Understanding Society

Working Paper Series

No. 2016 – 07

September 2016

Total survey error for longitudinal surveys

Peter Lynn[1] and Peter Lugtig[1,2]

[1]Institute for Social and Economic Research, University of Essex

[2]Department of Methodology and Statistics, Utrecht University

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Total survey error for longitudinal surveys

Peter Lynn and Peter Lugtig

Non-Technical Summary

When considering the design of a survey, it is beneficial to take into account all the different aspects of survey design and survey implementation that can affect the statistical error of estimates based on the survey data. These aspects may include things like the way the sample is selected, the way the questionnaire is designed, the success of the survey at persuading people to participate, and so on. Moreover, the effects of these different design aspects may not be independent. For example, the effect of questionnaire design on statistical error may depend on which kinds of people are in the sample. The need to understand the complex ways in which various design aspects combine to influence statistical error is reflected in the idea of "total survey error," but most writing and thinking to date about total survey error addresses issues relevant to one-time surveys and largely ignores issues that are unique to longitudinal surveys.

There are several aspects of survey error, and of the interactions between different types of error, that are distinct in the longitudinal survey context. Furthermore, error trade-off decisions in survey design and implementation are subject to some unique considerations. For these reasons, a framework for total survey error in the longitudinal survey context – lacking to date – is desirable. This article aims to show how uniquely longitudinal sources of error in surveys should be understood within the established total survey error framework. We also provide examples of studies of some of the interactions between errors that are unique to the longitudinal survey context and we make suggestions for further research that should help to improve design decisions for longitudinal surveys in the future.

Total survey error for longitudinal surveys

Peter Lynn and Peter Lugtig

Abstract

This article describes the application of the total survey error paradigm to longitudinal surveys. Several aspects of survey error, and of the interactions between different types of error, are distinct in the longitudinal survey context. Furthermore, error trade-off decisions in survey design and implementation are subject to some unique considerations. Previous literature on total survey error mostly fails to explicitly consider uniquely longitudinal issues. We aim to show how uniquely longitudinal sources of error in surveys should be understood within the Total Survey Error framework and we provide examples of studies of some of the unique interactions between errors.

Key words: between-wave intervals, data collection mode, following rules, questionnaire design, specification error, survey error components

JEL classifications: C81, C83

Author contact details: plynn@essex.ac.uk

Acknowledgements: A revised version of this article will be published as Chapter 13 in the book *Total Survey Error in Practice* (editors Paul P. Biemer, Edith D. de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker & Brady T. West), to be published by Wiley in 2017. Readers wishing to cite this article should cite the book version, not this Working Paper. The first author is funded by awards from the UK Economic and Social Research Council to the University of Essex for *Understanding Society*, the UK Household Longitudinal Study.

Total survey error for longitudinal surveys

Peter Lynn and Peter Lugtig

1. Introduction

In this chapter we describe the application of the total survey error paradigm to longitudinal surveys, by which we mean surveys that collect data on multiple occasions from the same sample elements. Longitudinal surveys are of great importance and considerable investment has been made in them by government agencies and academic funding bodies in recent years [2, 45, 67]. However, there are several aspects of survey error, and of the interactions between different types of error, that are distinct in the longitudinal survey context. Furthermore, error trade-off decisions in survey design and implementation are subject to some unique considerations.

Previous literature on total survey error mostly deals with survey design and implementation issues pertinent to cross-sectional surveys and does not explicitly consider uniquely longitudinal issues [3, 27]. Weisberg [74] describes several relevant features of longitudinal surveys but does not address their error structures. Smith [64] draws attention to this limitation and points out that the relationship between error structures at different waves may be at least as important as the nature of the error structures. He suggests “to extend TSE to cover multiple surveys including … longitudinal panels” (p.481).

For these reasons, a framework for total survey error in the longitudinal survey context – lacking to date – is desirable. This chapter aims to show how uniquely longitudinal sources of error in surveys should be understood within the Total Survey Error framework.

