Survey Research Methods (2025)
Vol. 19, No. 1, pp. 25-41
https://doi.org/10.18148/srm/2025.v19i1.8263
European Survey Research Association

Nonresponse and Dropout in an App-Based Household Budget Survey: Representativeness, Interventions to Increase Response, and Data Quality

Evelien Rodenburg[1] · Barry Schouten[2] · Bella Struminskaya[1]
[1]Utrecht University, Department of Methodology and Statistics
[2]Statistics Netherlands

Household budget surveys struggle with low response and participation rates, and lower data quality, in part due to a high respondent burden. App-assisted budget surveys may provide solutions to both these problems. This cross-country study carried out in the Netherlands, Luxembourg, and Spain, investigates the use of an app-based diary for collecting household expenditure data compared to a web-based method. We report the results of two randomized experiments: 1) using personalized feedback and 2) interviewer-assisted versus mail recruitment in terms of influence on response and participation rates. The app-based household budget survey yields slightly higher registration, activity, and completion rates compared to the web-based household budget survey that we use as a reference. We find disproportionate representation of certain groups in the app-based sample, but no substantial differences in the overall representativeness between the app-based and web-based samples. Providing households with personalized feedback does not affect registration or activity in the app. Using interviewers for recruitment does increase registration and activity rates, although this negatively affects the representativeness of the sample. Neither providing personalized feedback nor using interviewers for recruitment significantly affects dropout during the study or data quality. We also find no substantive differences between the quality of web-collected expenditure data and data collected in the app. Overall, using an app could be suitable for collecting expenditure data especially in combination with the use of interviewers for recruitment. However, this may come at a cost to representativeness.

Keywords: app-based surveys; household budget surveys; participant feedback; interviewers, data quality

1 Introduction

Diary studies carry a high respondent burden (Pettersen, 2005), which can lead to survey fatigue and dropout (Schmidt, 2014) and cause respondents to be less willing to put effort in the diary over time. Respondents are then more inclined to postpone filling out their diaries, which in turn leads to higher recall bias (Elevelt, Bernasco, Lugtig, Ruiter, & Toepoel, 2021). Moreover, diary studies like other surveys suffer from low response rates (Jäckle, Wenz, Burton, & Couper, 2022). For example, in the 2015 wave of the EU Household Budget Survey (HBS), the mean response rate was 50%, with the Netherlands yielding a response rate of only 17% (Eurostat, n.d. a). Thus, continuing to conduct diary studies in the same way may be no longer sustainable.

In recent years, smartphones have increasingly been used for data collection (Struminskaya et al., 2021). Using smartphones for the collection of expenditure data has potential benefits such as the replacement of manual entry of a purchase by pictures of receipts, data collection close to the time of an event potentially reducing recall bias (Elevelt et al., 2021), and a possibility to follow respondents over a longer period without increasing respondent burden (Elevelt et al., 2021; Elevelt, Lugtig, & Toepoel, 2019).

However, smartphone-based data collection can also lead to increased response burden as respondents are asked to

Supplementary Information The online version of this article (https://doi.org/10.18148/srm/2025.v19i1.8263) contains supplementary material.

Corresponding author: Bella Struminskaya, Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584CH Utrecht, The Netherlands (Email: b.struminskaya@uu.nl)

share more (personal) information (Elevelt et al., 2019). Additionally, coverage error may arise if groups within the target population are excluded because they do not own a smartphone or are unable to use one for data collection (Jäckle, Burton, Couper, & Lessof, 2019; Keusch, Struminskaya, Antoun, Couper, & Kreuter, 2019; Struminskaya et al., 2020; Wenz, Jäckle, & Couper, 2019). Lastly, participants may be unwilling to use an app for data collection due to privacy concerns (Jäckle et al., 2019; Keusch et al., 2019; Revilla, Couper, & Ochoa, 2019; Struminskaya et al., 2020; Wenz et al., 2019), or not able to use smartphones due to physical capabilities or a lack of storage on the phone (Jäckle et al., 2019).

Smartphone-based studies fielded in the general population have reported low participation rates (Jäckle et al., 2022; Kreuter, Haas, Keusch, Bähr, & Trappmann, 2020) and so far only the UK Understanding Society Panel has used an app to collect expenditure data from the general population (Jäckle et al., 2022). Because of the limited number of studies, not much is known about methods to increase response rates in smartphone-based studies or the quality of the expenditure data collected through apps. Thus, it remains unclear whether switching from a paper or web diary to an app improves the response rates and/or data quality of household budget surveys.

In this study, we aim to investigate the impact of several push-to-app design choices on registration, activity, and dropout in an app-based version of the Household Budget Survey. We conducted a randomized experiment in the general population of the Netherlands to investigate the effects of personalized feedback and using interviewers for recruitment on registration, activity, dropout, and data quality.

2 Background and Research Questions

In the following, we give a brief account of nonresponse in surveys and how app-based 'smart' surveys may assist in increasing response rates and/or data quality and formulate our research questions.

2.1 Nonresponse in Household Surveys

Although smartphones are used frequently in daily life (Bian and Leung, 2015), smartphone-based surveys have not yet been able to reverse the declining trend in survey response rates (De Bruijne and Wijnant, 2014; Jäckle et al., 2022; Wenz, Jäckle, Burton, & Couper, 2022). Nonresponse in household surveys can occur at different stages in the response process, such as initial contact or the interview request (Bethlehem, Cobben, and Schouten, 2011). Additionally, dropout can occur during the study period (Peytchev, 2009). Many different reasons may underlie nonresponse and dropout, such as the households' lack of interest in the survey topic (Groves, Presser, & Dipko, 2004), survey length (Deutskens, De Ruyter, Wetzels, & Oosterveld, 2004; Sahlqvist et al., 2011), or survey sponsorship (Keusch et al., 2019).

As a consequence of nonresponse, nonresponse bias can occur when survey respondents are systematically different from the nonrespondents (Bethlehem et al., 2011). Although low response rates do not necessarily result in large nonresponse biases, high response rates can reduce the risk of bias (Groves & Peytcheva, 2008). If responding households have different spending behaviour than nonresponding households, the accuracy of the collected data is compromised. Additionally, small samples due to nonresponse lead to wider confidence intervals and decreased statistical power (Lohr, 2019). Finally, nonresponse increases the cost of surveys and may lead to the need for more complex designs, which introduces new sources of error (Peytchev, 2013).

2.2 Methods to Increase Survey Response

Researchers deal with nonresponse in household surveys in different ways. One solution is to use postsurvey adjustments such as imputation and weighting (Toepoel and Schonlau, 2017). However, these adjustments rely on auxiliary information, which is not always easily available. Alternatively, surveys can incorporate methods to increase response rates into their designs, for example, by providing participants with monetary incentives which improve response rates (e.g., Laurie, 2007; Ryu, Couper, & Marans, 2006; Singer & Ye, 2013). For app-based studies, promising monetary incentives to participants can increase willingness to download a research app (Haas, Kreuter, Keusch, Trappmann, & Bähr, 2020; Keusch et al., 2019).

