-based panel participants, supporting H1.3. The average absolute difference in digital affinity between the app study samples is 7.0 percentage points. Comparing respondents to the baseline surveys, the differences between the two panels are again already there and larger, with two exceptions. Regarding willingness to do tasks for a survey on their smartphone and data security concerns, the differences between app users from the two panels are larger than the differences between the baseline survey respondents. The propensity models in Supplementary Appendix 4 again shows some differences between the panels. In the nonprobability panel, those who use a desktop/laptop to access the internet were more likely to participate in the app study. In the probability panel, those who use a tablet to access the internet, those who do not use their smartphone daily, and those with intermediate or advanced smartphone skills were more likely to use the app. In both panels, those with lower willingness to do tasks on a smartphone and those with higher levels of concern about data security were less likely to use the app.

RQ2: Are there differences between the panels in how participants use the app?

Table 4 shows the app usage behaviors of the probability-based and nonprobability panel participants. We find that the nonprobability panel participants click on the landing page of the app less frequently, are more likely to report eight or more direct debits or standing orders and less likely to report 18–35 purchases than the probability-based panel participants. However, we did not find any significant differences in the number of days participants used the app, mean number of purchases reported per day, and the number of reported small purchases. Based on these findings, H2.1 and H2.2, that the nonprobability panel participants are more likely to maximize their monetary compensation and minimize their effort than the probability-based panel participants, are not supported. The average absolute difference in app usage characteristics between the app study samples is 3.4 percentage points.

Table 4. Differences in App Use Behaviors Between the Probability and Nonprobability Samples

	Probability sample %	Nonprobability sample %	Delta	*p*
No. times clicked landing page				
1–35	31.6	37.3	5.6	
36–55	32.5	32.6	0.1	
56–218	35.9	30.1	–5.7	.029
No. days used app				
1–14	32.3	37.0	4.7	
15–25	37.0	33.3	–3.7	
26+	30.7	29.7	–1.1	.195
No. direct debits/standing orders				
0	26.9	21.6	–5.3	
1–7	39.2	36.3	–3.0	
8–21	33.9	42.2	8.3	.005
No. purchases				
1–17	31.4	36.3	4.9	
18–35	37.0	29.4	–7.6	
36+	31.6	34.3	2.7	.006
Mean no. purchases/day				
<1	20.2	19.1	–1.1	
1–2	57.6	56.1	–1.5	
3+	22.2	24.8	2.6	.466
No. small purchases (<£3)				
0	28.0	29.9	1.9	
1–3	37.0	35.3	–1.7	
4+	35.0	34.8	–0.2	.665
Average absolute difference			3.4	

NOTES.—$N = 408$ in the nonprobability sample, $N = 446$ in the probability sample. $p = p$-values from χ^2 tests for differences in distributions between the two samples.

RQ3: Are there differences between the panels in expenditure estimates, ie, the study's main outcome of interest?

Table 5 shows the expenditures that the probability-based and nonprobability panel participants reported in the app. Not all respondents reported expenditures in all categories. Therefore, the upper panel in table 5 examines the percentage of respondents who did not report any expenditure in a given category. The results indicate that the nonprobability panel members were more likely to report zero expenditures. For example, 11.2 percent of respondents in the probability panel reported zero expenditure on eating and drinking out, compared to 18.1 percent of the nonprobability panel members (a difference

Table 5. Differences in Expenditure Measured in the App Between the Probability and Nonprobability Samples