We outline how such a framework differs from the TSE framework in a cross-sectional context, discuss how it might be applied, and provide examples of studies of some of the

unique interactions between errors. The next section describes the aspects of longitudinal surveys that have distinct implications for the consideration of total survey error. The subsequent section outlines the nature of each major error source in the longitudinal context and how different sources can interact on longitudinal surveys. We then outline ways in which design and implementation decisions can be informed by total survey error considerations. We show that the decisions taken depend on the goals and nature of the survey, and that interactions between error sources may change over waves of the survey. Three examples show how the trade-off between TSE components can be studied in a longitudinal context. In the concluding discussion section, we draw attention to and discuss the interplay between TSE, other dimensions of survey quality (e.g. timeliness, relevance), and survey costs. We also highlight areas of research or practice that we believe deserve more attention regarding TSE for longitudinal surveys.

2. Distinctive Aspects of Longitudinal Surveys

Longitudinal surveys vary greatly in their objectives, study populations and design features [4, 32]. However, as pointed out by Lynn [45] and Smith [64] they share certain methodological considerations that are distinct from cross-sectional surveys and have implications for the consideration of total survey error. In this section we outline those issues.

Key *dependent variables* for analysts of longitudinal survey data are typically measures of change. Respondents typically answer the same questions at multiple time points, enabling the study of change at the level of the individual. When longitudinal surveys are representative of the population, individual rates of change can be aggregated to study trends at the population level. Because of this, longitudinal surveys are important sources of data for social scientists or policy-makers interested in studying dynamics.

Change estimates can take the form of either transition matrices, in the case where the underlying responses are categorical, or continuous distributions. Key descriptive estimates are therefore means or proportions of these measures of change, including for subgroups, while key analytical estimates tend to be parameter estimates from various forms of models predicting these changes.

Time plays a particularly important role in longitudinal surveys, contributing to the definition of both the survey sample and the study population, as well as forming an essential element of the measurement of change and associated aspects such as respondent recall. In data collection, consideration of the timing of each wave and the time interval between waves has multiple implications. Furthermore, data release and data analysis is a dynamic evolving process. Many surveys release data after each wave is complete. This gives rise to many opportunities for previously-released data to be revised in the light of updated information, including value-added data such as non-response weights and imputed values.

For recall questions, which are common in longitudinal surveys, responses can be *bounded* [57] by responses given at a previous wave. Furthermore, responses given at previous waves can be explicitly fed forward to the respondent through the use of *dependent interviewing* [29]

The responses given by sample members to survey questions may be different from the responses that they would have given had they not previously been interviewed as part of the survey, an effect known as *panel conditioning* [32, 73].

Unit non-response is not a simple dichotomy, but is instead a complex pattern determined by a combination of wave non-response and sample attrition [42]. Similarly, patterns of item missing data have two dimensions rather than one, between-items and between-waves. The dynamic and multi-dimensional nature of the non-response process introduces

extra challenges and complexities for adjustment methods to correct for associated errors [50, 71].

Finally, cost structures are quite distinct in the case of longitudinal surveys. The division between initial and ongoing costs is important. Different design decisions may have different cost implications for future waves. There may be legacy effects and precedent effects. In this respect, and others, there is an important distinction between time-limited studies (like SIPP, with its rotating panels) and those which are open-ended, possibly continuing for as long as funding can be obtained (like Understanding Society and LISS).

Three example surveys

1. **Understanding Society** is a multi-topic panel survey of almost 100,000 members of the UK population that started in 2009. All sample members aged 16+, and other adult members of their households, are interviewed annually and followed as long as they remain in the UK. Young persons aged 10 to 15 fill a self-completion questionnaire. Interviews are conducted primarily by Computer Assisted Personal Interviewing (CAPI), but with some CATI and, since wave 7, web too. The sample includes the 16,500 members of the **British Household Panel Study (BHPS)**, which began in 1991, plus a new sample of 77,300 persons. See Buck & McFall [7] or www.understandingsociety.ac.uk.

