Smart Survey Implementation

Grant Agreement Number: 101119594 (2023-NL-SSI)

Workpackage 2: Research Methodology

Deliverable M6: review stage

Version 1.3, 2023-10-30

Prepared by:

Hannah Bucher, University of Mannheim, Germany
Florian Keusch, University of Mannheim, Germany
Claudia de Vitiis, ISTAT, Italy
Fabrizio de Fausti, ISTAT, Italy
Francesca Inglese, ISTAT, Italy
Theun Pieter van Tienoven, Vrije Universiteit Brussel, Belgium
Danielle Mccool, Utrecht University, Netherlands
Peter Lugtig, Utrecht University, Netherlands
Bella Struminskaya, Utrecht University, Netherlands

ESSnet co-ordinator: Remco Paulussen

Workpackage Leader:

Peter Lugtig
Utrecht University
p.lugtig@uu.nl

INTRODUCTION TO DELIVERABLE

Smart surveys have emerged as a promising data collection method, bridging the gap between traditional survey techniques and modern technological advancements. The key characteristic of smart surveys is that they intelligently combine the use of asking questions (surveys through self-report) with smart features collected via sensors on smartphones, wearables and other devices. The goal of smart surveys is to improve data quality, reduce participant burden, provide more timely and more granular data, or combinations of these.

Over the past years, small-scale experiments have studied aspects around the design of smart surveys. This deliverable reports on the first stage of the Smart Survey Implementation (SSI) project; in particular on the question of how to develop an end-to-end research methodology for smart surveys, which is the overarching goal of Workpackage 2 of the SSI project. This deliverable has the goal to review smart surveys with the twin goal to learn about how to establish a successful methodology, and highlight gaps in our knowledge that will be addressed later in the SSI project.

This deliverable should be read in the context of other deliverables around the project coordination (workpackage 1), developing microservices for smart surveys (workpackage 3), the logistics of running smart surveys (workpackage 4), and ethical and legal issues (workpackage 5). This deliverable can, however, also be read as a standalone product.

Within the overall goal of developing a research methodology for implementing smart surveys from start to finish, we have identified four large pressing issues that prevent smart surveys from being implemented in the context of European Official Statistics data collection. These issues are:

1. How to successfully **recruit and retain** participants for smart surveys, taking into account difficult-to-reach groups in society.
2. How to use **machine learning models** to improve Human-Computer Interaction in smart surveys.
3. How to design smart surveys from a User Experience (UX) or **usability** perspective and involve respondents, and the **human-computer interaction** with sensor data after being processed by machine learning models.
4. How to integrate data from smart surveys with traditional survey methods by **estimating the mode effect** (that is, a difference due to the mode of administration being smart vs. traditional).

The rest of this deliverable is structured along these four main topics of the research methodology and separated into four chapters. Each chapter introduces the main problems that we face in the implementation of smart surveys, with a particular focus on the use cases of the European Time Use Surveys (TUS) and Household Budget Surveys (HBS). Later in the project, smart surveys around Time Use and Household Budget will be implemented in two platforms: the MOTUS platform that has been developed by Hbits, and the HBS-platform as developed by Statistics Netherlands. For an overview of how the apps look and feel, and how details on time use and household budget can be collected using a digital smartphone diary, we refer to the deliverable of Workpackage 1.

After reviewing findings from earlier ESSNET projects relevant for smart surveys and reviewing the wider literature from other national and international projects, each chapter will outline the main issues and what will be done in the SSI project to address and solve these issues.

In order to solve open questions, the project will carry out several small and larger field-tests in the period 2023-2025 that seek to test solutions in practice and provide evidence for best-practices using Randomized Controlled Trials.

In practice, it is quite likely that there are multiple successful methodologies for conducting smart surveys, that also depend on local circumstances. For example, in some countries, interviewers may play a big role in both recruitment and retainment for smart surveys (issue 1), but also in the usability of the app (issue 3), while other countries may for various reasons choose not to use interviewers. Countries may rely to a greater or lesser extent on traditional non-smart surveys in combination with smart surveys to produce official statistics (issue 4). Or, as a final example, the data available for training and re-training machine learning models in smart surveys may differ both between and within countries over time (issue 2). One of the final goals of this workpackage is to establish what combinations of smart survey designs work, and what types of combinations do not work. To account for differences between countries, we conduct field experiments and usability tests in multiple countries.

A final goal of this workpackage is to establish trade-offs between design features in smart surveys. One such trade-off is between recruitment and retainment (issue 1) and the mode-effect (issue 4). As an example, offering alternative data collection modes, such as web or paper diaries, next to smart surveys may potentially lead to higher response rates in the recruitment of surveys, but comes at the expense of differences in data across the modes (mode-effects: issue 4). The more alternative modes are offered, the more complex it will be become to estimate mode effects, and integrate data from multiple modes.

Another trade-off can be found between using machine learning models (issue 2) and usability of smart surveys (issue 3). One of the primary reasons for doing smart surveys is that we can measure things that respondents find impossible or very hard to answer (e.g. the start time of a time use activity or the exact expenses while shopping for groceries). When machine learning models work well, this should improve the usability of the response task for the respondent. For example, automaticaly classifying pictures from shopping receipts should lower burden for respondents and improve the quality of measurements. Should machine learning models however not perform well (e.g. because of low quality pictures or problems in classification of products) then the respondent may be presented with smart data that is ‘wrong’. When data from machine learning models requires manual correction by the respondents, this leads to usability problems (issue 3), and ultimately perhaps problems with retainment (issue 1).

It is the ultimate goal of this workpackage to also provide insights into these trade-offs by running field experiments that vary design aspects of smart surveys. An overview of the design of all tests will be published in summer 2024, with findings from all tests and recommendations for a research methodology for smart surveys released by the end of the project in spring 2025. This deliverable does not discuss trade-offs between design elements in detail, but instead focuses on earlier research into the issues of recruitment and retainment, machine learning, usability and the mode-effect in data integration in four separate chapters.

Utrecht, Mannheim, Rome, Brussels, 30 October 2023

The team of Workpackage 2 of the Smart Survey Implementation project

Chapter 1: Factors That Influence the Willingness to Participate in Smart Surveys: Implications for Recruitment Strategies in the SSI Project

1. INTRODUCTION

Smart surveys combine survey data collection with the collection of digital trace data through device sensors and applications (e.g., accelerometer, GPS, microphone, camera, etc.) (Bruno et al. 2022). Integrating the collection of these two data sources on one device, in many cases a smartphone, provides a powerful tool to gain new insights into the daily and social life of individuals since it allows linking information on attitudes and predispositions collected via self-report in surveys with behavioral data collected via sensors, apps, and other smart features of the devices (Struminskaya et al. 2020).

However, from the respondents' perspective, smart survey data collection constitutes a potential barrier to participation since it requires respondents to, for example, download an app on their smartphones to participate in the data collection process (Wenz, Jäckle, and Couper 2019a). Further, for collecting data, respondents must consent to collect digital data on their smartphones. These tasks pose challenges for participants in smart surveys: First, participation requires smartphone access. Second, it requires technical knowledge on how to download and install the app on the smartphone. Third, participation also requires the willingness to share data via the app.

With these tasks required for participants to take part in smart survey data collection, it is not surprising that participation rates for smart survey data collection are considerably lower than those achieved in more traditional data collection modes, such as face-to-face or web surveys (Keusch et al. 2022; Scherpenzeel 2017; Struminskaya et al. 2021). Smart survey data collection is particularly prone to bias arising from non-participation and non-consent (Keusch et al. 2019a). There is bias if sample members willing to participate in smart survey data collection differ from those who do not participate in terms of variables of interest measured in the data collection. For general population smart surveys that included an app, passive sensor data collection, and/or active data collection using features such as the camera, for example, to study expenditure, participation rates vary around 17% for an app download in the UK's Understanding Society Panel (Jäckle et al. 2019) or 24% in the app-based Household Budget Survey (HBS) field test in the Netherlands (Rodenburg et al. 2022).

Given the moderate to high non-participation and non-consent rates in smart survey data collection, a substantial challenge in utilizing smart survey data collection lies in increasing individuals' willingness to participate in the data collection process and decreasing potential

nonparticipation bias (Struminskaya et al. 2021). The research synthesis presented in this chapter organizes prior literature on factors that affect willingness to participate in smart surveys to inform decisions on survey design features in the SSI project. For this aim, we offer a comprehensive overview of prior research dedicated to the examination of factors affecting participation in smart surveys. We focus our review on the use of smartphones for smart survey data collection, because this is what the focus of the work in the SSI project will be.

Here, we concentrate on two critical dimensions: 1) the general willingness of individuals to participate in smart surveys using smartphone, encompassing actions such as downloading and installing survey applications, and 2) the specific tasks inherent to smart survey data collection that directly pertain to Time Use Surveys (TUS) and Household Budget Surveys (HBS). By delving into these dimensions, this synthesis aims to provide practical insights that can inform decision-making processes related to survey design within the SSI project, ultimately enhancing the effectiveness and quality of data collection in this context.

2. Studies Considered in this Review

Studies considered in this review come from multiple sources. Initially, we conducted extensive searches on Google Scholar, focusing on papers related to app data collection, app-based surveys, and smart surveys (search conducted in Sumer 2023). To complement these searches, we used AI-based literature searches through the Elicit platform (search conducted in 2023). Furthermore, we incorporated materials from previous ESSNET projects. We combined these sources into a database for this literature review which enhanced the comprehensiveness of our analysis and allowed us to focus specifically on the recruitment of respondents in the context of the SSI project. Taken together, we consider a total of 35 studies in this literature review.

Figure 1 gives an overview of the general participation rates for the Smart Surveys of the studies reviewed in this literature synthesis[1]. It further categorizes these participation rates according to employed recruitment strategies and sampling methods applied. We can see from this figure a substantial variation in participation rates among the utilized recruitment strategy, from over 80% participation rates for a university student sample (Assemi et al. 2018) to lower than 2% percent for respondents recruited via Facebook advertisement (Xu et al. 2016). We further see that many studies considered in this literature review are based on hypothetical willingness to participate in smart surveys (that is, do not require participants to use sensors or apps), so most of the studies are not measuring the actual participation in such surveys. Although we must be careful when transferring questions on hypothetical willingness to participate in smart surveys to actual participation behavior in smart surveys, we know from previous studies that hypothetical willingness to participate highly correlates with actual participation in smart surveys (Revilla, Couper, and Ochoa 2019). Hence, we chose to incorporate the findings of these studies into this review. However, it is evident that studies measuring hypothetical willingness to participate in smart survey data collection consistently yield higher participation rates compared to those based on observed behavior.

Our examination of actual participation in smart surveys reveals intriguing trends. When recruiting participants from established probability panels, the participation rates appear relatively stable and comparable, with the exception of one study that yielded remarkably high rates of around 35% (McCool et al. 2021). This is not surprising, given that people in panels have already committed to participate in survey data collection at an earlier point in time. However, a striking pattern emerges when we consider drawing new samples for studies out of the general population ("cold recruitment methods") for smart surveys — participation rates exhibit large fluctuations. This analysis sheds light on the wide range of participation rates in the context of smart surveys, emphasizing the need for further exploration and elucidation of the underlying factors contributing to these fluctuations.

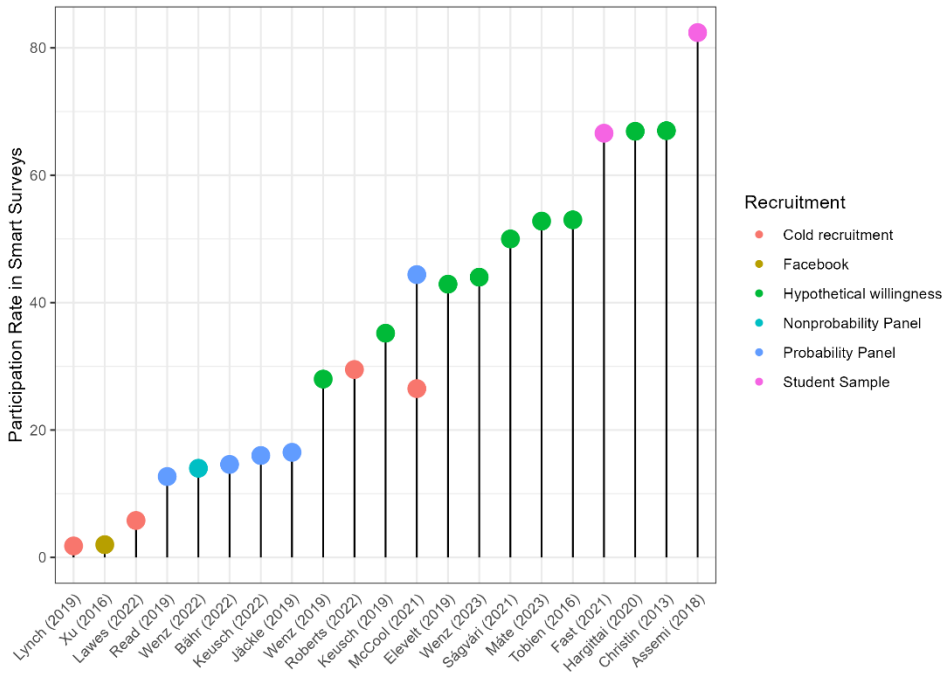

Figure 1: Participation Rates in Smart Survey Data Collection

3. Factors Influencing Participation in Smart Survey Data Collection

Prior research has predominantly classified factors influencing the willingness to participate in smart survey data collection into two categories. The first category comprises study characteristics that fall under the control of researchers, including aspects such as the design of the invitation to participate in the data collection process, providing vs. not providing a landing page, incentives, etc. The second category involves respondents' characteristics that cannot be directly influenced by researchers but significantly impact their willingness to participate in smart survey data collection, such as technical knowledge or privacy concerns (Keusch et al. 2023).

In this literature review, we aim to juxtapose these two factors. The objective of this comparison is to derive practical solutions for optimizing recruitment for the large field tests in the SSI project. This involves a specific focus on the distinct requirements of the TUS and HBS surveys, explicitly specifying which tasks, that is, the type of data collection (passive, that is, without actions by respondents vs. active, that is when actions by respondents are required), each respective finding pertains to. In presenting this review, we focus on studies that allow us to directly measure the influence of survey design features and respondent characteristics on recruitment success, either through an experimental study design or observational methods. We refrain from comparing features across studies because of potential confounding.

4. Survey Design Features

In this chapter, our primary focus is on survey design features. Survey design features refer to aspects of the study design that can be modified or adapted by researchers, and these modifications can have an impact on participation. We selected the features for inclusion in this review based on three key considerations: First, we drew upon insights derived from theoretical frameworks of survey participation in general. Second, we incorporated features unique to Smart Survey data collection. Third, our analysis encompasses features that are particularly relevant within the special context of SSI surveys, such as TUS and HBS.

Survey Sponsor

While the term survey sponsor is not strictly defined in the survey literature, it usually means the entity that is presented to the participant as who commissioned the study. When investigating the influence of the survey sponsor on participation rates in surveys in general, it is anticipated that the identity and perceived reputation of the sponsoring organization could play a role in shaping respondents' willingness to participate. Positive feelings towards a survey's sponsoring organization can be strengthened when the organization provides benefits to individuals (i.e., an unconditional incentive), prompting a desire to reciprocate. This reciprocal behavior can lead to intangible benefits such as a sense of social solidarity or fulfilling civic duty. When the survey sponsor is highly regarded, emphasizing their sponsorship during the survey request can boost participation rates (Groves et al. 2012; Klingwort, Bakker, and Toepoel 2023).

In the context of smart surveys, we assume trust in the survey sponsor plays a more important role than in non-smart surveys with regard to the decision to participate. Given the advanced technology and potential data integration involved, participants need assurance that their data will be handled ethically and securely. A reputable sponsor inspires confidence, increasing the willingness of individuals to participate in smart surveys (Hargittai et al. 2020; Keusch et al. 2023). Of course, people invited to a smart survey find reputable as a sponsor might heavily depend on the country and the context of the study (i.e., topic).

Empirically, the question of the influence of the survey sponsor on participation in smart surveys has been examined in several studies. However, to date, results have only been based on the investigation of hypothetical willingness to respond (Keusch et al. 2023; Christin, Buchner, and Leibecke 2013; Guo 2022; Hargittai et al. 2020; Struminskaya et al. 2020). There are no studies relying on actual participation behavior, which is connected to the difficulty to vary the sponsoring organization in the real-world settings. Table 1 summarizes the findings from existing literature. Note that all the studies summarized in this table included hypothetical questions about the participants' willingness to participate in smart surveys.

Table 1: Experimental evidence for sponsor of the survey

Study	Task	Study Feature	Respondents	Country	Measurement	Result
Guo et al 2022	Downloading and installing app	Governmental vs. commercial sponsor	Random sampled app users	China	Hypothetical willingness	Governemental sponsor prefered over commercial sponsor
Hargittai 2020	Downloading and installing app	number of sponsors mentioned	Members of a nonprobability online panel	USA	Hypothetical willingness	One sponsor achieved highest participation rate
Hargittai 2020	Downloading and installing app	Health agency vs. governmental vs university sponsor	Members of a nonprobability online panel	USA	Hypothetical willingness	Health agency prefered over governmental sponsor
Keusch et al 2023	Sensor data collection	University vs. statistical agency vs market research company sponsor	Members of a nonprobability online panel	Germany	Hypothetical willingness	University as sponsor prefered over statistical agency and commercial sponsor
Christin et al 2013	Sensor data collection	University vs. statistical agency vs market research company sponsor	Students	Germany	Hypothetical willingness	University as sponsor prefered over statistical agency and commercial sponsor
Christin et al 2013	Sensor data collection	Previous relationship with sponsor	Students	Germany	Hypothetical willingness	Previous relationship improved willigness to participate

We first examine findings from previous studies regarding the willingness to download and install an app. Research from China indicates that a governmental sponsor tends to elicit higher willingness to participate compared to a commercial survey sponsor (Guo 2022). Furthermore, one study reveals that mentioning multiple sponsors tends to result in lower willingness to participate compared to mentioning only one sponsor (Hargittai et al. 2020).

Concerning the willingness to share sensor or other passively collected data, it is observed that universities, as sponsors, perform comparatively better in three studies than statistical agencies or market research companies (Christin, Buchner, and Leibecke 2013; Keusch et al. 2023; Struminskaya et al. 2020). Noteworthy a prior relationship with the survey sponsor also has a positive impact on willingness to participate (Christin, Buchner, and Leibecke 2013).

Applying these findings to the question of what survey sponsor should be mentioned in the large field tests of the SSI projects, these results suggest that in countries where both universities and statistical agencies are involved in data collection, universities should be listed as the survey sponsor, especially if there is no pre-existing connection or trust with the statistical agency. This recommendation holds particular weight when sensor data are being collected (Christin, Buchner, and Leibecke 2013; Keusch et al. 2023). However, since previous findings are solely derived from hypothetical willingness to respond, future research is necessary to ascertain whether these findings also extend to actual participation in smart surveys. It also might be the case that the sponsor effect is affected by the type of organization that is fielding the study. Keusch et al. (2019), for example, found that in a study fielded by a market research agency the willingness to participate was higher for the market research agency than official statistics agency, both of which followed the highest willingness for university sponsor. On the other hand, Struminskaya et al. (2020) in a study within a probability-based panel housed at a university the willingness to participate in sensor-based data collection was higher for an official statistics office than market research agency, both of those also following the highest willingness associated with university sponsorship.

INCENTIVES

Incentives are traditionally applied in survey research to boost participation (Klingwort, Bakker, and Toepoel 2023). In the context of the SSI project, two questions regarding the development of effective incentive strategies emerge. First, an appropriate incentive strategy should be developed to generally increase participation rates. Second, there is a distinct need to deliberate on the utilization of incentives to specifically enhance the willingness of individuals to share passive data or grant access to their device's camera potentially for longer periods of time. Crafting incentives tailored to address this specific dimension of participation is integral to optimizing data collection within the SSI project.

Table 2: Experimental evidence on incentives

Study	Task	Study Features	Respondents	Country	Measurement	Result
Jackle et al 2019	Downloading and installing app	2 vs. 6 pounds	Respondents recruited from an existing probability panel	UK	Actual behavior	No difference
Haas et al 2020	Downloading and installing app	10 vs 20 euro	Respondents recruited from an existing probability panel	Germany	Actual behavior	Higher incentive led to higher participation rate
McCool et al 2021	Downloading and installing app	5+5 vs 0+10 vs 0+20 euro	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	0 + 20 resulted in highest participation rate, 5 + 5 lowest participation rate
Kesuch et al 2023	Sensor data collection	0 vs 10 vs. 20 euro	Respondents recruited from an existing nonprobability panel	Germany	Hypothetical willingness	10/20 euro compared to 0 Euros improved willingness to download/install app
Christin et al, 2013	Sensor data collection	0 vs. some incentive	Respondents recruited from anonprob online panel	Germany	Actual behavior	No difference
Haas et al 2020	Sensor data collection	1 euro per task vs 1 euro per task and additional 5 euro for activating all 5 tasks	Students	Germany	Actual behavior	additional bonus incentive of 5 euro did not increase participation in <> tasks

Concerning the goal of improving participation rates in smart surveys, previous research demonstrates that most studies employ post-paid (conditional) incentives (see Table 2). As a general trend, providing a monetary incentive tends to elicit a higher willingness to participate compared to instances where no incentive is offered (Keusch et al. 2022, 2023). Findings regarding whether a higher amount of incentive leads to higher downloading and installing rates of the app mostly show that higher incentives lead to higher downloading and installation rates (Haas et al. 2020; McCool et al. 2021), only one study considered in this review did not find a relation between the total amount of the incentive and

downloading and installing the app (Jäckle et al. 2019). Further, one interesting finding in this regard is that one study found a significant effect of increasing the initial incentive on respondents' willingness to participate in additional *smart tasks* of data collection (Haas et al. 2020). With this, the amount of the initial incentive may, next to participation rates also have an impact on the willingness to participate in sensor/app data collection.

Further, some studies investigated whether response rates were affected by offering additional incentives for participating in *smart tasks* of data collection, such as sharing passively collected data, next to the initial incentive. These studies showed mixed results. While one study found that most respondents are willing to take part in additional tasks of data collection for free (Christin, Buchner, and Leibecke 2013), another study found that offering respondents an incentive for each *smart task* completed increases their willingness to participate in sensor data collection (Keusch et al. 2023).

When relating these findings to the question of how to incentivize respondents in the large field tests of the SSI project, we would recommend using an initial incentive to improve general participation in data collection. Ideally, this would be an unconditional small monetary incentive, but we are aware that this might not be feasible for many NSIs. Further, the amount of the incentive may also have an impact on the willingness to participate in smart tasks of data collection. However, based on this review no clear recommendation can be made regarding whether additional incentives for different smart tasks of data collection should be employed to boost participation in these features of data collection. Another important point of investigation for the SSI project is the use of incentives to increase engagement. So far, only a few studies on this exist (e.g., Haas et al. 2020) and these do not provide definitive findings on which recommendations can be based. No studies have investigated whether incentives are related to privacy concerns (e.g., for people with high privacy concerns incentives do not work), which would be an avenue for the investigation in the SSI.

Duration/Length of Data Collection

Another aspect that has been studied in relation to participation in smart survey data collection is the duration of data collection. While there is extensive research investigating the effect of survey length on participation in non-smart surveys, there is limited previous research directly related to the duration of the entire data collection period in smart surveys (Keusch et al. 2019a; Máté et al. 2023; Ságvári, Gulyás, and Koltai 2021; Becker et al. 2015; see Table 3). However, from a theoretical perspective, we would expect shorter data collection duration to have a positive impact on overall participation in smart surveys. Shorter duration is associated with reduced respondent burden, as well as lower time and effort costs compared to longer data collection periods (Keusch et al. 2019a).

Table 3: Duration of data collection

Study	Task	Study Features	Respondents	Country	Measurement	Result
Keusch et al 2019	Sensor data collection	1 vs. 6 months	Members of a nonprobability online panel	Germany	Hypothetical willingness	Shorter duration improved willingness to share passive data
Mate et al 2023	Downloading and installing the app	1 vs. 6 months	Members of a nonprobability online panel	Hungary	Hypothetical willingness	Shorter duration improved willingness to install and download the app
Sagvari et al 2021	Passive data collection	1 vs. 6 months	Members of a nonprobability online panel	Hungary	Hypothetical willingness	Shorter duration improved willingness to share passive data
Becker et al 2015	Active usage of app (incl. smart features)	Time until dropout from the study	Registered users of a medical app	Germany	Actual behavior	Higher break-off rates for younger participants compared to older participants

When examining previous investigations of data collection duration, a consistent trend emerges, indicating that as the duration of data collection extends, the likelihood of participating in app-based data collection decreases (Keusch et al. 2019a; Máté et al. 2023; Ságvári, Gulyás, and Koltai 2021; Becker et al. 2015). However, these results are primarily based on studies comparing a one-month duration versus a six-month duration and stem from studies utilizing data from nonprobability online panels, which asked about hypothetical willingness to participate in smart surveys (Keusch et al. 2019a; Máté et al. 2023; Ságvári, Gulyás, and Koltai 2021).

Based on these findings, we suggest that the SSI project limits data collection periods to not more than two weeks, if possible, to boost participation rates in the large field tests.

Giving Respondents Control Over the Data Collection Process

Granting respondents autonomy over data collection is a crucial aspect directly linked to gathering sensor/app data in Smart Surveys. The general hypothesis is that providing

respondents with control over the data collection process, either by allowing them to temporarily deactivate data collection or by letting them choose the specific data they wish to share with researchers, leads to a higher willingness to share these data (Keusch et al. 2023; Schaewitz, Winter, and Krämer 2021; Struminskaya et al. 2020; Struminskaya et al. 2021). This approach enhances trust in the data collection process (Bemmann et al. 2022; Schnorf, Ortlieb, and Sharma 2014). A counter hypothesis would be that too much choice reduces the likelihood of participation ("choice overload").

Among previous investigations (see Table 4), divergent outcomes surface concerning distinct task types. Findings are mixed regarding respondents' willingness to partake in general tasks with limited control. While one study found a reduced inclination to partake in general tasks with limited control, such as GPS tracking, compared to tasks where they have a measure of control, such as taking photographs (Revilla, Couper, and Ochoa 2019), another study revealed completely different findings: While most of the respondents were willing to share their GPS location, only few of them were also willing to share personal photos (Struminskaya et al. 2021). Struminskaya et al. (2020; 2021) conducted a randomized experiment in which they either emphasized the ability of respondents to control data sharing in the request or provided no text about the possibility to review and revoke the measurement. Mentioning control significantly affected the actual sharing (2021) but not the hypothetical willingness (2020).