	Unweighted			Weight 1		Weight 2		Weight 3	
	PrS	Delta	*p*	Delta	*p*	Delta	*p*	Delta	*p*
% of respondents reporting zero expenditure									
Food and groceries	5.8	–0.2	.871	–1.2	.256	–1.1	.318	–1.2	.260
Eating and drinking out	11.2	**6.9**	**.001**	**6.6**	**.001**	**6.2**	**.002**	3.4	.056
Clothes and footwear	40.6	1.1	.667	0.7	.777	–0.8	.753	–2.3	.351
Transport and car	25.8	**5.1**	**.036**	2.8	.233	**6.1**	**.013**	**6.0**	**.015**
Child costs	85.7	–0.6	.739	2.5	.127	–0.9	.609	0.9	.615
Home improvements and household goods	46.9	**5.6**	**.033**	3.1	.227	3.3	.204	2.8	.279
Health expenses	70.9	–2.5	.281	–4.6	.050	–3.4	.144	**–8.6**	**.001**
Socializing and hobbies	45.7	**17.0**	**<.001**	**15.8**	**<.001**	**18.2**	**<.001**	**11.4**	**<.001**
Books, magazines, films, and music	64.6	3.1	.220	0.7	.769	1.2	.641	–2.7	.294
Games and toys	78.5	–3.5	.082	–2.9	.138	**–5.3**	**.011**	**–5.8**	**.006**
Haircuts, manicures, massages	60.3	**11.5**	**<.001**	**11.4**	**<.001**	**9.0**	**<.001**	**4.5**	**.044**
Holidays	79.8	**8.7**	**<.001**	**7.1**	**<.001**	**7.9**	**<.001**	**3.6**	**.021**
Gifts and donations	52.9	**7.9**	**.003**	**10.2**	**<.001**	**9.2**	**.001**	4.5	.079
Rent (not direct debit/standing order)	93.3	–1.1	.454	–2.6	.114	–3.3	.052	–2.0	.201
Bills (not direct debit/standing order)	80.9	**–11.1**	**<.001**	**–14.1**	**<.001**	**–12.6**	**<.001**	**–12.5**	**<.001**
Other	42.6	–4.1	.073	**–5.4**	**.020**	**–8.3**	**<.001**	**–7.4**	**.002**
Total value of direct debits	26.9	**–5.3**	**.018**	**–7.7**	**.001**	**–7.0**	**.002**	**–7.2**	**.001**

Continued

Table 5. Differences in Expenditure Measured in the App Between the Probability and Nonprobability Samples (*Continued*)

	Unweighted			Weight 1		Weight 2		Weight 3	
	PrS	Delta	*p*	Delta	*p*	Delta	*p*	Delta	*p*
Mean (excluding zero expenditure reports)									
Total value of purchases	1041.5	–161.5	.065	–79.1	.359	–156.2	.074	**–199.5**	**.024**
Food and groceries	262.7	–27.4	.418	–17.4	.605	–19.1	.572	–41.2	.225
Eating and drinking out	111.6	**33.3**	**.001**	**57.1**	**<.001**	**21.6**	**.023**	17.4	.064
Clothes and footwear	83.2	**54.7**	**<.001**	**32.6**	**<.001**	**21.6**	**<.001**	4.1	.471
Transport and car	210.2	–52.9	.371	–49.9	.398	–70.8	.233	–40.1	.497
Child costs	76.1	–6.7	.665	–3.5	.819	–21.4	.181	–25.9	.112
Home improvements and household goods	313.0	–151.1	.171	–105.5	.335	–168.1	.129	–211.6	.059
Health expenses	56.3	8.8	.413	**27.5**	**.018**	8.4	.436	–11.0	.311
Socializing and hobbies	69.2	–13.2	.081	–12.6	.095	**–18.4**	**.017**	**–17.4**	**.024**
Books, magazines, films, and music	21.9	0.6	.811	2.5	.310	1.7	.497	–2.0	.406
Games and toys	32.8	10.6	.080	**19.3**	**.004**	**17.9**	**.007**	**16.6**	**.010**
Haircuts, manicures, massages	43.3	–2.0	.834	–5.4	.571	–7.0	.463	–11.6	.230
Holidays	340.5	41.9	.460	10.5	.852	15.1	.789	–16.2	.773
Gifts and donations	53.9	7.5	.362	8.6	.294	4.6	.571	13.7	.098
Rent (not direct debit/standing order)	293.2	40.1	.442	29.2	.560	51.3	.347	8.8	.854
Bills (not direct debit/standing order)	184.5	–42.1	.333	–30.0	.487	15.3	.721	–18.4	.669
Other	135.0	–46.4	.205	–34.6	.342	–30.2	.406	–42.2	.248
Total value of direct debits	753.1	–35.4	.550	–54.1	.362	–36.5	.537	–20.5	.729

Continued

Table 5. Differences in Expenditure Measured in the App Between the Probability and Nonprobability Samples (*Continued*)