2. **The Longitudinal Internet Study for the Social Sciences (LISS)**. In 2007, 8.000 respondents were recruited using a mixed-mode design. Respondents complete questionnaires monthly through a Web Survey. The topics in the questionnaire change monthly, but several questionnaire modules are repeated annually. See www.lissdata.nl for all questionnaires and data [62].

3. **Survey of Income and Program Participation (SIPP)**. The SIPP collects data on economic well-being, education, assets, health insurance, food security and more. It has undergone several major changes throughout its history. In the 2008 panel of the SIPP, members of about 42.000 households including 105.000 respondents were interviewed every 4 months for 16 waves. Waves 1, 2 and 6 are conducted by Computer Assisted Personal Interviewing (CAPI), while the other waves are conducted primarily by telephone (CATI) [70]. See http://www.census.gov/sipp/.

3. Total Survey Error Components in Longitudinal Surveys

The conceptual components of TSE in longitudinal surveys are the same as those for cross-sectional surveys, yet errors arise in somewhat different ways and may have distinct forms. In the previous section, we have argued that longitudinal surveys differ from cross-sectional surveys in a number of important senses. In particular, estimates of interest are usually related to *change*, and *time* is an important element in most aspects of survey design and implementation. These differences have an impact both on the process that leads to errors arising, and on the likely nature of the resultant errors. In this section we present an overview of the cause and nature of each major type of survey error, focusing on those aspects that differ from the cross-sectional perspective. We focus on issues that are germane to all types of longitudinal surveys, including cohort studies, household panels, online panels, and others; where issues may be specific to particular types of longitudinal surveys, we point this out.

Coverage error. In cross-sectional surveys, coverage is a one-off problem that is linked to quality of the sampling frame that is used to draw a sample from the target population. When a longitudinal survey studies a stable population, such as a birth cohort, the cause of under-coverage is the same, but the nature of the resulting error could be different as it depends on the extent to which subsequent substantive change is different amongst the excluded sub-population rather than the extent to which current characteristics differ. Other longitudinal surveys study a dynamic population of which membership changes over time. For such surveys to correctly represent the changing population over time, a dynamic sampling method is needed to regularly add appropriate samples of new population entrants. The source and frequency of these additional samples can contribute to coverage error. In the case of a longitudinal survey of graduates of a training scheme, for example, the new population entrants would consist of new graduates since the time the

initial sample was selected. In the case of household panel surveys, the new entrants may be immigrants, new births and people leaving institutions to re-enter the household population. Typically, household panel surveys have procedures to routinely add samples of new births at each wave – by adding those born to existing sample members – while the addition of immigrants or people leaving institutions require special sampling procedures to be implemented periodically [46]. Unlike *Understanding Society*, the German Socio-Economic Panel adds to the sample at each wave all new members of the current household of any sample member. Over time, this results in a sample with a very different structure and leads to major complexities in estimating selection probabilities [63].

Undercoverage of immigrants can be a serious problem in longitudinal surveys. New immigrants usually have different characteristics and experiences than the previously resident population, and estimates of change become biased when they are under-represented. In *Understanding Society* for example, recent immigrants have been included in the sample by selecting an additional sample of addresses and carrying out a short screener at the time of wave 6 to identify people who have entered the country in the five years since the wave 1 sample was selected [49].

A final point regarding coverage is that overcoverage can occur, both in longitudinal surveys of static populations and those of dynamic populations. To avoid this, it is important to identify when sample members leave the study population (e.g. due to emigration, institutionalisation, or death). Otherwise, such people will be assumed to still be eligible for the survey. The likely consequence of this is that non-response adjustment procedures will lead to an over-representation of persons with characteristics similar to those of population leavers (see [58] for an example of this). Tracking procedures (see below) play an important role in ensuring that population exits are correctly identified.