Two methods that have not received as much attention, especially for app-assisted surveys, are the use of interviewers for recruitment and providing participants with personalized feedback about the study results or the data they provide. The use of interviewers has been widely studied for web and paper-based surveys. First, interviewers can persuade sampled persons and households to participate: unit nonresponse in interviewer-based surveys is lower than in self-administered surveys (e.g., Durrant et al., 2010; Lord et al., 2005). Second, interviewers can also provide clarifications if respondents have questions about both core and additional survey tasks or elements. For example, studies have shown that permission for linkage increases when interviewers ask for consent to data linkage compared to when consent is only asked through mail or web (e.g., Korbmacher and Schroeder, 2013; Sakshaug, Hülle, Schmucker,

& Liebig, 2017). Additionally, interviewers can assess the situation during the fieldwork and apply resources to where needed, for example, to track households who have moved since the previous wave of a longitudinal survey. Couper and Ofstedal (2009) found that the use of interviewers for data collection resulted in the greatest chance to locate sample members. When respondents have to engage with technologies that are less or more familiar to them such as download a research app, interviewers can help respondents walk through the steps of app download and installation, or better explain the goals of the study. For an app-based budget survey, Jäckle et al. (2022) found that response rates increased when interviewers were used for recruitment, although dropout during the study also increased for the group recruited through interviewers, suggesting that participants who were hesitant about using research apps could be persuaded/helped by the interviewers but then were less motivated to continue with the study on their own. Jäckle et al. (2022) focused on participation and selectivity, however, the quality of data provided by participants in the app for those recruited via interviewers vs. not using interviewers, remains an open question.

A disadvantage of both monetary incentives and using interviewers for recruitment is high cost (Wenz et al., 2022). Personalized feedback could be a less costly alternative. In line with the quantified-self paradigm, participants can be more motivated to share information about themselves if they can get insights into own behavior (Bietz et al., 2019), This can be especially relevant for app-based studies since commercial apps provide such feedback that can allow individuals to implement behavioral changes. However, research on providing feedback about study results to participants has found mixed effects on survey response and participation. Often, studies only provide general feedback about their results, which has been shown to not affect response rates in surveys (Edwards et al., 2009; Göritz & Neumann, 2016; Scherpenzeel & Toepoel, 2014). Personalized feedback, on the other hand, has been shown to increase response rates in web-based or paper-based surveys (Bälter, Bälter, Fondell, & Lagerros, 2005; Bälter, Fondell, & Bälter, 2012; Marcus, Bosnjak, Lindner, Pilischenko, & Schütz, 2007), especially in the case of non-salient survey topics (Marcus et al., 2007). For smartphone-based surveys, Struminskaya et al. (2020) found that promising survey participants feedback increased willingness to share sensor-collected data among respondents sampled from the general population of the Netherlands, but Wenz et al. (2022) found that although participants reacted positively to the feedback, it did not increase response or participation rates in an app-based budget survey. However, the sample Wenz et al.'s (2022) study came from a nonprobability-based panel, so the respondents were likely more motivated to participate than respondents sampled from the general population. Providing feedback to participants can have both negative and positive influence on data quality. On the one hand, feedback might motivate participants to engage with the study and report more accurately, on the other hand, it can introduce effects similar to panel conditioning where the behavior or attitudes can change as a result of participating in a study (e.g., Struminskaya and Bosnjak 2021). There has not been much research on this topic yet, but Wenz et al. (2022) found that there are very few differences between the feedback groups in their main outcome of interest, median total spending, which did not differ significantly between those who received feedback and those who did not.

2.3 Research Questions

In this study, we investigate response and participation in the app-based Household Budget Survey, focusing on the registration, activity, completion, and dropout rates. We define *registration* as completing the entry questionnaire in the app. The entry questionnaire followed directly after logging in the app for the first time and had to be completed before the diary could be started. After filling in the questionnaire, the household was directed to the diary overview screen.

Activity is defined as entering at least one purchase in the app diary. *Completion* is defined in terms of the last-seen activity in the app and had to be at least 14 days after completing the entry questionnaire. The household had to confirm and validate each day, also when not making a purchase. Consequently, the last-seen activity could simply be a day validation. Finally, *time of dropout* is the moment of last seen activity in the app for households that did not complete the survey (i.e., a household dropped out if last seen activity occurred before 14 days had passed).

We answer four research questions:

1. What is the representativeness in an app-based HBS of households at three stages of participation (a) registration in the app; (b) participating, that is, actively providing data; and (c) completing the two-week diary?
2. What is the impact of personalized feedback on response?
3. What is the impact of contact mode (face-to-face vs. mail) on response?
4. To what extent is measurement data quality improved by providing feedback or using either contact mode?

Our first question addresses the relationships between household characteristics and registration and activity rates, the representativeness of the sample. Previous research has suggested that using smartphones for data collection might reduce respondent burden (Elevelt et al., 2019), and since high respondent burden can decrease the probability that a household will want to participate in a survey, lowering

the burden can enhance representativeness. In terms of respondent characteristics, age and education level may be related to participation in app-based budget surveys (Jäckle et al., 2019). In addition to age and education, Struminskaya et al. (2021) showed that household size, ethnic background, urbanization, homeownership, and income are related to willingness to participate in app-based surveys,.

Our second question considers the effect of promising households personalized feedback (*insights*) about their spending on registration, activity, and dropout. Most of the studies that investigated the effect of personalized feedback on response in app-based surveys suggest that feedback may be an effective in increasing participation rates (Bälter et al., 2005; Bälter et al., 2012; Marcus et al., 2007; Struminskaya et al., 2020)

The third research question is about the effect of contact mode on registration, activity, and dropout. Existing research, albeit scarce, suggests that using interviewers for recruitment can increase participation rates in app-based surveys (Jäckle et al., 2022).

Finally, the fourth question turns to the influence of the two interventions, providing feedback and contact mode, on the quality of the data collected in the app. Regarding the quality of expenditure data collected through an app, Jäckle et al. (2022) reported similar reporting behaviours in an online diary and app diary. Furthermore, using an app rather than a web-diary might decrease respondent burden and recall bias (Elevelt et al., 2019; Elevelt et al., 2021), while personalized feedback may increase motivation throughout the study (Wenz et al., 2022).

In the following, we introduce the study design and data, which is followed by describing our analysis strategy and presentation of results.

3 Data

In our study we combine two data sets: the household expenditure data stemming from the Household Budget Survey app and administrative data from Statistics Netherlands. The administrative data can be linked on the individual household level and includes Tax Office administration on amount and type of household income.

3.1 Survey Design and Experimental Conditions

This study was embedded in a field test of a smartphone-app based Household Budget Survey, which was part of Eurostat's ESSnet Smart Surveys. The HBS app was developed by Statistics Netherlands (CBS), and all expenditure data was collected by CBS. Households were randomly sampled from population registers in three countries: the Netherlands, Luxemburg and Spain (in this paper, we focus on the Dutch data only). Per household one member of the household core was randomly selected and the invitation was addressed to this household member. The selected household member was encouraged to act as a contact person and to involve the other household members.

The app could be installed on multiple devices and by multiple persons within the household, using the same login information. After logging in, the households had to answer a short entry questionnaire before they could start entering expenditures. This questionnaire included questions about household characteristics such as household size and composition and asked whether the household would go on any holidays during the study period. The households were asked to enter all their expenditures for 14 days. Expenditures could be entered manually or by taking pictures of receipts by accessing a smartphone's camera function through the app. The app also collected paradata, which provided insights on households' behavior in the app. We obtain the exact times at which activity in the app was first and last seen for each household from these paradata. All households received an incentive with the invitation (5 euros). Households that completed the survey received 20 euros These incentives are the same as in the regular Dutch Household Budget Survey, which is conducted via web.