	Unweighted			Weight 1		Weight 2		Weight 3	
	PrS	Delta	*p*	Delta	*p*	Delta	*p*	Delta	*p*
Mean (including zero expenditure reports)									
Food and groceries	247.3	–25.3	.414	–13.4	.665	–15.3	.621	–36.1	.246
Eating and drinking out	99.1	19.5	.029	**39.5**	**<.001**	10.9	.217	11.0	.210
Clothes and footwear	49.4	**31.0**	**<.001**	**18.5**	**<.001**	**13.6**	**.001**	4.5	.256
Transport and car	156.0	–47.3	.292	–41.5	.354	–61.1	.175	–40.0	.372
Child costs	10.9	–0.5	.820	–2.3	.325	–2.6	.281	–4.2	.083
Home improvements and household goods	166.3	–89.3	.128	–62.6	.284	–94.1	.110	–115.3	.051
Health expenses	16.4	4.2	.184	**11.9**	**<.001**	4.6	.141	0.7	.830
Socializing and hobbies	37.5	**–16.7**	**.001**	**–15.8**	**.001**	**–19.2**	**<.001**	**–15.3**	**.002**
Books, magazines, films, and music	7.7	–0.5	.615	0.7	.468	0.3	.748	–0.2	.853
Games and toys	7.1	**3.8**	**.017**	**5.7**	**.001**	**6.5**	**<.001**	**6.4**	**<.001**
Haircuts, manicures, massages	17.2	–5.5	.171	–6.5	.111	–6.1	.135	–6.0	.137
Holidays	68.7	–24.7	.055	–23.0	.073	–24.9	.052	–15.1	.234
Gifts and donations	25.4	–1.3	.747	–2.3	.572	–3.2	.432	3.4	.404
Rent (not direct debit/standing order)	19.7	6.4	.190	**10.3**	**.039**	**14.8**	**.003**	6.7	.172
Bills (not direct debit/standing order)	35.2	7.8	.375	16.1	.070	**28.0**	**.002**	17.2	.053
Other	77.5	–23.0	.291	–14.5	.504	–8.7	.689	–17.3	.423
Total value of direct debits	550.5	12.4	.782	14.1	.753	23.7	.598	38.0	.398

NOTES.—$N = 408$ in the nonprobability sample, $N = 446$ in the probability sample. Weight 1 is based on socio-demographic variables only, weight 2 includes financial behaviors, and weight 3 includes digital affinity variables. Delta = percentage point difference between the nonprobability sample (not shown) and the probability sample estimate in column 1. In the upper panel, percent of respondents reporting £0 expenditure in the given category, $p = p$-values from χ^2 tests for differences in distributions between the two samples. In the second and third panel, $p = p$-value from tests of differences in means between the two samples. Numbers in bold are biases with p-value $< .05$.

of 6.9 percentage points, $p = .001$). There are six other categories where the nonprobability panel members were significantly more likely than the probability panel to report zero expenditure (transport and car; home improvements and household goods; socializing and hobbies; haircuts, manicures and massages; holidays; and gifts and donations). The nonprobability panel members were, however, less likely than the probability panel to report zero expenditure for bills and direct debits.

The middle panel in table 5 examines the means of reported (non-zero) expenditures. The unweighted results show significant differences in only two spending categories: nonprobability panel members reported higher mean spending on eating and drinking out (£111.6 in the probability sample, £144.8 in the nonprobability sample, a difference of £33.3, $p = .001$) and higher mean spending on clothes and footwear (£83.2 in the probability sample, £138.0 in the nonprobability sample, a difference of £54.7, $p < .001$). That is, although there are differences in whether or not respondents report spending in a given category between the samples, the estimates of mean expenditure are similar in all but two categories. The mean total expenditure, calculated by summing up all expenditures across categories, is also similar in the two samples.

The lower panel reports tests for differences when cases with zero reported expenditure in a category are included in the analyses. The results are broadly similar. The same patterns are observed for the expenditure categories where there are no significant differences when the zeros are excluded, with the exception of rent and bills where some differences appear when the zeros are included. The same patterns are also observed for clothes and footwear and health expenses. For eating and drinking out, only one of the estimates remains significant, whereas for socializing and hobbies and games and toys, all estimates become significant.

Based on these mixed findings, hypothesis (H3), that estimates from the nonprobability panel would suggest lower average expenditures, have to be rejected.

RQ4: Do the differences between panels in expenditure estimates remain after weighting?