Sampling error. When the population of interest is dynamic, undercoverage can be avoided by periodically adding additional samples, either of new population entrants or of

the entire population (which includes new population entrants)[1]. However, this can lead to systematic sampling error unless great care is taken to ensure that relative selection probabilities can be estimated. In particular, the inclusion probabilities of people in the additional sample will depend on whether they were already part of the population at the time when the original sample was selected (and/or when any previous additional sample was selected), or whether they have since entered the population. Lynn and Kaminska [51] outline how design weights can be produced when different types of additional samples are added to a longitudinal survey. How complicated this procedure is depends on the specific rules for including new sample members. The LISS panel tries to include immigrants by oversampling anyone born outside the Netherlands in the top-up samples. The total inclusion probabilities for those sample members depend on the moment an immigrant entered the population. This information is unavailable from the sampling frame, and information from the recruitment only gives information about respondents, not nonrespondents. Computing weights becomes extremely complicated.

Random sampling variance can also be affected by the choice of methods for adding additional samples to the survey over time. For example, if the initial sample involves a multi-stage (clustered) design and an additional sample is defined through a known linkage to existing sample members (as in the case of adding new births to household panel surveys, described above) then clustering effects will also be present in the additional sample. Another practice which can have unpredictable effects on sampling variance in longitudinal surveys is the assignment of sampling fractions based on the current value of a time-variant variable (e.g. over-sampling “poor” households or unemployed persons). This may have beneficial effects on the precision of estimates in the initial waves, but over time the association between the sampling strata and survey

[1] These are variously referred to in different contexts as refresher, refreshment, replenishment or top-up samples.

measures of interest may weaken to the extent that an equal-probability sample would have provided greater precision.

Nonresponse error. Nonresponse error in longitudinal surveys is the cumulative result of initial non-response, item- and wave-nonresponse and attrition. Because respondents have more opportunities to not give an answer to some questions, or drop out from the survey altogether than they have in a cross-sectional survey, it is very important to design procedures that minimize missing data and the resulting nonresponse error. A slowly decreasing sample size will decrease statistical power, and is likely to lead to biases in survey estimates.

There are at least two aspects of the process leading to unit non-response that are unique to the longitudinal survey context. The first is that sample mobility can lead to non-location being a major component of non-response at each wave subsequent to the first. To tackle this, tracking sample members between waves should be a major survey activity [1, 15). Despite best efforts, sample members who have moved since the last wave are more likely than others to be nonrespondents at the next wave. The second aspect of the nonresponse process that is unique to longitudinal surveys is that, from the second wave onwards, respondents know in detail the content and nature of the survey. The experience of having taken part previously is likely to be a major factor affecting willingness to take part again. Though there is some evidence that willingness to take part again is not very sensitive to the length of the interview [48] it is likely to be sensitive to the content of the interview. For this reason survey designers take great pains to ensure that sample members are left with a pleasant impression of participation. Both of these unique aspects of the nonresponse process are likely to result in a component of nonresponse that is strongly associated with some survey outcomes. Moving will tend to be associated with certain characteristics, experiences and events, including those which longitudinal surveys are often particularly interested in measuring, such as job changes and partnership formation or dissolution. Being unwilling to participate again as a result of a negative

experience of the previous interview may be associated, for example, with true scores on sensitive or embarrassing survey items.

Furthermore, when either of two consecutive measurements is missing, it becomes impossible to estimate change between waves. For this reason, even small proportions of nonresponse at each wave can lead to large errors and biases, especially when the reason for nonresponse is related to change. Wave nonresponse or attrition can be caused by failing to locate a sample member, noncontact, a refusal or inability of the sample member to participate. All these reasons potentially signal that something changed in the respondent's life. For example, nonlocation may be disproportionately associated with household moves, refusal with marital breakdown or other domestic disruption, inability to participate with deterioration in health, and so on. On the one hand it can be seen as a weakness of longitudinal surveys that nonresponse – at least from wave 2 onwards – may be more likely to be associated with survey measures than it is in cross-sectional surveys. On the other hand, paradoxically, it could be that the data are more likely to be missing at random than in the cross-sectional case, due to the wealth of auxiliary data available from wave 1 and other waves. Nevertheless, maximizing participation rates remains of paramount importance for longitudinal surveys and consistency in participant communications and fieldwork practices may help to induce the habit of participation.