The study contained three experiments: 1) contact mode, 2) personalized feedback; and 3) feedback on automated text extraction from receipts. Households were randomly assigned to the experimental conditions. All households participated in all experiments. The sampled households were randomly assigned to be invited to participate either by an interviewer at the door or by postal mail. Households who did not register after the initial invitation received a second invitation. Households in the interviewer condition were contacted by the interviewer halfway through the study period and were able to contact the interviewer with questions during the study. Households in the mail condition received an invitation letter and were able to contact a helpdesk if they had any questions. The personalized insights (feedback) the households received consisted of informative graphs and charts about spending behavior. Households were either informed about the insights in the invitation and could view them immediately after entering an expenditure in the app (immediate condition), or they were not informed about the insights in the invitation and could only see them after the study period ended (delayed condition). See Fig. 1 for screenshots of the personalized insights pages in the app. In this paper, we only analyze the personalized feedback and contact method conditions; feedback on automated text extraction from receipts condition is not discussed.

Initially, 4000 households were randomly sampled from the general populations in three countries the Netherlands

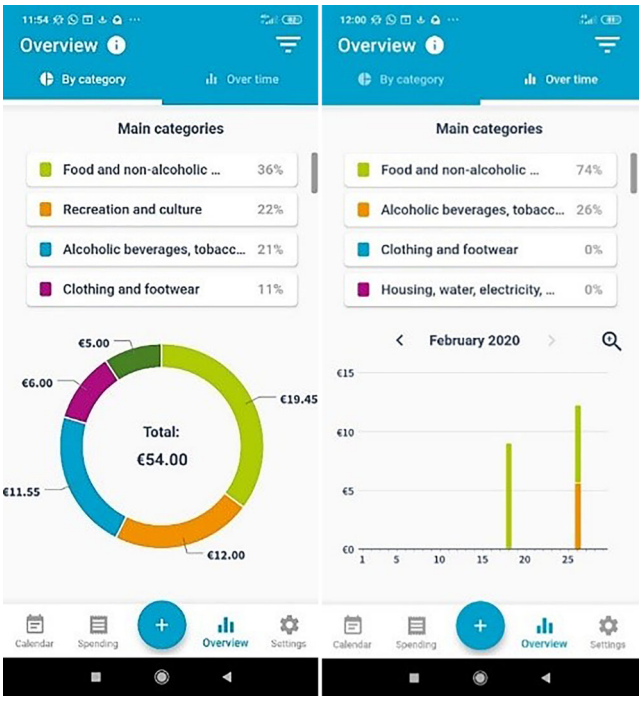

Fig. 1

Screenshots of the personalized in-app insights

(n = 1600), Luxembourg (n = 1600), and Spain (n = 800). The sample sizes were chosen such that differences of 5% in country completion rates between experimental conditions could be observed at type I and type II errors of, respectively, 5 and 20%. The Spanish sample size was smaller due to interviewer workload constraints. In this paper, we focus on the Dutch sample since the designs vary too much between the countries to perform a cross-country comparison. However, more importantly, availability of administrative data shared by all countries is very limited. Table 1 displays the sample sizes per condition. The sample was fielded in two consecutive months, September 2024 and October 2024. Ultimately, fewer households were fielded in the Netherlands (n = 1485). The sample size for the interviewer condition was lower than for the non-interviewer due to workload issues in October. The October sample was randomly subsampled. As a consequence, minimally observable differences were larger than the prescribed 5%.

Table 1

Sample sizes per experimental condition

	n
Feedback instant	748
Feedback delayed	737
Interviewer	685
No interviewer	800

3.2 Administrative Data

Household characteristics were obtained from the population registries. We chose household characteristics based on the weighting model used by Statistics Netherlands for the regular HBS. The choice of the household characteristics was further motivated by their expected relationship with smartphone abilities and access motivated by previous research (Struminskaya et al., 2021; Wenz et al., 2019). Variables age, gender type of household, migration background, education level, urbanization of the municipality, homeownership, and income (in standardized percentiles) were available. For households larger than one person, the socio-demographic characteristics of the selected reference person within the household were used. Descriptive statistics and further explanations of the administrative data variables are given in Table A1 in the Appendix.

4 Analysis Strategy

In this section, we discuss how we address each of the four research questions.

4.1 Representativeness of an App-based HBS (RQ 1)

We evaluate the first research question using logistic regression models and R-indicators. First, we compute regression models for registration and activity in the app. We include age, household size, migration background, education level, degree of urbanization of the municipality, homeownership, and income.

To analyze the representativeness of the registered and active households in the app-based samples, we calculate R-indicators. These indicators evaluate the estimated variance in estimated response propensities based on the regression models. The larger the variance, the smaller the indicators (see Schouten, Cobben, and Bethlehem (2009) for a detailed explanation). We formulate response models for registration and activity including all variables (see Table A1 in Appendix A). We further analyze representativeness on the level of the household characteristics using unconditional partial indicators (Pu) and conditional partial indicators (Pc) (see Schouten, Shlomo, and Skinner, 2011). Partial indicators isolate the contributions of individual variables to the variance of estimated response propensities without (unconditional) and with (conditional) adjustment for collinearity with the other included variables.

Finally, we calculate coefficients of variation (CV) as a measure of maximal absolute bias (De Heij, Schouten, & Shlomo, 2015).

4.2 Effects of Personalized Insights and Contact Mode on Participation (RQ 2 and 3)

To answer the second and third research questions, we perform two analyses per experimental condition. First, we calculate the differences in the registration and activity rates per condition (immediate vs. delayed, or self-administered vs interviewer). We then use Chi-Square Tests to test for associations between the insights condition and registration or activity. We use the combined app-based sample from all three countries for this analysis. If we find a significant association between personalized insights and registration or activity, we will calculate R-indicators to analyze the representativeness of the participating households in each of the insights conditions.

Second, we use a Kaplan-Meier survival curve to visualize dropout of active households during the study period. To assess the difference in the dropout rates during the study period between the households in the immediate insights condition and the delayed insights condition, we use a Log-Rank Test. We round moment of dropout to a full day, rounding up from and including 0.5. Households are censored if no dropout occurred before the end of the study. We assume that censoring time and survival time (i.e., time remaining in the study) are independent (Leung, Elashoff, & Afifi, 1997).

4.3 Effects of Personalized Insights and Contact Mode on Data Quality (RQ 4)

To analyze the effect of insights and contact mode on data quality, we need a set of measurement quality indicators. To our knowledge, there are no existing quality indicators for expenditure data. However, intuitively it makes sense that high quality expenditure data should be more diverse, that is, both small and large purchases are reported, and purchases are reported in multiple store types. We use the following quality indicators: 1) the number of entries per household (entries); 2) the difference between the maximum and minimum amount spent per entry per household (amount range); 3) the standard deviation of the amount of money spent per entry within a household (SD amount); 4) the difference between the maximum and minimum number of products bought per entry per household (products range); 5) the standard deviation of the number of products bought per entry within a household (SD products); and 6) the expected number of different store types in which a household entered a purchase (store types). We note that the Household Budget Survey app included scanning of receipts and automated text recognition and interpretation. We, therefore, did not evaluate data entry errors by respondents.