The results in table 5 suggest that none of the weights reduce the differences between panels in expenditure estimates. Focusing on total mean expenditure, the unweighted estimates are not significantly different between the two samples. Neither the first weight (based on socio-demographic characteristics), nor the second weight (which in addition includes financial behaviors) change this conclusion. However, using the third weight (which in addition includes digital affinity) suggests a significantly lower mean total expenditure in the nonprobability panel compared to the probability panel (–£199.5, $p = .027$).

Examining individual expenditure categories, the results are mixed. The differences between samples in mean expenditure on eating and drinking out, and clothing and footwear, become insignificant when weight 3 is used. On

Table 6. Overview of Research Questions, Hypotheses, and Whether the Hypotheses are Supported by the Results

Research question	Hypothesis	Supported?
RQ1: Do the app study participants in the probability-based panel have different characteristics than those in the nonprobability online panel?	Compared to those participants from the probability-based panel, app study participants from the nonprobability online panel. . .	
	H1.1: . . .are more likely to be middle-aged, female, and highly educated.	Yes
	H1.2: . . . have higher levels of digital skills.	Mostly
	H1.3: . . . are more willing and less concerned to participate in additional smartphone-based tasks.	Yes
	H1.4: . . . have greater interest in their finances and are more involved in behaviors related to their finances.	Yes
RQ2: Are there differences between the panels in how participants use the app?	H2.1: . . . are more likely to apply strategies to maximize their monetary compensation.	No
	H2.2: . . . are more likely to apply strategies to minimize their effort.	No
RQ3: Are there differences between the panels in expenditure estimates, that is, the study's main outcome of interest?	H3: Compared to estimates from a probability-based panel, estimates from the nonprobability panel suggest lower average expenditures.	No
RQ4: Do the differences between the panels in expenditure estimates remain after weighting?	Differences between probability and nonprobability samples will . . .	
	H4.1: . . . decrease when weighting for socio-demographic characteristics, . . .	No
	H4.2: . . . decrease further when adding financial behaviors, . . .	No
	H4.3: . . . and decrease further still when adding digital affinity measures related to the app use task.	No

the other hand, significant differences in mean expenditure emerge when weights are applied for home improvements and household goods, health expenses, socializing and hobbies, and games and toys.

These results do not support the hypotheses that differences between probability and nonprobability samples will decrease when weighting for socio-demographic characteristics (H4.1), decrease further when adding financial

behaviors (H4.2), and decrease further still when adding digital affinity measures related to the app use task (H4.3).

7. CONCLUSIONS

In this case study, we examined differences between app study data gathered based on a probability-based mixed-mode panel and a nonprobability online panel. Our case study is a "system comparison," testing for differences in data collected with the same tool in two different panels. The two panels differ in their sample selection mechanism, but also in other aspects such as their incentive scheme, survey modes, data collection frequency, respondent motivation, and relationship with the survey organization.

We found that app study participants from the two panels differed on socio-demographics, financial behavior, and digital affinity as well as their spending reported in the app. These findings show that different people participated in the two app study samples, and that the data from the two app study samples led to different conclusions regarding key substantive app study outcomes. Weighting did not reduce differences in mean expenditure between samples, even when the weighting scheme included extensive participant information from the baseline questionnaire.

In line with most of the research on nonprobability and probability-based panel data, our results indicate that the differences in the nonprobability and probability-based panel recruitment processes lead to differences in the data gathered (e.g., socio-demographic characteristics, financial behavior, digital skills and behavior, and substantive outcomes of interest) from these data sources. We examined two stages of selection into the app study, testing for differences between baseline survey respondents from the two panels (i.e., those eligible for the app study) and differences between the app users from both panels. In addition, we estimated propensity models examining whether the predictors of participation in the app study differed between panels. The results indicate that the differences between baseline respondents are in the same direction and larger than the differences between app users from the two panels, with two exceptions (willingness to do different tasks for a survey on a smartphone and data security concerns). In addition, in line with most research on nonprobability and probability-based panel data, our findings show that these differences in the data are difficult to eliminate by weighting. While most existing research only examined survey data, our findings indicate that the differences can also be found in app study data gathered from nonprobability and probability-based panel respondents.