Table 4: Control over the data collection process

Study	Task	Study Features	Respondents	Country	Measurement	Result
Keusch et al 2019a	Sensor data collection	Option to switch off sensor data collection	Members of a nonprobability online panel	Germany	Hypothetical willingness	Option to switch off data collection improved willingness to share data
Struminskaya et al 2021a	Different smart data collection tasks	Option to switch off data collection	Individuals who participated in at least one Statistics Netherlands Survey	Netherlands	Hypothetical willingness	Option to switch off data collection improved willingness to share data
Bemmann2021	Sensor data collection	Different control features	Students and Employees of IT companies	Germany	Hypothetical willingness	Option to disabling data logging improved willingness to share data the most
Schaewitz 2021	Sensor data collection	Option to switch off sensor data collection	Convenient sample of facebook users	Germany	Hypothetical willingness	Option to switch off data collection

						improved positive evaluation of app, but not willingness to share data
Revilla et al 2019	Different smart data collection tasks	Tasks that differ in control over data collection for respondents	Members of a nonprobability online panel	Spain	Hypothetical willingness	willingness is higher for tasks where respondents have control over the reporting of the results than for passive tracking behaviors
Schnorf et al 2014	Sharing google profile information	Option to adjust collected data	Members of nonprobability online panel	USA	Hypothetical willingness	Offering respondents more control increases trust in data collection,but only for respondents who care about the data they shared
Elevelt et al 2019	GPS and call data collection	Option to switch off sensor data collection	Respondents recruited from an existing probability panel	Netherlands	Actual behavior	74.7 percent took part in diary study and 55.4 percent shared their GPS data although they had the possibility to turn switch off data collection

It is worth noting that allowing respondents to temporarily suspend data collection has been observed to increase willingness and enhance participants' perceptions of the app itself (Keusch et al. 2019b, 2023; Schaewitz, Winter, and Krämer 2021; Bemmann et al. 2022). However, one study suggests that the value of control over data collection varies among different types of respondents (Schnorf, Ortlieb, and Sharma 2014), being a valuable feature to increase trust and participation in smart data collection tasks only for those respondents who are concerned about the data they share (Schnorf, Ortlieb, and Sharma 2014).

Regarding concerns about allowing respondents to temporarily disable passive data collection, previous research does not support this being a widespread issue: In a study conducted by Elevelt, Lugtig, and Toepoel (2019), 74% of participants continued sharing their GPS data even though they had the option to turn it off even if they had given initial consent.

Applying these findings to the surveys conducted within the SSI project, we recommend providing respondents with control over the data collection process. Allowing them to temporarily pause data collection can enhance trust in the data collection process and, consequently, improve respondents' perception of participating in the survey.

Transparency on the Data Collection Process

In addition to giving respondents control over the data collection process, another feature discussed in previous studies in the context of (successful) participant recruitment for smart surveys is the transparency of the data collection process. Here, the general assumption is that providing respondents with transparent information about the data collection process may increase trust in the smart survey. Furthermore, when setting up a smart survey, researchers must align with GDPR requirements for transparency. Thus, providing respondents with transparent information about the data collection process is, at least to a certain degree, obligatory (Kreuter et al. 2020).

Table 5: Transparency of the data collection process

Study	Task	Study Features	Respondents	Country	Measurement	Result
Farke et al 2021	Google data collection	Evaluation of google's transparency tool (MyActivity)	Members of a nonprobability online panel	USA	Evaluation after using the tool	Transparency tool decreases concerns about data collection practices
Van Kleek et al 2021	Using smart features in app	Testing different privacy information interfaces against each other	Students	UK	Discussion in Lab	Transparency increases willingness to share data for people who do not like to be tracked
Tsai et al 2011	Use of commercial purchase apps	Different levels of privacy policy	General population survey, recruitment via flyers	USA	Hypothetical willingness in survey + online shopping experiment	Privacy information increases likelihood to purchase

Regarding how transparent privacy information about data collection impacts respondents' willingness to participate in smart surveys, most studies (see Table 5) have found a positive effect of transparency on concerns related to smart data collection (Farke et al. 2021; Van Kleek et al. 2017). Concerning actual participation behavior in smart surveys, one study discovered that providing respondents with transparent privacy information increases the usage of smart data collection tasks (Tsai et al. 2011).

Based on these findings, we highly recommend providing respondents with transparent information about the data collection process in the surveys conducted within the SSI project.

App vs. Browser-Based Data Collection

When administering smart surveys, researchers have the choice to do so via a designated survey app that participants need to download to their smartphone or let participants complete thesurvey via a mobile web browser. A variant of the web-based approach is so called (progressive) web apps, which basically mimic the look and feel of a full-blown app but still open in the (mobile) browser of the participants device (see Buskirk and Andres (2012) for an overview of these approaches). One of the main advantages of the two web-based approaches is that survey participants do not have to download an app to their device – potentially reducing the participation burden for people who do not trust an app or do not (feel that they) have the technical skills to download an app. They also allow people who do not have a smartphone or do not want to use their smartphone for data collection to still participate in the survey using their desktop or laptop computer. Finally, web-based approaches are independent of the operating system of the device they run on, that is, other than for apps, not a separate version for Android and iOS needs to be deployed. The app-approach, on the other hand, allows a better integration of smart elements into the survey, taking advantage of the full range of smart features of a smartphone (e.g., sensors, camera). Apps also can operate for longer time periods without an active Internet connection. Finally, they do allow to actively contact the participants via messages directly from within the application, for example, by sending push notifications to remind participants to complete the diary. This feature makes apps very attractive for intensive longitudinal data collection, such as TUS and HBS.

One of the few studies that empirically compared app- and browser-based approaches is Roberts (2022). The study found that the sample assigned to the browser condition yielded a 4 percentage points higher response rate compared to the sample assigned to the app. As has been shown in other contexts, self-selection into one of the two approaches skewed the sample who uses the app to be younger and more tech savvy compared to the browser, as people could also use a browser on a desktop or laptop computer and were not bound to a smartphone (see section on respondent characteristics below). Interestingly, among all people who used a smartphone to complete the survey, respondents with the app

compared to responding on a mobile browser reported significantly lower subjective burden. Using the app was also positively correlated with higher likelihood to participate in future waves of the study.

Based on the few findings in the literature, we suggest using an app for data collection in TUS and HBS, which allows a seamless integration of smart survey features into the survey and more direct interaction directly through the app with the participant throughout the data collection period (assuming days of data collection >1). In addition, NSI will need to decide whether they offer participants a browser-based alternative to be more inclusive and allow people who do not want to or cannot download an app to participate. Ideally, this would be done with a progressive web app, to reduce potential measurement error due to different instruments being used.

5. RESPONDENTS' Characteristics

In this first part of the literature review, we focused on study design features that should be considered when setting up a smart survey. Altogether, this chapter shows that design decisions influence the willingness to participate in smart surveys. However, what should also be considered when setting up a smart survey is that respondents recruited for smart surveys bring with them predispositions and characteristics that inform their decision on whether to participate. While we cannot directly influence these characteristics, it is important to be aware of them as they substantially contribute to the willingness to participate. Further, these predispositions can be addressed in the invitation to participate in a smart survey. Therefore, the second part of this review provides an overview of respondents' characteristics that have been addressed in previous research to inform the decision-making process of participating in smart surveys.

SMARTPHONE USAGE BEHAVIOR

For smart surveys, a relevant respondent characteristic that can influence the participation decision is how familiar the respondents are with the technology that is to be used for data collection. Thus, it is not surprising that a substantial body of research has explored the connection between general smartphone usage behavior and engagement in smart surveys that use smartphones for data collection. Participation in smart surveys with smartphones requires meeting two technical prerequisites: First, having a smartphone available to take part in the survey, and second, possessing the skills to navigate the survey process. Smartphone usage behavior thus plays a pivotal role in determining participation in smart surveys (Wenz, Jäckle, and Couper 2019b; Keusch, Wenz, and Conrad 2022).

Having a smartphone available for downloading and installing the survey app is a necessary condition for participation in a smart survey. This criterion systematically excludes individuals who do not own a smartphone from participating. While smartphone ownership is on the rise, certain demographic subgroups may remain inadequately reached, and which groups are underrepresented depends on the country. Some evidence (see Table 6) suggests that individuals who can be reached via smartphone differ systematically from

those who do not own such a device (Couper, Antoun, and Mavletova 2017; Couper 2017; Keusch et al. 2021). Therefore, the effect of collecting data via smart surveys on the overall quality of survey data remains an open question (Keusch et al. 2023).

Table 6: Smartphone usage behavior

Study	Task	Characteristics	Respondents	Country	Measurement	Result
Hargittai 2020	Downloading and installing app	General internet usage skills	Members of nonprobability online panel	USA	Hypothetical willingness	Individual with more general Internet usage skills are more likely to download and install the app
Jackle et al 2019	Participation in app based data collection	Access to mobile technologies	Members of probability panel	UK	Actual behavior	Using the Internet everyday and owning a smartphone positive affect participation in app based study
Jackle et al 2019	Participation in app based data collection	Ability to use mobile technologies	Members of probability panel	UK	Actual behavior	Frequency of mobile device use positive affect participation in app based study
Jackle et al 2019	Participation in app based data collection	Willingness to use mobile technologies	Members of probability panel	UK	Actual behavior	Willingness to download an app positive affect participation in app based study
Oyibo et al 2022	Downloading and installing the app	Perceived ease of use and compatibility with smartphone skills	Members of nonprobability online panel	Canada	Hypothetical willingness	Perceived ease of use and compatibility with smartphone skills do not affect likelihood to download and install the app
Revilla et al 2023	Different smart tasks of data collection	Frequency of Internet usage	Members of nonprobability online panel	Spain	Hypothetical willingness	Frequency of Internet usage does not have an impact of willingness to perform <> tasks of data collection
Wenz et al 2023	Participation in app based data collection	Frequency of Smartphone use	Members of probability panel	USA	Hypothetical willingness	Frequency of Internet usage does not have an impact of willingness to participate in app based data collection
Wenz et al 2023	Participation in app based data collection	Number of Smartphone activities	Members of probability panel	USA	Hypothetical willingness	Number of smartphone activities does have an impact of willingness to participate in app based data collection

Wenz et al 2019	Different smart tasks of data collection	Device familiarity	Members of probability panel	UK	Hypothetical willingness	Number of smartphone activities does have an impact of willingness to participate in app based data collection, frequency of use and self-reported smartphone skills not
Wenz et al 2019	Different smart tasks of data collection	Mobile device specifications	Members of probability panel	UK	Hypothetical willingness	Smartphone contract specification does not have an impact on willingness to participate in app based data collection

Even if individuals own a smartphone, their unfamiliarity with the tasks required for participating in a smart survey, such as downloading and installing apps, taking photos, or using sensors, can increase the burden of participation. This trend is also observed in studies examining how familiarity with smartphones relates to participation in smart surveys. Previous studies have explored various aspects of smartphone usage and their association with willingness to participate in smart surveys. These investigations suggest that both the frequency of smartphone usage (Jäckle et al. 2019; Revilla, Couper, and Ochoa 2019; Wenz and Keusch 2023) and the diversity of its utilization (Keusch, Wenz, and Conrad 2022; Revilla, Couper, and Ochoa 2019; Jäckle et al. 2019; Wenz, Jäckle, and Couper 2019b; Struminskaya et al. 2020; Struminskaya et al. 2021) have an effect on the willingness to participate in Smart Surveys. Additionally, the perception of the smart survey itself plays a role; when perceived as easy to complete, it positively influences willingness to participate (Oyibo and Pelegrini Morita 2022).

Regarding practical implications of these findings for the SSI projects, we recommend reducing the participation burden for people less familiar with smartphones and those who have general lower digital literacy by simplifying the process of installing and downloading the app to the largest possible extent (i.e., considering legal and ethical requirements for consent). Providing a progressive web app in addition or instead of the app would accomplish such a goal. Furthermore, it should be explicitly stated in the survey invitation for the large field tests that participation in this survey does not require specific knowledge.

Privacy Concerns

One attitudinal barrier to participation in smart surveys poses concerns about privacy, potentially reducing the likelihood of participation in a smart survey. Especially about the collection of sensor data (Keusch et al. 2020), concerns about privacy may prevent individuals from participating in data collection. Findings from previous studies (see Table 7) consistently show that privacy concerns are a significant predictor of nonparticipation in smart surveys (Keusch et al. 2020; Oyibo and Pelegrini Morita 2022; Wenz and Keusch 2023; Revilla, Couper, and Ochoa 2019; Wenz, Jäckle, and Couper 2019b). These concerns matter

especially for individuals who previously perceived a violation of their privacy in the online realm (Wenz and Keusch 2023), due to a lacking trust in the data collection organization.

Table 7: privacy concerns

Study	Task	Characteristics	Respondents	Country	Measurement	Result
Keusch et al 2020a	Different smart tasks of data collection	general privacy concerns	Findings from four different surveys	Germany and Austria	Hypothetical willingness	High general privacy concerns go ahead with higher concerns regarding all different <> tasks of data collection
Oyibo et al 2020	Downloading and installing the app	Privacy concerns, perceived trust and risk of the app	Members of nonprobability online panel	Canada	Hypothetical willingness	Privacy concerns impacts willingness to participate in app-based data collection for all users, trust especially important for low experienced app users
Wenz et al 2023	Participation in app based data collection	Security concerns regarding research apps	Members of probability panel	USA	Hypothetical willingness	Security concerns regarding research apps have a negative impact on willingness to participate
Wenz et al 2023	Participation in app based data collection	Perveiced privacy violation	Members of probability panel	USA	Hypothetical willingness	Perveiced privacy violation online have a negative impact on willingness to participate
Wenz et al 2023	Participation in app based data collection	Trust in organization not to share data	Members of probability panel	USA	Hypothetical willingness	Trust in organization has a positive impact on willingness to participate

Wenz et al 2019	Different smart tasks of data collection	Privacy and security concerns	Members of probability panel	UK	Hypothetical willingness	Security concerns have an impact on the willingness to perform all different tasks of <> data collection
Revilla et al 2019a	Different smart tasks of data collection	Lack of trust	Members of nonprobability online panel	Spain	Hypothetical willingness	Lack of trust has an impact on the willingness to perform all different tasks of <> data collection

With privacy concerns potentially acting as a barrier to participation in smart surveys, it is important for the SSI project to consider possible privacy concerns that participants in our studies may have and try to address them in the survey invitations. A key task of the large field tests will be to identify how to best offer respondents clear and transparent but at the same time concise information regarding privacy in the survey invitation.

SOCIODEMOGRAPHIC CHARACTERISTICS

Among the most studied respondents' characteristics about participation in smart surveys are sociodemographics. Since sociodemographic variables are often available for both respondents and nonrespondents, a lot of studies report sociodemographic differences between individuals participating in smart surveys and those who do not.

Within the domain of sociodemographic characteristics, some trends emerge that can consistently be found in previous studies: younger individuals and those with higher levels of education display a heightened likelihood of engaging in smart survey data collection. Further, sample members with an immigration background are less likely to participate in smart surveys. These patterns are discernible in both theoretical willingness surveys (Christin, Buchner, and Leibecke 2013; Hargittai et al. 2020; Keusch et al. 2023) and analyses of actual participation in app-based data collection tasks (Keusch et al. 2022; McCool et al. 2021; Lynch et al. 2019). The influence of gender yields less homogeneous findings. While one study finds a higher likelihood of participation among male participants (Keusch et al. 2023), another study contradicts this by identifying a greater predisposition for participation among female individuals (Jäckle et al. 2019). Furthermore, additional sociodemographic characteristics have been analyzed in previous studies in relation to participation in smart surveys. Table 8 provides an overview of all sociodemographic features and the effect on survey participation in smart surveys.

Table 8: Sociodemographic characteristics

Study	Task	Characteristics	Respondents	Country	Measurement	Result
Hargittai et al 2020	Downloading and installing app	Education	Members of nonprobability online panel	USA	Hypothetical willingness	Individuals with higher education more willing to install the app
Jackle et al 2019	Participation in app based data collection	Gender	Members of probability panel	UK	Actual behavior	Women more likely to participate
Keusch et al 2022a	Participation in app based data collection	Age	Members of probability panel	Germany	Actual behavior	Older people are less likely to participate in app based data collection
Keusch et al 2022a	Participation in app based data collection	Citizenship	Members of probability panel	Germany	Actual behavior	German citizens are more likely to participate in app based data collection
Keusch et al 2022a	Participation in app based data collection	Community size	Members of probability panel	Germany	Actual behavior	No difference
Keusch et al 2022a	Participation in app based data collection	Marital status	Members of probability panel	Germany	Actual behavior	No difference
Keusch et al 2022a	Participation in app based data collection	household size	Members of probability panel	Germany	Actual behavior	No difference
Keusch et al 2022a	Participation in app based data collection	presence of children	Members of probability panel	Germany	Actual behavior	No difference
Keusch et al 2022a	Participation in app based data collection	household income	Members of probability panel	Germany	Actual behavior	No difference
Keusch et al 2022a	Participation in app based data collection	employment status	Members of probability panel	Germany	Actual behavior	No difference
Keusch et al 2022a	Participation in app based data collection	Welfare receipt	Members of probability panel	Germany	Actual behavior	No difference
Keusch et al 2023a	Sensor data collection	Gender	Members of nonprobability panel	Germany	Hypothetical willingness	Male respondents are more willing to participate in sensor data collection

Keusch et al 2023a	Sensor data collection	Age	Members of nonprobability panel	Germany	Hypothetical willingness	Older respondents are less willing to participate in sensor data collection
Keusch et al 2023a	Sensor data collection	Education	Members of nonprobability panel	Germany	Hypothetical willingness	High educated respondents are more willing to participate in sensor data collection
Keusch et al 2020a	Downloading and installing app	Gender	Findings from four different surveys	Germany and Austria	Hypothetical willingness	No difference
Keusch et al 2020a	GPS data collection	Gender	Findings from four different surveys	Germany and Austria	Hypothetical willingness	Women are more concerned about sharing GPS data
Lynch et al 2019	Participation in app based data collection	Household size	Random Sample	USA	Actual behavior	One household member respondents are more likely to participate in app data collection
Lynch et al 2019	Participation in app based data collection	Employment status	Random Sample	USA	Actual behavior	Employed respondents are more likely to participate in app data collection
Lynch et al 2019	Participation in app based data collection	Age	Random Sample	USA	Actual behavior	Older respondents are less likely to participate in app data collection
McCool et al 2021	Participation in app based data collection	Age	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	Older respondents are less likely to participate in app data collection
McCool et al 2021	Participation in app based data collection	Immigration	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	Immigrated respondents are less likely to participate in app data collection
McCool et al 2021	Participation in app based data collection	Education	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	High educated respondents are more likely to participate in app data collection
McCool et al 2021	Participation in app based data collection	Marital status	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	Divorced/widowed respondents are more likely to participate in app data collection
McCool et al 2021	Participation in app based	Household size	Split-half recruitment from existing	Netherlands	Actual behavior	Single household member respondents are less likely to participate

	data collection		panel and fresh register sample			in app data collection
McCool et al 2021	Participation in app based data collection	Income	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	High income respondents are more likely to participate in app data collection
McCool et al 2021	Participation in app based data collection	Possessing a car	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	No difference
McCool et al 2021	Participation in app based data collection	Possession of driving licence	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	Respondents with driving licence are more likely to participate in app data collection
McCool et al 2021	Participation in app based data collection	Rural vs. urban area	Split-half recruitment from existing panel and fresh register sample	Netherlands	Actual behavior	No difference

6. SUMMARY

The aim of this literature review is to provide a comprehensive overview of the state of research concerning factors influencing participation in smart surveys. We focus on aspects relevant to the specific needs of the large field tests of the SSI project, aiming to inform decisions regarding which design features should be considered in the survey invitation, contact with the respondents and data collection for these specific surveys. To achieve this goal, we differentiate between factors influencing the general decision to download and install the app and participation in "smart" data collection tasks, whenever such a distinction can be derived from previous studies. We also provide clear recommendations, whenever previous research permits, on how these design features should be addressed in the SSI surveys. Our intention is to assist all countries involved in the large field tests in finding the most suitable design for their purposes.

In summary, our review allows for the following suggestions for the large field tests with TUS and HBS in the SSI project:

- NSIs that collaborate with universities should leverage the high trust that the general public has in universities by announcing them as a (co-)**sponsor** of the study.
- An unconditional small monetary **incentive** should be used to increase the likelihood of participation. If this is not possible, then a conditional incentive should be provided both for starting the survey (e.g., downloading the app) plus for continuous participation throughout the study period. What incentive amounts should be used would be an avenue for the investigation in the SSI.

- The **length of data collection** should be limited to probably not more than two weeks, if possible, to boost participation rates in the large field tests.
- If GPS tracking or other forms of passive data collection is used, respondents should have **control over the data collection process** by allowing them to temporarily pause data collection.
- **Transparent information about the data collection process** (I.e., what data are collected and analyzed) should be provided to potential respondents to increase the likelihood of participation.
- An **app-based approach** is best suited to incorporate the smart features of the SSI and allows for direct reminders during the data collection period. Using a **progressive web app** could help reach members of the general population who are not willing to or able to download apps.
- Reducing participation burden for people less familiar with lower **digital literacy** by simplifying the process of installing and downloading the app to the largest possible extent (i.e., considering legal and ethical requirements for consent) will be key for recruitment success.
- **Privacy concerns** reduce the willingness to participate in smart surveys. How to best enhance trust in the data collection organization and alleviate privacy-related concerns in the invitation process will be one major research question to be addressed in the large field test of the SSI project.

Our literature search provided just 1 report on empirical tests about the use of **interviewers** (Rodenburg et al. 2022) when recruiting for smart surveys (as opposed to recruiting via letters). This study did show that recruitment rates can be doubled (from 15% to 30%) when interviewers are used, but at this point we know too little to evaluate the effect of interviewers on recruitment and retainment. Of course, there is a large body of survey methodological literature that clearly shows that face-to-face surveys yield higher response rates than self-administered ones. Whether these findings translate to smart surveys, will be tested as part of the large field tests in some countries.

Chapter 2: Machine Learning in smart surveys

1. INTRODUCTION

Machine learning (ML) algorithms play an important role in smart surveys but how machine learning is to be used in this context is the key question. The level to which automation can be brought to replace the direct acquisition of information or replace manual processes without degrading data quality and/or increasing respondent burden, is the crucial point in the use of ML.

One goal of Workpackage 2 of the SSI project is to develop methodological standards around the use of these machine learning models. Key questions are:

1) Under what circumstances results from ML models can be used directly as statistical data, and under what circumstances data should be fed back to respondents?

2)What to do when the quality of the machine learning outcome is too low?

3) When should respondents be asked to provide new input (a picture or open text) because no meaningful information could be extracted?

Underlying all the above processes, are the training datasets used in the ML.

4) How and when should training datasets be updated or improved?

Case studies are the ML methods used in HBS and TUS: In HBS, Optical Character Recognition (OCR), is used to classify text from the receipt images taken by the respondents. In TUS, geolocation data and contextual data (e.g. Open Street Map) are used to predict activities and associate them with one (or more) HETUS time usage categories.

In sensor data applications, models seldom reach 100% accuracy. Certain population subgroups or certain survey statistics may require manual inspection. In most ML classification problems, it takes little effort to achieve close to 80% accuracy, but it is increasingly difficult to push for the last 20%. This is a significant challenge for official statistics that require high precision and accuracy. Acceptable error rates are usually agreed between survey teams and their end users, typically less than 5% (Benedikt et *al*., 2020).

In such cases to improve the accuracy of the ML models, human interventions (respondent, coder) must be envisaged to assign correct labels. The new labelled item is used to retrain the model to make it more up to date. Over time, the machine learns from humans and becomes more and more accurate.

Furthermore, ML methods require continuous updating. Updating can be done fully automated through online learning or semi-automated through active learning. Retraining is ideally done based on incoming datasets while preserving the privacy of the respondents. In practice, when respondents provide data for which processing performance falls below specified thresholds, then this data should be used for retraining ML model.

Active learning is the subset of ML in which a learning algorithm can query a user interactively to label data to obtain the desired outputs. In active learning, the algorithm selects the subset of examples to be labelled from a set of unlabelled data. This algorithm represents a key component in Human-in-the-Loop where human and machine intelligence combine to create more accurate models.

The problem in the use of ML in a survey then becomes: 1) Build the automation part, 2) Design a mechanism whereby machine alerts human when it needs input and 3) Design an efficient UI to facilitate human machine interaction (Benedikt et *al*., 2020).

This chapter reads a review of the experiences gained on the above topics in the ESSNet Smart Surveys and other projects (paragraph 2.2 and 2.3) and outlines the gaps and some possible solutions in paragraph 2.4.

2. LESSONS LEARNED FROM THE TRUSTED SMART SURVEYS PROJECT

The ESSNet Trusted Smart Surveys was structured in two workpackages. Workpackage 2 (WP2) was the empirical and applied component and identified the needs, (further) designed and tested smart survey solutions in four main survey areas: Consumption, Time use, Health and Living Conditions. Workpackage 3- (WP3), following a top-down approach, had the objective of defining a framework both architectural and methodological for the smart surveys and develop Proof of concepts for some elements of the framework. Both workpackages (WPs) addressed ML issues, from a practical and theoretical point of view.

TRUSTED SMART SURVEYS - WORKPACKAGE 2

Workpackage 2 tested existing solutions for four different survey topics, as well as TUS and HBS. The focus in Health has been on physical activity tracking, which may be linked to the European Health Interview Survey (EHIS). The focus in Living Conditions has been on indoor climate, which may be linked to Statistics on Income and Living Conditions (SILC), again to EHIS and also to housing surveys conducted in most countries[2].

The four pilots included some of the main smart features highlighted in WP2 deliverables, as described in the following table 9, where the smart features and the methodological elements of the WP3 Proof-of-concepts (PoC) are mapped on the pilots.

Table 9: Smart features, PoC methodological elements and pilots.

	Consumption	Time use	Health	Living conditions
Device intelligence	YES	YES	Not in this pilot	Not in this pilot
Internal sensors	YES	MAYBE (location)	NO	NO
External sensors	NO	NO	YES	YES
Public online data	NO	MAYBE (location)	MAYBE	YES
Personal online data	MAYBE (bank transactions)	NO	YES (personal devices)	YES (personal devices)
Big data linkage	NO	NO	MAYBE (data on health care)	YES (data on dwellings)

Active-passive data	Receipt validation	In part	Activity reports	Living conditions reports
Machine learning	OCR and classification	In part	Transformation of sensor data	Transformation of sensor data

What emerges is that ML algorithms for sensor data processing are used mainly in HBS, Health and Living Conditions, the procedures implemented in HBS being the most mature. In the following the main results and recommendations are presented in a schematic form.