We standardize all quality indicators, with the exception of entries. To standardize the amounts and numbers of products, we divide the amount per household by the average amount, and we divide the number of products per household by the average number of products. We use the standardised amounts and numbers of products to calcu-

Table 2

Registration, activity, and completion rates for the 2020 regular HBS and the 2021 app-assisted HBS

			95% C.I.	
	N	%	lower	upper
2020 Web-based				
Registered	16,520	21	21	21
Active	13,483	17	17	17
Complete	10,420	13	13	13
Active/registered	–	82	81	82
Complete/active	–	77	77	78
2021 App-based				
Registered	292	20	18	22
Active	246	17	15	19
Complete	208	14	13	16
Active/registered	–	84	80	88
Complete/active	–	85	79	90

Table 3

Logistic regression models for predicting registration and activity using household characteristics. Both parameters estimates (Coef.) and average marginal effects (AME) are given as well as corresponding standard errors and p-values for the parameter estimates

	Registration				Activity			
	AME	Coef.	S.E.	*p*	AME	Coef.	S.E.	*p*
Intercept		–2.13*	0.50	<0.001		–2.32*	0.56	<0.001
Age (ref = 18–24)								
25–34	–0.00	–0.01	0.40	0.975	0.03	0.22	0.45	0.619
35–44	0.02	0.12	0.41	0.766	0.04	0.34	0.46	0.450
45–54	–0.02	–0.15	0.41	0.714	0.01	0.12	0.46	0.797
55–64	0.03	0.19	0.40	0.636	0.05	0.36	0.44	0.412
65–74	0.01	0.07	0.42	0.870	0.04	0.32	0.47	0.495
≥75	–0.09	–0.75	0.48	0.115	–0.07	–0.76	0.55	0.168
Household size (ref = 1 person)								
2 persons	–0.00	–0.02	0.18	0.928	–0.02	–0.16	0.19	0.414
3 persons	–0.02	–0.16	0.24	0.510	–0.03	–0.27	0.26	0.289
4 or more persons	0.02	0.14	0.22	0.545	–0.02	–0.18	0.24	0.445
Origin (ref = Dutch)								
Non-western immigrant	–0.07	–0.52	0.28	0.068	–0.08	–0.69*	0.32	0.032
Western immigrant	–0.04	–0.31	0.23	0.168	–0.07	–0.63*	0.26	0.016
Education level (ref = lower education)								
Secondary or lower tertiary	0.09	0.59*	0.26	0.026	0.07	0.58*	0.29	0.044
Higher tertiary	0.16	0.99*	0.27	<0.001	0.14	0.98*	0.30	<0.001
Unknown	–0.02	–0.19	0.27	0.480	–0.02	–0.25	0.30	0.417
Urbanization (ref = very strong)								
Strong	–0.00	–0.03	0.17	0.862	0.00	0.03	0.18	0.864
Moderate	–0.00	–0.01	0.29	0.977	–0.01	–0.06	0.31	0.850
Little	–0.02	–0.11	0.26	0.676	0.01	0.04	0.27	0.876
Homeownership (ref = owner)								
Rent	–0.03	–0.18	0.18	0.272	–0.04	–0.32	0.19	0.101
Income (ref = 0–20)								
20–40	0.07	0.55	0.29	0.054	0.03	0.28	0.32	0.373
40–60	0.07	0.59*	0.29	0.040	0.06	0.57	0.31	0.623
60–80	0.09	0.67*	0.29	0.018	0.06	0.58	0.31	0.060
80–100	0.15	1.08*	0.28	<0.001	0.14	1.07*	0.31	<0.001
Unknown	0.04	0.38	0.60	0.522	0.05	0.44	0.68	0.521

*p<0.05

late the amount range, SD amount, products range, and SD products. Because we intend to measure the variation in the data, the amount range, SD amount, products range, and SD products are set to missing for households with only one entry. Because longer study durations give households more opportunities to report purchases in more store types, we correct store types for the difference in study durations. For this, we calculate an expected number of store types using the following formula:

$$\sum_{m=1}^{M}(1-(1-p_{h,m})^{a_h}),$$

where m is the store type, $p_{h,m}$, is the probability of a purchase entered by a household being in a certain store type, and a_h is the number of purchases entered by a household.

Table 4

R indicators, CVs, and Variable-level partial for household characteristics for the 2021 samples

	Registered						Active					
		95% C.I. (bootstrapped)			95% C.I. (bootstrapped)			95% C.I. (bootstrapped)			95% C.I. (bootstrapped)	
	R	lower	upper	CV	lower	upper	R	lower	upper	CV	lower	upper
Overall	0.77	0.71	0.79	0.56	0.44	0.68	0.78	0.73	0.81	0.64	0.50	0.78
Variable-level partial												
Age	0.06			0.02			0.05			0.02		
Household size	0.03			0.00			0.02			0.01		
Origin	0.03			0.01			0.04			0.02		
Education level	0.09			0.05			0.09			0.04		
Urbanization	0.00			0.00			0.01			0.00		
Homeownership	0.04			0.01			0.04			0.01		
Income	0.07			0.03			0.07			0.03		

Table 5

Difference in registration and activity rates between the two personalized insights conditions

	Registered			Active		
	%	95% C.I. (bootstrapped)		%	95% C.I. (bootstrapped)	
		lower	upper		lower	upper
Immediate	19	15.9	21.9	16	13.5	19.0
Delayed	21	18.4	23.8	18	15.0	20.0
χ^2	0.90			0.48		
P	0.344			0.489		

We use two five-stage hierarchical logistic regression models to analyze the relationships between the quality indicators and the experimental conditions. The dependent variable of each hierarchical regression model is the experimental condition (insights = 1/0 or contact mode = 1/0 where 1 is for face-to-face and 0 is for mail), and the quality indicators and household characteristics (age and household size) are included as predictors. We enter age and household size at stage one to control for differences in spending behavior due to household type. We then add one quality indicator at each subsequent stage based on the model fit, until all quality indicators are included in the model. We use the Akaike Information Criterion (AIC) to assess model fit. Because the correlations between amount range and amount SD (r = 0.98), and products range and products SD (r = 0.91) are very high, we only add either the range indicator or the SD indicator. Hence, four quality indicators are included at the final stage of the hierarchical logistic regression models.

4.4 Bootstrapping Confidence Intervals

To reflect the uncertainty in the data, we calculate bootstrapped confidence intervals (CIs) around means and proportions. There are multiple approaches to calculate CIs with bootstrapped samples (DiCiccio & Efron, 1996; Efron & Tibshirani, 1994; Jung, Lee, Gupta, & Cho, 2019). Since the sample sizes per country are quite small (Jung et al., 2019) and the distributions for the bootstrapped statistics are all approximately normal (DiCiccio & Efron, 1996), percentile CIs are appropriate for the means and proportions of interest and are thus used in this study.

