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 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 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.

### Coverage (different section)

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>

## 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, is important to researchers and policy-makers alike (Browning et al., 2003). Declining response rates and data that don’t align well with aggregate measures have become a problem for researchers (Crossley & Winter, 2014). 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).

* Respondents are generally quite good at being aware of when and how they have spent money, and the task is rarely complex. Thus, of the three benefits that smart surveys offer, expenditure research can be said only to benefit from a reduction in burden. This can be done either by offloading 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 . As with the mobility case, previous efforts to decrease the burden have involved moving the data collection online, incorporating decision rules to attempt to prevent motivated misreporting (Eckman, 2022). (French et al., 2008; Riegler, 2015; Sekula et al., 2005) <Evidence for differences in selection and measurement in moving non-smart -> smart>

Talk about scanning only vs scanning plus diary vs smartphone diary only

Add difficulty of the respondent to be sufficiently precise in describing the product/quantities

## 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 (DeLeeuw, 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 (DeLeeuw, 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

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? (Groves, 2006; Groves et al., 2006)
* Registers (Groves, 2006; Groves et al., 2006)

## 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 designing experiments to control for

* Auxiliary information for controlling for the confounding of selection and measurement error (also here registers?)
* Re-interview design (Biemer & Lyberg, 2003; Elliott et al., 2000)

### Unit nonresponse

Differences in unit nonresponse across different modes, factors affecting unit nonresponse (survey length, complexity, mode preference), differential strategies for reducing unit nonresponse (follow-ups, incentives) – for the last, it’s notable that apps can provide immediate feedback as a potential incentive.

* Systematic selection error of the estimator of the sample mean – depends on difference in means of resp and non resp individuals as well as the selection rate(Bethlehem et al., 2011)
* Response rates are not directly related to the size of nonresponse bias (Bethlehem, 1988; Groves & Peytcheva, 2008)

### Item nonresponse

Specific questions/sections where resp. more likely to skip or provide incomplete info, effects of mode on item nonresponse (more in self-administered vs. interviewer-administered… to what degree are apps like interviewer-administered? This has probably been investigated wrt web surveys that perform checks at entry etc.) Techniques for minimizing item non-response (here apps may have an advantage, but how to know where the tipping point is for too much app guidance leading to attrition?) Maintaining motivation is even more important in diary research.

* More impactful differences because of patterns of missingness that are more likely to be present with apps than traditional

## Mode effects due to measurement error

Differentiate between sources of measurement error. Suggestion to distinguish social desirability bias from satisficing bias—low level of involvement of respondent who approximates results for effort/interest -- (partial non-response, rounding or misunderstanding of instructions due to rapid reading). Smart surveys can have a major impact on the second – low level of smartness, search bar for an input field, undesirable effects (points where responses accumulate) which in part fall into category of satisficing bias. Few differences were found between mobile web and PC web for social desirability bias (Couper et al., 2017), but there the user is still responsible for contributing the entire response. We may find evidence of reduced social desirability bias with passive measurement <cite the website tracking fake news study by …Oriol?>.

The fact that slider/dropdown questions produce more response errors in mobile than on web could mean that we find measurement error that is induced by the UI/UX of the app (Couper et al., 2017)

Measurement error – can differ both in the amount of random error and systematic effect

Key components resp, interviewer, questionnaire. Deliberately or unintentionally provide incorrect info. Ambiguity as source of meas. Err.

Evaluate measurement equivalence

* Dillman et al, 2014 resp in interview may interpret question differently from web survey
* Two types – mode effects that shift the response distribution (differences in estimates between modes, but not between correlations), vs change in question-answer process which can produce answers not equivalent between modes (Hox et al., 2017)
* Type IV/V designs like switching modes of administration
* effect of systematic measurement error differing estimates of means and otals, random measurement error variance and reduce correlations (biemer & lyberg 2003 vs alwin 2007)?
* Klausch thesis introduction – if you combine modes where one is assumed to be a benchmark, like survey vs smart survey, means the overall measurement error bias of mixed mode design is increased
* For e.g. mobility, where someone goes may be like sensitive questions. For sensitive questions, measurement effect between web and nonweb (Burkill et al., 2016)
* Improvements to satisficing from e.g. HBS lookup where using search field, resp. are able to find the categorization for expenditures which are very extensive, but necessary for harmonization (EUROSTAT., 2003) very quickly.
* “Mixing modes can reduce the total survey measurement error, depending on the relative magnitudes of the errors in the two modes and whether the errors are in the same or opposite direction” (Tourangeau, 2017)

## 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.

Manually handling the input in paper questionnaires, especially when categorizing …. <Budget and time use activities citation here?>

-- Studies comparing web-based surveys that directly populate databases reduce scope for manual error <citation>

But then the error comes from the respondent’s misunderstanding or the interaction with the user interface?