Table 10: Main results and recommendations from Trusted Smart Surveys project

HBS
Within the ESSnet, functional tests, usability tests and field tests have been linked to the Household Budget Survey app developed by CBS. ML routines play a crucial role in receipt processing and must be viewed as a micro-service that requires separate maintenance and coordination. Overall, the most demanding new features of a smart HBS are ML routines for product classification and the creation of rich product lists.
Lessons learned and recommended next steps
- Product search algorithms are effective but depend on the richness and form of language of product lists. Creation of such lists facilitates also non-app implementations and requires an investment in time. - The involvement of respondents in both product search and receipt scans has been one of the main focal points during development. It is advisable to perform usability tests. - Automated In-app OCR and NLP of receipt scans are feasible but should be supplemented by the option of in-app respondent editing. - Improvement of ML approaches for classification, exploiting active and online learning options of ML models

Time use
WP2.2 addressed the use of MOTUS for Time Use and Mobility smart surveys, to evaluate to what degree MOTUS's current and future development stages will allow for implementation in different countries and over various domains. The pilots were qualitative with small study populations (a few tens of persons) but featured elaborated scopes covering a wide variety of substantive and methodological matters. The methodological focus of the first pilot was on functionality, while in the second pilot, the thematic interest shifted entirely to (passenger) mobility, and methodologically the aim was to evaluate certain aspects of MOTUS usability, how well MOTUS functions. In this pilot several surveys were combined with a diary and some smarter elements like notifications and a geofence (implying tracking of respondent and inclusion of a geolocation service) were introduced. However, none of these pilots produced explicit results about the use and results of ML models.
Lessons learned
Although the basic features of MOTUS like organizing a survey are fully operational, smarter features still need to gain maturity before they can be deployed in genuine research. For example, retrieving in the backend the geographical coordinates resulting from the tracking necessary to enable the geofence microservice (developed in TF INNO HBS-TUS) was not possible yet.

Health

For the use case of 'health', the attention was focussed on measuring physical activity by means of an accelerometer, as sensor measurement would potentially lead to higher quality data than diary interviews. The accelerometer used in the pilots is the thigh-worn activPAL (algorithms from the activPAL software suite were used to interpret the data in terms of length and intensity of physical activity). The device stores the data that are downloaded once the device is sent back to the National Statistical Institute (NSI). Small scale feasibility pilots were performed in three countries.
Lessons learned
For machine learning procedures, two variables are relevant: the activities performed, and the intensity with which they are performed. For all pilots, the algorithms developed by activPAL were used. However, if objective measurement with accelerometers is implemented, 'official statistics' algorithms need to be developed as NSIs cannot be dependent on commercial algorithms that are not transparent and can change without informing NSI. Some initial effort was undertaken to train own algorithms, but they were not good enough. In the lab phase, training machine learning algorithms got a success to recognize laying, sitting, standing, walking, running and bicycling, the activities that were performed in the lab. However, the algorithms did not generalize well to the activities in the free living week: the physical activity diaries that could have informed the machine learning algorithms (Van Hoek et al. 2022), were on the one hand very imprecisely filled in, and on the other hand not detailed enough. There are already international research collaboration platforms that work on the development of machine learning algorithms for thigh worn accelerometers (ProPASS, propassconsortium.org).
Recommendation: Develop proprietary algorithms to generate the requested variables was advisable for the project in the context of official statistics.

Living conditions
Living conditions are related to national health surveys, the European Health Interview Survey (EHIS) and to the Statistics on Income and Living Conditions survey (SILC). In this use case, living conditions are limited to indoor environment quality (IEQ). NSI's are in a unique position to combine the subjective measurements in surveys, knowledge about respondents' illness and health and registry knowledge on buildings with the objective measurements that sensors can provide. This project was hence more about new data than about new methods of data collection.
Lessons learned
Data quality: There are some concerns about the quality of the data that have been measured. Satisfying information on measurement precision and reliability were not received from the distributer, nor on the calibration algorithms that are used for the sensors. However, the data analysis shows interpretable results with at least face validity. The consensus among the experts is that (cheap) sensors can give an impression of the relative variation over time. The question is, of course, if that is good enough for our purposes. Precise definition of those purposes will help answer this question.

TRUSTED SMART SURVEYS - WP3

WP3 addressed the topics of machine learning and the quality of produced data in two activities:

Table 11: Activities around machine learning in Turst Smart Surveys project.

Methodological sub task 3.1.1
The activity of the methodological sub task 3.1.1 aimed at developing a robust smart survey methodology[3]. It explores design requirements for TSS_u in contrast to traditional paper-based or online surveys. The main problems addressed were related to: sensor data from a variety of devices

that are not standardized in structure, format or availability; innovative ways of handling sensor data including ML algorithms (microservices); sources of error in TSS_u and error management. Deliverable 3.1 Preliminary framework reports a review of methodological issues, useful for SSI project as well.

Proofs-of-Concept on methodology and machine learning

A modular prototype element for an essential aspect of the architecture for the smart surveys was designed and developed. A Generalized Machine Learning Component (GMLC) for data provided by the same type of sensor (e.g. accelerometer, gyroscope, thermometer), divided into different software modules to be applied to different contexts and survey needs.

The GMLC was developed on a cross-survey component performing multi-class supervised classification tasks, although the extension to regression tasks can be implemented. The generalization of the process is realised considering different modules that perform a specific function in the pipeline. Modules can be interposed into the process in a different order and new modules can be added to fit new surveys.

ActivPAL and iLog datasets[4] constituted a base for developing a component in terms of modularity. Both data collected from a diary and sensors by Health pilot (Workpackage 2.3) and in SmartUnitn(Two) surveys were used to build a generalised pipeline. The use of data collected through different devices (wearables for Health pilot and smartphones for SmartUnitn (Two) surveys) have implications on the results. In fact, while for the activPAL use case, the performance of the physical activity classifier is good (accuracy of 87.6%), for the iLog use case, the performance of the "mean of transport" classifier is lower (accuracy of 61.6%). These differences are due to the good quality of the data collected in a controlled environment, such as the laboratory for the annotation of the labels, and also to the wearable instrument activPAL that measures the acceleration with a good sampling rate, and a systematic collection of the associated acceleration signal.

Lesson learned

- The Proofs-of-Concept showed that it is possible to develop a generalized smart data tool to transform signals into statistical variables; a ML component is generalizable in the sense that it can perform its function in several contexts and in different phases of the survey.
- The trade-off between the level of generalization and the quality of model is a crucial issue of this component.
- ML applications can perform better involving human interaction in the functions. Keeping experts, or respondents themselves, in the loop can improve model accuracy and reduce data errors. The machine learning should be widened in terms also of online learning and active learning, including the respondent involvement
- Concerning the (re)training of machine learning algorithms further investigation is needed. Situations are different in each country and we have to know when we should retrain algorithms. How often and in which phase of the survey algorithms should be retrained are open issues.
- During the data collection, it is necessary to monitor the collected data. In cases in which the quality of the data is not satisfactory, data could be improve using one of two kinds of approaches: active approach (e.g. notification mechanisms to the user that require his action) or passive approach (e.g. through a centralized edit and imputation phase)

3. LESSONS LEARNED FROM OTHER ESSNET PROJECTS AND WIDER LITERATURE

HBS PROJECTS - RECEIPT PROCESSING AND PRODUCTS CLASSIFICATION

In this section, we report excerpts from various works carried out as part of two projects: the ESSnet project - @HBS>An app-assisted approach for the Household Budget Survey (2020); the project 2020-NL-INNOV (@HBS2). The aim is to highlight in a concise and schematic form the most relevant aspects that emerged, and which concern the use of ML in a smart survey.

The ESSnet project - @HBS>An app-assisted approach for the Household Budget Survey (2020) addressed the question of modernising the Household Budget Survey (HBS) data collection process. The project investigated the entire end-to-end data collection process developing, in particular, a proof of concept for a system to process scanned receipts, develop Optical Character Recognition, and automated coding.

In the work of Benedikt et *al*. (2020) the processing of shopping receipts is described focussing on how data science techniques and Human-in-the-Loop AI can be applied to automate this process. Relevant information such as shop names, dates, purchased items and prices are extracted from receipts and products are classified to their 5-digit Classification of Individual Consumption by Purpose (COICOP).

The question of how automation affects the quality of output forms the core of the work. In response to this the authors propose to design an automation pipeline that comprises the following steps: Scanning, Image processing, Optical Character Recognition (OCR), Natural Language Processing (NLP), Automated classification. Table 12 shows the problems encountered in the various pipeline steps and lessons learned.

Table 12: Pipeline steps in smart surveys with potential problems

Pipeline Steps	Issues	Lessons learned
Receipt scanning	Human factor	Respondents are not necessarily tech-savvy and do not know how the images are going to be automatically processed. ***Without specific instructions, they may make common mistakes.***
	Technological	Depending on the quality of the mobile device, photos may be too low resolutions, which negatively affects OCR accuracy. ***A high quality image may be too large, which takes time to send and may increase respondent burden.***
Image processing	Quality of the paper receipts	Paper receipts still need to be repaired, such as faded receipts and removing shadows caused by wrinkles on receipts. ***Some level of human intervention will be needed to decide whether image processing is required to improve OCR.***

Optical Character Recognition (OCR)	Receipts are not standardised, difficult to infer the meta-data	Further data parsing to infer the meta-data. *Developing data parsing methods that should work for any receipt, from any shop, in any country and any language is technically challenging.*
Natural Language Processing (NLP)	Misspelled words due to characters being wrongly recognised	NLP to correct such errors. NLP Module can be embedded in both the OCR and the classification modules.
Automated classification: Supervised Machine Learning (ML) models used to automatically classify items to COICOP codes.	Erroneous ML classification	**Human-in the loop/Active learning** To match human judgements on receipt items to COICOP classification, a supervised ML text classification approach should use features created from the text descriptions of receipt items. A good supervised classifier should learn rules to allocate the data into provided COICOP categories. ***In case of rare and unseen products, the re-labelled data can be used to retrain the models and make them more up-to-date -Active learning.*** ***In case of ambiguous items the coder has to contact the respondent for clarification, which increases respondent burden, workload and processing time.*** To mitigate this problem it is necessary to include a 'Usual Purchases' page in the questionnaire, asking respondents what kind of product they preferably buy, so that can be imputed.

The goal of project 2020-NL-INNOV (@HBS2) was to complete the HBS app, introduce it for use in the project members' countries and to develop further the app based on the feedback received from national pilot surveys.

The deliverable 1.2 – Report on the action (Schouten, 2022) contains important conclusions and recommendations concerning both the development of the app and the use of ML algorithms in the receipt processing steps - receipt scan text extraction and product classification.

The @HBS2 project highlighted that in-app respondent editing of extracted receipt texts is beneficial and reduces distances to true values, but also the need for further optimization of the receipt processing and for refinement of the pipeline for cross-country implementation.

In the following, examining the other deliverables produced, we report more specific issues concerning the receipt processing pipeline, in particular the classification step, ML models and training development. Also, recommendations and lessons learned are highlighted.

Actions still needed for the optimization of the receipt-processing pipeline (Oerlemans & Schouten, 2022 - Deliverable 1.3) are listed below considering above all the involvement of the HBS respondents.

Table 13: receipt processing steps in smart surveys

Receipt processing steps	Involvement of HBS respondents
1) Scanning (in-app)	**Respondents make pictures of paper receipts, or possibly digital receipts.**
2) OCR and language processing (in-app)	Procedures embedded in the app provide a first attempt to extract products and corresponding prices. An OCR score is computed per supposed product-price line and averaged over all lines. **A lower threshold can be provided and when the average score is smaller, the respondent is redirected to step 1.**
3) Editing of product-price extraction (in-app)	**The respondent can edit results and the results along with the scan are submitted to the backend**
4) OCR and language processing (in-house)	The receipt is processed and a new set of products – price couples is derived. Results are overruled by the in-app results, when **respondents indeed checked and possibly edited results.** Respondent editing can be monitored by in-app paradata and by the absence of products with a zero price. A price zero occurs by default when OCR cannot find or read a price.
5) Classification (in-house)	The selected set of product-price couples are classified to COICOP through a *mix of machine learning predictions, string matching and manual evaluation*

Below we list the Machine earning algorithms used in various steps of receipt processing pipeline and lessons learned (van Hoek et al., 2022 Deliverable 2.3 – Receipt processing).

Table 14: Machine learning algorithms used in receipt processing

Steps / ML Algorithms
Image Processing
– Multi-Region Convolutional Neural Network (MRCNN) *Description:* Trained on about 300 images, the aim is to remove the background from a photo.
Optical Character Recognition (OCR)
– Pretrained Tesseract (in-app) *Description:* Guide the process the right rotation of image and return a feedback of the quality of photo to respondent. – Pretrained Tesseract (in-house) *Description:* The OCR extracts text from the pre-processed image by NLP pipeline.
Classification
Model selected respect a quality metric from ML classifier models: Logistic Regression, Multinomial Naive Bayes, Decision Tree, Random Forest, Support Vector Classifier, FastText *Description:* Classification between the textual description of the item on the receipt and the first four hierarchical levels of the COICOP taxonomy. **The classifier has been trained with over 200,000 product descriptions.**
Lessons learned on the use of ML
ML algorithms need to be re-trained for adapting to changes underlying the phenomenon for which they are to perform their tasks - changes over time in the lists of products. The Introduction of new products implies the addition of new words not present in the training phase of the models this deeply affects the performance of the models.

ML algorithms must be chosen according to the needs of the survey in terms of the quality of intermediate process outputs and final output statistics.
Some algorithms are more interpretable (Logistic Regression, Decision Tree), or are more accurate (Random Forest, FastText) or more suitable to reside in-app due to lightness and speed of execution.

Concerning classification development, issues and lessons learned are listed below (Oerlemans & Schouten, 2022 - Deliverable 1.3; Oerlemans, de Wolf & Schouten, 2022 - Deliverable 2.1)

Table 15: Classification issues in receipt processing

Classification issues (Oerlemans & Schouten, 2022)
Classification is very strongly **country-specific.**
The specific nature of receipt 'language' makes it hard to impossible to transfer a trained classification procedure in one country to another. Printed products texts usually contain no common vocabulary. They may contain abbreviations, punctuations, quantities, shop names, brand names and/or references to bio/ecological production.
Products printed on paper or digital receipts show dynamics over time. Product dynamics in stores operating on a national scale should be accounted for in classification routines. They tend to have large revenues and occur frequently in HBS. Keeping all receipts texts ultimately deteriorates performance of classification and train sets need to be refreshed from time to time.
ML training and Human-in-the-loop (Oerlemans, de Wolf & Schouten, 2022)
ML models and training for classification of products. (Classification of products can be in-house and in-app as well depending on how the ML models are trained)
ML models can be trained on - **Annotated receipts => Active and online learning, making sure that models are retrained.** - EAN/GTIN product descriptions as typically available in scanner data => countries without scanner data or limited scanner data. - Receipt texts directly obtained from shops => under investigation but very promising.
Inspection of classified receipts - Human-in-the-loop ML
OCR and classification return indicators of accuracy. Based on lower thresholds to these indicators, it can be decided to flag processed receipts for inspection. This may be **country-dependent as ML models for classification depend heavily on the quality of training and retraining of models.** **The extent to which OCR and receipt classification contain human check is still an open decision.**

Table 16: Summary of lessons learned from ESS projects on HBS

Lessons learned on classification step, ML and Recommendations
- Perform learning and retraining only in between waves of the survey. - Manual data entry of products that are not recognized can be annotated by respondents. - **Products unknown to ML classification may be gathered to inform manual checks and active learning.** - **Online learning is warranted to account for the dynamics in products over time.** ***Methodology can adopt a mix of online learning and active learning approaches. These approaches combine survey-independent external information from stores (such as delivery of printed texts) and observed products texts in the HBS that have a low maximum classification probability.***

Lessons learned from ESS collaboration extracted from different deliverables
About ML and OCR/text extraction routines
Country machine learning routines must be available to all so that overseas purchases can also be processed - it would be beneficial to exchange machine learning routines for COICOP classification. (Deliverable 1.2 – Report on the action)
OCR/text extraction routines may be tailored per country (Deliverable 3.3 – Field test analyses)
About Classification – shop and product lists
Shop list is very helpful (Deliverable 2.2 – Product/shop guidelines): - **in designing the appropriate ML procedures**, in particular the language processing step after OCR - for better user experience and optimal manual registration of expenditure search - to assist processing of scanned receipts - to pre-classify products into main categories for scanned receipts - to prepare/anticipate digital receipts in the near future.
Product list is an important ingredient in manual data entry. Respondents can type products that are then matched real-time to the product list entries. Exceptions are when products do not (closely) match to any of the product names in the list or the respondent deems all proposed matches as unfit. In those cases, the respondent still has the opportunity to provide a categorization her/himself. **Product lists must be elaborated to be linked to respondent answers through string matching.** (Deliverable 1.2 – Report on the action) **Purpose of product list is the Reduction of respondent burden and improvement of data quality.** **(**Deliverable 2.2 – Product/shop guidelines)

The project identified two future actions concerning receipt processing.

- Optimize and harmonize code for classification based on a mix of machine learning and string matching.
- Formalize active and online learning procedures.

STUDIES ON GEOTRACKING AND TRANSPORT MODES PREDICTION

The potential of automatic transport mode prediction in app-based surveys using mobile device location sensors and ML algorithms has been demonstrated by several studies. In some

studies, then, diaries for the active collection of the position provided by the interviewees were used with the aim of determining whether or not collecting this information in a more passive way would have allowed a deeper and more accurate inference.

A rich literature review of these studies is reported in the work of Smeets, Lugtig & Schouten (2019). In this section we focus on highlighting crucial aspects that emerge from studies that are placed in the official context of the survey (large scale survey, general population, sample, etc). To this end we consider the study conducted in *internal projects of national statistical institutes* with reference to travel surveys.

The **pilot for the use of smartphone-based travel studies launched by Statistics Netherlands** (CBS) aims to test the feasibility of, at least partly, automatic transport modes prediction in a representative national travel surveys, to study some questions related to the accuracy of transport mode prediction (e.g. the effect on accuracy of collapsing different transport modes into broader categories, of including some and additional features).

The pilot design involved tracking/monitoring respondents for a week via an app that collected location data - latitude and longitude coordinates on a per-second basis when the user was moving and on a per-minute basis when the user was still. An automatic diary of stops and trips was generated and respondents were asked to give context about the stops (purpose) and trips (mode of transportation). Stop decisions are based only on time-location data. The app does not use motion sensors and does not use geo-location data from open online databases. The restriction to location sensors was chosen to reduce the amount of battery use and to limit local storage of sensor data. Consultation of geo-location data is more complex and may lead to privacy issues, and was ignored in the proof-of-concept (Smeets, Lugtig & Schouten, 2019; McCool et *al*, 2021).

In the following tables, we extract, first the main and more general lessons learned from this study and then, we highlight issues and suggestions about the quality of data and ML accuracy.

Table 17: Summary of lessons learned from travel surveys

Lessons learned (general)
Smartphone-based travel studies lead to more precise measurements of distance and time travelled.

To produce high-quality official statistics (for example, which age groups travel with what transport mode when), respondents still need to actively label trips, as in a diary-based study.
Even if full automatization of transport mode prediction is not feasible, the results of this research could lead to a **reduction in response burden in the future**. *(Respondent burden can be reduced by relying on the passive data itself or by adopting a verification approach in which respondents confirm the expected pattern or correct it. (McCool et al., 2021).*
Reduction of respondent burden – how?
Through an algorithm able to accurately predict transport modes, either by fully predicting transport modes or by giving respondents suggestions they only need to confirm. A classification algorithm can be used **as an imputation method** to predict the transport modes of those trips that were collected, but not labelled. Ideally, future iterations of the app would refrain from asking respondents to manually select transport modes from a long list, but would instead automatically classify the transport mode of different trips. **Improvement of prediction accuracy can be realized** using the fitted model to calculate expected probabilities for different modes and, based on that, only present the top three to the respondent. Alternatively, only asking respondents to label trips for which the algorithm is less than 70% certain about the prediction.

To classify transport modes, different types of supervised ML algorithms can be used. These range from linear models, such as multinomial logistic regressions, to convolutional neural networks. The most popular methods are, in order, rule-based algorithms (including decision trees), Random Forests, Support Vector Machines, and Bayesian Networks. For the prediction of transport modes different ML algorithms were used and compared in their performance.

Table 18: Summary of lessons learned from ESS projects on transport mode prediction

Lessons learned on the accuracy of ML for transport modes prediction
Steps for improving ML models: (Smeets, Lugtig & Schouten, 2019).

- **Data cleaning and pre-processing** of the raw data before it can be used to engineer features - **Selection of features and context location data** (location of bus stops, train stations)
ML model - Random Forest was the best model explored.
The model suggests that it is likely not accurate enough to distinguish between the nine most commonly used transport modes and to also flag erroneously recorded trips. The category User error is likely to be underestimated. Respondents could not delete trips or label them as erroneously recorded, but had to make the effort to leave a comment in the 'other transport modes' text field.
Suggestions
- More data on the rarer modes, especially scooter, tram, and metro, would likely increase prediction accuracy, as would further improvement of the app. - Better separation of trips into different segments with a single transport mode is also likely to improve the model accuracy - **splitting algorithms using context location features** (e.g. location of public transport systems or road networks).
Acceptable balance between accuracy and the number of transport modes that need to be distinguished.

To obtain good estimates of transport mode usage, it is necessary to be able to correctly place a stop between each transport mode change and define, a priori, how to define a stop.

Table 19: Summary of lessons learned from ESS projects on stop predictions**Lessons learned on *Stop decisions and preliminary conclusions***
Stop detection based on time-location sensor data is relatively robust. An improvement of stop detection will only be possible through linkage of geo-locations and/or employment of motion sensors. (McCool, Lugtig & Schouten, 2018?)
Future research could abandon the paradigm of first separating stops and trips, then engineering features, and then employing a ML algorithm. With sufficient data, a convolutional neural network adapted for structured data should be able to pick up on features such as speed and points of interest from the raw location data itself and predict transport modes. Then instead of estimating how many trips a person average made per day and how long those trips were, even more precise statistics could be calculated.

The presence of measurement errors (outlier, noise) and a high proportion of missing data, is a common trait of location data generated passively. Incomplete data can occur at multiple levels, and for multiple reasons, some related to the physical surroundings and others related to the device, the user, or the interaction between the two (McCool, Lugtig & Schouten, 2018?).

Below, starting from several works, we highlight some issues and lessons learned related to data quality that can have a significant impact on the accuracy of ML algorithms. Furthermore, we report some methods used to deal with missing data.

Table 20: Summary of lessons learned from travel surveys on measurement errors

Measurements errors (outlier, noise, missing) issues and lessons learned

Pre-processing of the raw data for treating measurement errors (outlier, noise) is an important step for the accuracy of ML model (McCool et al., 2021). Methods to filter likely measurement errors in GPS data include discarding single points with a too wide or omitting data points that would lead to an unrealistic high speed. Further pre-processing in the form of smoothing the data to remove random noise (Savitzky-Golay filter, Kalman filter).
Identification of the issues underlying missingness and measurement is an important step in assessing data quality (McCool et al., 2021).
Understanding the composition of the missing data is integral to making the correct decisions about its content. The composition can involve the length of the component gaps, the overall sparsity of the data, or the time at which the gaps begin or end (McCool, Schouten & Lugtig, 2023).
Aggregation of Individual mobility trajectories data (difficult to measure and often with missing data for long periods) without accounting for the missingness leads to erroneous results, underestimating travel behavior (McCool, Schouten & Lugtig, 2023).

Table 21: Summary of lessons learned on missing data in travel studies

Methods for dealing with missing data
Method that combines a top-down ratio segmentation method with simple linear interpolation (McCool et al., 2021). Method designed for relatively short gaps, but evaluated also for longer gaps. In this approach, the linear interpolation imputes missing data while the segmentation method transforms the set of location points to a series of lines - segments.
Dynamic Time Warping-Based Imputation (DTWBI) to imputing travel behavior characteristics in human trajectory data (McCool, Schouten & Lugtig, 2023). Method developed for more general use in the imputation of time series data. Because the method makes use of patterns within the temporal characteristics of the data, it is useful to evaluate its potential as a mechanism for correcting for long gaps in trajectory data.
Multi-stage model (Bähr et al., 2020). Complex model for analysing and controlling the error source in the missingness processes that implies availability of: Paradata (i.e. information on the device, battery level, display state, the state of the mobile network connection); Contextual data (i.e spatial, temporal, etc.) and other type of information (socio-demographic).

4. SUMMARY: GAPS IN OUR KNOWLEDGE AND POSSIBLE SOLUTIONS

Regarding HBS domain and classification issues, the reports of the Smart Surveys project and of the project 2020-NL-INNOV (@HBS2) - in particular the deliverable 2.3 - underline the need to improve the ML classification accuracy. Retraining mechanisms, online learning and active learning (AL) has not been implemented but only discussed, considering that retraining phases of the models can be carried out between waves of the survey.

For HBS survey, a fundamental task to improve products classification, concerns the measures that must be adopted to ensure that the level of accuracy of the ML algorithms over time remains constant and at pre-established levels. Such actions become necessary as the cases of unlabeled products increase.

In HBS, active learning ("sequential design" in statistics) may be the most appropriate ML algorithm, since the situations where the classification procedure fails have to be managed during the data collection phase. The involvement of the respondent is necessary to collect labels that train a more accurate model.

In an interactive learning procedure, it is necessary to develop the decision-making process. One gap in our knowledge is how to deal with the problem of automatically determining when to start asking queries or to stop. What are the requirements that a stopping criterion must have? How aggressive or conservative should be the behaviour of stopping methods? Perhaps, we need to look at stopping methods that are more widely applicable, more robust respect to data set changes and that provide user-adjustable stopping behaviour.

In a survey context, two aspects must be taken into account that may be in conflict with each other, the burden on respondents and the high level of classification accuracy. Therefore, a stopping criterion must find the right trade-off between annotation and ML performance.

The pilot for the use of smartphone-based travel studies launched by Statistics Netherlands provides useful information about the geotracking domain and the methodological issues faced in transport modes prediction for the official statistics. Exploiting GPS data is the common element with the TUS domain on which the SSI project should focus. In fact, one of the main objectives is the development of microservices that exploit geolocation data for supporting the TUS respondents in providing the daily activities.

Regarding the geotracking case study, the gaps in our knowledge concern, especially, how to prevent/deal with measurement errors (outliers, noise, missing data) in location data (GPS) and how to choose the features and the contextual data that are functional for the required prediction, both the daily activities for TUS or the transport modes. All these choices have a significant impact on the quality/accuracy of the predictions. Even though we reported experiences mainly on transport mode, similar issues affect the quality of location data and the accuracy of the prediction of stops and trips in the TUS context.

The idea for the pursuing of the objectives of Task 2.2 is to address the above issues through the formalization of methodological strategies functional to developing different smart tasks aiming to:

- improve accuracy in the product classification algorithm, through active learning, online learning and retraining;
- improve the accuracy of the prediction based on location data (GPS), handling missing data, using App logs and contextual data on location and proximity;
- define metrics or tools to evaluate quality of ML prediction.

Chapter 3: Human Computer Interaction and Usability: A Review of Practice and Theory

1 INTRODUCTION

Smart surveys (i.e., platforms or applications to conduct studies that make use of smart features) are considered part of the answer to the increasing challenges of the production of official statistics. The challenges largely align with the principles of the European Statistics Code of Practice (CoP) such as reduce respondent burden (principle 9), improve on cost efficiency (principle 10), and maintain or improve on the quality of the data (cfr. principles 12 and 13) (Eurostat, 2018). Smart features collect data through a smart device from smart options such as employment of in device-sensors, linkage to external sensor systems or public online data (e.g., data on weather, traffic …), or data donations. Smart features are enabled through microservices.