4.5 Software and Packages

All analyses are conducted in R version 4.0.3 (R Core Team, 2020). We use the 'margins' package to calculate average marginal effects (Leeper, 2018). R-indicators, par-

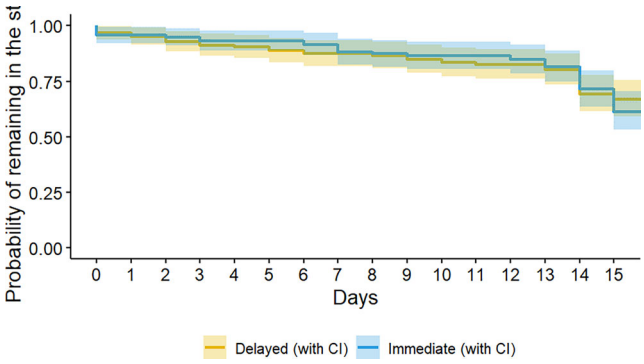

Fig. 2

Kaplan-Meier survival curves for dropout in the personalized insights conditions

tial R-indicators, and CVs for the 2021 samples are calculated using R code made available by the RISQ project (http://www.risq-project.eu/) (De Heij et al., 2015). We use the 'survival' package for the survival analysis and log-rank tests (Therneau & Grambsch, 2000; Therneau, 2022); weighted means, quartiles, and standard deviations for the 2020 quality indicators are calculated using the 'Hmisc' package (Harrell Jr., 2021).

5 Results

In the following, we present the results for each of the research questions. We start by considering response to the app-based HBS. We then move to the impact of the two experimental conditions, personalized feedback and contact mode, on participation. Finally, we discuss data quality relative to the two experimental conditions.

5.1 How Representative is Response to the App-Based HBS?

Let us first look at registration, activity, and completion rates. Table 2 displays the three rates for the 2021 app-based survey and for the regular 2020 web-based survey. The regular HBS is conducted every five years, the last edition being 2020. Apart from the reference year, the two surveys also differ in design. The regular survey has a one month reporting period with one week of reporting all expenditures and three weeks of larger expenditures only. The HBS app study included two weeks of full reporting. The regular survey included recurrent expenditure questionnaires whereas the app study did not. The app-based rates are a mix of the two experimental conditions, i.e. with or without an interviewer and with or without feedback. The regular HBS is purely self-administered and provides no personalized feed-

Table 6

Log-rank test for difference in survival probabilities between the two personalized insights conditions

	Observed dropout (%)			Expected dropout[a] (%)		
	%	95% C.I. (bootstrapped)		%	95% C.I. (bootstrapped)	
		lower	upper		lower	upper
Immediate	39	31.0	46.6	37	31.0	42.7
Delayed	33	25.5	42.3	36	30.3	41.7
χ^2	0.43					
P	0.511					

[a]Expected dropout is the dropout that is expected if the null hypothesis is true, i.e. there is no difference in survival probability

Table 7

Difference in registration and activity between the two contact mode conditions

	Registered			Active		
	%	95% C.I. (bootstrapped)		%	95% C.I. (bootstrapped)	
		lower	upper		lower	upper
Interviewer	26	22.5	28.9	24	20.5	26.6
Mail	16	13.4	18.6	12	9.5	14.0
χ^2	20.6			36.5		
P	<0.001			<0.001		

Table 8

Differences in representativity of the registered and active households between the two contact mode conditions

	Registered						Active					
	R			CV			R			CV		
		95% C.I. (bootstrapped)			95% C.I. (bootstrapped)			95% C.I. (bootstrapped)			95% C.I. (bootstrapped)	
	PU	lower	upper	PC	lower	upper	PU	lower	upper	PC	lower	upper
Interviewer	0.71	0.64	0.78	0.56	0.42	0.70	0.71	0.64	0.78	0.61	0.46	0.76
Variable-level partial												
Age	0.07			0.04			0.07			0.04		
Household size	0.04			0.02			0.04			0.02		
Origin	0.05			0.02			0.05			0.02		
Education level	0.10			0.04			0.10			0.04		
Urbanization	0.02			0.02			0.01			0.02		
Homeownership	0.06			0.01			0.06			0.00		
Income	0.09			0.03			0.10			0.04		
Mail	0.81	0.74	0.87	0.60	0.40	0.80	0.83	0.77	0.89	0.75	0.48	1.01
Variable-level partial												
Age	0.04			0.02			0.04			0.02		
Household size	0.04			0.03			0.03			0.02		
Origin	0.02			0.01			0.03			0.01		
Education level	0.07			0.04			0.07			0.04		
Urbanization	0.02			0.01			0.02			0.01		
Homeownership	0.03			0.01			0.03			0.01		
Income	0.05			0.02			0.04			0.01		

back. Nonetheless, it is relevant to compare the two. The rates are relatively comparable, however, conditional completion rates are higher for the app-based study (Table 2).

Next, we perform logistic regressions employing all available auxiliary variables. We display the parameter estimates in Table 3. The strongest selection in app registration is on educational level and income. There is furthermore an underrepresentation for household reference persons of 75 years and older and/or with immigration background. The representation of those being active in the diary is very similar; only for immigration background the representation gets weaker.

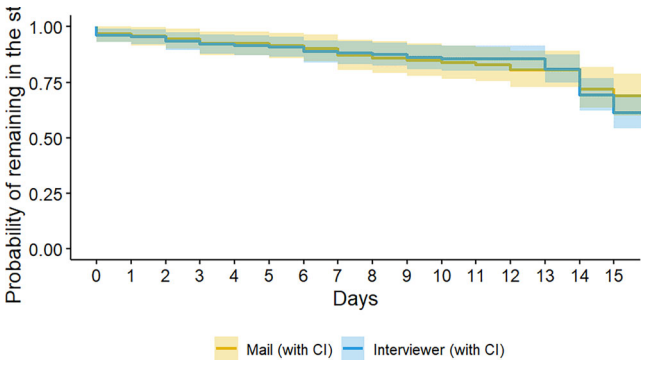

Fig. 3

Kaplan-Meier survival curves for dropout in the contact mode conditions

We analyze the representativeness of the registered and active samples, in addition, using R-indicators. Table 4 shows R-indicators (R), unconditional partial R-indicators (Pu), conditional partial R-indicators (Pc), and coefficients of variation (CV). All these values require different interpretations: For the R-indicators, values closer to 1 indicate a more representative sample; Pu can range from –0.5 to 0.5, whereas negative values indicate underrepresentation and positive values indicate overrepresentation; Pc can range from 0 to 0.5. For Pu and Pc, values further away from zero indicate larger effects of the household characteristic. Differences between the Pu and Pc values indicate that certain household characteristics show collinear response behavior with other household characteristics and therefore do not have a separate impact on representative

Table 9

Log-rank test for difference in survival probabilities between the two contact mode conditions

	Observed dropout (%)			Expected dropout[a] (%)		
	%	95% C.I. (bootstrapped)		%	95% C.I. (bootstrapped)	
		lower	upper		lower	upper
Interviewer	39	32.5	46.7	36	30.5	41.6
Mail	31	23.8	40.8	36	30.5	42.3
χ^2	0.94					
P	0.332					

[a]Expected dropout is the dropout that is expected if the null hypothesis is true, i.e. there is no difference in survival probability

response (Schouten et al., 2011). Lower CV values indicate lower maximal bias.

We conclude that an app-based HBS creates substantial variation in subgroup response rates. In particular, it does so for one of the most relevant household characteristics, the level of income. We also conclude that representativeness is not affected by drop-out. However, as the CVs show, the risk of bias increases due to the drop-out.

