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

# Introduction

Over the past two decade, 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 them (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., First 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 (Luiten et al., 2020; Stedman et al., 2019). The causes behind the falling response rates aren’t clear, 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 [“wasn’t relevant”], burden [“no time”], and interest [“too boring”] (Couper et al., 2007; Singer & Couper, 2017; Tait et al., 1995). Unsurprisingly, these same aspects are also well-represented in surveys of what respondents find bothersome in surveys (Husebø et al., 2018; Johnston, 2014; 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 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, despite both researcher interest and impending necessity.

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 equality. 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 and data integration. Finally, in Section 7, we synthesize the findings from the literature, provide recommendations, and suggest experimental methods for closing the gaps in existing literature.

# 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; Peterson et al., 2017; Toepoel & Lugtig, 2015).

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 and 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, while conceptually easy to distinguish, is sometimes fuzzy. In this way, it may be useful to describe different levels of ‘smartness’ that a survey may have. A survey that a user accesses by typing in a website to the browser in their phone, then fills in questions in the same way as they would if they were answering in any other mode, would not be considered a smart survey. On the other hand, the addition of search bar for an input field for, say, the food you last ate could be considered a smart feature as it makes use of a smartphone’s capacity to serve up search results to reduce a user’s total effort in typing out a full phrase. This is a low level of smartness, but if it is compared to an entry blank on a pen and paper survey, it is easy to identify that it is present.

A high level of smartness for a similar question might involve scanning the barcode or taking a picture of the food you ate in order to answer a similar question. 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 a number of 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.

<Consequence: more smart = more measurement differences>

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). At the Second International Conference for New Survey Methods in Transport, differences in reporting were identified between days incorporating interviewer assistance and not, and between recorded behavior and road sensors (Ampt et al., 1985). Ashley et. al. (2009) discuss trip underreporting across multiple studies that rely on varied modes for comparison against the self-administered travel diary. Although the particulars differ, they share the commonality that the data from self-administered travel diaries consistently shows fewer trips than other sources.

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 (Bricka et al., 2009; Sarasua & Meyer, 1996; Yalamanchili et al., 1999). This worked quite well at capturing user locations, but because it was an external sensor, linkage with data from the respondent was not simple, and methods varied across studies. Although some proponents posed 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 (Bähr et al., 2020; Gong et al., 2014; Nguyen et al., 2020; Sadeghian et al., 2021).

At the same point in time, the increase in Internet penetration and web familiarity led other researchers to experiment with bringing travel diaries online (Adler et al., 2002; Arentze et al., 2001). 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). Much like with the GPS loggers, the data collected from the enhanced web-based travel diaries was often quite different.

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 (Berger & Platzer, 2015; Cottrill et al., 2013; Greaves et al., 2015; Nitsche et al., 2014). 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 of purchases made, or retrospective surveys on past average behaviors, 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 not very good at giving estimates outside of the highest categorization level, such as “food” or “clothing” when asked in this way, which placed limits on what these surveys could be used for, 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.

Unfortunately, expenditure diaries are imperfect for their task, as respondents often face difficulties in recalling and documenting irregular and small expenditures. 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. In this way, it can be taxing not only in the total amount of time required to complete the diary, but in the overall cognitive effort required for a researcher to determine what level of specificity is sufficient, or to convert between units to report total weight or volume of certain purchases.

Fortunately, respondents are generally quite good at being aware of when and how they have spent money – in other words, 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 smart features, such as by replacing a paper diary with an app-based diary to 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.

Data generated under these new conditions, however, are at risk for 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

Short history of time use research and how it has evolved. Discuss app from TSS1. Issues that make it a good fit for smart surveys

# 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. This article therefore relates 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 1 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.

For the purposes of this article, briefly describing each of these sources of nonsampling error is sufficient to provide context for the following sections. 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. 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 traditional diaries, you do not see item non-response>

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.

The following chapter uses the TSE framework as a basis for discussion of the differences that are known to arise within each of these components between different modes of survey administration. Chapter 5 presents what is currently known specifically on the comparison between smart surveys and their non-smart counterparts.

# 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 discussed in Chapter 3: 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 in fact 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, within the field of mixed-mode design, mode effects not contributing to an increase total survey coverage of these are frequently seen as nuisance elements to be avoided or corrected for against some gold standard measurement (Burton & Jäckle, 2020; Klausch et al., 2013). This view is at least partially 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. 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 paper 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 (de Leeuw et al., 2015b; Salant & Dillman, 2008). 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 don’t 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 & Edwards, 2016; Dillman et al., 2014). 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 lend itself to seeing differences at the level of specification rather than measurement.