A large variety of smart features exists. For example, in transport research, applications use smartphone location data to predict travel mode (Smeets, Lugtig, & Schouten, 2019) or to measure the frequency and duration of use of certain geographically defined spaces (Fenton, Glorieux, Letesson, & Minnen, 2020).

In this Smart Survey Implementation (SSI) project, the focus shifts more towards smart features that are supportive of data collection as opposed to collect data as primary purpose. The two central use cases are Time Use Surveys (TUS) and Household Budget Surveys (HBS). Both consist of keeping a diary either with daily activities or with expenses and the provisioned smart features – Geo-tracking and Optical Character Recognition (OCR) – facilitate and relieve these dairy registrations. For example, based on geo-points designated as work, school, home[1], tentative diary entries can be suggested from the online activity classification list (OACL). Similar, based on OCR and machine learning processing tentative entries can be suggested from the Classification of Individual Consumption According to Purpose (COICOP) list.

Given the specificity of the smart features used in this SSI project, this deliverable will focus on testing the usability as part of the multifaceted field of HCI in smart survey applicationsfor TUS and HBS.

In previous ESSnet projects – SOURCE^{TM} (Minnen, Nagel, & Sabbe, 2020), Smart Surveys and CRŒSS (Minnen, Olsen, & Sabbe, 2022) – it has already been investigated which platforms and applications are able to carry out complex studies such as TUS and HBS. Various reports concluded that the MOTUS platform (developed by the Vrije Universiteit Brussel [VUB] and currently owned by hbits) and the @HBS app (owned by CBS), among others (e.g.,

applications from SBB and Insee), are technically and functionally capable of this.[2] However, next to technical and functional capability, the usability of these platforms and applications at the user-end are equally relevant in smart surveys. Firstly, because full participation consists of a series of sequential tasks (i.e., installing application, creating profile, completing questionnaires, keeping diaries). Secondly, because new challenges arise regarding (communication about) privacy and consent, and data security. Thirdly, because the diary registration requires complex actions (i.e., activity registration with context questions, expenditure registration with many details) and classifications (i.e., OACL for TUS, COICOP for HBS). Fourth, because the assumption is that adding smart features (e.g., receipt scanning, GPS tracking) to present tentative entries (i.e., the tentative state refers to being captured in a microservice and presented in the front office of the core environment) to respondents reduces complexity, registration burden, and improves data quality.

In this ESSnet project –SSI, workpackage 2.3 – focusses on HCI and usability. HCI relates to the design and use of computer technology with a focus of the interaction between humans (users) and computers. It encompasses various components, ranging from UI and UX experiences, usability, and cognitive and physical ergonomics. As will be outlined below, previous ESSnet projects focussed on UI. This ESSnet mainly focusses on usability testing, in the light of the aforementioned CoP, to assess to what extent these smart surveys and smart features support participants to complete the expected tasks (Stehrenberg & Giannakouris, 2021).

In what follows, this chapter provides an overview of the tests performed in the previous ESSnet projects and which were focussed on functionalities and the UI. This is the starting point. Next, the attributes of usabiliyt relevant for this ESSnet project will be discussed. Finally, a suggestion for the methodology used to make this assessment is given based on literature on usability testing.

Note that the current ESSnet project also includes the HBS application from Statistics Norway (SSB) and the TUS application from Statistics France (Insee). However, the reports of the previous ESSnet projects only discussed the MOTUS platform and the @HBS application. A new report by SSB will be released in the coming weeks after the delivery due date of this report. The overview of previous practices therefore only relates to the last two applications mentioned.

Previous Practices

The functionality and usability of smart surveys in terms of the User Interface (UI) and User Experience (UX) have been tested for the MOTUS platform that has been used for a Time Use Survey (TUS) and a Household Budget Survey (HBS), and the Dutch @HBS application. The testing of TUS on the MOTUS platform and the @HBS application took part within the Essnet Smart Survey project (Volk, Knapp, & Sommer, 2020; Volk, Knapp, Sommer, & Zins, 2021). The testing of HBS on the MOTUS platform took place within the CRŒSS-project (Knapp, Richter, Sommer, & Brecht, 2022). TUS on the MOTUS platform has also been tested in the German and Hungarian context (Hagymásy, József, Keresztes, Vámos, & Vida, 2022; Knapp, Rödel, Sommer, & Volk, 2021).

MOTUS is a platform first designed by VUB, and since 2018 continued by hbits, that allows complex studies – including TUS and HBS – to be designed in its back-office application and pushed to its front office application for respondents to participate (Minnen, Rymenants, Glorieux, & van Tienoven, 2023). The @HBS has been developed by Statistics Netherlands (CBS) and is a cross-platform application for HBS (Schouten, Bulman, Järvensivu, Plate, & Vrabic-Kek, 2020).

The tests have been conducted by the German Federal Statistical Office (Destatis) and consisted of an evaluation through written feedback and interviews conducted with internal colleagues and external test persons who used one of the smart survey applications (MOTUS platform or @HBS app) for one of the studies (TUS or HBS). Test persons were asked to install the app on their smartphone and participate in the assigned study. The TUS study on the MOTUS platform consisted of a household questionnaire, an individual questionnaire and a two-day time-diary. The HBS test study on the MOTUS platform consisted of a seven-day expenditure-diary and the HBS study on the @HBS application consisted of an (at least) two-day expenditure-diary. All tests were conducted in German.

The focus of the tests was to evaluate the HCI of the applications in terms of the functionality of the applications, that is, can the applications be used to conduct complex studies like TUS and HBS. The reports establish that both applications can successfully administer a TUS (via the MOTUS platform) and/or an HBS (via the MOTUS platform or the @HBS application). Nevertheless, the reports make several suggestions for improving the applications in terms of functionality, interface, information provision, and user experience.

SUGGESTIONS FOR IMPROVEMENTS

General improvements to the *functionality* related on the one hand to the removal of confusion about whether you can swipe to switch screens, whether you can tap on certain icons, and suggestive colour codes of certain buttons, and on the other hand to the appearance of a numpad when numerical values need to be entered and an improvement of the default values and autocorrect function. Specific improvements to the functionality related to the sorting activities and a copy/edit function (in TUS on MOTUS platform) and improvements on the search function and quality of the product and shop lists (in HBS on @HBS application).

Comments on the *interface* related to the shape and positioning of icons and buttons, the visibility of entry fields, the default settings of calendars, and the space on the screen occupied by long activity or product names. As far as the testing of the user experience went, suggestions were made to include more and customizable reminders and in-app feedback (for TUS on the MOTUS platform). Overall, the TUS was considered too time consuming due to additional queries for every entry in the time-diary.

Finally, the lack of *information provision* in the applications turned out to be a relatively major stumbling block. Test subjects repeatedly indicated that it was not clear what exactly was expected of them, or what the level of detail was to be recorded in the diaries. The

latter appeared to influence the user-friendliness of the search function of the activity list and the product and store list. Specifically for TUS, it was not always clear when a secondary activity had to be registered, when there was 'presence of others', or what exactly the time tracker for ongoing activities entailed. For HBS, it was not always clear how to handle complex tasks such as adding returns, registering deposits, or dealing with missing product or store codes. Also note that the receipt scanning function was not implemented in the German version of the @HBS and MOTUS application and therefore not evaluated.

Since the MOTUS platform works with an 'empty' application that is fed by studies designed in the back office, the look and feel of the app is comparable between, for example, a TUS and an HBS study. The diary component is central to both studies. Many of the improvements suggested during the testing of TUS on the MOTUS platform in 2020 had already been implemented before HBS was tested in 2021. At this time, the authors cannot determine the extent to which the @HBS application has considered the proposed improvements that emerged from the 2021 testing.

In general, it can be concluded that the MOTUS platform and the @HBS application are *smart* in the sense that they are technically and functionally capable of offering complex studies such as TUS and HBS to participants via online applications. Yet smart tools, smart features, or microservices only work if they equate with a high degree of usability and user-friendly experiences and interactions with the applications. As such, it is important for online applications for complex studies and related smart features to be *clever* as well. Cleverness then refers to the usability of the applications and smart features for users that do not necessarily have deep knowledge or understanding of the applications. Clever applications and smart features are presented in a way that they enable users to participate and meet their cognitive processes and expectations.

2 HCI: THE CONCERN OF USABILITY

HCI covers anything that touches upon the relation and the interaction between humans and computers. As mentioned earlier, this can therefore concern UI (i.e., visual and functional aspects of use), UX (i.e., emotional aspects of use), cognitive ergonomics (i.e., perception, memorability), physical ergonomics (i.e., look and feel of devices), or usability. Usability refers to the ease of use and the quality of the user's experience with a platform or application.

Despite the wide prevalence of smartphones and mobile applications and continuous technological modernisations amongst the many identified challenges and limitations, usability remains the main concern (Garcia-Lopez, Garcia-Cabot, Manresa-Yee, De-Marcos, & Pages-Arevalo, 2017). Usability relates to the intersection between system and user or the task and experience in the context of use. The ISO 9241-11 norm is most commonly used as the definition of usability, where usability is "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (Weichbroth, 2020). Indeed, the usability attributes that contribute

to the quality of use most often mentioned are efficiency, satisfaction, learnability and effectiveness (Weichbroth, 2018). Additionally, memorability, simplicity, comprehensibility, error, accuracy, time taken, features, safety, attractiveness, cognitive load, and communicability are mentioned as usability attributes in several studies (Harrison, Flood, & Duce, 2013; Hussain & Kutar, 2009; Kronbauer, Santos, & Vieira, 2012; Lew & Olsina, 2013; Zhang & Adipat, 2005). Table 22 provides an overview of the most common adopted usability attributes for mobile settings following a metareview by Weichbroth (2020) of 39 eligible studies conducted between 2001 and 2018.

Overall, usability and its attributes are not easy to define because is it always associated with the product/application in question (Weichbroth, 2020). *Generally, etymologically 'usability' breaks down into 'use' and 'ability' and thus refers to the ability to use an application for its intended purpose(s).*

Table 22. Most common adopted usability attributes (Weichbroth, 2020)

Attribute	**Share in metareview**	**Elements**
Efficiency	70%	Complete task with speed and accuracy
Satisfaction	66%	Comfort, pleasure, perceived level of fulfilment
Effectiveness	58%	Complete task in given context
Learnability	45%	Interact with newly encountered system and achieve proficiency
Memorability	23%	Remember how to use the application
Cognitive load	19%	Mental activity for instruction/presentation (extraneous), task complexity (intrinsic), integrate new information with prior knowledge (germane)
Errors	17%	Occurrence and applications ability to recover
Simplicity	13%	Easy to understand and navigate
Ease of use	9%	Level of effort needed
Others: *navigation, operability, usefulness, attractiveness, comprehensibility, aesthetics, accessibility, accuracy, adaptability, consistency, interaction, learning performance, training, understandability, user error protection*	<9%	

3 ATTRIBUTES OF USABILITY

Within the SSI project, where the call asks to arrive at an end-to-end solution and given the wide range of usability attributes a focus is required and, therefore, a bottom-up approach to usability is considered. For this approach, three crucial elements of online (smart) surveys are identified: a) recruitment and retainment of participants (see also WP2.1), b) sharing personal data by participants (see also WP2.1 and WP5), and c) participants' ability to complete complex tasks. The bottom-up approach then implies that high levels of usability (i.e., an application's high performance on the attributes of usability) will positively associate with these elements. For each of the three elements, five attributes are identified.

a. *Recruitment and retainment.* Applications and smart features need to be easy to use, efficient, and provide a satisfactory user experience. High levels of usability and a good match with expectations of the HCI are crucial to attract and retain participants. Recruitment here, relates to the ease of use to install the application as well as recruitment of participants to new stages of complex studies (i.e., continuation from survey to diary stage). Focus should lie on the following, interrelated elements:
    - Participants should feel *engaged* in the survey by its design, appeal, intuitive use, clear navigation (i.e., both in terms of the expected tasks as well as the use of the application), et cetera.
    - It should be *accessible* to a wide range of diverse participants with different capabilities both in terms of design as well as task processes.
    - It should provide participants with *clear instructions* at relevant times (i.e., in the invitation letter, when proceeding to next stages in the study, etc.) and 'locations' throughout the study to have them understand what to expect and what to do.
    - It should be *time efficient* and not requiring excessive effort.
    - If participants encounter errors or have questions during the use of the application and smart features, there should be *feedback* and *error handling* options available.
b. *Sharing personal data.* Complex studies such as TUS and HBS that require respondents to record activities or expenditures in diaries give rise to the participants feeling of disclosing personal information. The usability and HCI of applications and smart features play a crucial role in reducing or even eliminating these feelings altogether. Sharing relates to participants allowing smart surveys to collect, retrieve, map, or merge their data. Focus should lie on the following, interrelated elements:
    - Applications with high levels of usability convey professionalism, legitimacy, and trustworthiness, which establishes *trust and credibility* with participants.
    - Usability considerations are closely tied to *security and privacy* which should be established by clear privacy policies, secure data transmissions, and conformity to (EU) regulations (e.g., GDPR download).

- Similarly, usability considerations relate to *transparent communication* which reassures participants and help them make informed decisions about sharing personal data.
- Participants are more likely to provide information if the application is designed to collect data in a concise and efficient manner. *Data collection efficiency* is one of the main drivers to develop smart features (e.g., diary recordings based on OCR scanning or GPS tracking).
- Related, however, is giving participants a sense of *user control* over their information; what they provide, how it will be used, how to edit, delete or commit.

c. *Complete complex tasks.* Typically, TUS and HBS studies are complex because they exist of a sequence of tasks and because the so-called diary phase (e.g., when recording activities or expenditures) involves complex step-by-step actions. Usability and HCI therefore play a crucial role in participants' ability to complete these complex studies and tasks. Focus should lie on the following, interrelated elements:
   - A *clear and intuitive interface* that stems from a logical organization, easy navigation and intuitive controls. This has been addressed mainly in the Smart Survey and CRŒSS-project (see above), but presentation also relates to the device type, quality of the camera (i.e., for OCR).
   - The tasks should have a clear *task flow and guidance* such as step-by-step instructions, visual cues, progress indicators all to minimize confusion and improve participants' ability to complete the task (i.e., 'know what to do next'). This has been partially addressed by the Smart Survey and CRŒSS-project but should be considered when using smart features such as making diary recordings based on OCR scanning or GPS tracking.
   - When doing complex tasks, participants will make mistakes. High levels of usability anticipate this and are characterized by *error prevention* mechanisms.
   - When asking participants to complete complex tasks, providing clear instructions or tutorials increases their necessary knowledge and skills to do complex tasks. This *training* is done through usability-focused training materials (i.e., online or in-app training or assistance).
   - Finally, *feedback and support,* such as encouragements, positive feedback, tips, acknowledging (partial) task completion, or helping options, enhances participants' confidence and improves their ability to complete complex tasks. Note that when it comes to feedback and support it is important to consider the trade-off between supporting the respondent and overwhelming the respondent. Toggle switches to turn on/off certain feedback of support might be considered, but also increase the cognitive load of the application (i.e., more settings are not always better).

Within this SSI-project, three givens are important to consider when assessing the usability attributes of each of these elements. Firstly, the platforms and applications involved already exist and the microservices are designed to be integrated with these applications. Secondly,

the microservices are aimed to be integrated with different platforms/servers. Thirdly, for TUS and HBS, NSIs typically use population samples. These considerations affect usability testing in a number of ways. Different user groups with different levels of digital literacy and task with different levels of complexity will put emphasis on different usability attributes. Additionally, not all processes that might improve usability, such as user centric designs, personalization (for discussion see e.g., Zanker, Rook & Jannach 2019), customization (for comparison of adaptable UIs, adaptive UIs and complex UIs, see Zangh, Qu, Chao & Duffy, 2020), gamification (see for a review, e.g., Oliviera & Paula, 2020), are easily implemented when platforms and user interfaces already exist and shareability of microservices is a key aspect.

4 USABILITY EVALUATION

A large variety of usability evaluation methods exist of which *user testing methods* are the most recognized (Bastien, 2010). Other methods include inspection methods, inquiry methods, and analytical modelling methods (Weichbroth, 2020). Within user testing methods, one of the primary tools used to assess the usability is *the think aloud (TA) protocol* (Boren & Ramey, 2000). Others include question-asking protocol, performance measures, log analysis, eye tracking, and remote testing (Weichbroth, 2020) (see Figure 2).

As a usability testing protocol, TA protocols are most commonly used, because they allow 'observing' what a user is thinking because a user verbally articulates the struggles or experienced difficulties when doing a task (Olmsted-Hawala, Murphy, Hawala, & Ashenfelter, 2010). In the strict sense of the protocol, these verbalizations draw on the participants' short term memory and are simulated by the test administrator by probes such as 'keep talking' and 'uhum?' (Ericsson, 2017). These verbalizations are called level 1 and level 2 verbalizations and these data reveal what information a user needed and in what order. To arrive at these data the protocol is as follows:

- Collect and analyze only level 1 and 2 verbal data.
- Give detailed *initial* instructions for thinking aloud.
- Remind participants to think aloud at regular intervals.
- Otherwise, do not intervene!

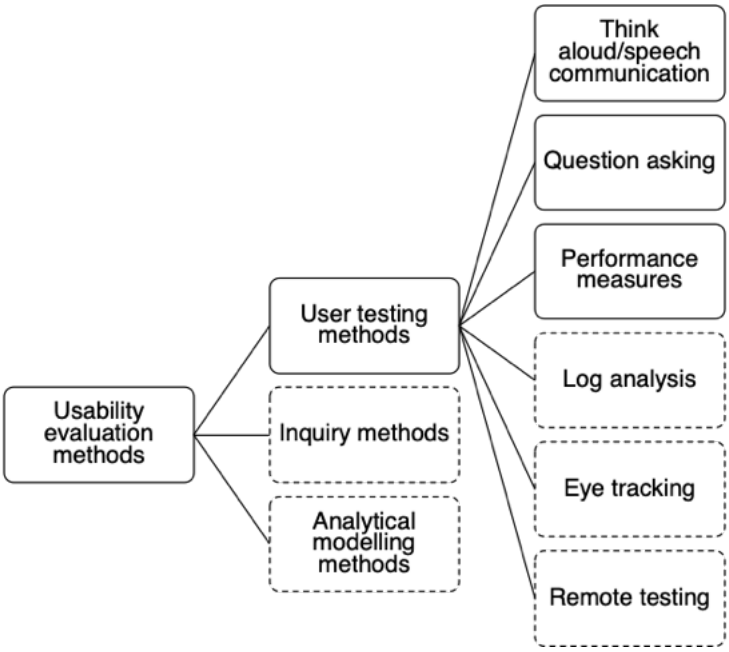

Figure 2. Overview of usability evaluation methods

However, behavior in usability testing can be highly variable (i.e., more unexpected), tasks are more complex and there are more experimental unknowns, which challenges the effectiveness of the strict, non-intrusive verbal protocol (Deffner, 1990). As a result, others argue that information obtained from participants' long term memory is equally relevant because that can provide explanations, coherency and design or revision ideas (Dumas & Redish, 1999; so-called level 3 verbalizations). These verbalizations are triggered by probes such as 'Why did you click on that orange tab?'.

As a reaction, a *speech communication* approach is proposed as an alternative. This approach acknowledges that communication between, e.g., a participant and an administrator, requires "back channels" from the listener (i.e., being an active listener) (Boren & Ramey, 2000). It is less non-intrusive as the TA protocol but still allows collecting level 1 and 2 verbalizations. Krahmer and Ummelen (2004) show that the speech communication protocol resulted in more tasks being completed and participants being less 'lost' compared to the TA protocol. Given the complexity of the systems to be tested within the SSI context, speech communication seems a promising protocol, which is partially in line with earlier protocols used in testing the @HBS app (see Giesen et al., 2019, who propose live observations of tasks and immediate interviewing afterwards).

SPEECH COMMUNICATION PROTOCOL

Table 23 shows the considerations when setting up the speech communication protocol. These considerations need to be addressed when setting up the research protocol for the small-scale experiments within the SSI project.

Table 23. Consideration for speech communication protocol (Weichbroth, 2020)

Area	Elements	Considerations
Setting the stage	System/product/application	Participants must understand that it is about the system not about them.
	Primary speaker	Participants must understand that they are the important contributor (i.e., expert).
	Primary listener	The practitioner is the learner and listener. (Other roles: host, technical support, should be played small.)
Data-collection	Acknowledgement tokens	Use acknowledgment tokens carefully to play role of engaged listener.
	Choice of acknowledgement tokens	Carefully choose tokens because they might affect available responses (see Drummond & Hopper, 1993).
	Frequency of acknowledgement	Acknowledgement should follow the flow of current communication.
	Keep participant talking	Complexity, for one thing, might lead to participants stop talking. Reminders should be unobtrusive, and imperatives are to be avoided (i.e., because of authoritarian nature).
Interactions	Technical issues	Either silently fix the problem or explicitly suspend the test and interrupt.
	Restarting participants	Interactions are needed when: a) the participant thinks the task is complete when it is not, b) the participant sidesteps an important functionality, c) the participant is stuck.
	Other communication	Interactions are needed when: a) the participant asks a question about the task, b) the participant asks a question that suggest an unexpected approach to the task, c) the participant is unusually 'chatty'.
Proactive interventions	Clarifying an unclear comment	Prompts should not influence the participant's response.
	Probing for more information	Preferably done post-test debriefing.

EXAMPLES OF ASSIGNMENTS FOR TA

Within the context of this SSI project, the usability testing of HCI will focus mainly on the usability attributes that relate to the completion of complex tasks (see above), whereby the support of smart features is a central element. In concrete terms, this would mean that in the TA experimental setting participants are given assignments that are related to the interaction with these smart features. When it comes to scanning tickets to facilitate the registration of expenses in the HBS diary, examples include assignments such as:

- scan a good receipt and add expenses to diary
- scan a moderate receipt and edit expenses before adding to diary
- scan a bad receipt (which generates no data) and proceed

Similarly, in the case of using geo tracking to facilitate the registration of activities in the TUS diary, assignments such as:

- enable geo tracking
- add a designated geo location
- use geolocation to edit tentative activity and add it to the diary

Although the focus is on testing the use of the smart features of the applications, room is also reserved for assignments that are an essential part of an end-to-end solution for TUS and HBS studies and may be specific to National Statistical Institutes (NSIs) or other stakeholders. This could, for example, concern:

- read the invitation letter and follow up (i.e., download and install the application)
- continue from the questionnaire to the diary
- invite household members to participate via the application
- report bug or error

OTHER USABILITY TESTING OPTIONS

It is likely that not all usability attributes and smart features can be tested in the experimental settings assumed by TA. Attributes such as learnability and memorability require longer use of the app than is possible in experimental settings. This also applies to testing the usability of geo-tracking. After all, for this a participant has to move about. At the same time, it might be desirable to test the HCI in terms of Active Learning (AL). This involves, for example, assessing the extent to which the user improves the ML algorithm for COICOP classification by adjusting the generated, tentative data. Longer use of the applications is also desirable for this.

For this reason, in addition to TA, the options to measure usability through asking questions after longer, independent and autonomous use of the application (see Giesen et al., 2019, who propose live observations of tasks and immediate interviewing afterwards) and performance measures (e.g., based on paradata) are also retained (see Figure 2). By asking questions, specific questions can be asked about user experience of certain aspects of the application after use. Performance measures, on the other hand, allow the 'use' (i.e., AL of the ML algorithm) to be measured in an objective manner.

5. CONCLUSION

The SSI project involves different platforms and applications that are smart in the sense that they are technically capable and functional solutions to conduct complex studies and can be extended by smart features that alleviate the complexity of the study and potentially further

increase the reliability and accuracy of the data. Yet the true test of these smart solutions and smart features lies in the ability to use the platforms and applications for the intended purposes. Usability tests might reveal performance on attributes of usability that might improve recruitment and retainment of participants, participants' willingness to share personal data, and participants' ability to complete complex tasks.

[1] Designation can be done either by respondents themselves or by linking geo-points to public data such as OpenStreetMap.

[2] Note that TUS was not only tested on the MOTUS platform but also using the native, Austrian Time Use application (see https://www.statistik.at/zve). Since the latter application is no longer part of the SSI-project, findings will not be summarized here.

Chapter 4: Combining smart and traditional survey methods: Mode effects and other data integration considerations

1 Introduction

Over the past two decades, sensor arrays and machine intelligence have moved from the exclusive domain of technophiles to become so mundane that we often take them for granted. Most people have a smartphone, and more people than ever report that they feel comfortable interacting with smartphones (Couper et al., 2018; Keusch, Wenz, et al., 2022). The sensors contained in a typical smartphone, such as cameras, accelerometers, GPS receivers, ambient light sensors, or gyroscopes, have become embedded in users' everyday life tasks with the goal of making things easier, faster, and more accurate (Khan et al., 2013). Users have become accustomed to the ways in which these devices can improve their experience.

In the same two decades, response rates to a broad range of long-running surveys have declined, requiring institutions such as National Statistical Institutes (NSIs) to expend more resources to achieve comparable sample sizes (Stedman et al., 2019; Luiten et al., 2020). The causes behind the falling response rates are unclear, although the sheer number of requests for participation and increase in surveyors from the commercial space has been proposed as a factor (Dillman, 2015). In the past, surveys following up with non-responders have suggested issues of salience/relevance, burden, and lack of interest (Tait et al., 1995; Couper et al., 2007; Singer & Couper, 2017). Unsurprisingly, these same aspects are also well-represented in surveys of what respondents find bothersome in surveys (Johnston, 2014; Husebø et al., 2018; Mayer, 2019). The places where surveys fail to perform, such as in asking repetitive questions, requiring heavy time investment, and precise and accurate measurement of things like time or space, are exactly the places in which sensors and algorithms shine. This fortuitous overlap has not gone unnoticed by survey researchers, and the last decade has been marked by an increase in "smart surveys" seeking to augment existing methodology by using the tools readily at hand within smartphones (Couper et al., 2018; Link et al., 2014; Struminskaya, Lugtig, et al., 2020).

Although these smart surveys can be deployed in isolation, researchers whose current surveys might make use of some of the theoretical benefits, are interested in integrating results from smart surveys with historical data sources and ongoing, established surveys. In addition, there may be a need to continue traditional surveys for specific sub-groups in the population. This presents an unfortunate conundrum, as the format of data gathered by

sensors are often very different from data acquired via survey questions and these data can require considerable cleaning and processing before it can even be directly compared (Harding et al., 2021; Kaplan et al., 2020; Keusch et al., 2023; McCool et al., 2021). Incomplete coverage of smartphones, combined with the potential for a differential self-selection bias between smart surveys and traditional surveys complicate the matter further (Stone et al., 2023; Wenz & Keusch, 2023). At the moment, there exists no comprehensive methodology proposing steps for the integration of data arising from smart and traditional survey methodologies.