5.2 What is the Impact of Personalized Feedback on Response?

To analyze the effect of personalized insights, we calculated the registration rates and activity rates for the immediate and delayed insights groups (Table 5). The registration and activity rates are higher in the delayed insights group, but the bootstrapped CIs for the two conditions overlap. Furthermore, Chi-Square Tests of Independence show no significant associations between personalized insights and registration, or personalized insights and activity. Because there are no significant associations between personalized insights and registration or activity, we do not analyze the representativeness of the registered and active samples in the immediate and delayed insights conditions.

The Kaplan-Meier curves for the two insights conditions show that dropout is gradual up to day 13 (Fig. 2). Around day 13 respondents probably drop out more because they believe their reporting period has ended. At a first glance, there seems to be no large differences in the rates at which households in the two insights conditions drop out. A Log-Rank Test shows that the observed dropout is slightly higher in the delayed insights condition, but again the bootstrapped CIs for the two conditions overlap (Table 6). The Log-Rank Test further shows that there is no significant difference in the survival probability at any point in the study between the immediate and delayed insights conditions.

Thus, to answer our second research question: There are no significant effects of personalized insights on registration or activity in the app. Furthermore, the results show that promising immediate personalized insights does not strongly reduce dropout during the study.

5.3 What is the Impact of Contact Mode on Response?

To evaluate the effect of the contact mode on registration, activity, and dropout, we again first look at the response rates under the two conditions. Both the registration rate and the activity rates are higher in the interviewer condition compared with the mail condition (Table 7). A Chi-Square Test also shows significant associations between contact mode and registration, and contact mode and activity.

Because there is a significant association between contact mode and registration, and between contact mode and activity, we analyze the representativeness of the registered and active households in the interviewer and mail groups (Table 8). For the registered households, the representativeness is better in the mail group compared to the interviewer group. The confidence intervals for the two R-indicators still overlap, however. In the interviewer group, age, income and immigration background have larger effects on representativeness whereas the other variables have similar contributions. These findings remain true for the active households.

Fig. 3 shows the Kaplan-Meier survival curves of the dropout in the two contact mode conditions for the active households. Similar to Fig. 3, the largest drops are towards day 13. A Log-Rank Test shows no difference in survival probability at any point in the study between the interviewer and mail conditions (Table 9). This means that there is no difference in the rate at which households drop out of the survey between the interviewer and mail groups.

To answer our third research question, interviewer-based recruitment increases registration and activity in the HBS app, but the registered and active households in the mail group do give a better representation of the general population. However, the much lower participation rates in the

Table 10

Logistic regression analyses to show the relationships between personalized insights and data quality

	Model 1			Model 2			Model 3			Model 4			Model 5		
	Coef.	S.E.	*p*	Coef.	S.E.	*p*	Coef.	S.E.	*p*	Coef.	S.E.	*p*	Coef.	S.E.	*p*
Intercept	0.69	0.50	0.168	0.79	0.52	0.123	0.85	0.52	0.100	0.90	0.53	0.090	–0.92	0.53	0.086
Age (ref = 18–24)															
25–34	–0.80	0.53	0.129	–0.93	0.54	0.085	–0.94	0.54	0.082	–0.92	0.54	0.088	–0.92	0.54	0.088
35–44	–0.88	0.52	0.088	–0.94	0.53	0.077	–0.93	0.53	0.079	–0.91	0.53	0.087	–0.92	0.53	0.084
45–54	–0.77	0.52	0.141	–0.79	0.54	0.141	–0.79	0.53	0.140	–0.78	0.53	0.143	–0.80	0.54	0.137
55–64	–0.98	0.53	0.065	–1.10*	0.54	0.042	–1.10*	0.54	0.043	–1.08*	0.55	0.049	–1.09*	0.55	0.047
65–74	–1.05	0.56	0.062	–0.92	0.58	0.110	–0.91	0.58	0.113	–0.89	0.58	0.121	–0.90	0.58	0.120
≥75	–0.94	0.63	0.134	–1.00	0.65	0.127	–0.99	0.65	0.131	–0.98	0.65	0.136	–0.99	0.66	0.130
Household size (ref = 1 person)															
2 persons	0.45*	0.21	0.036	0.60*	0.23	0.009	0.60*	0.23	0.009	0.61*	0.23	0.008	0.61*	0.23	0.008
3 persons	0.09	0.23	0.685	0.17	0.24	0.470	0.19	0.24	0.436	0.20	0.24	0.413	0.20	0.24	0.419
4 or more persons	0.01	0.22	0.967	0.09	0.23	0.709	0.10	0.23	0.681	0.11	0.24	0.635	0.11	0.24	0.638
Quality indicators															
Amount SD	–	–	–	–0.10	0.05	0.057	–0.10	0.05	0.057	–0.10	0.05	0.066	–0.10	0.05	0.066
Products SD	–	–	–	–	–	–	–0.05	0.06	0.357	–0.05	0.06	0.351	–0.05	0.06	0.363
Store types	–	–	–	–	–	–	–	–	–	–0.02	0.05	0.686	–0.04	0.06	0.539
Entries	–	–	–	–	–	–	–	–	–	–	–	–	0.01	0.01	0.640
Model AIC	1033.1			928.1			929.22			931.1			932.8		
Log-likelihood	–506.54			–453.04			–452.61			–452.53			–452.42		

* $p < 0.05$

Table 11

Logistic regression analyses to show the relationships between contact mode and data quality

	Model 1			Model 2			Model 3			Model 4			Model 5		
	Coef.	S.E.	*p*	Coef.	S.E.	*p*	Coef.	S.E.	*p*	Coef.	S.E.	*p*	Coef.	S.E.	*p*
Intercept	0.57	0.73	0.438	0.48	0.74	0.512	0.43	0.75	0.567	0.62	0.78	0.424	0.58	0.79	0.461
Age (ref = 18–24)															
25–34	0.04	0.80	0.965	–0.15	0.81	0.857	–0.15	0.81	0.856	–0.12	0.81	0.879	–0.09	0.82	0.909
35–44	0.20	0.79	0.800	0.16	0.80	0.843	0.13	0.81	0.873	0.11	0.81	0.887	0.14	0.81	0.861
45–54	0.33	0.80	0.681	0.26	0.80	0.744	0.26	0.80	0.745	0.21	0.81	0.799	0.23	0.81	0.776
55–64	–0.04	0.78	0.956	–0.20	0.79	0.800	–0.22	0.79	0.785	–0.23	0.80	0.775	–0.20	0.80	0.805
65–74	0.15	0.81	0.851	0.10	0.82	0.907	0.09	0.82	0.911	0.10	0.82	0.905	0.12	0.83	0.882
≥75	0.33	0.85	0.702	0.08	0.87	0.930	0.06	0.87	0.945	0.02	0.87	0.985	0.05	0.88	0.952
Household size (ref = 1 person)															
2 persons	–0.44	0.30	0.135	–0.42	0.32	0.190	–0.45	0.33	0.169	–0.43	0.33	0.194	–0.43	0.33	0.188
3 persons	0.07	0.34	0.826	0.14	0.36	0.689	0.13	0.34	0.728	0.14	0.36	0.706	0.13	0.36	0.723
4 or more persons	–0.54	0.33	0.097	–0.51	0.35	0.144	–0.54	0.36	0.128	–0.50	0.36	0.166	–0.50	0.36	0.166
Quality indicators															
Products SD	–	–	–	0.11	0.11	0.305	0.12	0.11	0.288	0.11	0.11	0.300	0.11	0.11	0.314
Entries	–	–	–	–	–	–	0.01	0.02	0.658	0.02	0.02	0.350	0.02	0.02	0.344
Store types	–	–	–	–	–	–	–	–	–	–0.08	0.09	0.361	–0.09	0.09	0.348
Amount SD	–	–	–	–	–	–	–	–	–	–	–	–	0.02	0.08	0.778
Model AIC	498.7			456.92			458.78			459.88			461.8		
Log-likelihood	–239.33			–217.46			–217.36			–216.94			–216.90		

* $p < 0.05$

mail condition lead to higher risks of bias, so that a mixed picture emerges.