## Mode effects due to frame coverage

The original concern of smartphone coverage was that smartphone coverage may have been limited by key demographics, but it may now be the case that access to the web on non-smartphone devices may be limited by key demographics.

Online surveys are done on many different devices. Mixed-device surveys. Online surveys not yet adapted to small screens. Variability in display of questions. No long matrix questions. Measurement differences can be minimal. (Toepoel & Lugtig, 2015)

<Stats from recent Eurobarometer on ownership and mention secondary divide using them is important, not just ownership -- Keusch study>

An increasing percentage of households now have Internet access only through their mobile devices (Peterson et al., 2017).

Antoun et al (Antoun et al., 2019) found coverage to be the largest source of difference between PC and Smartphone mode effects, with significant differences on eight of nineteen different measures across technology, lifestyle and political dimensions.

Differences in coverage between web-only and web + mail response options have been demonstrated in the past, even when penetration is high (Bandilla et al., 2014; de Leeuw, 2018). (and it’s reasonable to assume that this effect will be similar for smartphones)

Omissions, erroneous inclusions, duplications (?)

People are usually still contacted through the same routes, making the frames the same (is this always true?), but when the mode of administration is via an app, suddenly some portion of the sample frame is ineligible.

* Maybe change to selection error and combine?
* I think here it’s relevant because the frames are potentially quite different because smartphone
* Web resp are somewhat higher educated and more affluent -> transfer to app? (Couper, 2007)
* Registers (Groves, 2006; Groves et al., 2006)
* There are known differences between Android and iPhone owners, where iPhone users are have more disposable income and report differential spending habits (Götz et al., 2017; Keusch, Bähr, et al., 2022)

## 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; Roberts, Herzing, Sobrino Piazza, et al., 2022; Struminskaya, Toepoel, et al., 2020). 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 (Struminskaya et al., 2021; Wenz et al., 2019; 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 (Felderer & Herzing, 2023; Roberts, Herzing, Sobrino Piazza, et al., 2022). 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 was truth versus nonresponse by the addition of location measurements, by indicating through notifications or design that several hours have not been accounted for in time use surveys, or by prompting for missing details such as quantity of an item purchased in mobility 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 the answer 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). Unexpected technical challenges are rare in pen and paper diaries, but their discussion is prominent 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 (Antoun et al., 2019; Couper et al., 2017). 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 off of 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 the largest benefits may be in these categorization implementations, it may be a wider consideration. The interpretation of questions depends on question wording and layout, and this can have a meaningful impact on measurement error (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. (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 (Burkill et al., 2016; Kreuter et al., 2008). 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 error. 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.

# 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 in the wider public, as discussed in Section 4.2. 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 excluding one of either Android or iOS, which has been shown to create coverage bias in important outcome variables (Keusch et al., 2023).

Evidence for differences in selection and measurement in moving non-smart -> smart in HBS (French et al., 2008; Riegler, 2015; Sekula et al., 2005)

Coverage and participation bias in budget app (Jäckle et al., 2019)

Unit-nonresponse: Panel participants who reported using their smartphone for more discrete tasks were 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, which may reflect a potential for non-response bias along this dimension (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 (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; Chatzitheochari & Mylona, 2021; Struminskaya, Lugtig, et al., 2020). Classifying lack of behavior as missing due to undercoverage or true absence of behavior is common whenever we offload the response from active to passive, whether this is with web-tracking, sensors, or (Bosch & Revilla, 2022; Courtney et al., 2023).

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 (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.

A number of comparisons using food diaries have also been conducted, Photo-based food diaries had a small underreporting mean bias versus a moderate underreporting from paper diaries using the same participants (Costello et al., 2017) Smartphone-based measurement of food intake was as accurate as paper-based food records (Hutchesson et al., 2015) Image-based food diary compared against known truth demonstrated underreporting (average reported to actual ratio of 81%), but at a comparable level to paper diaries (54% to 94%) (Boushey et al., 2017)

Comparisons against self-report data outside of the diary format have also demonstrated meaningful measurement mode effects. Logged smartphone usage data are different from self-report data, but show better correlations with psychosocial contexts (Jones-Jang et al., 2020)

As an example, user experience and perception of the survey instrument can be quite different. Many of them relate to the interaction with the smartphone itself Respondents report feeling less connected to their time use when entering it as checkboxes with an app, versus the inclusion of a greater level of detail via traditional diary methods (Frąckowiak et al., 2022) The extra features such as tooltips and linked help sheets can provide guidance on questions that may enable users to better assess the pragmatic intent of a question, which is often the source of difficulty for respondents who do not understand a survey question (Schwarz, 2012). Users can trust the smart elements too much, disregarding their own intuition, such as using the defaults provided in an app where otherwise they would be required to estimate it themselves. (Bucher Della Torre et al., 2017)

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 is very similar to what is found in physical activity studies comparing sensor readings to objective measurements (Parmenter et al., 2022).