While the usage of smartphone-acquired sensor data is certainly a new challenge, the field of survey methodology has contended with similar issues in the past. Mixed-mode design, in which a survey is delivered across multiple platforms (e.g., via telephone and face-to-face), has been used for decades to improve low response rates, and adjust for issues of selection and coverage (de Leeuw & Hox, 2008; Klausch, 2014; Schouten et al., 2021b). Lessons learned on mode effect estimation and data integration of other disparate modes can provide a framework for smart surveys, although the larger differences between traditional and smart data means a higher onus on researchers to demonstrate measurement equivalence. This review uses the Total Survey Error (TSE) framework to investigate and describe potential areas for differences to arise between smart and traditional modes of administration (Biemer & Lyberg, 2003).

This literature review aims to accomplish the following:

1. Identify, classify, and quantify sources of error that may pose risks for the integration of smart surveys with traditional survey methods
2. Establish patterns of similarity between smart/traditional survey integration and previous research on mixed mode surveys
3. Provide an overview of methods to disentangle the various sources of error

In Section 2, we describe and outline different examples of smart surveys that provide concrete examples for the sections that follow. In Section 3, we briefly describe and review the literature on Total Survey Error to provide the necessary vocabulary for following sections. In Section 4, we present relevant literature on mixed mode survey methodology and its relationship to the question at hand. In Section 5, we present the literature describing initial findings on mode effects in smart surveys. In Section 6, we present results on estimation methodology. Section 7 follows with relevant findings on data integration. Finally, in Section 8, we synthesize the findings from the literature, provide recommendations, and suggest experimental methods for closing the gaps in existing literature.

2 SMART SURVEYS

SMARTPHONES AND APPS

Well before apps gained their current level of prevalence, researchers were investigating the usage of smartphones and other mobile devices independently of their capacity to provide complementary external data to the survey. Couper et. al (2017) offers a comprehensive review of the literature on web surveys completed on mobile devices. Important considerations included differences in coverage, non-response, break-offs, and how best to design web surveys to accommodate the new device (Pearce & Rice, 2013; Toepoel & Lugtig, 2015; Peterson et al., 2017).

The primary difference between completing a web survey on a smartphone and using an app on a smartphone is the length of time a person will need to interact with the device. Aspects such as coverage and differential non-response remain pertinent to smart surveys. Although smartphone penetration has increased in the United States and Europe, the differences between who has them and who does not have remained.

LEVELS OF SMARTNESS

The difference between a web survey conducted on a smartphone and a smart survey is sometimes not immediately clear. In this way, it may be useful to describe different levels of 'smartness' that a survey may have. A survey asking a respondent about the last item they purchased in a store, even if accessed and completed on a smartphone or tablet, would not be considered a smart survey if the respondent answers by filling in a text field as they might on any other mode. On the other hand, the simple addition of a search bar to an input field could be considered a smart feature, as it makes use of the device's capacity to interpret input, retrieve data from a stored list, and display the resulting options, reducing a user's total effort in typing out a complete and well-formatted word.

A high level of smartness for a similar question might involve scanning the barcode or taking a picture of the item you purchased in order to provide an answer. Although the gradation is not clear-cut, a high level of smartness tends to involve device sensors as its smart features because these offer an extended set of tools for meeting the goals of smart surveys: reducing burden and measuring concepts that respondents are unlikely to know or cannot measure. Schouten et. al (2021a) list several smart features that smart surveys may have: device intelligence, internal sensors, external sensors, access to public online data, access to personal online data, or linkage consent. Often, fully-developed smart surveys will employ combinations of many of these at once. Each of these features is likely to contribute its own sources of differential measurement. As a consequence, the smarter the survey, the more measurement differences researchers are likely to encounter.

Following are three examples of smart survey types that are currently in use among NSIs. These surveys share a common history as complicated pen-and-paper diaries that often required interviewer assistance. Their high burden and the presence of questions that are difficult to measure or recall have made them ideal targets for novel methodologies over

the years, which allows investigation into the impact of using differing combinations of smart features.

MOBILITY

Surveys looking to measure people's travel behavior identified shortcomings from the beginning (Clarke et al., 1981). The goal of these studies is to reliably measure travel behavior for a sample within a given geographic area, including aspects of the travel such as mode of transportation, precise start and stop times for each trip, and addresses for visited places, which is accomplished by asking respondents to record this information in diaries spanning varied lengths of time depending on the study (Axhausen, 1995). Past studies have identified differences in reporting between days incorporating interviewer assistance and not, and between recorded behavior and road sensors (Ampt et al., 1985; Ashley et al., 2009).

The mobility survey represents the first of the included surveys to incorporate smart features. Early in the 90s, researchers began to make use of standalone GPS receivers for the purposes of recording all trips (Sarasua & Meyer, 1996; Yalamanchili et al., 1999; Bricka et al., 2009). While this worked quite well, leading some proponents to pose the GPS logger as a complete solution that would eliminate the need for respondent involvement altogether (Wolf et al., 2001), the capacity for accurately determining trip purpose, transportation mode, and the identification of missing data has yet to prove itself as accurate as user input (Gong et al., 2014; Bähr et al., 2020; Nguyen et al., 2020; Sadeghian et al., 2021).

At the same point in time, other researchers experimented with bringing travel diaries online (Arentze et al., 2001; Adler et al., 2002). This allowed for the introduction of different smart features, including machine intelligence that added checks to the data entry stages that prevented impossible or unlikely entries, and linkage to personal data to decrease respondent burden (Hoogendoorn-Lanser et al., 2015).

In the early 2010s, smartphones began to come with embedded GPS technology and other sensors that made it feasible for them to record user locations, and researchers began to develop smart surveys for mobility behavior that made use of these features (Cottrill et al., 2013; Nitsche et al., 2014; Berger & Platzer, 2015; Greaves et al., 2015). Here, too, the specific smart features differed per app: some made use of additional device sensors, fusing the GPS records with accelerometer data (Prelipcean et al., 2018), and some integrated the machine-based check mechanisms with user feedback (Greaves et al., 2015). Soon, recommendations began to emerge for how best to make use of all possible smart features in order to improve data quality and reduce user burden (Harding et al., 2021).

As the travel diary became increasingly smart, it introduced new avenues that could account for previous sources of error, as well as new avenues for error to occur. While GPS coordinates could help to reduce recall error for respondents, the sensor could also fail in a number of ways that pen-and-paper studies were unlikely to fail. Determining the reasons for different outcomes between surveys with and without smart features requires considering each of these levels independently.

EXPENDITURE

Expenditure data, often gathered in the form of recall or diary studies, has seen declining response rates and data that don't align well with aggregate measures (Crossley & Winter, 2014). While early research into expenditure involved either maintaining daily diaries, or retrospective surveys, independently these modes were both lacking. People face difficulties in estimating the amount of money spent on consistent but irregular purchases, such as grocery shopping or transportation costs, which made retrospective surveys a poor choice for documenting daily behavior (Crossley & Winter, 2014; Sekula et al., 2005). People were also limited in their capacity to specify beyond basic levels of categorization, such as "food" or "clothing" , and when restricted to shorter time periods, tended to "telescope" their answers by including responses occurring before the specified period (Crossley & Winter, 2014). The alternative, paper diary studies, allowed for categorization into different products, but this level of extensive reporting could only be carried out for a brief length of time and the quality of the collected data decreased even over the two-week timespan often requested. Because of this, many of the larger expenditures such as healthcare costs, appliance purchases, or rent were very difficult to capture with the diary method. The current methodology employed by NSIs therefore deploys to each household both a diary for daily expenditures and a face-to-face survey asking about the larger line items that would be missed with the diary (EUROSTAT., 2003). Expenditure research must already contend with the concept of data integration with its two complementary sources.

There have been multiple efforts to improve the quality of the data generated, but recent work suggests that the intensive burden of having to report all expenditures by writing down amounts and details is a hindrance to both nonresponse and measurement quality (Wenz, 2023). Issues of diary fatigue, where reported expenditure declines over the measurement period, are quite common (Brzozowski et al., 2017; Silberstein & Scott, 2011). Additionally, as expenses are shared at a household level, obtaining a clear picture for households of two or more people requires either extrapolation or collaboration. Lastly, in some countries, respondents must provide detailed itemization of all purchases, for example to be able to distinguish between meat, vegetables and hygiene products, all of which may be purchased at the same store. This granularity is crucial for classification purposes, for example, to be able to assess the impact of taxing different expenditure categories differently.

Fortunately, respondents are generally quite good at being aware of large and regular purchases – in other words, when the task that is required is central to the respondents. Thus, of the three benefits that smart surveys offer, expenditure research benefits most from a reduction in burden. This can be done with the introduction of minimally smart features and an app-based diary that can assist with the product input, prompting respondents for the necessary specifics such as type and quantity. As with the mobility case, previous efforts to decrease the burden have involved moving the data collection online, allowing for the incorporation of decision rules to attempt to prevent motivated misreporting (Eckman, 2022). More advanced smart surveys can offload laborious tasks onto the available sensors, by taking pictures of the receipts to automatically fill in line items (Jäckle et al., 2019; Wenz, 2023), or by using geolocation to offer reminders when

people are in areas where they are likely to make purchases. The opportunities afforded by the addition of one or more smart features are significant, and are expected to lead to both richer data as well as more of it.

Data generated under these new conditions, however, are at risk of being quite different from data gathered without these benefits. This is exacerbated by the existing complexities required to integrate the large-purchase face-to-face surveys with the diary. While NSIs have abundant macro-level consumption data, such as national and bank account data, the household budget survey is often the sole micro-level source, making it crucial that this adjustment and integration process is carried out with care.

TIME USE

Most contemporary Time Use Diaries (TUDs) use research protocols, including the Harmonised European Time Use Survey (HETUS) and American Time Use Study (ATUS), have origins that can be traced to the work of Szalai (1972). Over the years, the scope of TUDs has broadened (Frazis & Stewart, 2007). The emphasis has shifted to garnering more intricate categorizations of time spent, aiming to enable comparisons both within households and between them (Bauman et al., 2019), resulting in a decline in response rates for a field where this was already an issue (Abraham et al., 2006; Elevelt et al., 2019).

However, this increase in scope has been accompanied by an increased respondent burden. For instance, the hierarchical coding system in HETUS, which comprises roughly a hundred distinct activities nested under nine overarching categories (Eurostat, 2009), can be cumbersome for respondents to navigate. Additionally, the solicitation of supplementary details like co-presence, enjoyment, mobility, and tech utilization often yields incomplete responses (Abraham et al., 2006). As the demands on the data grew, longer periods of collection became necessary (Glorieux & Minnen, 2009; Frazis & Stewart, 2012).

TUDs originally emerged as a methodology designed to combat problems with overestimation and underestimation associated with recall of daily activities as people tend to be more accurate the closer in time they are to the described period (Schwarz, 2012). Despite this, biases persist, even in the diary format, where respondents tend to overestimate certain tasks (Harms et al., 2019; Sullivan et al., 2020), and underestimate others (Kelly et al., 2015).

As with mobility and expenditure, new technology offered researchers a chance for secondary data streams to augment their methodology. Over the last ten years, the paper diary has been paired with external smart devices, such as cameras (Gershuny et al., 2020), accelerometers (Harms et al., 2019; Gershuny et al., 2020), and GPS units (Millward & Spinney, 2011), often with the goal of validation against an objective instrument. A single point of interaction combining the data streams of each smart feature with the diary itself can improve the data quality by 1) allowing the respondent to integrate all available information in his or her answer, remaining the single source of truth, and/or 2) reducing the burden on respondents by completing fields with the captured data.

In addition to these high-smart features, TUDs on smart phones may see the greatest benefit in the short term from the introduction of in-app features that involve no sensor data at all. For example, automated look-up of activities as the user types should assist with improving categorization (Minnen et al., 2014; Rinderknecht et al., 2022), error-checking rules can improve per-activity full completion rates (Chatzitheochari et al., 2018) and in-app reminders can be improve diary fatigue and drop-out (Lev-On & Lowenstein-Barkai, 2019; Chatzitheochari & Mylona, 2021).

The transition to smart survey is not without issue, however. Researchers have demonstrated that the physical interaction with the input mechanisms can be cumbersome, especially as the amount of auxiliary information increases (Sullivan et al., 2020). Efforts to offload tasks from the machine to the respondent must be careful not to get in the way of the user. Lastly, and most centrally to the task at hand, the new streams of data may collect more accurate but incompatible information, such as when comparing ground truth to a coarser gradation of time (Gershuny et al., 2020).

Smart TUDs may offer unique opportunities to researchers that have no equivalent in their traditional counterparts. Visualizations such as tempograms and transition diagrams, currently in use by researchers for analysis purposes (Kolpashnikova et al., 2021), can also be generated on demand to provide user feedback, potentially increasing engagement. And apps, not confined to the limits of a page, can be designed to reduce or eliminate language barriers by using pictograms to represent activities (Daum et al., 2018). This opens entirely new areas to applied researchers. Indeed, while the potential advantages of bringing time use research onto smart surveys are substantial, it is this very transformative nature that underscores the imperative for a deliberate and nuanced approach to their integration with established survey methodologies.

3 Total Survey Error

Total Survey Error (TSE) is a paradigm in which the varied ways that error can permeate through a survey can be described, and provides a basis for their joint and independent evaluation for contribution to the overall quality of the survey estimates (Biemer, 2010; Groves & Lyberg, 2010). Total survey error, conceptually, describes the difference between a parameter as it might be measured within a population, and the estimate of the same parameter as it might be measured by a survey (Biemer & Lyberg, 2003, p. 36). If the objective is to compare a smart survey against its non-smart counterpart, we are ultimately interested in the comparison of each of these against the population. While adaptations of the TSE framework have been proposed for big data, found data, and metered data (Amaya et al., 2020; Biemer & Amaya, 2020; Bosch & Revilla, 2022), none have been proposed for smart surveys. We therefore relate the scheme as presented by Biemer and Lyberg (2003) to the case studies at hand. This version offers sufficient flexibility to categorize and demonstrate the differences in potential error sources between smart surveys and traditional surveys. Figure 3 is a graphical overview of the categorization levels: within total survey error, we distinguish between sampling error, caused by the process by which the

sample is drawn from the population, and nonsampling error, of which we distinguish five categories.

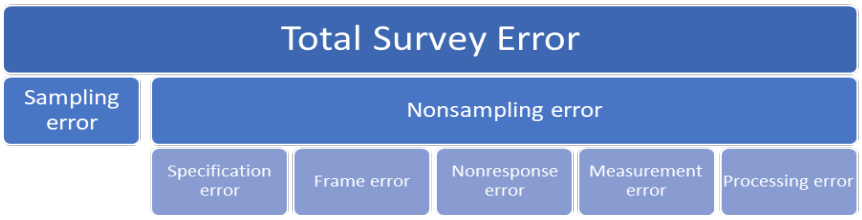

Figure 3. Total survey error framework

Given that NSIs are likely to employ consistent sampling procedures for both smart and non-smart surveys, integration concerns arising from this aspect are minimal. We therefore focus on differences arising within nonsampling error and describe each of the five categories in brief. Specification error arises when there is a mismatch between the parameters of interest for the researchers and the information that the survey will capture. Frame error, also referred to as coverage, results from the failure of the sampling frame to adequately represent the population. Nonresponse error comes from a sampled person's failure to respond to the survey instrument, either completely, or in part. Measurement error arises when a respondent answers in a way that differs from the truth, whether intentionally or not. Finally, processing error comes from processing, coding, editing, or working with the data.

The categories of nonresponse error benefit from an additional structural layer. Nonresponse can be called unit nonresponse if the sampled person does not respond to any part of the survey, or item nonresponse for when a sampled person has some response, but it is incomplete. Traditional diaries often make detection of item non-response quite difficult. Importantly, for diary studies, which are intensive and longitudinal, item nonresponse may be more complex (Lynn & Lugtig, 2017). While traditional item nonresponse is often conceptualized as questions left unanswered, the existence of patterns occurring over time, such as response that decreases over time or ends prematurely suggest the need for a third category of nonresponse or a classification of differential item response patterns.

In addition to the longitudinal aspects of diary survey methods, there are additional considerations specific to smart surveys making use of passive data collection. Bosch and Revilla (2022) note two important deviations for passively-collected data from actively-collected data: it is difficult to distinguish missing data from absence of behavior, and similarly difficult to categorize missing data as either item nonresponse or measurement error. In their adaptation of the TSE framework to Big Data, Amaya et al. (2020) address this by assessing the concept of missing data error in place of nonresponse error, noting that the confounding can be abated when the generation mechanism of the missingness is identifiable.

4. MIXED-MODE SURVEYS AND MULTI-SOURCE STATISTICS

A survey may be administered through one or more methods, including face-to-face, paper-based, telephone, or via a smartphone app. The choice of mode by which a survey is administered is known to influence the accuracy of the data collected (de Leeuw, 2018). When the same survey content is assessed by researchers by differing modes of response, the survey design is considered to be mixed-mode, as distinct from single-mode (de Leeuw et al., 2015a). Each mode of administration in a mixed-mode survey will accumulate error within the non-sampling error components: specification error, frame error, nonresponse error, measurement error, and processing error. When differences in error exist between different modes, we speak of mode effects.

To some degree, mode effects represent the desirable element of conducting mixed-mode surveys. A telephone-based survey is limited in its coverage by default to persons who possess a telephone and web-based surveys will encounter coverage errors related to Internet access, but the development of a survey design that incorporates both modes will have greater coverage of the total population, assuming that th A survey may be administered through one or more methods, including face-to-face, paper-based, telephone, or via a smartphone app. The choice of mode by which a survey is administered is known to influence the accuracy of the data collected (de Leeuw, 2018), and when the same survey content is assessed in differing modes, the survey design is considered to be mixed-mode (de Leeuw et al., 2015a). Each mode of administration in a mixed-mode survey will accumulate error within the non-sampling error components, which, when this differs between modes, is called a mode effect.

To some degree, mode effects represent the desirable element of conducting mixed-mode surveys. A telephone-based survey is limited in its coverage by default to persons who possess a telephone and web-based surveys will encounter coverage errors related to Internet access, but the development of a survey design that incorporates both modes will have greater coverage of the total population, assuming that the two modes differ in their coverage error. Most researchers who employ mixed-mode designs make use of this fact in order to improve coverage and response (de Leeuw, 2018). On the other hand, mode differences that do not contribute to an overall decrease in coverage/nonresponse errors are frequently seen as nuisance elements to be avoided or corrected for against some gold standard measurement (Klausch et al., 2013; Burton & Jäckle, 2020). This view is at odds with the goals of smart surveys, which often seek to combine the benefits of both active and passive measurements precisely because of the lack of a gold standard.

In their book Mixed-Mode Official Surveys, Schouten et al. (2020) devote a chapter to the discussion of smart devices as an emerging new mode, noting that the new types of data "challenge the comparability of response with and without" the data (2021a, p. 223). The task of combining data generated by smart and non-smart surveys may ultimately bear greater resemblance to combining data from different sources if the variables arising from traditional surveys and smart surveys differ in their level of aggregation or frequency, corresponding to situations 7 or 8 respectively as discussed by Waal, Delden, and Scholtus

(2020). We can therefore contrast the mixed-mode paradigm with the multi-source paradigm in which the existence of differential error between data sources can provide a method by which to compensate for the disadvantages of each (De Broe et al., 2021). Although the perspectives between mixed-mode and multi-source statistics differ, the methodology for the estimation of differences between the two is very similar, and so this section condenses literature out of both disciplines. We will assess the relevant literature on mode/source differences at each level of nonsampling error within the TSE framework.

Mode effects due to specification

There has been relatively little attention paid explicitly to the concept of specification error as it relates to mixed-mode survey design, although the importance of proper concept specification as the "backbone" of survey quality has been repeatedly emphasized (Salant & Dillman, 2008; de Leeuw et al., 2015b). Specification is the process by which the concepts of interest are translated into a variable that can be measured by the survey instrument, and specification error the mismatch between the two. Careful alignment of theory and questions by involving everyone in the process, along with a pretesting stage, can identify specification error (de Leeuw et al., 2015b). Regardless of whether the operationalization has been sound, survey modes that do not differ in their presentation of the question are unlikely to elicit differences here -- except perhaps longitudinally (Lynn & Lugtig, 2017). In this way, the unified mode approach, in which all modes have questions phrased as similarly as possible, limits the introduction of mode specification effects (Dillman et al., 2014; Dillman & Edwards, 2016). The line between mode specification effect and mode measurement effect is not always clear in the data. In their chapter on Mixed-Mode Research, Hox et al. (2017) note the potential for instruments to "reflect different constructs across modes," in the worst-case scenario of mode measurement effects.

Unlike in the mixed-mode domain, the difficulties arising from mode specification effect come up regularly in multi-source literature, both because the data sources under consideration may be created independently of each other, and because the collecting instrument may limit the ways that the concept can be operationalized (Zhang, 2012). Here, too, there is confounding with measurement effect, but often the presence of clear differences in the operationalization of a concept lends itself to seeing differences at the level of specification rather than measurement.

Mode effects due to frame coverage

Coverage error results from the sampling frame representing the population insufficiently. Consequently, mode effects can arise within coverage if the capacity of two modes to adequately represent the population differs in some way. If researchers make use of the same sampling frame for the invitation procedure, coverage differences are unlikely as a letter will reach both groups with the same probability.

A sampled person receiving such a letter may make the choice to respond or not differentially based on whether they are invited to participate in a smart survey, but so long as they have the same capacity to respond, this distinction is one of nonresponse rather

than coverage. The situation in which a mode effect due to frame coverage exists is therefore one in which a differential capacity to respond to the survey instrument exists between two modes.

Coverage was a primary concern in early research as smartphone ownership differed across key demographics. Today, many researchers operate under the assumption that the rapid uptake of mobile telephone ownership has rendered this issue moot, and indeed, the most recent report from the Eurobarometer covering this topic indicates that 96% of Europeans report having access to a mobile phone (E-communications and the digital single market: Report, 2021). However, there remains ample evidence demonstrating that both smartphone ownership and smartphone usage are unevenly distributed within the population (Klingwort & Schnell, 2020; Keusch et al., 2023).

If the only avenue offered to survey participants is via a smartphone app, this is likely to lead to coverage differences and may in fact contribute to the largest source of difference between modes (Antoun et al., 2019). Prior research has demonstrated meaningful coverage differences between web-only and web+mail response options, even when penetration is high (Bandilla et al., 2014; de Leeuw, 2018). The situation is likely to be similar in a smartphone-only condition, with a bias towards higher education and more affluent persons in smartphone-only response (Couper, 2007).

An additional concern arises when researchers must make decisions on whether to develop their smart survey within a single ecosystem. Developing apps that can be deployed both to Android and iOS is more expensive and may come at the cost of feature loss, but known differences between Android and iPhone owners make this risky (Götz et al., 2017; Keusch, Bähr, et al., 2022). Similar to the push to add mixed mode response to other surveys to mitigate coverage error, it remains important for this concern to ensure that all sampled persons are capable of responding to a version of the survey instrument.

Mode effects due to nonresponse

Nonresponse error has been frequently addressed within mixed-mode survey design, often in the context of increasing response rates by adding new modes, with the ultimate goal of decreasing total nonresponse error, at the putative cost of increasing measurement error (Sakshaug et al., 2010). Unlike between specification error and measurement error, there is a distinct boundary between the concepts of nonresponse and measurement error, allowing researchers to disentangle the two sources by experimental design.

Unit nonresponse

A primary concern with unit-nonresponse is the biasing impact arising from the differences in patterns of data between the people who respond to a survey and those who do not. In the context of any one mode, this comparison is with regards to the population, but in the context of the comparison of smart and non-smart surveys, the concern is whether non-responders to each mode differ from the population in the same way. One primary cause of this non-response bias is likely to arise from privacy concerns.

Where the smart features are more invasive, such as with a web-tracking app or GPS mobility-tracking app, participants are much more likely to report being not at all willing to complete these data collection tasks on a smartphone (Wenz et al., 2019). At least in the mobility case, this appears to be a meaningful distinction for people between the automated reporting of an app-based system, and the completion of a paper diary, with self-reported willingness to participate differing upwards of 20% in some countries (Verzosa et al., 2021). Multiple studies investigating reasons for non-participation in app-based studies have indicated that privacy concerns play a critical role (Kreuter et al., 2020; Struminskaya, Toepoel, et al., 2020; Roberts, Herzing, Sobrino Piazza, et al., 2022). On the other hand, with traditional diaries, non-participation is often due to the perceived effort involved (Verzosa et al., 2021). Taken together, the differential reasons for non-response across differing methodologies indicates a high likelihood of finding mode effects that are due to non-response.

While the interplay of privacy versus effort may be the most salient aspect to respondents, there are additional factors that may influence nonresponse. Here, a primary tool for assessing these differences are studies investigating hypothetical willingness to participate in smart surveys. These are embedded within non-smart surveys, from which we infer that stated non-willingness to a future smart survey implies a difference in response between the two modes. Willingness to participate differs across sociodemographic groups as well as attitudinal measures (Wenz et al., 2019; Struminskaya et al., 2021; Wenz & Keusch, 2023). A frequent finding is that surveys involving smartphones decrease the probability of response in older persons, whereas this demographic tends to show increased response probability on traditional surveys (Roberts, Herzing, Sobrino Piazza, et al., 2022; Felderer & Herzing, 2023). Other areas with a strong potential for differential nonresponse include respondent's IT literacy (Felderer & Herzing, 2023), differences in education (de Bruijne & Wijnant, 2014; Felderer & Herzing, 2023), and differences in employment status (Roberts, Herzing, Sobrino Piazza, et al., 2022).

Item nonresponse

Current household surveys often contain specific questions or sections where respondents are more likely to skip them or provide incomplete information. This is in fact one of the primary motivations for many researchers to include smart features in their survey: establishing whether a day with no recorded trips or expenses was true or not. There are several ways to limit item nonresponse in smart surveys: by using location measurements, by using notifications to respondents, or by prompting for missing details such as quantity of an item purchased in budget surveys.

Researchers conducting web surveys have investigated ways to reduce the consequences of item nonresponse by introducing smart features that check and validate a user's entries, which is in some ways analogous to the presence of an interviewer by the completion of a survey, who can direct respondents to complete missing items (Conrad et al., 2005). A similar strategy of employing edit checks at the moment of answer entry has been successfully used by interviewers to reduce underreporting (Lugtig & Jäckle, 2014). A potential concern in this area is whether there may exist a tipping point for respondents

where “checks” as a smart feature are concerned (Peytchev & Crawford, 2005). For example, in requiring respondents to enter all auxiliary information on daily activities (location, participants, enjoyment, etc.), respondents may be disincentivized to enter more activities than necessary, which would result in paper diaries being simultaneously less complete in activity context, but more complete in covering the breadth of the activities (Chatzitheochari et al., 2018).

A further source of differential item nonresponse between smart surveys and their traditional counterparts is the impact that the device may have on missing data. This is especially true when the smart feature under consideration is a sensor, as this requires the sensing device to be charged and functional, which can prove challenging (Struminskaya, Lugtig, et al., 2020). Such unexpected technical challenges are rare in pen and paper diaries, but common in research involving smartphones. The interaction with the survey instrument on the smartphone, if not optimized, has been shown to produce more missing data, and have higher breakoff rates. (Mavletova & Couper, 2015; Roberts, Herzing, Manjon, et al., 2022). While this is unlikely to be the case with smart surveys that are specifically designed with smartphone interfaces in mind, aspects inherent to these devices, such as screen size or internet connectivity, pose unique challenges for smart surveys that do not exist with pen-and-paper surveys.