The only data quality aspect for which we did not find evidence of differences between the nonprobability and probability-based panel was behavior in using the spending app. This finding is contrary to the argument that nonprobability panel participants try to maximize their monetary incentive at the

expense of data quality. However, this finding is in line with some of the scarce existing literature on response behavior in surveys, which is inconclusive regarding the question of whether nonprobability panel participants answer questions less conscientiously than probability-based panel respondents.

Since our study is the first to compare app study data based on nonprobability and probability-based panels, more research is needed to verify our findings and to further explore which features of nonprobability and probability-based panels lead to these differences in the gathered data. This is particularly important since our study is unable to establish whether the probability versus nonprobability sampling procedure is responsible for the differences we found in the app data. For example, while in our experience it is typical to use comparatively lower and in-kind incentives in nonprobability panels, applying the same incentive schemes in both app studies may have led to different results.

The use of single imputation with modal categories for missing items is a limitation. Robustness checks showed that using the imputed values produced more conservative estimates than complete case analyses. As the rate of missingness is very low (ranging from 0 to 1.8 percent in the Innovation Panel and from 0 to 0.2 percent in the access panel), the imputation method is unlikely to affect the conclusions from this study.

In addition, the extent to which our findings are generalizable beyond spending apps and beyond the specific panel survey studies (*Understanding Society* Innovation Panel and Lightspeed UK) also remains to be explored in future research. Moreover, since our study in 2018, app technology has progressed and smartphone use has increased (for an expenditure app with enhanced design features, see Toepoel et al. 2020). In addition, future research should examine which sample source for gathering app study data leads to more accurate estimates relative to benchmark data for the intended target population. Based on our study, we conclude that it remains critical to be cautious about app studies conducted in a nonprobability panel, since results may not translate to app studies conducted in a probability-based panel setting.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Ansolabehere, S., and Rivers, D. (2013), "Cooperative Survey Research," *Annual Review of Political Science*, 16, 307–329.

Ansolabehere, S., and Schaffner, B. F. (2014), "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison," *Political Analysis*, 22, 285–303.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., and Weimer, D. L. (2003), "The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples," *Political Analysis*, 11, 1–22.

Brüggen, E., van den Brakel, J., and Krosnick, J. (2016), *Establishing the Accuracy of Online Panels for Survey Research. Statistics Netherlands Discussion Paper 2016-04*, The Hague: Statistics Netherlands. Available at https://www.cbs.nl/en-gb/background/2016/15/establishing-the-accuracy-of-online-panels-for-survey-research

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., and Lavrakas, P. J. (2014), "Online Panel Research: History, Concepts, Applications and a Look at the Future," in *Online Panel Research: A Data Quality Perspective*, eds. M. Callegaro, J. Baker, A. Göritz, J. A. Krosnick, and P. J. Lavrakas, Chichester: John Wiley and Sons, 132.

Chan, P., and Ambrose, D. (2011), "Canadian Online Panels: Similar or Different," Vue, 16–20.

Chang, L., and Krosnick, J. A. (2009), "National Surveys via RDD Telephone Interviewing versus the Internet: Comparing Sample Representativeness and Response Quality," *Public Opinion Quarterly*, 73, 641–678.

Cornesse, C., and Blom, A. G. (2020), "Response Quality in Nonprobability and Probability-Based Online Panels," *Sociological Methods and Research*, 52, 879–908.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., de Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., and Wenz, A. (2020), "A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research," *Journal of Survey Statistics and Methodology*, 8, 4–36.

d'Ardenne, J., and Blake, M. (2012), *Developing Expenditure Questions: Findings from Focus Groups. IFS Working Paper W12/18*, London: Institute for Fiscal Studies. Available at https://ifs.org.uk/publications/developing-expenditure-questions-findings-focus-groups

Dassonneville, R., Blais, A., Hooghe, M., and Deschouwer, K. (2020), "The Effects of Survey Mode and Sampling in Belgian Election Studies: A Comparison of a National Probability Face-to-Face Survey and a Nonprobability Internet Survey," *Acta Politica*, 55, 175–198.

Duffy, B., Smith, K., Terhanian, G., and Bremer, J. (2005), "Comparing Data from Online and Face-to-Face Surveys," *International Journal of Market Research*, 47, 615–639.