5.4 How is Data Quality Affected by Feedback and Contact Mode?

To answer research question 4, we model the data quality indicators listed in Sect. 4.3 for personalized feedback and for contact mode.

For personalized insights, we obtain the best fitting model when we only include amount SD as a quality indicator (Table 10). However, we want to know the relationships between insights and all quality indicators. Thus, the remaining quality indicators are still included based on the AIC. We find no significant relationships between the quality indicators and the insights condition (Table 10). However, we do find significant relationships between household size and the insights conditions for all 5 models, and between age and the insights condition in models 2, 3, 4, and 5. Households that consist of two persons are more likely to be in the immediate insights condition compared with single-person households. Households with a reference person between 55 and 64 years old are more likely to be in the delayed insights condition compared with households with a reference person who is 18–24 years old.

Because there are no significant relationships between the quality indicators and personalized insights, we analyze the relationships between the quality indicators and contact mode for all households. We obtain the best fitting model when we only include products SD as a quality indicator (Table 11). However, again we include the remaining quality indicators based on the AIC because we are interested in the relationships between all quality indicators and contact mode. There are no significant relationships between the quality indicators and contact mode (Table 11). Furthermore, we find no significant relationships between the household characteristics and contact mode.

Hence, there are no significant relationships between personalized insights and data quality, or between contact mode and data quality. Additionally, personalized insights are related to the household characteristics age and household size in our combined app-based sample.

6 Discussion

Despite smartphone apps being increasingly used for data collection, there are still a lot of questions about the use of apps to collect expenditure data. This study addresses several of those research questions. First, we investigate the relationships between household characteristics and registration and activity rates for the app-based Household Budget Survey, as well as representativeness of the samples. Second and third, we study the effects of personalized insights and contact mode on registration, active use of the app (activity), and dropout. Finally, we analyze the effects of personalized insights and contact mode on data quality.

We found registration, activity, and completion rates for the app-based sample that resemble those of the regular survey when no interviewers are used to recruit respondents. We found that education level and income were associated with registration and activity, and origin was related to activity. These findings are consistent with previous research (Jäckle et al., 2019; Struminskaya et al., 2021). Using R-indicators, we further showed that the app-based samples were not representative of the general population.

Contrary to our expectations, we found that promising households personalized insights about their expenditures did not increase registration and activity rates. These findings are in contrast with previous research (Bälter et al., 2005; Bälter et al., 2012; Marcus et al., 2007; Struminskaya et al., 2020), but they are in line with findings from a study on an app-based budget survey (Wenz et al., 2022). In line with our expectations, promising households immediate personalized insights did not increase dropout.

We further found that using interviewer assistance for recruitment strongly increased registration and activity rates for the app-based Household Budget Survey and did not increase dropout. These findings are partially in line with Jäckle et al. (2022), although they found that dropout was higher when interviewers were used for recruitment compared with an invitation by mail. However, there are differences in the way in which interviewers were used in our study compared with the Jäckle et al.'s (2022) study, which may partially explain why dropout did not increase compared with the mail group in our study. Since there was a significant effect of contact mode on registration and activity, we analyzed the representativeness of the interviewer and mail groups for the households that registered and were active in the app. We found that the mail group was more representative for both the registered and active samples, but does pose a larger risk of bias due to the much lower response rates.

Finally, we investigated the quality of the expenditure data. We used six quality indicators that we developed specifically for this purpose. We found that personalized feedback and contact mode were not related to the quality indicators we used in this study. Although we expected to find higher data quality when adding feedback or interviewer contacts, the absence of these findings might be a good sign for adaptive survey designs: Researchers can decide to use different data collection modes and interventions for different household types without this substantially affecting the quality of the expenditure data.

Two main limitations can affect the generalizability of results of our study. First, the results presented here are from one country, the Netherlands. While data has been collected in Luxembourg and Spain as well, the design differences were too substantial to draw strong conclusions and we had to exclude data from the other two countries. Second, the recruitment fieldwork was conducted in the aftermath of COVID-19 pandemic, making the interviewer contacts less extensive as potential participants were more cautious with face-to-face contacts. Interviewer debriefing revealed that interviewers had felt they had insufficient options to get rapport.

There are still some questions that need to be answered in future studies. First, we did not investigate how respondent burden differs between app-based and web-based budget surveys. Household budget surveys have become shorter over time to compensate for the decreasing response rates. If app-based surveys can help decrease respondent burden, the duration of budget surveys can be increased allowing us to collect more diverse expenditure data. Second, future research should investigate ways in which the representativeness of app-based samples can be improved. Furthermore, we were limited in our analysis of representativeness as statistical power was limited. Finally, given the limitations interviewers felt at the door, future research should explore more extensively the role of interviewers in recruiting and motivating households to stay active in the app in app-based studies.

Note Ethical approval for this study was obtained by the Ethics Review Board of the Faculty of Social & Behavioural Sciences at Utrecht University. The data used for this study are stored on the secure environment at Statistics Netherlands. Contact CBS to get access to the data. The scripts are stored on a closed repository. To access the scripts used for this study please contact the authors.

References

Bälter, K.A., Bälter, O., Fondell, E., & Lagerros, Y.T. (2005). Web-based and mailed questionnaires: "A comparison of response rates and compliance". *Epidemiology*, *16*(4), 577–579.

Bälter, O., Fondell, E., & Bälter, K.A. (2012). Feedback in web-based questionnaires as incentive to increase compliance in studies on lifestyle factors. *Public Health Nutrition*, *15*(6), 982–988.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. New Jersey: John Wiley & Sons Inc.

Bian, M., & Leung, L. (2015). Linking loneliness, shyness, smartphone addiction symptoms, andpatterns of smartphone use to social capital. *Social Science Computer Review*, *33*(1), 61–79.

Bietz, M., Patrick, K., & Bloss, C. (2019). Data donation as a model for citizen science health research. *Citizen Science: Theory and Practice*, *4*(1), 6.

Couper, M.P., & Ofstedal, M.B. (2009). Keeping in contact with mobile sample members. In P. Lynn((Ed.), *Methodology of longitudinal surveys* (pp. 183–203). New Jersey: Wiley.

De Leeuw, D. (2005). To mix or not to mix data collection modes in surveys. *Journal of OfficialStatistics*, *21*(2), 233–255.