Mode effects due to measurement error

Measurement error in surveys can stem from various sources, either from the respondent, such as with social desirability bias and satisficing bias, or the survey instrument itself, impacting its usability. Because smart surveys will differ from non-smart surveys in the interaction between these critical elements, we can expect this to contribute to the overall mode effect in a meaningful way, despite research indicating few differences between mobile and PC web surveys (Couper et al., 2017; Antoun et al., 2019). Mode effects might either shift the overall response distribution or modify the question-answer process, producing non-equivalent responses between different modes (Hox et al., 2017). An important distinction is that smart surveys are designed to offload some portion of the response generation process onto the user. We may also see the goal as mixing the benefits from each mode, in this way reducing the total survey measurement error (Tourangeau, 2017).

A respondent’s low level of involvement can lead to rushed responses, misunderstandings, or approximations. Previous mixed-mode research has indicated that careless reporting is fairly consistent across paper, web and smartphone surveys (Magraw-Mickelson et al., 2022). Similarly to their capacity to help with missing data, edit check rules that discourage or disallow reporting of improbable events have been used in web surveys to successfully reduce the measurement error commonly encountered in pen-and-paper diary studies (Conrad et al., 2005). Respondents often struggle to understand the intention of the researchers when answering questions, and this is made more difficult in situations in which their knowledge of the topic differs from that of researchers. For example, being able to search through the various categories for a specific item in the Household Budget Survey, or activity in the Time Use Survey are likely to produce categorizations that are more complete.

While categorization implementations may represent a considerable benefit, the total impact of its usage can be seen as a facet of a broader consideration. The interpretation of questions depends on question wording and layout, and this can have a meaningful impact on measurement error (Scherpenzeel & Saris, 1997; Kasprzyk, 2005). These may differ by necessity between smart and traditional surveys or may simply be interpreted differently due to the context. This is especially relevant because we know that respondents in interviews interpret questions differently from respondents in web surveys (Dillman et al., 2014), and an app that provides feedback may sometimes behave more like the former than the latter.

Interestingly, there is also potential for a decrease in social desirability bias with passive input, as shown by Keusch, Bach, et al. (2022) in their web-tracking study. Self-administered modes have long been known to reduce social desirability bias in sensitive answers relative to interview and face-to-face modes (Kreuter et al., 2008; Burkill et al., 2016). Passive smart features may be able to reduce this even further, by virtue of being outside of the cognitive purview of respondents. On the other hand, not all features that offer information to the user are likely to be taken advantage of, and in this case, may only serve to complicate procedures (Conrad et al., 2006).

UI/UX elements, like sliders or dropdowns, are known to induce more errors on mobile compared to web (Couper et al., 2017). These effects could be amplified in the case of an app, considering the length of involvement expected from the user, making design decisions that reduce measurement error in the mode a necessity. Differing physical characteristics of a smartphone can produce differing response quality of response (Wenz, 2021). This could lead to greater variability in response in smart surveys but might also induce bias due to existing relationships between personal characteristics and the particular device someone owns (Keusch, Bähr, et al., 2022).

Mode effects due to processing error Because traditional surveys are susceptible to a variety of human-induced errors in data entry and coding, it may be that smart surveys have the potential to reduce processing errors. In paper questionnaires, the task of interpreting the respondent's answers and aligning them with the proper categorization falls on the researcher. Where smart features can provide the tools to allow respondents to categorize activities themselves, this task is removed from the researcher and placed on the respondent (Ng & Sarjeant, 1993). The potential tradeoff here is one of reduced processing error for increased measurement error, in the case that the user is not always aware of the specific goals of the researcher. We expect neither mode to be perfect, but for there to be systematic differences between modes.

A concern for smart surveys is that some portion of the processing may not be visible. This is especially true in the case in which commercial entities are involved with the processing of the data and the algorithms used for processing are not freely available. For example, neither Google nor Apple share their proprietary algorithms by which the locations are generated on their respective mobile operating systems.

Aschauer et al. (2021) describe the processing steps of a combined travel/time use/expenditure diary, including the extensive validation process following plausibility

checks. Part of this processing involves the retrospective manual analysis of missing or unlikely data as part of a preparatory step before contacting users for validation. Here, sensor measurements would impact this processing and validation step.

5. MODE EFFECTS OF SMART FEATURES

As noted in prior sections, few studies exist considering mode effects in smart surveys versus traditional surveys. However, prior studies have investigated the impact that individual smart features may have on the collected data, which can be used to estimate effect sizes and directionality. Existing literature focuses on three sources of error in particular: coverage differences, non-response differences, and measurement differences.

The consideration of smart features with respect to coverage differences is similar to that of smartphone ownership and usage in the wider public, as discussed in Section 4. Because smart features intentionally take advantage of some distinct functionality, each additional feature increases the demands on the respondent's device, decreasing the available pool of respondents and in so doing, effectively reducing the frame relative to the original sampling frame. This impact is unlikely to be unevenly distributed: many people continue to use damaged phones for many years because the cost of repair or of a new phone is too high (Schaub et al., 2014), older phones and cheaper phones, both more likely to be owned by older persons, often cannot be upgraded past a certain version of their operating system, which may leave users unable to install an app built under more recent framework requirements (Mosesso et al., 2023). Some smart surveys may be developed to take advantage of features specific to one operating system or another – often focusing on either Android or iOS – but this practice may lead to coverage bias with respect to important outcome variables (Keusch et al., 2023).

Evidence for a combined difference in selection and measurement has been identified in the transition of HBS from non-smart to smart, with differential response across both personal characteristics, including employment, immigration status, children, age, and education (French et al., 2008; Riegler, 2015). Jäckle et al. (2019) identified similar differences in coverage and participation for a household budget survey conducted via app. Mode effects with respect to participant nonresponse has also been documented, with panel participants who reported using their smartphone for more discrete tasks being more likely to agree to take pictures of themselves, receipts, their house, or their surroundings when asked within the confines of an otherwise non-smart survey (Struminskaya et al., 2021).

Non-response error may differ across smart features. Active smart features may increase item non-response on sensitive topics if users consider them more invasive than purely textual responses, as Whatnall et al (2023) found when asking participants who had already reported their weight to take a picture of their scale. Conversely, passive smart features may decrease item non-response on sensitive topics, as the reactivity of changing behaviors that are being monitored lessens over a period of time (Keusch, Bach, et al., 2022). One of the largest impediments to the use of smart features is the increase in item non-response associated with passive measurements (Bähr et al., 2020; Struminskaya, Lugtig, et al., 2020;

Chatzitheochari & Mylona, 2021). Classifying lack of behavior as missing due to item nonresponse or true absence of behavior is common whenever we offload the response from active to passive, whether this is with web-tracking, or sensors (Bosch & Revilla, 2022; Courtney et al., 2023).

In a similar category to features providing passive measurement are those features like tooltips and linked help sheets, which can provide guidance on questions. The goal of these features is to enable users to better assess the pragmatic intent of a question, which remains an obstacle for respondents providing accurate responses (Schwarz, 2012).

Where specific comparisons have been made between surveys with smart elements and traditional surveys, they have been compared on the basis of measurement differences. Wenz (2023) looked at a comparison of a household budget app either with or without scanned receipts, in comparison with the national budget. After using inverse probability weighting to match the sample composition of the app and diary, both the high-smart and low-smart app underestimated expenditures as compared to the diary benchmark.

Chatzitheochari et al. (2018) report on the usage of hard and soft checks as implemented in a travel diary survey offered on the web, respectively requiring or suggesting certain actions from the user to reduce incomplete data and found that this increased increase full completion rates by 30-50% over the paper diaries with no such features. A corresponding decrease in the number of activities was also noted, however, which may indicate that the overall completion rate increases at the cost of additional information that proved difficult for the user to encode. A similar relationship between the number of recorded events in paper versus app-based diaries is reported in a media-specific time use diary (Lev-On & Lowenstein-Barkai, 2019). In this study, Lev-On & Lowenstein-Barkai (2019) found a significant and large difference between the number of recorded viewings, with respondents to paper-based diaries tending to report approximately the same number of items as available lines on the diary.

This may be due to the fact that user experience and perception of the survey instrument can differ significantly between app-based and paper diaries. Respondents report feeling less connected to their behaviors when entering it as checkboxes with an app, versus the required inclusion of a greater level of detail via traditional diary methods (Frąckowiak et al., 2022). In fact, users can trust the smart elements too much, disregarding their own intuition. Users have been found to be more likely to use the defaults provided in an app than to generate their own response (Bucher Della Torre et al., 2017), which would be a requirement with a traditional survey.

When comparisons are made between traditional diary studies and app-based mobility studies, they often demonstrate large differences in in rates, distances, and lengths of trips (Greaves et al., 2015; Gillis et al., 2023). The straightforward interpretation of this difference is that sensor data reduce the overall measurement error relative to non-smart methods, but Bradley et al. (Bradley et al., 2018) posit the interpretation that this reflects a difference in “soft refusals” or non-response bias between the two methods.

On the other hand, differences may also arise not only between sensors and self-reported data, but between multiple sensors. A study combining two different sensors measuring alcohol levels in comparison with a daily retrospective survey on the previous day's alcohol usage encountered large amounts of discrepancy not only between the sensors and self-report measures, but between the sensors themselves (Courtney et al., 2023). This has also been shown in physical activity studies where sensors are compared against each other (Parmenter et al., 2022).

Food diary research has been at the forefront of implementing image capture data streams into diary studies, which may provide insight into expectations for budget and time use research. For instance, photo-based food diaries demonstrated a small underreporting mean bias compared to the moderate underreporting observed in the same participants' paper diaries (Costello et al., 2017). Other studies have suggested that smartphone-based measurement of food intake was only as accurate as paper-based food records (Hutchesson et al., 2015), and that both methods still suffer from underreporting when compared against known truth (Boushey et al., 2017).

Overall, the existing research indicates that researchers should expect to find meaningful differences between modes with or without smart features.

6. Estimation Methodology

Estimating total mode effects between different survey modes boils down to a straightforward principle: any existing difference between sample means or variances between two different modes indicates the presence of some sort of mode effect. This means that any study categorizing its analyses by mode implicitly offers an approach for estimating this effect. Unfortunately, the total mode effects aggregate multiple sources of error and therefore don't offer clear insights for correcting for differential responses between modes (Vannieuwenhuyze & Loosveldt, 2013). Pinpointing differences between specific sources of error is a vital step in the integration process, despite the fact that some may represent beneficial processes, and others not. For example, consider a data collection design in which smart survey nonrespondents are followed up with paper surveys. It may be desirable that smart and non-smart surveys attract different types of respondents, which would result in coverage or nonresponse differences between the smart- and non-smart survey. Measurement differences in such a situation are, however, undesirable when the goal is to integrate the data from the smart and non-smart survey.

In her review of the current literature on data integration, Salvatore (2023) found that propensity scores, missing data, and regression estimators were commonly mentioned by researchers, indicating an emerging sensitivity to the need to consider these issues. On the other hand, because finely parceling out total error into its component pieces requires access to some known standard against which to compare, and most survey research is conducted precisely to solve for some unknown quantity, assessment of mode effects generally requires the use of a specific research design. This is usually only possible in the contexts of experimental research within probability samples whose properties are more

predictable, or when some external source of data may be used to provide validation (Klausch et al., 2013).

Experimental research on the estimation of mode effects comprises a relatively small proportion of all literature on the topic, but ostensibly offers the strongest properties for establishing the extent of mode effects. Tourangeau (2017) suggests three main strategies for disentangling the two sources of mode differences: 1) direct assessment of measurement error by comparing reports from different sources to a gold standard, 2) rendering the mode groups comparable statistically with weighting or regression and 3) estimating the errors using modeling techniques (often CFA or LCM).

Probability samples are often used as a basis in these experimental designs in order to reduce or remove the impact of frame error (van den Brakel, 2013). Non-response bias can be estimated based on correlations between demographic characteristics and survey response, leaving the remainder to be considered as measurement error (van den Brakel & Renssen, 2005; van den Brakel, 2008). Although there is limited literature in which differences are estimated in this way between smart and traditional diary surveys, Premkumar et al (2023) employed this method to estimate measurement difference between an app-based diary and recall survey, finding differences of up to 26% between the two measurements.

A second type of experimental methodology similarly involves splitting the sample followed by random assignment to a mode but goes further by then following up by assessing the same participants with repeated measurements within a single mode (Schouten et al., 2013). This has the benefit of not relying on correlations between survey responses and known demographic profiles, but increases the costs significantly, and introduces a small possibility of differential memory effects between modes (Klausch et al., 2015). Here, too, a consideration must be made for non-response between waves, which is solved for by Klausch et al. (2015) by imputation of unit non-response after the follow-up. This is similar to a method which considers longitudinal random allocation in a panel in which the modes are switched over time, and the differences estimated using a latent measurement model (Cernat, 2015; Cernat et al., 2016). This same experimental design can be estimated in different ways, with some research indicating that the performance is dependent upon the amount of error and benchmark choice (Klausch et al., 2017). A more complex variation on the longitudinal method involves a crossover design, allowing for the estimation of differences between modes while also accounting for changes due to memory effect or attitudinal changes between measurement moments (Antoun et al. 2019).

Finally, coverage error can be estimated external to the survey under consideration question by obtaining measurements of smartphone ownership within an established probability-based survey, which can be further broken out across relevant facets such as smartphone OS (Keusch et al., 2023).

There are also options for estimating measurement effects outside experimental design. One potential method makes use of observed variables on both survey modes that are insensitive to measurement differences, such as demographic characteristics. (Vannieuwenhuyze et al., 2010, Vannieuwenhuyze & Loosveldt, 2013). A particular variant

of this method, called propensity score matching, first estimates the selection effect given a set of observed covariates, then matches individuals from both modes. The remaining differences between modes are assumed to be the measurement effect (Morgan & Harding, 2006; Stuart, 2010; Lugtig et al., 2011; Capacci et al., 2018; Rosenbaum, 2021). This is often accomplished with regression methods but can be extended with greater levels of complexity (Jäckle et al., 2010). Adding self-reported mode preference to the propensity score models has been proposed as one such extension to address some issues with the method (Vandenplas et al., 2016).

An extension of this method has precedence in multi-source research as well, such as estimating the amount to which survey and register variables differ on key measured variables using overlapping variables within each source. Here options for the estimation include latent class modeling (Guarnera & Varriale, 2016; Oberski, 2017), hidden Markov models (Pavlopoulos & Vermunt, 2015), or Multiple Imputation Latent Class Modelling (MILC) (Boeschoten & Oberski, 2017).

Where there are overlaps between respondents and variables, it is possible to use Structural Equation Modeling to build models of the relationships in the data under both modes and use fit indices to establish which mode should serve as a benchmark on a per-question basis (Bakker, 2012; Scholtus et al., 2015).

7. DATA INTEGRATION

Integration under small differences

When mode specification/measurement and coverage/nonresponse effects are estimated to be negligible, the data may be integrated as-is. Supporting literature for this comes from studies comparing differences in smartphone-completed and PC-completed web surveys which can be treated as a single data source when measurement and coverage error are estimated to be sufficiently low, as is often the case when the instrument has been optimized for smartphone usage, but made available on the web (de Leeuw, 2018; de Leeuw & Hox, 2018).

Here, the primary concern may be related to small differences in recording that arise between smart and traditional surveys. For example, an app-based Time Use Diary may allow for smaller time-window increments than a paper-based Time Use Diary, requiring that the data be aggregated to the same time scale (Chatzitheochari & Mylona, 2021). These will be application-specific but may be categorized as general harmonization procedures. When these differences become too large, or, in other words, when a meaningful mode measurement effect arises, such harmonization will still be a necessary component, but will be insufficient on its own to correct for the bias. In this case, integration under medium or large differences will be more appropriate.

Integration under moderate coverage/non-response only differences

Previous research has suggested that coverage differences may pose the largest source of error in smartphone-based surveys (Antoun et al., 2019). In situations where the survey

instruments are very similar between a smart survey and its traditional counterpart, such as might be expected in the comparison of an app-based diary format with limited smart features, the methodology currently used for adjusting for differences in selection within mixed-mode surveys can be directly employed.

Here the most common methods employed involve adjustment by postsurvey weighting, making use of available demographic variables (Bethlehem, 1988; Dzikiti, 2019). Researchers should be cautious here, both because the full necessary set of demographic variables may not be available to account for coverage differences (Antoun et al., 2019), and because non-response effects may be in the Missing Not at Random context (Andridge & Little, 2011)

Small area estimation can be used to further improve these estimates and reduce the variance (Rao & Molina, 2015). This is not limited to geographic areas, although it can be used to good effect in this context, but can be deployed in any situation in which there are distinct and related small categories (Boonstra et al., 2008).

Integration under moderate differences involving mode measurement effect

The literature reviewed in previous sections suggests that many smart surveys, especially those with smart features making use of machine intelligence or data linkage, but lacking aspects of passive measurement, are likely to result in moderate measurement differences when compared to a traditional survey. This requires a greater level of consideration when integrating the datasets

A natural extension of the weighting procedure used to adjust for selection effects involves reweighting through some mechanism to calibrate unit response propensity in addition to the measurement effect, but as distinct elements. Most methods accomplish this through the selection of a benchmark mode (Buelens & van den Brakel, 2015; Vannieuwenhuyze et al., 2014). When reinterviewing is possible, this offers a mechanism of disentangling measurement error that can be used in the integration process (Buelens & Van den Brakel, 2017; Klausch et al., 2017). Different models are available here, but frequently either Structural Equation Models or IRT-approaches are used (Mariano & Elliott, 2017).

A similar approach involves handling mixed-mode measurements as treatment effects, and then handling them within the causal modelling framework. In this way counterfactual potential outcomes (“what if this participant had completed the other version of the survey?”) can be estimated with regression for each mode, and the overall estimates combined to produce a final estimate (Suzer-Gurtekin et al., 2012; Park et al., 2017; Suzer-Gurtekin & Valliant, 2018).

Integration under large differences

Some smart surveys may lend themselves to larger differences than others, such as when passive measurements are used not to augment a direct response from a participant, but to replace it, as is the aim of some highly-smart surveys on travel behavior, in which measurements of distance or the number of stops would optimally be algorithmically calculated (Lawson et al., 2023). Matters of extreme time scale differences may pose similar

problems. Here, research that arises from the field of multi-source methodology may provide more comprehensive solutions.

Similar to the counterfactual methods described above, responses under the other survey method can be seen as missing. In this way, multiple imputation can be used to generate a response for the other method in order to combine two data sources. (Kolenikov & Kennedy, 2014; Park et al., 2016).

When there is a clear preference for one mode to be used as a benchmark, it may be preferable to integrate the modes by using inequality restrictions that make use of the features of one source to impose constraints on the estimation (Boonstra et al., 2011). This can also be accomplished with bootstrapping rather than regression methods, which can be beneficial when the data linkage involves high levels of complexity (Chipperfield, 2020).

Special cases of integration

Qualitative methods can be used that make use of human analysis to combine disparate sources of data in a way that can make use of the benefits of both without requiring this to be algorithmically deterministic. Resch et al. (2020) demonstrate this method for combining eDiary and sensor measurements using what they term a "visual analytics approach" in which they use one mode to provide context to another. Similarly, this method has been used to integrate camera and accelerometer measurements of time usage with paper diary methods by displaying the captured images to the respondent to support a retrospective "what did you do yesterday" face-to-face interview (Harms et al., 2019).

8. Conclusion and Recommendations

The integration of smart survey methodologies with traditional survey techniques presents a rapidly evolving frontier in survey research. The versatility and pervasiveness of smartphones offer researchers an expansive toolset, providing more dynamic and real-time data collection. However, this advancement is not without challenges.

Previous research has demonstrated the utility of smart surveys as a methodology for addressing the limitations of traditional survey methodology by substantially reducing respondent burden and enriching the data. The most significant obstacle to this natural progression arises from the concerns of how the new data streams from the smart surveys can be integrated with the old. Even subtle differences between modes of data collection can potentially lead to differential measurement, and the differences in data format that may arise from passive data collection or sensor measurements may not be easily reconcilable. To this end, next steps in this area must involve the intentional estimation of mode effects to establish where the vulnerabilities are greatest. Depending on the degree and location of differences, various integration techniques may be appropriate, ranging from direct integration to complex methodologies.

Researchers developing smart surveys must proactively consider sources of error when introducing smart features, while being aware that neither mode may serve as a suitable benchmark for the other. Developing standardized guidelines for data integration applicable

to all smart surveys may not be appropriate at this point, as considerations are likely to vary not only across surveys themselves, but across the selection of smart features that are deployed. Here it will be helpful to maintain a collaborative approach that seeks to involve researchers who are familiar with the individual surveys with the developers who will bridge the gap between traditional and smart surveys, as nuances on both ends are likely to dictate the requirements for successful data integration.

List of abbreviations

- AL ACTIVE LEARNING
- CBS Centraal Bureau voor de Statistiek – Statistics Netherlands
- CRŒSSCROss-domain data collection platform for the ESS
- COICOP Classification of Individual Consumption According to Purpose
- CoP Code of Practice
- Destatis Statistisches Bundesamt Deutschland
- ESS European Statistical System
- HBS Household Budget Survey
- HCI Human Computer Interaction
- Insee Institut national de la statistique et des études économiques
- ML Machine Learning
- MOTUS Modular Online Time Use Survey
- NSI National Statistical Institute
- OACL Online Activity Classification List
- OCR Optical Character Recognition
- SSB Statistik sentralbyrå – Statistics Norway
- SSI Smart Survey Implementation
- SOURCE™ Software Outreach and Redefinition to Collect E-data Through MOTUS
- TA Think Aloud
- TOR Tempus Omnia Revelat
- TUS Time Use Survey
- UI User Interface
- UX User Experience
- VUB Vrije Universiteit Brussel

REFERENCES

REFERENCES CHAPTER 1

Assemi, Behrang, Hamed Jafarzadeh, Mahmoud Mesbah, and Mark Hickman (2018) "Participants' Perceptions of Smartphone Travel Surveys." *Transportation Research Part F: Traffic Psychology and Behaviour* 54 (April): 338–48. https://doi.org/10.1016/j.trf.2018.02.005.

Becker, Stefan, Christopher Brandl, Sven Meister, Eckhard Nagel, Talya Miron-Shatz, Anna Mitchell, Andreas Kribben, Urs-Vito Albrecht, and Alexander Mertens.(2015) "Demographic and Health Related Data of Users of a Mobile Application to Support Drug Adherence Is Associated with Usage Duration and Intensity." Edited by Pal Bela Szecsi. *PLOS ONE* 10 (1): e0116980. https://doi.org/10.1371/journal.pone.0116980.

Bemmann, Florian, Maximiliane Windl, Jonas Erbe, Sven Mayer, and Heinrich Hussmann (2022) "The Influence of Transparency and Control on the Willingness of Data Sharing in Adaptive Mobile Apps." *Proceedings of the ACM on Human-Computer Interaction* 6 (MHCI): 1–26. https://doi.org/10.1145/3546724.

Bruno, Mauro, Massimo De Cubellis, Fabrizio De Fausti, Claudia De Vitiis, Francesca Inglese, Giuseppina Ruocco, Monica Scannapieco, et al. (2022) "ESSnet Smart Surveys. Workpackage 3 Development of a Conceptual Framework, Reference Architecture and Technical Specifications for the European Platform for Trusted Smart Surveys Deliverable 3.4 Final Report." Deliverable 3.4 Final Report.

Buskirk, Trent D. and Charles Andres (2012) "Smart Surveys for Smart Phones: Exploring Various Approaches for Conducting Online Mobile Surveys via Smartphones." Survey Practice 5 (1). https://doi.org/10.29115/SP-2012-0001.

Christin, Delphine, Christian Buchner, and Niklas Leibecke (2013) "What's the Value of Your Privacy? Exploring Factors That Influence Privacy-Sensitive Contributions to Participatory Sensing Applications." In *38th Annual IEEE Conference on Local Computer Networks - Workshops*, 918–23. Sydney, Australia: IEEE. https://doi.org/10.1109/LCNW.2013.6758532.

Couper, Mick P. (2017) "New Developments in Survey Data Collection." *Annual Review of Sociology* 43 (1): 121–45. https://doi.org/10.1146/annurev-soc-060116-053613.

Couper, Mick P., Christopher Antoun, and Aigul Mavletova (2017) "Mobile Web Surveys: A Total Survey Error Perspective." In *Total Survey Error in Practice*, edited by Paul P. Biemer, Edith Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, and Brady T. West, 1st ed., 133–54. Wiley. https://doi.org/10.1002/9781119041702.ch7.

Elevelt, Anne, Peter Lugtig, and Vera Toepoel (2019) "Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process?" *Survey Research Methods*, April, 195–213 Pages. https://doi.org/10.18148/SRM/2019.V13I2.7385.

Farke, Florian M., David G. Balash, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. (2021) "Are Privacy Dashboards Good for End Users? Evaluating User Perceptions and Reactions to Google's My Activity (Extended Version)." arXiv. https://arxiv.org/abs/2105.14066.

Groves, R. M., S. Presser, R. Tourangeau, B. T. West, M. P. Couper, E. Singer, and C. Toppe (2012) "Support for the Survey Sponsor and Nonresponse Bias." *Public Opinion Quarterly* 76 (3): 512–24. https://doi.org/10.1093/poq/nfs034.

Guo, Yuanyuan (2022) “Does User Preference Matter? A Comparative Study on Influencing Factors of User Activity Between Government-Provided and Business-Provided Apps.” *Frontiers in Psychology* 13 (June): 914528. https://doi.org/10.3389/fpsyg.2022.914528.

Haas, Georg-Christoph, Frauke Kreuter, Florian Keusch, Mark Trappmann, and Sebastian Bährn (2020) “Effects of Incentives in Smartphone Data Collection.” In *Big Data Meets Survey Science*, edited by Craig A. Hill, Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg, 1st ed., 387–414. Wiley. https://doi.org/10.1002/9781118976357.ch13.

Hargittai, Eszter, Elissa M. Redmiles, Jessica Vitak, and Michael Zimmer (2020) “Americans’ Willingness to Adopt a COVID-19 Tracking App.” *First Monday*, October. https://doi.org/10.5210/fm.v25i11.11095.

Jäckle, Annette, Jonathan Burton, Mick P. Couper, and Carli Lessof (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* Vol 13 (April): 23–44 Pages. https://doi.org/10.18148/SRM/2019.V1I1.7297.

Keusch, Florian, Sebastian Bähr, Georg-Christoph Haas, Frauke Kreuter, and Mark Trappmann (2023) “Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey.” *Sociological Methods & Research* 52 (2): 841–78. https://doi.org/10.1177/0049124120914924.