Adjustment error. Error can be introduced in the process of adjusting the data to correct for differential inclusion probabilities and coverage and nonresponse errors. As with cross-sectional surveys, weighting or imputation methods are typically used. The complications of adjusting longitudinal survey estimates are related to the fact that the population of interest is often dynamic and to the complex patterns of non-response. Consequently, eligibility status (population membership) can change over time and is not always known with certainty [51]. Dynamic sampling over time can also lead to difficulties in estimating selection probabilities [40]. A specific feature of imputation in the longitudinal context is

that it can in principle take into account both earlier and later responses. Depending on survey policy, this can result in imputed values being revised in subsequent data releases as more information becomes available. Consequently, the structure of imputation error is likely to differ from that which would arise with simple cross-sectional imputation, and may change as more waves of data are collected.

Specification error. In order to study change over time, it is important that relevant questionnaire items remain the same over time. In practice however, some concepts of interest change. For example, the LISS panel has an annual module on social integration and leisure, in which several questions are asked about Internet use. Between 2007 and 2013, topics on 'online dating', 'social networks' and 'online gambling' were added to the questionnaire, while several other questions were reworded to keep up with changing Internet use in the population. Similarly, the show cards used on BHPS as respondent prompts in connection with questions on income sources have had to be revised several times over the years to include new welfare benefits and reflect changes to the names of others. Such changes on the one hand introduce doubts about the comparability of measurements over time. If the questionnaire is altered, it becomes harder to meaningfully compare data over time. On the other hand, not changing the questionnaire over time may lead to larger specification errors. In that case, data between waves can be compared to estimate change, but the change measures risk becoming irrelevant.

Measurement error. Estimates of change can have rather different measurement error properties from analogous estimates relating to cross-sectional measures. In particular, error in a measure of change will depend on the correlation between errors at each wave if the change is inferred from the answers to independently asked questions, while it will depend on a more complex relationship between errors if dependent interviewing is used. In cross-sectional surveys, problems with the validity of the survey question will lead to biased estimates, whereas problems with the reliability lead to a lower precision, and attenuated estimates of measures of association. In longitudinal surveys, low validity and

reliability of survey questions can have different effects. Consistent biases due to validity problems can produce unbiased estimates of change. Problems with reliability however, lead to attenuated correlations over time and an overestimation of change [8, 30, 29].

To better understand the properties of dependent Interviewing, the BHPS conducted a study which involved three experimental designs – proactive and reactive forms of dependent interviewing, and a standard questionnaire with independent questions (control) – combined with validation through linkage to administrative data and employer records. It was found that dependent interviewing reduced underreports of income received in the year preceding the interview but did not lead to overreports, leading to a net reduction of measurement error in sources and amounts of income received [53, 54]. Longitudinally, the reduction in measurement error at each wave led to a small increase in the consistency of reports over time and a reduction of measurement error in change estimates [44]. More consistent and better estimates of transitions were also found with respect to aspects of employment, such as occupation and industry [52]..

When respondents implicitly or explicitly remember questions from earlier waves (and the answers they gave), the assumption of independent measurement errors over time may be violated. This is a form of panel conditioning. For example, there is evidence that respondents remember and recognize filter questions and on purpose underreport to avoid follow-up questions [21]. This can lead to underestimation of change, if respondents avoid mentioning changes in their life in order to avoid additional questions. But it could also lead to overestimation of change if, for the same reason, respondents avoid mentioning circumstances that applied previously.