Dutwin, D., and Buskirk, T. D. (2017), "Apples to Oranges or Gala versus Golden Delicious? Comparing Data Quality of Nonprobability Internet Samples to Low Response Rate Probability Samples," *Public Opinion Quarterly*, 81, 213–239.

Erens, B., Burkill, S., Couper, M. P., Conrad, F., Clifton, S., Tanton, C., Phelps, A., Datta, J., Mercer, C. H., Sonnenberg, P., Prah, P., Mitchell, K. R., Wellings, K., Johnson, A. M., and Copas, A. J. (2014), "Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison with a Probability Sample Interview Survey," *Journal of Medical Internet Research*, 16, e276.

Gelman, A., Goel, S., Rothschild, D., and Wang, W. (2016), "High-Frequency Polling with Non-Representative Data," in *Political Communication in Real Time: Theoretical and Applied Research Approaches*, eds. D. Schill, R. Kirk, and A. E. Jasperson, New York: Routledge, pp. 117–133

Gittelman, S. H., Thomas, R. K., Lavrakas, P. J., and Lange, V. (2015), "Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples," *Journal of Advertising Research*, 55, 368–379.

Greszki, R., Meyer, M., and Schoen, H. (2014), "The Impact of Speeding on Data Quality in Nonprobability and Freshly Recruited Probability-Based Online Panels," in *Online Panel Research: A Data Quality Perspective*, eds. M. Callegaro, J. Baker, A. Göritz, J. A. Krosnick, and P. J. Lavrakas, Chichester: John Wiley and Sons, pp. 238–262.

Hillygus, D. S., Jackson, N., and Young, M. (2014), "Professional Respondents in Nonprobability Online Panels," in *Online Panel Research: A Data Quality Perspective*, eds. M. Callegaro, J. Baker, A. Göritz, J. A. Krosnick, and P. J. Lavrakas, Chichester: John Wiley and Sons, pp. 219–237.

Jäckle, A., Burton, J., Wenz, A., Read, B., Hanson, T., and Xu, D. (2019), *Understanding Society: The UK Household Longitudinal Study: Spending Study 2, User Guide*, Colchester: University of Essex.

Jäckle, A., Burton, J., Wenz, A., and Read, B. (2018), *Understanding Society the UK Household Longitudinal Study: Spending Study 1 User Guide*, Colchester: University of Essex.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., and Gimenez, A. (2016), *Evaluating Online Nonprobability Surveys: Vendor Choice Matters; Widespread Errors Found for Estimates Based on Blacks and Hispanics*, Washington, DC: Pew Research Center. Available at https://www.pewresearch.org/methods/2016/05/02/evaluating-online-nonprobability-surveys/

Keusch, F., Batinic, B., and Mayerhofer, W. (2014), "Motives for Joining Nonprobability Online Panels and Their Association with Survey Participation Behavior," in *Online Panel Research: A Data Quality Perspective*, eds. M. Callegaro, J. Baker, A. Göritz, J. A. Krosnick, and P. J. Lavrakas, Chichester: John Wiley and Sons, pp. 171–191.

Kocar, S., and Baffour, B. (2023), "Comparing and Improving the Accuracy of Nonprobability Samples: Profiling Australian Surveys," *Methods, Data, Analyses*, 17, 171–206.

Lee, S. (2006), "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys," *Journal of Official Statistics*, 22, 329–349.

Legleye, S., Charrance, G., Razafindratsima, N., Bajos, N., Bohet, A., and Moreau, C; the FECOND Research Team (2018), "The Use of a Nonprobability Internet Panel to Monitor Sexual and Reproductive Health in the General Population," *Sociological Methods and Research*, 47, 314–348.

Loosveldt, G., and Sonck, N. (2008), "An Evaluation of the Weighting Procedures for an Online Access Panel Survey," *Survey Research Methods*, 2, 93–105.

MacInnis, B., Krosnick, J. A., Ho, A. S., and Cho, M.-J. (2018), "The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension," *Public Opinion Quarterly*, 82, 707–744.

Malhotra, N., and Krosnick, J. A. (2007), "The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples," *Political Analysis*, 15, 286–323.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, 9, 403–425.

Mercer, A. W., Kreuter, F., Keeter, S., and Stuart, E. A. (2017), "Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference," *Public Opinion Quarterly*, 81, 250–271.

Pasek, J. (2016), "When Will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence," *International Journal of Public Opinion Research*, 28, 269–291.