Keusch, Florian, Sebastian Bähr, Georg-Christoph Haas, Frauke Kreuter, Mark Trappmann, and Stephanie Eckman (2022) “Non-Participation in Smartphone Data Collection Using Research Apps.” *Journal of the Royal Statistical Society Series A: Statistics in Society* 185 (Supplement_2): S225–45. https://doi.org/10.1111/rssa.12827.

Keusch, Florian, Mariel M. Leonard, Christoph Sajons, and Susan Steiner (2021) “Using Smartphone Technology for Research on Refugees: Evidence from Germany.” *Sociological Methods & Research* 50 (4): 1863–94. https://doi.org/10.1177/0049124119852377.

Keusch, Florian, Bella Struminskaya, Christopher Antoun, Mick P Couper, and Frauke Kreuter (2019a) “Willingness to Participate in Passive Mobile Data Collection.” *Public Opinion Quarterly* 83 (S1): 210–35. https://doi.org/10.1093/poq/nfz007.

———(2019b) “Willingness to Participate in Passive Mobile Data Collection.” *Public Opinion Quarterly* 83 (S1): 210–35. https://doi.org/10.1093/poq/nfz007.

Keusch, Florian, Bella Struminskaya, Frauke Kreuter, and Martin Weichbold (2020) “Combining Active and Passive Mobile Data Collection: A Survey of Concerns.” In *Big Data Meets Survey Science*, edited by Craig A. Hill, Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg, 1st ed., 657–82. Wiley. https://doi.org/10.1002/9781118976357.ch22.

Keusch, Florian, Alexander Wenz, and Frederick Conrad (2022) “Do You Have Your Smartphone with You? Behavioral Barriers for Measuring Everyday Activities with Smartphone Sensors.” *Computers in Human Behavior* 127 (February): 107054. https://doi.org/10.1016/j.chb.2021.107054.

Klingwort, Jonas, Jeldrik Bakker, and Vera Toepoel (2023) “Survey Design Features With Potential for High Response Rates: A Meta-analytical Approach.” Milano.

Kreuter, Frauke, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann (2020) “Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges Around Privacy and Informed Consent.” *Social Science Computer Review* 38 (5): 533–49. https://doi.org/10.1177/0894439318816389.

Lynch, Joann, Jeffrey Dumont, Elizabeth Greene, and Jonathan Ehrlich (2019) “Use of a Smartphone GPS Application for Recurrent Travel Behavior Data Collection.” *Transportation Research*

Record: Journal of the Transportation Research Board 2673 (7): 89–98. https://doi.org/10.1177/0361198119848708.

Máté, Ákos, Zsófia Rakovics, Szilvia Rudas, Levente Wallis, Bence Ságvári, Ákos Huszár, and Júlia Koltai (2023) "Willingness of Participation in an Application-Based Digital Data Collection Among Different Social Groups and Smartphone User Clusters." *Sensors* 23 (9): 4571. https://doi.org/10.3390/s23094571.

McCool, Danielle, Peter Lugtig, Ole Mussmann, and Barry Schouten (2021) "An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges." *Journal of Official Statistics* 37 (1): 149–70. https://doi.org/10.2478/jos-2021-0007.

Oyibo, Kiemute, and Plinio Pelegrini Morita (2022) "Factors Influencing the Willingness to Download COVID-19 Contact Tracing Apps: The Moderating Effect of Persuasive Design and Smartphone Usage Experience." *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* 11 (1): 163–69. https://doi.org/10.1177/2327857922111033.

Revilla, Melanie, Mick P. Couper, and Carlos Ochoa (2019) "Willingness of Online Panelists to Perform Additional Tasks." *Methods* data (July): 29 Pages. https://doi.org/10.12758/MDA.2018.01.

Rodenburg, Evelien, Struminskaya, Bella, and Barry Schouten (2022) Nonresponse and Dropout in an App-Based Household Budget Survey: Representativeness, Interventions to Increase Response, and Data Quality. Presentation at the 3rd MASS workshop, Utrecht.

Ságvári, Bence, Attila Gulyás, and Júlia Koltai (2021) "Attitudes Towards Participation in a Passive Data Collection Experiment." *Sensors* 21 (18): 6085. https://doi.org/10.3390/s21186085.

Schaewitz, Leonie, Stephan Winter, and Nicole C. Krämer (2021) "The Influence of Privacy Control Options on the Evaluation and User Acceptance of Mobile Applications for Volunteers in Crisis Situations." *Behaviour & Information Technology* 40 (8): 759–75. https://doi.org/10.1080/0144929X.2020.1723703.

Scherpenzeel, Annette (2017) "Mixing Online Panel Data Collection with Innovative Methods." In *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung*, edited by Stefanie Eifler and Frank Faulbaum, 27–49. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15834-7_2.

Schnorf, Sebastian, Martin Ortlieb, and Nikhil Sharma (2014) "Trust, Transparency & Control in Inferred User Interest Models." In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, 2449–54. Toronto Ontario Canada: ACM. https://doi.org/10.1145/2559206.2581141.

Struminskaya, Bella, Peter Lugtig, Florian Keusch, and Jan Karem Höhne (2020) "Augmenting Surveys With Data From Sensors and Apps: Opportunities and Challenges." *Social Science Computer Review*, December, 089443932097995. https://doi.org/10.1177/0894439320979951.

Struminskaya, Bella, Peter Lugtig, Vera Toepoel, Barry Schouten, Deirdre Giesen, and Ralph Dolmans (2021) "Sharing Data Collected with Smartphone Sensors." *Public Opinion Quarterly* 85 (S1): 423–62. https://doi.org/10.1093/poq/nfab025.

Tsai, Janice Y., Serge Egelman, Lorrie Cranor, and Alessandro Acquisti (2011) "The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study." *Information Systems Research* 22 (2): 254–68. https://doi.org/10.1287/isre.1090.0260.

Van Kleek, Max, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J. Weitzner, and Nigel Shadbolt (2017) "Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5208–20. Denver Colorado USA: ACM. https://doi.org/10.1145/3025453.3025556.

Wenz, Alexander, Annette Jäckle, and Mick P. Couper (2019a) "Willingness to Use Mobile Technologies for Data Collection in a Probability Household Panel." *Survey Research Methods* Vol 13 (April): 1–22 Pages. https://doi.org/10.18148/SRM/2019.V1I1.7298.

———(2019b) “Willingness to Use Mobile Technologies for Data Collection in a Probability Household Panel.” *Survey Research Methods* Vol 13 (April): 1–22 Pages. https://doi.org/10.18148/SRM/2019.V1I1.7298.

Wenz, Alexander, and Florian Keusch (2023) “Increasing the Acceptance of Smartphone-Based Data Collection.” *Public Opinion Quarterly*, June, nfad019. https://doi.org/10.1093/poq/nfad019.

Xu, Runhua, Remo Manuel Frey, Elgar Fleisch, and Alexander Ilic (2016) “Understanding the Impact of Personality Traits on Mobile App Adoption Insights from a Large-Scale Field Study.” *Computers in Human Behavior* 62 (September): 244–56. https://doi.org/10.1016/j.chb.2016.04.011.

REFERENCES CHAPTER 2

Bähr, S., Haas, Georg-C., Keusch, F., Kreuter, F., & Trappmann M., (2020) Missing Data and Other Measurement Quality Issues in Mobile Geolocation Sensor Data. Social Science Computer Review. DOI: 10.1177/0894439320944118

Benedikt, L., Joshi, C., Nolan, L., de Wolf, Nick. & Schouten, B., (2020) Optical Character Recognition and Machine Learning Classification of Shopping Receipts - @HBS>An app-assisted approach for the Household Budget Survey

Cerasti E., De Cubellis M., De Fausti f., De Vitiis C., Guandalini A., Inglese F. Bruno M, Ruocco G, Aracri R.M., Meise N., van Etten J., Cotton F. (2020) ESSnet Smart Surveys - Deliverable 3.2: Proof-of-Concept (WP3)

De Cubellis M., De Fausti f., De Vitiis C., Inglese F. Bruno M, Ruocco G, Meise N., van Etten J., Cotton F., Scannapieco M, Meise N., Maślankowski J., Jug M., Joeri van Etten, Rob Warmerdam (2020) ESSnet Smart Surveys - Deliverable 3.4: Final report (WP3)

Hoek, S. van, Windmeijer, D., Luiten, A., Bolte, J., & Schouten, B. (2022). Comparing activity trackers to investigate physical activity. CBS working paper, accessible at https://www.cbs.nl/engb/background/2022/08/comparing-activity-trackers-to-investigate-physical-activity

Lugtig, P., Roth, K., & Schouten, B., (2022) Nonresponse analysis in a longitudinal smartphone-based travel study. Survey Research Methods (2022) Vol. 16, No. 1, pp. 13-27 doi:10.18148/srm/2022.v16i1.7835 European Survey Research Association

Luiten A., Schouten B., Lusyne P. (2020) ESSnet Smart Surveys - Deliverable 2.12: End report

de Groot J., Oerlemans T. , Rodenburg E., Schouten B. Lope Mariscal A. C., Martin Bernia E., Poch J. , Horcajo Garcia T., Balsa Criado V., Gauche C., Osier G., ESSnet Smart Surveys - Deliverable 2.1: Consumption (WP2.1)

Luiten A., Toepoel V., Schouten B., Cierpiał-Wolan M., Kapica K., Szlachta P., Lusyne P., van der Beken H., (2020) ESSnet Smart Surveys - Deliverable 2.1: Health-measuring physical activity (WP2.3)

Luiten A., Schouten B.,, Blanke K., Knapp D., Volk J., Lusyne P., (2020) ESSnet Smart Surveys - Deliverable 2.1: Living conditions - Measuring indoor environment (WP2.4)

Lusyne P., Delloye J. (2020) ESSnet Smart Surveys - Deliverable 2.1: Time Use (WP2.2)

McCool, D., Lugtig, P., Mussmann, O., & Schouten, B., (2021) An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. Journal of Official Statistics, Vol. 37, No. 1, 2021, pp. 1–23, http://dx.doi.org/10.2478/JOS-2021-xxxx

McCool, D., Schouten, B., & Lugtig, P. (2023) Dynamic Time Warping-based imputation of long gaps in human mobility trajectories.

McCool, D., Lugtig, P., & Schouten, B. (2022) Maximum interpolable gap length in missing smartphone-based GPS mobility data. *Transportation*, 1-31.

McCool, D., Lugtig, P., & Schouten, B. (2018) Preliminary analyses of the CBS verplaatsingen app field test data (AVA and EVA). Internal report
Oerlemans, T. & Schouten, B., (2022) - Deliverable 1.3 Governance plan - Project 2020-NL-INNOV (@HBS2)
Oerlemans, T., de Wolf, N. & Schouten, B., (2022) Deliverable 2.1 HBS app - Project 2020-NL-INNOV (@HBS2)
Schouten, B. (2022) - Deliverable 1.2 Report on the action - Project 2020-NL-INNOV (@HBS2)
Smeets, L., Lugtig, P. & Schouten, B., (2019) Automatic Travel Mode Prediction in a National Travel Survey. Discussion Paper
van Hoek, S., de Wolf, N., van den Heuvel, G., Bos, J. & Schouten, B. (2022) - Deliverable 2.3 – Receipt processing - Project 2020-NL-INNOV (@HBS2)
Vrabič Kek, B. Oerlemans, T. & Schouten, B. (2022) - Deliverable 2.2 - Product/shop guidelines - Project 2020-NL-INNOV (@HBS2)

REFERENCES CHAPTER 3

Bastien, J. C. (2010). Usability testing: a review of some methodological and technical aspects of the method. *International journal of medical informatics, 79*(4), e18-e23.
Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE transactions on professional communication, 43*(3), 261-278.
Deffner, G. (1990). *Verbal Protocols as a Research Tool in Human Factors: Symposium Abstract.* Paper presented at the Proceedings of the Human Factors Society Annual Meeting.
Drummond, K., & Hopper, R. (1993). Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on language and Social Interaction, 26*(2), 157-177.
Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*: Intellect books.
Ericsson, K. A. (2017). Protocol analysis. *A companion to cognitive science*, 425-432.
Eurostat. (2018). *European statistics code of practice: for the national statistical authorities and Eurostat (EU statistical authority)*. Luxembourg: Publications Office of the European Union.
Fenton, G., Glorieux, A., Letesson, Q., & Minnen, J. (2020). *Centre-ville, piétonnisation et modes de vie*. Retrieved from Brussels:
Garcia-Lopez, E., Garcia-Cabot, A., Manresa-Yee, C., De-Marcos, L., & Pages-Arevalo, C. (2017). Validation of navigation guidelines for improving usability in the mobile web. *Computer Standards & Interfaces, 52*, 51-62.
Hagymásy, T. n., József, G., Keresztes, T. s., Vámos, V. r., & Vida, B. z. (2022). *Testing of the HCSO MOTUS application, time use survey methodological report*. Retrieved from Budapest:
Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science, 1*, 1-16.
Hussain, A., & Kutar, M. (2009). Usability metric framework for mobile phone application. *PGNet, ISBN, 2099*, 978-971.
Knapp, D., Richter, C., Sommer, A., & Brecht, P. (2022). Deliverable D4.2: Report on the usability tests for the “MOTUS-HBS”-App.
Knapp, D., Rödel, E., Sommer, A., & Volk, J. (2021). *Pretestbericht zu den Anwendungen der deutschen Zeitverwendungserhebung 2022: App- und Web-Anwendung*. Retrieved from Wiessbaden:
Krahmer, E., & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE transactions on professional communication, 47*(2), 105-117.

Kronbauer, A. H., Santos, C. A., & Vieira, V. (2012). *Smartphone applications usability evaluation: a hybrid model and its implementation.* Paper presented at the Human-Centered Software Engineering: 4th International Conference, HCSE 2012, Toulouse, France, October 29-31, 2012. Proceedings 4.

Lew, P., & Olsina, L. (2013). *Relating user experience with MobileApp quality evaluation and design.* Paper presented at the Current Trends in Web Engineering: ICWE 2013 International Workshops ComposableWeb, QWE, MDWE, DMSSW, EMotions, CSE, SSN, and PhD Symposium, Aalborg, Denmark, July 8-12, 2013. Revised Selected Papers 13.

Minnen, J., Nagel, E., & Sabbe, K. (2020). *SOURCE TM: Software Outreach and Redefinition to Collect E-data Through MOTUS. Towards a Modular Online Time Use Survey.* Brussels & Bonn: Statistics Belgium, Destatis, hbits CV & Vrije Universiteit Brussel.

Minnen, J., Olsen, J., & Sabbe, K. (2022). *CROESS: Establishing a Cross-domain data collection platform for the ESS (European Statistical System)*. Brussels & Bonn: Satistics Belgium, Destatis, hbits VC & Vrije Universiteit Brussel.

Minnen, J., Rymenants, S., Glorieux, I., & van Tienoven, T. P. (2023). Answering current challenges of and changes in producing official time-use statistics using the data collection platform MOTUS. *Journal of Official Statistics*, Online First.

Oliveira, K. W., & Paula, M. M. (2020). Gamification of online surveys: A systematic mapping. *IEEE Transactions on Games*, *13*(3), 300-309.

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). *Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Schouten, B., Bulman, J., Järvensivu, M., Plate, M., & Vrabic-Kek, B. (2020). *Report on the action @HBS*. Retrieved from https://ec.europa.eu/eurostat/documents/54431/11489222/1+Report+on+the+action.pdf

Smeets, L., Lugtig, P., & Schouten, B. (2019). *Automatic Travel Mode Prediction in a National Travel Survey*. Retrieved from The Hague:

Stehrenberg, S., & Giannakouris, K. (2021). Trusted Smart Surveys: Solutions for the European Statistical system - An overview of the objectives and the main challenges. *Proceedings 63rd ISI World Statistics Congress, 11*, 455-460.

Volk, J., Knapp, D., & Sommer, A. (2020). Testing the @HBS-App. Test report on usability.

Volk, J., Knapp, D., Sommer, A., & Zins, S. (2021). *Testing the MOTUS-App. Test report on usability*. Retrieved from

Weichbroth, P. (2018). *Usability attributes revisited: a time-framed knowledge map.* Paper presented at the 2018 Federated Conference on Computer Science and Information Systems (FedCSIS).

Weichbroth, P. (2020). Usability of mobile applications: a systematic literature study. *Ieee Access, 8*, 55563-55577.

Zanker, M., Rook, L., & Jannach, D. (2019). Measuring the impact of online personalisation: Past, present and future. *International Journal of Human-Computer Studies*, *131*, 160-168.

Zhang, D., & Adipat, B. (2005). Challenges, methodologies, and issues in the usability testing of mobile applications. *International journal of human-computer interaction, 18*(3), 293-308.

REFERENCES CHAPTER 4

Abraham, K. G., Maitland, A., & Bianchi, S. M. (2006). Nonresponse in the American time use survey. Public Opinion Quarterly, 70(5), 676–703. Adler, T., Rimmer, L., & Carpenter, D. (2002). Use of Internet-Based Household Travel Diary Survey Instrument. *Transportation Research Record*, *1804*(1), 134–143.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: adapting the TSE framework to big data. *Journal of Survey Statistics And*. https://academic.oup.com/jssam/article-abstract/8/1/89/5728725

Ampt, E. S., Richardson, A. J., & Brög, W. (1985). *New Survey Methods in Transport: Proceedings of 2nd International Conference, Hungerford Hill, Australia, 12-16 September 1983*. VSP.

Andridge, R. R., & Little, R. J. A. (2011). *Proxy pattern-mixture analysis for survey nonresponse*. scb.se. http://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/proxy-pattern-mixture-analysis-for-survey-nonresponse.pdf

Antoun, C., Conrad, F. G., & Couper, M. P. (2019). Simultaneous estimation of multiple sources of error in a smartphone-based survey. *Journal of Surveying Engineering*. https://academic.oup.com/jssam/article-abstract/7/1/93/4924615

Arentze, T., Dijst, M., Dugundji, E., Joh, C.-H., Kapoen, L., Krygsman, S., Maat, K., & Timmermans, H. (2001). New Activity Diary Format: Design and Limited Empirical Evidence. *Transportation Research Record*, *1768*(1), 79–88.

Aschauer, F., Hössinger, R., Jara-Diaz, S., Schmid, B., Axhausen, K., & Gerike, R. (2021). Comprehensive data validation of a combined weekly time use and travel survey. *Transportation Research Part A: Policy and Practice*, *153*, 66–82.

Ashley, D., Richardson, T., & Young, D. (2009). Recent information on the under-reporting of trips in household travel surveys. *Australasian Transport Research Forum (ATRF), 32nd, 2009, Auckland, New Zealand*, *32*. https://trid.trb.org/view/1149551

Axhausen, K. W. (1995). *Travel diaries: An annotated catalog* (2nd ed). Institut für Strassenbau und Verkehrsplanung. https://rosap.ntl.bts.gov/view/dot/13806

Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., & Trappmann, M. (2020). Missing Data and Other Measurement Quality Issues in Mobile Geolocation Sensor Data. *Social Science Computer Review*, 0894439320944118.

Bauman, A., Bittman, M., & Gershuny, J. (2019). A short history of time use research; implications for public health. BMC Public Health, 19(Suppl 2), 607.

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*(1), 8–17.

Bandilla, W., Couper, M. P., & Kaczmirek, L. (2014). The effectiveness of mailed invitations for web surveys and the representativeness of mixed-mode versus internet-only samples. *Survey Practice*, *7*(4), 1–9.

Berger, M., & Platzer, M. (2015). Field Evaluation of the Smartphone-based Travel Behaviour Data Collection App "SmartMo." *Transportation Research Procedia*, *11*, 263–279.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Biemer, P. P. (2001). *Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing*. http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/nonresponse-bias-and-measurement-bias-in-a-comparison-of-face-to-face-and-telephone-interviewing.pdf

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, *74*(5), 817–848.

Biemer, P. P., & Amaya, A. (2020). Total error frameworks for found data. In *Big Data Meets Survey Science* (pp. 131–161). Wiley. https://doi.org/10.1002/9781118976357.ch4

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. John Wiley & Sons.

Boeschoten, L., & Oberski, D. (2017). Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). *Journal of Official Statistics*. https://dspace.library.uu.nl/handle/1874/363295

Boeschoten, L., Oberski, D., & de Waal, A. G. (2016). *Latent Class Multiple Imputation for multiply observed variables in a combined dataset*. ine.es. http://www.ine.es/q2016/docs/q2016Final00047.pdf

Bonnel, P., Bayart, C., & Smith, B. (2015). Workshop Synthesis: Comparing and Combining Survey Modes. *Transportation Research Procedia*, *11*, 108–117.

Boonstra, H. J. H., Blois, C. J. de, & Linders, G.-J. M. (2011). Macro-integration with inequality constraints: an application to the integration of transport and trade statistics: Macro-integration with inequality constraints. *Statistica Neerlandica*, *65*(4), 407–431.

Boonstra, H. J. H., van den Brakel, J., Buelens, B., Krieg, S., & Smeets, M. (2008). Towards small area estimation at Statistics Netherlands. *Metron. International Journal of Statistics*, *66*(1), 21–49.

Bosch, O. J., & Revilla, M. (2022). When survey science met web tracking: Presenting an error framework for metered data. *Journal of the Royal Statistical Society. Series A, , 185*(Suppl 2), S408–S436.

Boushey, C. J., Spoden, M., Delp, E. J., Zhu, F., Bosch, M., Ahmad, Z., Shvetsov, Y. B., DeLany, J. P., & Kerr, D. A. (2017). Reported energy intake accuracy compared to doubly labeled water and usability of the mobile Food Record among community dwelling adults. *Nutrients*, *9*(3), 312.

Bradley, M., Greene, E., Spitz, G., Coogan, M., & McGuckin, N. (2018). The millennial question: Changes in travel behaviour or changes in survey behaviour? *Transportation Research Procedia*, *32*, 291–300.

Bricka, S., Zmud, J., Wolf, J., & Freedman, J. (2009). Household Travel Surveys with GPS: An Experiment. *Transportation Research Record*, *2105*(1), 51–56.

Brzozowski, M., Crossley, T. F., & Winter, J. K. (2017). A comparison of recall and diary food expenditure data. *Food Policy*, *72*, 53–61.

Bucher Della Torre, S., Carrard, I., Farina, E., Danuser, B., & Kruseman, M. (2017). Development and evaluation of e-CA, an electronic mobile-based food record. *Nutrients*, *9*(1), 76.

Buelens, B., & van den Brakel, J. A. (2015). Measurement error calibration in mixed-mode sample surveys. *Sociological Methods & Research*, *44*(3), 391–426.

Buelens, B., & Van den Brakel, J. A. (2017). Comparing Two Inferential Approaches to Handling Measurement Error in Mixed-Mode Surveys. *Journal of Official Statistics*, *33*(2), 513–531.

Burkill, S., Copas, A., Couper, M. P., Clifton, S., Prah, P., Datta, J., Conrad, F., Wellings, K., Johnson, A. M., & Erens, B. (2016). Using the web to collect data on sensitive behaviours: A study looking at mode effects on the British National Survey of Sexual Attitudes and Lifestyles. *PloS One*, *11*(2), e0147983.

Burton, J., & Jäckle, A. (2020). Mode effects. *ISER: Understanding Society Working Paper Series*, *2020–05*. https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2020-05.pdf

Capacci, S., Mazzocchi, M., & Brasini, S. (2018). Estimation of unobservable selection effects in on-line surveys through propensity score matching: An application to public acceptance of healthy eating policies. *PloS One*, *13*(4), e0196020.

Cernat, A. (2015). The Impact of Mixing Modes on Reliability in Longitudinal Studies. *Sociological Methods & Research*, *44*(3), 427–457.

Cernat, A., Couper, M. P., & Ofstedal, M. B. (2016). Estimation of Mode Effects in the Health and Retirement Study Using Measurement Models. *Journal of Survey Statistics and Methodology*, *4*(4), 501–524.

Chambers, R., Chipperfield, J. O., Davis, W., & Kovacevic, M. (2009). *Inference Based on Estimating Equations and Probability-Linked Data*. https://ro.uow.edu.au/cssmwp/38/

Chatzitheochari, S., Fisher, K., Gilbert, E., Calderwood, L., Huskinson, T., Cleary, A., & Gershuny, J. (2018). Using new technologies for time diary data collection: Instrument design and data quality findings from a mixed-mode pilot survey. *Social Indicators Research*, *137*(1), 379–390.

Chatzitheochari, S., & Mylona, E. (2021). Data quality in web and app diaries : a person-level comparison. *Electronic International Journal of Time Use Research*, *16*(1), 19–34.

Chen, B. (2012). A balanced system of U.s. industry accounts and distribution of the aggregate statistical discrepancy by industry. *Journal of Business & Economic Statistics: A Publication of the American Statistical Association*, *30*(2), 202–211.

Chipperfield, J. O. (2020). Bootstrap inference using estimating equations and data that are linked with complex probabilistic algorithms. *Statistica Neerlandica*, *74*(2), 96–111.

Chipperfield, J. O., & Chambers, R. L. (2015). Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data. *Journal of Official Statistics*, *31*(3), 397–414.

Clarke, M., Dix, M., & Jones, P. (1981). Error and uncertainty in travel surveys. *Transportation*, *10*(2), 105–126.

Conrad, F. G., Couper, M. P., Tourangeau, R., & Galesic, M. (2005). *Interactive Feedback Can Improve the Quality of Responses in Web Surveys*. researchgate.net. https://www.researchgate.net/profile/Roger-Tourangeau-2/publication/228689449_Interactive_feedback_can_improve_quality_of_responses_in_web_surveys/links/53eb6c290cf23b8116a9bda4/Interactive-feedback-can-improve-quality-of-responses-in-web-surveys.pdf

Conrad, F. G., Couper, M. P., Tourangeau, R., & Peytchev, A. (2006). Use and Non-use of Clarification Features in Web Surveys. *Journal of Official Statistics*, *22*(2). https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/use-and-non-use-of-clarification-features-in-web-surveys.pdf

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future Mobility Survey: Experience in Developing a Smartphone-Based Travel Survey in Singapore. *Transportation Research Record*, *2354*(1), 59–67.

Couper, M. P. (2007). Issues of representation in eHealth research (with a focus on web surveys). *American Journal of Preventive Medicine*, *32*(5 Suppl), S83-9.

Couper, M. P., Antoun, C., & Mavletova, A. (2017). Mobile web surveys. In *Total Survey Error in Practice* (pp. 133–154). John Wiley & Sons, Inc.

Couper, M. P., Gremel, G., Axinn, W., Guyer, H., Wagner, J., & West, B. T. (2018). New options for national population surveys: The implications of internet and smartphone coverage. *Social Science Research*, *73*, 221–235.

Couper, M. P., Peytchev, A., Strecher, V. J., Rothert, K., & Anderson, J. (2007). Following up nonrespondents to an online weight management intervention: randomized trial comparing mail versus telephone. *Journal of Medical Internet Research*, *9*(2), e16.

Courtney, J. B., Russell, M. A., & Conroy, D. E. (2023). Acceptability and validity of using the BACtrack skyn wrist-worn transdermal alcohol concentration sensor to capture alcohol use across 28 days under naturalistic conditions - A pilot study. *Alcohol (Fayetteville, N.Y.)*, *108*, 30–43.