Another manifestation of panel conditioning with negative consequences is that because of participating in the panel survey, respondent may start genuinely to think or behave differently, again leading to biased estimates of change as the behaviour of panel members is no longer representative of that of the study population [73]..

However, panel conditioning need not always have a negative impact on measurement. For example, respondents may learn to prepare for the interview by having relevant documents such as pay-slips to hand, or may acquire trust in the interviewer and the survey, leading to more honest and open answers. For example, Uhrig [69] found that apparent social desirability bias in reported height and bodyweight appeared to reduce with repeated interviewing.

Processing error. Processing errors are not often studied in either cross-sectional or longitudinal surveys. Such errors may be small, and are often thought to contribute to variance, but not to bias. In longitudinal surveys, advantage can be taken of the fact that respondents have typically been asked the same questions previously. Dependent Interviewing, as discussed above, does this in order to reduce measurement error. In similar fashion, both coding and editing can be carried out in ways that make use of responses provided previously. In the BHPS for example, industry and occupation codes are checked longitudinally for consistency [38]. Although such edit checks can be helpful to reduce measurement error, they also risk introducing new processing error and, specifically, artificially inflating the between-wave correlation of errors. One specific issue associated with dependent interviewing concerns the editing of previous wave verbatim responses before feeding them forward into the script for the next interview. It is important that the respondent should recognize their responses, so language should not be changed [59]. For example, the verbatim description of a job should be fed forward, rather than the description of the occupation to which it was subsequently coded, even if the aim is to verify that the same code still applies. But editing of spelling, use of upper case, and obvious typographical errors may be helpful to the interviewer.

4. Design of Longitudinal Surveys from a Total Survey Error Perspective

In this section we will outline how the nature and interactions of components of Total Survey Error in longitudinal surveys affect decisions on the design and implementation of these surveys. We identify five key design features of longitudinal surveys that influence the nature and interactions of Total Survey Error. It is important to keep in mind that the research questions of interest, as well as the nature of the population under study should largely determine the specific design of the longitudinal study. For example, the rationale of a birth cohort study is often the study of diverging life trajectories among a sample that is in some respects initially relatively homogenous. Birth cohort studies are mainly designed to study 'age' effects, keeping 'cohort' and 'period' effects constant. Household panel studies typically have much wider aims: apart from looking at development that is associated with age, household panels can also compare different birth cohorts ('cohort' effects) to see whether development differs over generations. The effects of changes in society (e.g. policy changes at a particular time point) can also be assessed when the sample is representative of the whole population. Although household panel studies offer advantages in terms of the types of empirical questions that can be answered with the data, they also have disadvantages. For example, the sample size of specific age cohorts will be small and the extent of age-tailored questioning will be limited.

1. Is the panel study fixed-time or open-ended?

An important feature of longitudinal survey design is whether or not it is known in advance for how long each sample element will be followed. Rotating panels, such as the SIPP, have in advance decided on the number and frequency of waves conducted in each rotation of the panel. In the case of SIPP, respondents are interviewed for 10 waves, over an average period of 32 months. The limited time-span of the survey will ensure that specification error will not increase much over the course of the panel survey. Similarly,

the cumulative level of nonresponse over the entire course of the panel can be anticipated. Coverage error will remain limited and the need for additional samples can be anticipated and planned in advance. With a rotating panel design, coverage error in cross-sectional estimates can in principle be avoided by appropriate weighting of population entrants in each panel.

In contrast to rotating panel surveys, many longitudinal surveys have no fixed end-date for respondent participation. Typically, the continuation of a longitudinal survey depends on funding, and can only be planned a few years ahead. This makes it much harder to take decisions on the design of the survey if the goal is to minimize total survey error. For some design decisions, both statistical and cost implications may depend on the number of waves remaining and therefore cannot be estimated if the number of further waves is unknown.