De Bruijne, M., & Wijnant, A. (2014). Improving response rates and questionnaire design for mobileweb surveys. *Public Opinion Quarterly*, *78*(4), 951–962.

De Heij, V., Schouten, B., & Shlomo, N. (2015). *RISQ Manual 2.1*

Deutskens, E., De Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: an experimental study. *Marketing Letters*, *15*(1), 21–36.

DiCiccio, T.J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–228.

Durrant, G.B., Groves, R.M., Staetsky, L., & Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, *74*(1), 1–36.

Edwards, P.J., Roberts, I., Clarke, M.J., DiGuiseppi, C., Wentz, R., Kwan, I., & Pratap, S. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.MR000008.pub4.

Efron, B., & Tibshirani, R.J. (1994). *An introduction to the bootstrap*. CRC press.

Elevelt, A., Lugtig, P., & Toepoel, V. (2019). Doing a time use survey on smartphones only: what factorspredict nonresponse at different stages of the survey process? *Survey Research Methods*, *13*(2), 195–213.

Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, S., & Toepoel, V. (2021). Where you at? Using GPSlocations in an electronic time use diary study to derive functional locations. *Social ScienceComputer Review*, *39*(4), 509–526.

Eurostat *Household budget survey 2015 wave EU quality report*

Eurostat *Household budget surveys – overview*

Göritz, A.S., & Neumann, B.P. (2016). The longitudinal effects of incentives on response quantity inonline panels. *Translational Issues in Psychological Science*, *2*(2), 163.

Groves, R.M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. *Public Opinion Quarterly*, *72*(2), 167–189.

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

Haas, G.-C., Kreuter, F., Keusch, F., Trappmann, M., & Bähr, S. (2020). Effects of incentives in smartphone data collection. In C. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov & L.E. Lyberg (Eds.), *Big data meets survey science: A collection of innovative methods* (pp. 387–414). New Jersey: Wiley.

Harrell, F. Jr (2021). *Hmisc: Harrell Miscellaneous. R package version 4.6-0*

Jäckle, A., Burton, J., 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*(1), 23–44.

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 SurveyStatistics and Methodology*. https://doi.org/10.1093/jssam/smac006.

Jung, K., Lee, J., Gupta, V., & Cho, G. (2019). Comparison of bootstrap confidence interval methods for GSCA using a Monte Carlo simulation. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.02215.

Keusch, F., Struminskaya, B., Antoun, C., Couper, M.P., & Kreuter, F. (2019). Willingness toparticipate in passive mobile data collection. *Public Opinion Quarterly*, *83*(S1), 210–235.

Korbmacher, J.M., & Schroeder, M. (2013). Consent when linking survey data with administrativerecords: the role of the interviewer. *Survey Research Methods*, *7*(2), 115–131.

Kreuter, F., Haas, G.C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting survey andsmartphone sensor data with an app: opportunities and challenges around privacy and informed consent. *Social Science Computer Review*, *38*(5), 533–549.

Laurie, H. (2007). *The effect of increasing financial incentives in a panel survey: an experiment on the British household panel survey, wave 14 (no. 2007-05)*. ISER Working Paper Series.

Leeper, T.J. (2018). *margins: marginal effects for model objects*. R package version 0.3.25.

Leung, K.M., Elashoff, R.M., & Afifi, A.A. (1997). Censoring issues in survival analysis. *Annual Review of Public Health*, *18*(1), 83–104.

Link, M.W., Murphy, J., Schober, M.F., Buskirk, T.D., Childs, H.J., & Tesfaye, L.C. (2014). Mobile technologies for conducting, augmenting and potentially replacing surveys: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*(4), 779–787.

Lohr, S.L. (2019). *Sampling: design and analysis*. CRC Press.

Lord, V.B., Friday, P.C., & Brennan, K. (2005). The effects of interviewer characteristics on arrestees' responses to drug-related questions. *Applied Psychology in Criminal Justice*, *1*(1), 36–55.

Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys: a field experiment on nonresponse in web surveys. *Social Science Computer Review*, *25*(3), 372–383.

Pettersen, H. (2005). *Survey Design and Sample Design in Household Budget Surveys (Chapter XXV)*. Household Sample Surveys in Developing and Transition Countries, United Nations Statistics Division, Studies in Methods, Series F, Vol. 96 (pp. 557–570). United Nations Department of Economic and Social Affairs.

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

Peytchev, A. (2013). Consequences of survey nonresponse. *The ANNALS of the American Academyof Political and Social Science*, *645*(1), 88–111.

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Revilla, M., Couper, M.P., & Ochoa, C. (2019). Willingness of online panelists to perform additional tasks. *Methods, Data, Analyses*, *13*(2), 29.

Ryu, E., Couper, M.P., & Marans, R.W. (2006). Survey incentives: Cash vs. in-kind; face-to-face vs. mail; response rate vs. nonresponse error. *International Journal of Public Opinion Research*, *18*(1), 89–106.

Sahlqvist, S., Song, Y., Bull, F., Adams, E., Preston, J., & Ogilvie, D. (2011). Effect of questionnaire length, personalisation and reminder type on response rate to a complex postal survey: randomised controlled trial. *BMC Medical Research Methodology*, *11*(1), 1–8.

Sakshaug, J.W., Hülle, S., Schmucker, A., & Liebig, S. (2017). Exploring the effects of interviewer and self-administered survey modes on record linkage consent rates and bias. *Survey Research Methods*, *11*(2), 171–188.

Scherpenzeel, A., & Toepoel, V. (2014). Informing panel members about study results: effects of traditional and innovative forms of feedback on participation. In M. Callegaro, R.P. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick & P.J. Lavrakas (Eds.), *Online panel*

research: a data quality perspective (pp. 192–213). New Jersey: Wiley.

Schmidt, T. (2014). Consumers' recording behaviour in payment diaries–empirical evidence from Germany. In *Survey methods: insights from the field*.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*, 101–113.

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, *27*(2), 231–253.

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 112–141.

Struminskaya, B., & Bosnjak, M. (2021). Panel conditioning: Types, causes and empirical evidence of what we know so far. In: P. Lynn (Ed.) Advances in Longitudinal Survey Methodology. John Wiley & Sons. pp. 272–301. https://doi.org/10.1002/9781119376965.ch12

Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2020). Understanding willingness to share smartphone-sensor data. *Public Opinion Quarterly*, *84*(3), 725–759.

Struminskaya, B., Lugtig, P., Toepoel, V., Schouten, B., Giesen, D., & Dolmans, R. (2021). Sharing data collected with smartphone sensors: willingness, participation, and nonparticipation bias. *Public Opinion Quarterly*, *85*(S1), 423–462.

Therneau, T.M. (2022). *A package for survival analysis in R*. R package version 3.3-1.

Therneau, T.M., & Grambsch, P.M. (2000). *Modeling survival data: extending the cox model*. New York: Springer.

Toepoel, V., & Schonlau, M. (2017). Dealing with nonresponse: strategies to increase participation and methods for postsurvey adjustments. *Mathematical Population Studies*, *24*(2), 79–83.

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), 1–22.

Wenz, A., Jäckle, A., Burton, J., & Couper, M.P. (2022). The effects of personalized feedback on participation and reporting in mobile app data collection. *Social Science Computer Review*, *40*(1), 165–178.