Crossley, T. F., & Winter, J. K. (2014). Asking households about expenditures: what have we learned? In *Improving the measurement of consumer expenditures* (pp. 23–50). University of Chicago Press.

Daum, T., Buchwald, H., Gerlicher, A., & Birner, R. (2018). Times Have Changed: Using a Pictorial Smartphone App to Collect Time–Use Data in Rural Zambia. Field Methods. https://doi.org/10.1177/1525822X18797303

De Broe, S., Struijs, P., Daas, P., van Delden, A., Burger, J., van den Brakel, J., ten Bosch, O., Zeelenberg, K., & Ypma, W. (2021). Updating the paradigm of official statistics: New quality criteria for integrating new data and methods in official statistics. *Statistical Journal of the IAOS*, *37*(1), 343–360.

de Bruijne, M., & Wijnant, A. (2014). Mobile response in web panels. *Social Science Computer Review*, *32*(6), 728–742.

De Cubellis, M., De Fausti, F., De Vitiis, C., Guandalini, A., Inglese, F., Meise, N., Rocci, F., & Varriale, R. (2019). Task 3.1. 1 Smart Survey Methodology. ESSnet Smart Surveys, 11.

de Leeuw, E. D. (2018). Mixed-mode: Past, present, and future. *Survey Research Methods*. https://ojs.ub.uni-konstanz.de/srm/article/view/7402

de Leeuw, E. D., & Hox, J. (2008). Mixing Data Collection Methods: Lessons from Social Survey Research1, 2. *Advances in Mixed Methods Research: Theories and Applications*, 138.

de Leeuw, E. D., & Hox, J. J. (2018). Internet surveys as part of a mixed-mode design. In *Social and Behavioral Research and the Internet* (pp. 45–76). Routledge.

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2015a). Mixed-mode Surveys. In *International Handbook of Survey Methodology*. Routledge.

de Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2015b). The cornerstones of survey research. In *International Handbook of Survey Methodology*. Routledge.

Dillman, D. A. (2015, September 18). *On Climbing Stairs Many Steps at a Time: The New Normal in Survey Methodology*. Faculty and Graduate Student Seminar Series, The School of Economics Washington State University. https://web.archive.org/web/20181221174204if_/http://ses.wsu.edu:80/wp-content/uploads/2015/09/DILLMAN-talk-Sept-18-2015.pdf

Dillman, D. A., & Edwards, M. L. (2016). Designing a mixed-mode survey. In *The SAGE Handbook of Survey Methodology* (pp. 255–268). SAGE Publications Ltd.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons.

Dzikiti, L. N. (2019). *Comparing approaches for combining data collected from multiple complex surveys, adjusting for clustering and stratification* [University of Pretoria]. http://hdl.handle.net/2263/73137

E-communications and the digital single market: Report. (2021). European Comission. https://doi.org/10.2759/572426

Eckman, S. (2022). Underreporting of purchases in the US consumer expenditure survey. *Journal of Survey Statistics and Methodology*, *10*(5), 1148–1171.

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

Elliott, M. R., Little, R. J. A., & Lewitzky, S. (2000). Subsampling callbacks to improve survey efficiency. *Journal of the American Statistical Association*, *95*(451), 730.

EUROSTAT. (2003). *Household budget surveys in the EU : methodology and recommendations for harmonisation - 2003* (Vol. 1–1 online resource (1 PDF-bestand.)). Office for Official Publications of the European Communities.

Felderer, B., & Herzing, J. M. E. (2023). What about the less IT literate? A comparison of different postal recruitment strategies to an online panel of the general population. *Field Methods*, *35*(3), 219–235.

Filipponi, D., & Guarnera, U. (2017). ST2_1 Overlapping numerical variables without a benchmark: Integration of adminis-trative sources and survey data through Hidden Markov Models for the production of labour statistics. *Cros-Legacy.Ec.Europa.Eu*. https://cros-legacy.ec.europa.eu/system/files/st2_1.pdf

Frąckowiak, M., Rogowski, Ł., & Sommer, V. (2022). Hopes and challenges of creating and using a smartphone application. Working on and working with a digital mobile tool in qualitative sociospatial research. *Qualitative Research: QR*, 146879412210989.

Frazis, H., & Stewart, J. (2007). Where does the time go? Concepts and measurement in the American Time Use Survey. In Hard-to-measure goods and services: Essays in honor of Zvi Griliches (pp. 73–97). University of Chicago Press.

Frazis, H., & Stewart, J. (2012). How to think about time-use data: What inferences can we make about long- and Short-Run time use from time diaries? Annals of Economics and Statistics, 105/106, 231.

Gershuny, J., Harms, T., Doherty, A., Thomas, E., Milton, K., Kelly, P., & Foster, C. (2020). Testing self-report time-use diaries against objective instruments in real time. *Sociological Methodology*, *50*(1), 318–349.

Giesen, D., S. Theunissen, M. Nyholt, B. Vrabič, M. Zgonec, M. Järvensinu & A. Niemelä (2019). Testing Materials Test August-September 2019. ONS, Stat Austria, Stat Finland, Stat Netherlands, Stat Slovenia, the
University of Essex.

Gillis, D., Lopez, A. J., & Gautama, S. (2023). An evaluation of smartphone tracking for travel behavior studies. *ISPRS International Journal of Geo-Information*, *12*(8), 335.

Glorieux, I., & Minnen, J. (2009). How many days? A comparison of the quality of time-use data from 2-day and 7-day diaries. *Electronic International Journal of Time Use Research*, *6*(2), 314–327.

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. *Procedia - Social and Behavioral Sciences*, *138*, 557–565.

Götz, F. M., Stieger, S., & Reips, U.-D. (2017). Users of the main smartphone operating systems (iOS, Android) differ only little in personality. *PloS One*, *12*(5), e0176921.

Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., & Crane, M. (2015). A Web-Based Diary and Companion Smartphone app for Travel/Activity Surveys. *Transportation Research Procedia*, *11*, 297–310.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly*, *70*(5), 720–736.

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879.

Guarnera, U., & Varriale, R. (2016). Estimation from contaminated multi-source data based on latent class models. *Statistical Journal of the IAOS*, *32*(4), 537–544.

Harding, C., Faghih Imani, A., Srikukenthiran, S., & Miller, E. J. (2021). Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys. *Transportation*. https://link.springer.com/article/10.1007/s11116-020-10135-7

Harms, T., Gershuny, J., Doherty, A., Thomas, E., Milton, K., & Foster, C. (2019). A validation study of the Eurostat harmonised European time use study (HETUS) diary using wearable technology. BMC Public Health, 19(Suppl 2), 455.

Hoogendoorn-Lanser, S., Schaap, N. T. W., & OldeKalter, M.-J. (2015). The Netherlands Mobility Panel: An Innovative Design Approach for Web-based Longitudinal Travel Data Collection. *Transportation Research Procedia*, *11*, 311–329.

Hox, J., de Leeuw, E. D., & Klausch, T. (2017). Mixed-mode research. In *Total Survey Error in Practice* (pp. 511–530). John Wiley & Sons, Inc.

Husebø, A. M. L., Morken, I. M., Eriksen, K. S., & Nordfonn, O. K. (2018). The patient experience with treatment and self-management (PETS) questionnaire: translation and cultural adaption of the Norwegian version. *BMC Medical Research Methodology*, *18*(1), 147.

Hutchesson, M. J., Rollo, M. E., Callister, R., & Collins, C. E. (2015). Self-monitoring of dietary intake by young women: online food records completed on computer or smartphone are as accurate as paper-based food records but more acceptable. *Journal of the Academy of Nutrition and Dietetics*, *115*(1), 87–94.

Hwang, S., Yalla, S., & Crews, R. (2017). Processing uncertain GPS trajectory data for assessing the locations of physical activity. In Big Data for Regional Science (1st Edition, pp. 131–142). Routledge.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, *86*(1), 4–29.

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), 22.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *Revue Internationale de Statistique [International Statistical Review]*, *78*(1), 3–20.

Johnston, K. A. D. R. (2014). Researching the respondents. *Market & Social Research*, *22*(1), 39–46.

Jones-Jang, S. M., Heo, Y.-J., McKeever, R., Kim, J.-H., Moscowitz, L., & Moscowitz, D. (2020). Good news! Communication findings may be underestimated: Comparing effect sizes with self-reported and logged smartphone use data. *Journal of Computer-Mediated Communication: JCMC*, *25*(5), 346–363.

Kang, J. D. Y., & Schafer, J. L. (2007). Rejoinder: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *22*(4), 574–580.

Kaplan, R. L., Kopp, B., & Phipps, P. (2020). Contrasting stylized questions of sleep with diary measures from the American time use survey. In *Advances in Questionnaire Design, Development, Evaluation and Testing* (pp. 671–695). Wiley. https://doi.org/10.1002/9781119263685.ch27

Kasprzyk, D. (2005). *Measurement Error in Household Surveys: Sources and Measurement*. https://econpapers.repec.org/paper/mprmprres/d7a25d262708428ba7a6236903ef5b0a.htm

Kelly, P., Thomas, E., Doherty, A., Harms, T., Burke, Ó., Gershuny, J., & Foster, C. (2015). Developing a method to test the validity of 24 hour time use diaries using wearable cameras: A feasibility pilot. PloS One, 10(12), e0142198.

Keusch, F., Bach, R., & Cernat, A. (2022). Reactivity in measuring sensitive online behavior. *Internet Research*, *33*(3), 1031–1052.

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2023). Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey. *Sociological Methods & Research*, *52*(2), 841–878.

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., Trappmann, M., & Eckman, S. (2022). Non-participation in smartphone data collection using Research apps. *Journal of the Royal Statistical Society. Series A, , 185*(Supplement_2), S225–S245.

Keusch, F., Wenz, A., & Conrad, F. (2022). Do you have your smartphone with you? Behavioral barriers for measuring everyday activities with smartphone sensors. *Computers in Human Behavior*, *127*. https://doi.org/10.1016/j.chb.2021.107054

Khan, W. Z., Xiang, Y., Aalsalem, M. Y., & Arshad, Q. (First 2013). Mobile Phone Sensing Systems: A Survey. *IEEE Communications Surveys & Tutorials*, *15*(1), 402–427.

Klausch, T. (2014). *Informed design of mixed-mode surveys: Evaluating mode effects on measurement and selection error*. Utrecht University.

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions. *Sociological Methods & Research*, *42*(3), 227–263.

Klausch, T., Hox, J., & Schouten, B. (2015). Selection error in single- and mixed mode surveys of the dutch general population. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *178*(4), 945–961.

Klausch, T., Schouten, B., Buelens, B., & Van Den Brakel, J. (2017). Adjusting measurement bias in sequential mixed-mode surveys using re-interview data. *Journal of Survey Statistics and Methodology*, *5*(4), 409–432.

Klingwort, J., & Schnell, R. (2020). Critical Limitations of Digital Epidemiology: Why COVID-19 Apps Are Useless. Survey Research Methods, 14(2), 95–101.

Knottnerus, P. (2016). On new variance approximations for linear models with inequality constraints. *Statistica Neerlandica*, *70*(1), 26–46.

Kolenikov, S., & Kennedy, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology*, *2*(2), 126–158.

Kolpashnikova, K., Flood, S., Sullivan, O., Sayer, L., Hertog, E., Zhou, M., Kan, M.-Y., Suh, J., & Gershuny, J. (2021). Exploring daily time-use patterns: ATUS-X data extractor and online diary visualization tool. PloS One, 16(6), e0252843.

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges Around Privacy and Informed Consent. *Social Science Computer Review*, *38*(5), 533–549.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*. https://academic.oup.com/poq/article-abstract/72/5/847/1833162

Lawson, C. T., Krans, E., Rentz, E. (green), & Lynch, J. (2023). Emerging trends in household travel survey programs. *Social Sciences & Humanities Open*, *7*(1), 100466.

Lev-On, A., & Lowenstein-Barkai, H. (2019). Viewing diaries in an age of new media: An exploratory analysis of mobile phone app diaries versus paper diaries. Methodological Innovations, 12(1), 205979911984444.

Link, M. W., Murphy, J., Schober, M. F., Buskirk, T. D., Hunter Childs, J., & Langer Tesfaye, C. (2014). Mobile Technologies for Conducting, Augmenting and Potentially Replacing SurveysExecutive Summary of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. *Public Opinion Quarterly*, *78*(4), 779–787.

Lugtig, P., & Jäckle, A. (2014). Can I just check...? Effects of edit check questions on measurement error and survey estimates. *Journal of Official Statistics*, *30*(1), 45–62.

Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(5), 669–686.

Luiten, A., Hox, J., & de Leeuw, E. D. (2020). Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics*, *36*(3), 469–487.

Lynn, P., & Lugtig, P. J. (2017). Total survey error for longitudinal surveys. In *Total Survey Error in Practice* (pp. 279–298). John Wiley & Sons, Inc.

Magraw-Mickelson, Z., Wang, H. H., & Gollwitzer, M. (2022). Survey mode and data quality: Careless responding across three modes in cross-cultural contexts. *International Journal of Testing*, *22*(2), 121–153.

Mariano, L. T., & Elliott, M. N. (2017). An item response theory approach to estimating survey mode effects: Analysis of data from a randomized mode experiment. *Journal of Survey Statistics and Methodology*, *5*(2), 233–253.

Mavletova, A., & Couper, M. P. (2015). A meta-analysis of breakoff rates in mobile web surveys. *Mobile Research Methods: Opportunities and Challenges of Mobile Research Methodologies*, 81–98.

Mayer, A. (2019). "your survey is biased": A preliminary investigation into respondent perceptions of survey bias. *Survey Practice*, *12*(1), 1–8.

McCool, D., Schouten, J. G., & Lugtig, P. (2021). An app-assisted travel survey in official statistics. Possibilities and challenges. *Journal of Official Statistics*, *37*(1), 149–170.

Millward, H., & Spinney, J. (2011). Time use, travel behavior, and the rural–urban continuum: results from the Halifax STAR project. Journal of Transport Geography, 19(1), 51–58.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects. *Sociological Methods & Research*, *35*(1), 3–60.

Mosesso, L., Maudet, N., Nano, E., Thibault, T., & Tabard, A. (2023, June 5). Obsolescence Paths: living with aging devices. *ICT4S 2023 - International Conference on Information and Communications Technology for Sustainability*. https://hal.science/hal-04097867/

Mushkudiani, N., Daalmans, J., & Bikker, R. (2018). Solving large-data consistency problems at Statistics Netherlands using macro-integration techniques. *Statistica Neerlandica*, *72*(4), 553–573.

Mushkudiani, N., Daalmans, J., & Pannekoek, J. (2014). Macro-Integration for Solving Large Data Reconciliation Problems. *AJS; American Journal of Sociology*, *43*(1), 29–48.

Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, *130*, 114–122.

Ng, J. C. N., & Sarjeant, P. M. (1993). Use of direct data entry for travel surveys. In *Transportation Research Record*. trid.trb.org.

Nguyen, M. H., Armoogum, J., Madre, J.-L., & Garcia, C. (2020). Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(4), 395–412.

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., & Maurer, P. (2014). Supporting large-scale travel surveys with smartphones – A practical approach. *Transportation Research Part C: Emerging Technologies*, *43*, 212–221.

Oberski, D. L. (2017). Estimating error rates in an administrative register and survey questions using a latent class model. *Total Survey Error in Practice*. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119041702.ch16

Park, Seho, Kim, J. K., & Stukel, D. (2017). A measurement error model approach to survey data integration: combining information from two surveys. *Metron*, *75*(3), 345–357.

Park, Seunghwan, Kim, J. K., & Park, S. (2016). An imputation approach for handling mixed-mode surveys. *The Annals of Applied Statistics*, *10*(2), 1063–1085.

Parmenter, B., Burley, C., Stewart, C., Whife, J., Champion, K., Osman, B., Newton, N., Green, O., Wescott, A. B., Gardner, L. A., Visontay, R., Birrell, L., Bryant, Z., Chapman, C., Lubans, D. R., Sunderland, M., Slade, T., & Thornton, L. (2022). Measurement properties of smartphone approaches to assess physical activity in healthy young people: Systematic review. *JMIR MHealth and UHealth*, *10*(10), e39085.

Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology*, *41*(1), 197–214.

Pearce, K. E., & Rice, R. E. (2013). Digital divides from access to activities: Comparing mobile and personal computer internet users. *The Journal of Communication*, *63*(4), 721–744.

Peterson, G., Griffin, J., LaFrance, J., & Li, J. (2017). Smartphone participation in web surveys. In *Total Survey Error in Practice* (pp. 203–233). John Wiley & Sons, Inc.

Peytchev, A., & Crawford, S. (2005). A typology of real-time validations in web-based surveys. *Social Science Computer Review*, *23*(2), 235–249.

Prelipcean, A. C., Gidófalvi, G., & Susilo, Y. O. (2018). MEILI: A travel diary collection, annotation and automation system. *Computers, Environment and Urban Systems*, *70*, 24–34.

Premkumar, P. S., Ganesan, S. K., Pandiyan, B., Krishnamoorthy, D. K., & Kang, G. (2023). Smartphone diary application in household surveys: Integration of high frequency temporal data in large-scale data collection. *Field Methods*, 1525822X231195525.

Rao, J. N. K., & Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Resch, B., Puetz, I., Bluemke, M., Kyriakou, K., & Miksch, J. (2020). An Interdisciplinary Mixed-Methods Approach to Analyzing Urban Spaces: The Case of Urban Walkability and Bikeability. *International Journal of Environmental Research and Public Health*, *17*(19). https://doi.org/10.3390/ijerph17196994

Rinderknecht, R. G., Max Planck Institute for Demographic Research, Rostock, Germany, Doan, L., Sayer, L. C., Department of Sociology, University of Maryland, United Stated of America, & Department of Sociology, University of Maryland, United Stated of America. (2022). MyTimeUse: An online implementation of the day-reconstruction method. Journal of Time Use Research, 1, 23–50.

Roberts, C., Herzing, J. M. E., Manjon, M. A., Abbet, P., & Gatica-Perez, D. (2022). Response burden and dropout in a probability-based online panel study – A comparison between an app and browser-based design. *Journal of Official Statistics*, *38*(4), 987–1017.

Roberts, C., Herzing, J. M. E., Sobrino Piazza, J., Abbet, P., & Gatica-Perez, D. (2022). Data privacy concerns as a source of resistance to complete mobile data collection tasks via a smartphone app. *Journal of Survey Statistics and Methodology*, *10*(3), 518–548.

Rosenbaum, P. R. (2021). *Design of observational studies* (2nd ed.). Springer Nature.

Sadeghian, P., Håkansson, J., & Zhao, X. (2021). Review and evaluation of methods in transport mode detection based on GPS tracking data. *Journal of Traffic and Transportation Engineering (English Edition)*, *8*(4), 467–482.

Sakshaug, J. W., Yan, T., & Tourangeau, R. (2010). Nonresponse error, measurement error, and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Public Opinion Quarterly*, *74*(5), 907–933.

Salant, P., & Dillman, D. A. (2008). *How to conduct your own survey*. John Wiley & Sons.

Salvatore, C. (2023). Inference with non-probability samples and survey data integration: a science mapping study. *Metron*, *81*(1), 83–107.

Sarasua, W., & Meyer, M. (1996). New technologies for household travel surveys. *Conference Proceedings 10: Conference on Household Travel Surveys: New Concepts and Research Needs*, 170–182.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313.

Schaub, F., Seifert, J., Honold, F., Müller, M., Rukzio, E., & Weber, M. (2014, April 26). Broken display = broken interface': the impact of display damage on smartphone interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '14: CHI Conference on Human Factors in Computing Systems, Toronto Ontario Canada. https://doi.org/10.1145/2556288.2557067

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions. *Sociological Methods & Research*, *25*(3), 341–383.

Scholtus, S., Bakker, B. F. M., & van Delden, A. (2015). Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. *Cbs.Nl*. https://www.cbs.nl/-/media/imported/documents/2015/46/modelling_measurement_error.pdf

Schouten, B., van den Brakel, J., Buelens, B., Giesen, D., Luiten, A., & Meertens, V. (2021a). Multi-Device Surveys. In *Mixed-Mode Official Surveys* (pp. 223–249). Chapman and Hall/CRC.

Schouten, B., van den Brakel, J., Buelens, B., Giesen, D., Luiten, A., & Meertens, V. (2021b). *Mixed-Mode Official Surveys: Design and Analysis*. CRC Press.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, *42*(6), 1555–1570.

Schwarz, N. (2012). *Why researchers should think "real-time": A cognitive rationale*. dornsife.usc.edu. https://dornsife.usc.edu/assets/sites/780/docs/schwarz_why_real-time_dec_2010_pri.pdf

Sekula, W., Nelson, M., Figurska, K., Oltarzewski, M., Weisell, R., & Szponar, L. (2005). Comparison between household budget survey and 24-hour recall data in a nationally representative sample of Polish households. *Public Health Nutrition*, *8*(4), 430–439.

Silberstein, A. R., & Scott, S. (2011). Expenditure diary surveys and their associated errors. In *Measurement Errors in Surveys* (pp. 303–326). John Wiley & Sons, Inc.

Singer, E., & Couper, M. P. (2017). Some Methodological Uses of Responses to Open Questions and Other Verbatim Comments in Quantitative Surveys. *Methods, Data, Analyses*, *11*(2), 20.

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The end of the (research) world as we know it? Understanding and coping with declining response rates to mail surveys. *Society & Natural Resources*, *32*(10), 1139–1154.

Stone, A. A., Schneider, S., Smyth, J. M., Junghaenel, D. U., Couper, M. P., Wen, C., Mendez, M., Velasco, S., & Goldstein, S. (2023). A population-based investigation of participation rate and self-selection bias in momentary data capture and survey studies. *Current Psychology* . https://doi.org/10.1007/s12144-023-04426-2

Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J. K. (2020). Augmenting Surveys With Data From Sensors and Apps: Opportunities and Challenges. *Social Science Computer Review*, 0894439320979951.

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*(Suppl 1), 423–462.

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.

Sullivan, O., Gershuny, J., Sevilla, A., Walthery, P., & Vega-Rapun, M. (2020). Time use diary design for our times - an overview, presenting a Click-and-Drag Diary Instrument (CaDDI) for online application. Journal of Time Use Research, 1–17.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21.

Suzer-Gurtekin, Z. T., & Valliant, R. (2018). *Mixed-Mode Surveys: Design, Estimation and Adjustment Methods*. https://books.google.nl/books?hl=nl&lr=&id=rWhvDwAAQBAJ&oi=fnd&pg=PA409&ots=Tk_ON-uATi&sig=FctKwYIkLcuA7lhMOe1c5pP4qg4

Szalai, A. (1972). *The use of time: daily activities of urban and suburban populations in twelve countries*. https://trid.trb.org/view/1150665

Tait, A. R., Reynolds, P. I., & Gutstein, H. B. (1995). Factors that influence an anesthesiologist's decision to cancel elective surgery for the child with an upper respiratory tract infection. *Journal of Clinical Anesthesia*, *7*(6), 491–499.

Toepoel, V., & Lugtig, P. (2015). Online surveys are mixed-device surveys. Issues associated with the use of different (mobile) devices in web surveys. *Methods, Data, Analyses*. https://doi.org/10.12758/MDA.2015.009

Tourangeau, R. (2017). Mixing modes. In *Total Survey Error in Practice* (pp. 115–132). John Wiley & Sons, Inc.

Tuba Suzer-Gurtekin, Z., Heeringa, S. G., & Valliant, R. (2012). *Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys*. asasrms.org. http://www.asasrms.org/Proceedings/y2012/Files/306174_74031.pdf

van Delden, A., Scholtus, S., & Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics*, *32*(3), 619–642.

van den Brakel, J. A. (2008). Design-Based Analysis of Embedded Experiments with Applications in the Dutch Labour Force Survey. *Journal of the Royal Statistical Society. Series A, , 171*(3), 581–613.

van den Brakel, J. A., & Renssen, R. H. (2005). Analysis of Experiments Embedded in Complex Sampling Designs. *Surv. Methodol.*, *31*(1), 23.

van den Brakel, Jan A. (2013). Design-based analysis of factorial designs embedded in probability samples. *Survey Methodology*, *39*, 323+.

Minnen, J., Glorieux, I., van Tienoven, T.P., Daniels, S., Weenas, D., Deyaert, J., Van den Bogaert, S., & Rymenants, S. (2014). Modular Online Time Use Survey (MOTUS)-Translating an existing method in the 21 st century. *Electronic International Journal of Time Use Research*, *11*(1). https://jtur.iatur.org/home/article/eba78532-c38f-4eff-b54e-fe6271ba4ccf

Vandenplas, C., Loosveldt, G., & Vannieuwenhuyze, J. T. A. (2016). Assessing the Use of Mode Preference as a Covariate for the Estimation of Measurement Effects between Modes. A Sequential Mixed Mode Experiment. *Methods, Data, Analyses*, *10*(2), 24.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045.

Vannieuwenhuyze, J. T. A., & Loosveldt, G. (2013). Evaluating Relative Mode Effects in Mixed-Mode Surveys:: Three Methods to Disentangle Selection and Measurement Effects. *Sociological Methods & Research*, *42*(1), 82–104.

Vannieuwenhuyze, J. T. A., Loosveldt, G., & Molenberghs, G. (2014). Evaluating mode effects in mixed-mode survey data using covariate adjustment models. *Journal of Official Statistics*, *30*(1), 1–21.

Verzosa, N., Greaves, S., Ho, C., & Davis, M. (2021). Stated willingness to participate in travel surveys: a cross-country and cross-methods comparison. *Transportation*, *48*(3), 1311–1327.

Waal, T., Delden, A., & Scholtus, S. (2020). Multi-source statistics: Basic situations and methods. *International Statistical Review = Revue Internationale de Statistique*, *88*(1), 203–228.

Wenz, A. (2021). Completing web surveys on mobile devices does screen size affect data quality? In *Sozialwissenschaftliche Datenerhebung im digitalen Zeitalter* (pp. 101–121). Springer Fachmedien Wiesbaden.

Wenz, A. (2023). *Quality of expenditure data collected with a mobile receipt scanning app in a probability household panel*. understandingsociety.ac.uk. https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2023-02.pdf

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

Wenz, A., & Keusch, F. (2023). Increasing the Acceptance of Smartphone-Based Data Collection. *Public Opinion Quarterly*, *87*(2), 357–388.

Whatnall, M. C., Kolokotroni, K. Z., Fozard, T. E., Evans, T. S., Marwood, J. R., Ells, L. J., & Burrows, T. L. (2023). How is online self-reported weight compared with image-captured weight? A comparative study using data from an online longitudinal study of young adults. *The American Journal of Clinical Nutrition*, *118*(2), 452–458.

Wolf, J., Guensler, R., & Bachman, W. (2001). Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record*, *1768*(1), 125–134.

Yalamanchili, L., Pendyala, R. M., Prabaharan, N., & Chakravarthy, P. (1999). Analysis of Global Positioning System-Based Data Collection Methods for Capturing Multistop Trip-Chaining Behavior. *Transportation Research Record*, *1660*(1), 58–65.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63.