Pennay, D. W., Neiger, D., Lavrakas, P. J., and Borg, K. (2018), *The Online Panels Benchmarking Study: A Total Survey Error Comparison of Findings from Probability-Based Surveys and Non-Probability Online Panel Surveys in Australia. CSRM & SRC Methods Paper No. 2/2018*, Canberra: Australian National University, Centre for Social Research and Methods. Available at https://csrm.cass.anu.edu.au/research/publications/online-panels-benchmarking-study-total-survey-error-comparison-findings

Scherpenzeel, A. C., and Bethlehem, J. G. (2011), "How Representative Are Online Panels? Problems of Coverage and Selection and Possible Solutions," in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, eds. M. Das, P. Ester, and L. Kaczmirek, New York: Routledge, pp. 105–132.

Schonlau, M., van Soest, A., and Kapteyn, A. (2007), "Are 'Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring?," *Survey Research Methods*, 1, 155–163.

Schonlau, M., van Soest, A., Kapteyn, A., and Couper, M. P. (2009), "Selection Bias in Web Surveys and the Use of Propensity Scores," *Sociological Methods and Research*, 37, 291–318.

Schonlau, M., Zapert, K., Simon, L. P., Sanstad, K. H., Marcus, S. M., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S. H. (2004), "A Comparison between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey," *Social Science Computer Review*, 22, 128–138.

Smyk, M., Tyrowicz, J., and van der Velde, L. (2021), "A Cautionary Note on the Reliability of the Online Survey Data: The Case of Wage Indicator," *Sociological Methods and Research*, 50, 429–464.

Sohlberg, J., Gilljam, M., and Martinsson, J. (2017), "Determinants of Polling Accuracy: The Effect of Opt-In Internet Surveys," *Journal of Elections, Public Opinion and Parties*, 27, 433–447.

Sparrow, N. (2006), "Developing Reliable Online Polls," *International Journal of Market Research*, 48, 659–680.

Steinmetz, S., Bianchi, A., Tijdens, K., and Biffignandi, S. (2014), "Improving Web Survey Quality: Potentials and Constraints of Propensity Score Adjustments," in *Online Panel Research: A Data Quality Perspective*, eds. M. Callegaro, J. Baker, A. Göritz, J. A. Krosnick, and P. J. Lavrakas, Chichester: John Wiley and Sons, pp. 273–298.

Sturgis, P., Kuha, J., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Lauderdale, B. E., and 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.

Suffield, M., Hasbrouck, H., Coulter, A., Jäckle, A., Burton, J., Crossley, T. F., Couper, M. P., and Lessof, C. (2018), *Understanding How People Think about Their Daily Spending. Understanding Society Working Paper 2018-02*, Colchester: University of Essex.

Szolnoki, G., and Hoffmann, D. (2013), "Online, Face-to-Face and Telephone Surveys—Comparing Different Sampling Methods in Wine Consumer Research," *Wine Economics and Policy*, 2, 57–66.

The American Association for Public Opinion Research (2023), *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (10th ed.), Alexandria, VA: AAPOR.

Toepoel, V., Lugtig, P. J., and Schouten, J. G. (2020), "Active and Passive Measurement in Mobile Surveys," *The Survey Statistician*, 82, 14–26.

University of Essex, Institute for Social and Economic Research (2021), *Understanding Society: Innovation Panel, Waves 1-13, 2008-2020* [data collection] (11th ed.), UK Data Service. SN: 6849, DOI: 10.5255/UKDA-SN-6849-14.

University of Essex, Institute for Social and Economic Research (2022), *Understanding Society: Spending Study 2, 2018-2019* [data collection]. UK Data Service. SN: 8909, DOI: 10.5255/UKDA-SN-8909-1.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015), "Forecasting Elections with Non-Representative Polls," *International Journal of Forecasting*, 31, 980–991.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., and Wang, R. (2011), "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples," *Public Opinion Quarterly*, 75, 709–747.

Zindel, Z. (2022), "Social Media Recruitment in Online Survey Research: A Systematic Literature Review," *Methods, Data, Analyses*, 42, 1–42.

Journal of Survey Statistics and Methodology, 2024, 00, 1–30
https://doi.org/10.1093/jssam/smae026
Article