2. Who to follow over time?

Longitudinal surveys often face decisions about which types of persons to follow over time. These decisions can relate either to the possibility of not continuing to follow certain existing survey respondents or to the possibility of adding in new sample members (see the earlier discussion of additional samples in the context of sampling error) who will then need to be followed. Such decisions clearly have implications for both sampling error and costs, but they may also have implications for other types of error. For example, a survey may be faced with a choice between selecting additional samples via a known relationship to existing sample members or via an independent selection method (see [46], for an example). The first method may introduce a larger sampling variance for any given sample size (due to dependency between sample elements) and may perhaps also suffer some additional under-coverage (if not all population entrants have a link with existing population members), but on the other hand it may achieve higher participation rates, and hence possibly lower nonresponse error, than the second method, and the unit cost per interview may also be lower. In other words, the choice of sampling method has implications for the

structure of the sample of people to be followed. And the structure of the sample in turn affects an interaction between coverage, sampling and nonresponse errors, and costs. Such considerations are to some extent behind the variation in following rules between different national household panel surveys [63].

Another important consideration is the possibility that units may leave and re-enter the study population. For example, many surveys, including household panels, birth cohorts and retirement panels, limit their focus to persons residing in the household population of a particular country. But people may leave the country, only to return some time later, or may enter a residential institution such as a nursing home or prison, but exit again at a later date. Attempting to track all sample members when they leave the study population, just in case they return later, can be complex and costly, especially when only a small proportion are expected ever to return to the study population. However, failing to do so can introduce coverage error, while tracking in a low-cost way may be relatively unsuccessful and therefore lead to nonresponse error.

3. Should the survey use interviewers or be self-administered?

The decision whether or not to use interviewers to administer the survey has consequences for many different error sources [16, 17, 19]. Interviewers are generally good at building rapport with respondents, leading to a lower nonresponse error over time, as compared to self-administered surveys. This beneficial effect however comes at the price of increased costs, to which we turn later. Interviewers can also influence measurement error, though this influence can be either beneficial or detrimental, depending primarily on the nature of the survey questions. For complicated factual questions, the interviewer may help the respondent to come to an answer by explaining concepts and definitions (for example about life histories or income), leading to lower measurement error. But for sensitive questions, respondents may be subject to a greater

social desirability bias if an interviewer is present leading, for example, to greater under-reporting of alcohol consumption, drug use, and extreme attitudes [36, 68]. With face-to-face interviewing, the interviewer's appearance and behaviour during the interview can additionally affect respondents' answers [20]. In general, interviewers introduce interviewer effects on measurements, reducing the variance between respondents interviewed by the same interviewer.

If feasible, respondents in face-to-face longitudinal surveys are typically assigned to the same interviewer at each wave. In that situation, interviewer effects may be stable over time, thus not affecting change estimates. The other, more important, reason for assigning the same interviewer to the respondent over time is that wave-on-wave response rates may tend to be higher (though this effect is not universal and may not, on average, be strong: [10, 11, 55]. There are some situations, however, in which it is either necessary or desirable to switch interviewer. When a respondent moves to a different area, a new interviewer will be assigned to save costs. Sample members who refuse to participate at one wave may also have a higher propensity to respond at the next wave when they are assigned to a different interviewer [55]. The decision of whether to assign the same or a different interviewer therefore involves a trade-off between nonresponse error, measurement error, and costs.

Careful choice of survey mode is extremely important, because the mode used to administer the questionnaires drives many other design features, and has large effects on both nonresponse and measurement error. Many cohort and panel studies have traditionally been conducted by face-to-face interviewers. Both the BHPS/Understanding Society and SIPP started as a face-to-face survey. In recent years, panel surveys have started experimenting with mixing survey modes, both within and across waves. SIPP has decided to use CATI interviewing after wave 2, while Understanding Society has experimented with different methods to move to a self-administered web survey for a proportion of the sample [31, 75]. LISS has taken a different approach. Like most

probability-based online panels [5, 6], LISS used multiple modes in the recruitment of respondents, but all further interviews are solely conducted by Web.