We expect neither mode to be perfect, but for there to be systematic differences between modes

More important than in traditional mixed-mode settings

* May pull from differences between processing less aggregated sources? GPS trackers -> travel diary -> recall; processing of diary studies versus other
* Part of the processing may not be visible
* (Aschauer et al., 2021) Describes processing steps of a combined travel/time use/expenditure diary, including the extensive validation after plausibility checks – describes three types of nonresponse i. errors detected during data processing with no additional observed info, ii. Errors which can only be detected by means of follow-up assessment and iii. errors non-detectable due to unwillingness to disclose. Adding sensor measurements impacts the certainty with which you can impute i and the ability to identify type ii potentially before validation step.

Notes:

* Variants
  + Mode choice bias vs real measurement differences (2012
  + Sequential vs concurrent
  + Single vs multiple per resp.
* Mode effects are evaluated relative to some benchmark mode
* Diary vs recall (“on a typical day…”) as comparison?
  + If so: time use diary sleep vs recall sleep (Kaplan et al., 2020)
* Confounding (Vannieuwenhuyze & Loosveldt, 2013)
  + Mode-specific response distributions are impacted by selection *and* measurement
  + Just looking at a difference between estimates from two surveys with different modes provides a combined estimate of the total source of mode effect
* Role of auxiliary variables in estimation -- exogeneity / not affected by measurement effects within the multiple modes. Often what’s available is weakly related to response mechanisms (source?)
* MEPS experiment
* 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)
  + Direct comparison (not simple) – no frame data exists on the smart survey question topics

<Small paragraph on coverage, but probably this is the other WP2 deliverable>

# Mode effects in smart surveys

TSS1 – WP2 deliverable 2.1 consumption interview vs none

* Diary vs app comparisons
  + Combined mode effect
    - (Wenz, 2023): Compares data from scanned receipts, scanned receipts plus direct entry with national budget survey, uses inverse probability weighting to match the sample composition of the app to the diary (could be considered reducing mode selection effect)
      * Distribution from SR+DE aligns with LCF for median total expenditure, but receipts alone underestimate expenditure (101.30 vs 122.80 vs 70.10)
      * Percentage zero expenditures is higher in the two-week period from scanned-only
      * SR for food and groceries is not significantly different from the benchmark, but SR + DE overestimates
  + Mode measurement effect
  + Mode selection effects
    - Coverage and participation bias in budget app (Jäckle et al., 2019)

# Estimation Methodology

## Experimental

* Probability samples with experiments for relative differences (probability sample, assigned a data collection mode, test hypotheses about differences)
  + Design- based analysis of factorial designs embedded in probability samples (van den Brakel, Jan A., 2013)
  + Design-based Analysis of Embedded Experiments with Applications in the Dutch Labour Force Survey (van den Brakel, 2008)
  + Analysis of Experiments Embedded in Complex Sampling Designs (van den Brakel & Renssen, 2005)
* Split sample – random assignment to a mode, then follow-up repeated measurement with one mode
  + Disentangling mode-specific selection and measurement bias in social surveys (Schouten et al., 2013)
  + Selection error in single- and mixed mode surveys of the Dutch general population (Klausch et al., 2015) -> imputation for unit nonresponse after the followup
* 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)

## 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 (Vannieuwenhuyze et al., 2010)
  + Evaluating Relative Mode Effects in Mixed-Mode Surveys: Three Methods to disentangle Selection and Measurement Effects (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., 2015)
  + 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)
* ???(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