Add:

When comparisons are made between traditional diary studies and app-based mobility studies, there large differences in in rates, distances, and lengths of trips (Gillis et al., 2023; Greaves et al., 2015). 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.

Chatzitheochari et al. (Chatzitheochari et al., 2018) report on the usage of hard and soft checks, respectively requiring or suggesting certain actions from the user in order to reduce incomplete data, which increase full completion rates by 30-50% over the paper diaries with no such features.

# Estimation Methodology

The estimation of total mode effects between modes is very simple in that any existing difference between sample means or variances between two different modes indicates the presence of some sort of mode effect. In this way, every study that separates their analyses by mode has provided a method for estimation of the total mode effect. Unfortunately, because this mode effect is representative of the summative process of all modes of error, this is of limited use in accounting for or adjusting for the differential responses between varying modes. Any useful estimation methodology must therefore provide a method for distinguishing some (combination of) error(s) from some other (J. T. A. Vannieuwenhuyze & Loosveldt, 2013)

.

In her survey of the current literature on mobile devices, data integration, and privacy contains, Salvatore (Salvatore, 2023) found that propensity scores, missing data, and regression estimators were commonly mentioned by researchers, indicating that researchers are sensitive 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 an explicit design choice. 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 suggests three main strategies for disentangling the two sources of mode differences: 1) direct assessment of measurement error by comparing survey reports 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). (Tourangeau, 2017)

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, Jan A., 2013). Non-response bias can be estimated on the basis of correlations between demographic characteristics and survey response, leaving the remainder to be considered as measurement error (van den Brakel, 2008; van den Brakel & Renssen, 2005). Although literature in which differences are estimated in this way between smart and traditional surveys, Premkumar et al (Premkumar et al., 2023) have demonstrated the usage of this method in a household budget survey. <\*\*findings?\*\*>

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. (Klausch et al., 2015) by imputation of unit non-response after the follow-up.

## Experimental

* Longitudinal random allocation in a panel – quasi simplex model
  + The impact of mixing modes on reliability in longitudinal studies. (Cernat, 2015)
  + Estimation of mode effects in the health and retirement study using measurement models. (Cernat et al., 2016)
* Estimating coverage bias using a survey to ask about smartphone-specific ownership questions and the outcome variables of interest
  + Coverage Error in Data Collection Combining Mobile Surveys With Passive Measurement Using Apps: Data From a German National Survey (Keusch et al., 2023)
* Crossover design with two sequences of modes Split sample, two waves switching between modes, creating three groups reference, covered and responding, and partialling out the errors.
  + Simultaneous estimation of multiple sources of error in a smartphone-based survey (Antoun et al., 2019)
* Re-interview design (Biemer & Lyberg, 2003; Elliott et al., 2000)

## Non-experimental

Less here in places where the error sources aren’t sorted

* Comparing mixed-mode to single mode to separate selection and measurement bias (observed variables insensitive to survey mode)
  + A method for evaluating mode effects in mixed mode surveys (J. Vannieuwenhuyze et al., 2010)
  + Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to disentangle Selection and Measurement Effects (J. T. A. Vannieuwenhuyze & Loosveldt, 2013)
* Estimate classification error rates Measurement error categorical variables with overlapping datasets - Latent class analysis
  + Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC) (Boeschoten & Oberski, 2017)
  + Measuring temporary employment. Do survey or register data tell the truth? (Pavlopoulos & Vermunt, 2015)
  + Estimating error rates in an administrative register and survey questions using a latent class model (Oberski, 2017)
* MTMM
  + The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies. (Scherpenzeel & Saris, 1997)
* Prediction of a “true” values of numeric variables of interest, given multiple data sources covering only subsets of the population – latent variable
  + Estimation from contaminated multi-source data based on latent class models. (Guarnera & Varriale, 2016)

### Specification effects

* Measurement error assuming “Overlapping Numerical Variables with a Benchmark”
  + Combining official and Google Trends data to forecast the Italian youth unemployment rate (Naccarato et al., 2018)
* Reconciling high frequency and low frequency data – quadratic minimization
  + Solving large-data consistency problems at Statistics Netherlands using macro-integration techniques (Mushkudiani et al., 2018)
  + Macro-Integration for Solving Large Data Reconciliation Problems (Mushkudiani et al., 2014)
* Administrative sources and surveys using HMM Overlapping numerical variables without a benchmark
  + Overlapping numerical variables without a benchmark: Integration of adminis-trative sources and survey data through Hidden Markov Models for the production of labour statistics (Filipponi & Guarnera, n.d.)