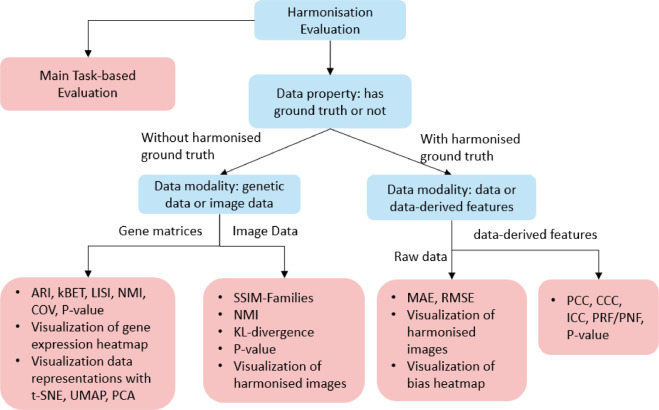
* \*\*Mixed-mode measurements as treatments in causal modelling framework counterfactual.. potential outcomes through regression, then overall estimates are combos of real answers and predicted answers
  + Suzer-Gurtekin, Z. T., Heeringa, S., & Vaillant, R. (2012). Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys. Proceedings of the JSM, Section on Survey Research Methods, 4711-2.
* \*\*Multiple imputation of non-observed answers
  + Kolenikov, S., & Kennedy, C. (2014). Evaluating three approaches to statistically adjust for mode effects. Journal of survey statistics and methodology, 2(2), 126-158.
  + Park, S., Kim, J. K., & Park, S. (2016). An imputation approach for handling mixed-mode surveys. The Annals of Applied Statistics, 10(2), 1063-1085.
* \*\*Reweighting of survey response – calibration correction of survey response to fixed-mode distributions
  + 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., and J. van den Brakel (2017). “Comparing two inferential approaches to handling measurement error in mixed-mode surveys,” Journal of Official Statistics, 33(2), 513-531.
* \*\* Covariate adjustments to correct for measurement effects
  + Vannieuwenhuyze, J.T., Loosveldt, G., Molenberghs, G. (2014) Evaluating mode effects in mixed-mode survey data using covariate adjustment models. J. Off. Stat. 30(1), 1–21
* \*\*IRT-based Bayesian hierarchical model/latent traits
  + 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.
* \*\*Re-interview after observational study and adjustment to benchmark mode
  + 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.
* \*\* Combine eDiary and sensor measurements using visual analytics approach (mixed-methods approach, quantitative and qualitative)
  + (Resch et al., 2020)
* Reducing variance estimates with small area estimation, which can be used not only for geographic areas, but distinct small categories that are related
  + Small Area Estimation. (Boonstra et al., 2008)
  + Towards small area estimation at Statistics Netherlands (H. J. Boonstra et al., 2008)
* Variance estimation in Combining several sources of aggregate data - Macro-integration by inequality restriction/balancing
  + Macro-Integration with Inequality Constraints: An Application to the Integration of Transport and Trade Statistics (H. J. H. Boonstra et al., 2011)

Bootstrap inference using estimating equations and data that are linked with complex probabilistic algorithms (Chipperfield, 2020)

# Synthesis

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

Additional ideas:



* Maybe somewhere mention combined MAED? (Aschauer et al. 2019 which combines mobility, expenditure, and time usage?)
* Nielsson boxes

# 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.

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.

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.

Bethlehem, J. G., Cobben, F., & Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Wiley.

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

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

Boonstra, H. J., 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.

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

Browning, M., Crossley, T. F., & Weber, G. (2003). Asking consumption questions in general purpose surveys. *Economic Journal (London, England)*, *113*(491), F540–F567.

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

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/

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.

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.

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 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., & 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.

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

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.

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.

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

French, S. A., Shimotsu, S. T., Wall, M., & Gerlach, A. F. (2008). Capturing the spectrum of household food and beverage purchasing behavior: a review. *Journal of the American Dietetic Association*, *108*(12), 2051–2058.

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.

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.

Groves, R. M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167–189.

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.

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.

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

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., 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.

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

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., 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.

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.

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.

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

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.

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.

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

Riegler, M. P. R. (2015). *Integrating the Web Mode in*. cros-legacy.ec.europa.eu. https://cros-legacy.ec.europa.eu/system/files/Presentation%20S17CP2.pdf

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.

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.

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.

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.

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.

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.

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.

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. 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. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045.

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

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 Prior Research

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