### Measurement effects

* Propensity score matching resp. using covariates, difference in survey estimates is measurement effect, assuming covs. explain selection
  + Estimating nonresponse bias and mode effects in a mixed-mode survey (Lugtig et al., 2011)
  + Estimation of unobservable selection effects in on-line surveys through propensity score matching: An application to public acceptance of healthy eating policies (Capacci et al., 2018)
* Estimate measurement error Overlapping units and variables – SEM
  + Estimating the Validity of Administrative Variables (Bakker, 2012)
  + Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables (Scholtus et al., n.d.)
  + Latent Class Multiple Imputation for multiply observed variables in a combined dataset (Boeschoten et al., 2016)
* Classification error (generic, maybe other error types) with audit data
  + (van Delden et al., 2016) Accuracy of Mixed-Source Statistics as Affected by Classification Errors.
* Linking two data sources at individual level – estimating quality of probabilistic linkage
  + Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data (Chipperfield & Chambers, 2015)
  + Inference based on estimating equations and probability-linked data. (Chambers et al., 2009)
* Variance estimation in Combining several sources of aggregate data - Macro-integration by inequality restriction/balancing
  + On New Variance Approximations for Linear Models with Inequality Constraints (Knottnerus, 2016)
  + A Balanced System of Industry Accounts for the U.S. and Structural Distribution of Statistical Discrepancy (Chen, 2012)

### Distinguishing selection and measurement mode effect

* Interview/re-interview to disentangle selection and measurement bias
  + Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing (Biemer, 2001)
* Using aux. vars “unconfoundedness assumption” from causal inference theory
  + Propensity score weighting
  + Matching
    - ???(Morgan & Harding, 2006; Rosenbaum, 2021; Stuart, 2010)
  + Imputation
    - ???(Kang & Schafer, 2007; Schafer & Kang, 2008)
  + Regression estimation
    - ???(Imbens, 2004)
* ???(J. T. A. Vannieuwenhuyze & Loosveldt, 2013)

### Selection effects

* Weighting/regression-based inference for generic selection effects in non-experimental settings
  + Assessing the effect of data collection mode on measurement (Jäckle et al., 2010)
* Mode preference as covariate for explaining selection effects
  + Assessing the use of mode preference as a covariate for the estimation of measurement effects between modes: a sequential mixed mode experiment. (Vandenplas et al., 2016)

# data integration

**Integration under small differences**

When the mode 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 medium 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 medium differences involving mode measurement effect**

Following from the literature reviewed in Section 5, many smart surveys without passive measurement are expected to fall within this category in which some amount of differential measurement necessitates more care in integrating the sets of data.

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; J. T. A. Vannieuwenhuyze et al., 2014). When reinterview is possible, this offers a mechanism of disentangling measurement error that can be used in the integration process (Klausch et al., 2017). <\*\*reread\*\*>(Buelens & Van den Brakel, 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 (Seho Park et al., 2017; Suzer-Gurtekin & Valliant, 2018; Tuba Suzer-Gurtekin et al., 2012).

**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; Seunghwan 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. (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.

# Synthesis

Tie in with the French TUS experiment from TSS2: households complete both paper and app-based diaries, split on order of completion.

# References

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.

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.

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 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

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

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. (n.d.). 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.

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.

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

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.

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

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.

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.

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.

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.

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

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. (n.d.). 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.

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

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+.

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.

# Appendix A

Table 1 Prior Research

| Error source(s) | Authors | Research design | Findings |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

| INTEGRATION METHODOLOGY | Study | Methodology | Comments |
| --- | --- | --- | --- |
| Two surveys on same outcome with some overlapping variables | (Seho Park et al., 2017) | Counterfactual prediction of outcomes under opposite survey using shared indicator variables, followed by a composite estimator | Also provides variance estimation for the combined estimator, methods for both continuous and categorical vars. |
| Correcting for differential coverage | (Keusch et al., 2023) | Weighting estimates on the basis of data relating sociodemographic information and smartphone ownership/specific OS | Weighting leaves remaining bias in many measures |
| Pooling estimates | (Dzikiti, 2019) | Separate estimates from each mode obtained, then a weighted linear combination |  |
|  |  |  |  |