Changing survey modes during the life of a panel can affect both nonresponse and measurement error. Some respondents may dislike the new mode more than the old one, or see it as an opportunity to stop participating. This can lead to higher nonresponse [47]. Because measurement errors differ between survey modes [35] the change of survey mode can also lead to a change of measurement error within individual respondents [18]. At the sample level, a change of survey modes will thus tend to lead to over-estimation of change.

Modes can also be mixed within a wave. This is done to target respondents with a mode so that either TSE and/or survey costs are minimized. For example, in Understanding Society, starting at wave 7, a web survey is being offered to respondents who did not participate at the previous wave, based on experimental evidence that suggested that response rates would be boosted amongst this group by this approach [31]. The goal of such targeting is to reduce nonresponse error at a limited cost. It may however come at the expense of increased measurement errors in estimates of change at the individual level. Another effect of allowing variation in the mode of response within each wave is that the proportion of responses in each mode may vary over waves, leading to net measurement effects and potentially inconsistent estimation of gross change.

An advantage of longitudinal surveys is that contact details of respondents (for example, email addresses and phone numbers) can be collected at every wave, leading to more possibilities to mix survey modes both within and across waves. Despite the potential negative impacts on measurement error when modes change between waves, there are substantial cost savings to be made when a longitudinal survey switched from interviewer administration to self-administration. Longitudinal studies can potentially combine the great advantage of interviewer-based surveys (low nonresponse) with the advantage of low costs and measurement capabilities of web-based surveys. However, little is known about

how to do this in such a way that both nonresponse errors and measurement errors remain limited.

4. How long should between-wave intervals be?

Many longitudinal studies collect data annually. Understanding Society is an example of such a study. Other examples include SOEP in Germany [72] and HILDA in Australia [76]. The EU-SILC collects data annually in all 27 EU member states: a rotating panel design is used in 24 countries and a simple (perpetual) panel in three countries [28]. Other panel studies collect data more frequently. The SIPP collect data on the same respondents every four months, while the LISS does so every month. Many national labour force surveys collect data quarterly using a rotating panel design [23, 24]. Collecting data more frequently increases the organizational demands of running the panel survey. Checking and editing data in time to feed it forward for the next wave, whether for the purposes of dependent interviewing or sample management, is more challenging the shorter the time available for the task. Related to this is the survey mode. Internet-administered surveys, such as LISS, can be much more timely than interviewer-administered surveys. While organizational and cost issues are an important aspect of choosing the duration of between-wave intervals, survey errors are also very important.

A great advantage of collecting data more frequently is that measures of change, and the timing of such change, can be estimated more accurately. Within the BHPS, respondents are for example asked to report sources of income received in each of the previous 12 months. Recall error may make the resulting data less accurate than for example in SIPP, where the reference period is only 4 months [26]. Lugtig, Glasner & Boevé [43] showed that shortening the reference and recall periods from one year to one month, greatly reduces underreporting of non-salient events.

Attempting to collect continuous histories of rapidly-changing phenomena may not be worthwhile unless the interval between waves is short enough to keep recall error to an

acceptable level. Another advantage of collecting data more frequently is that more questions can be asked, or, alternatively, that long questionnaires can be broken up in shorter parts. A downside of collecting data very frequently, is that panel conditioning effects may become larger. Recognition of this is one of the reasons why the U.S. Consumer Expenditure Survey is moving from a rotating panel design with four waves at 3-month intervals to one with two waves at 12-month intervals [22].

Collecting data more frequently may also come at the expense of increased nonresponse (error). With every wave that is fielded, respondents have a chance to drop out from the panel. An extreme example of this is a diary study, in which respondents are asked to record data in real-time. Such studies put a heavy burden on respondents, and can lead to higher nonresponse [66].. As with the choice of survey mode, there is a trade-off between measurement and nonresponse error with the choice of between-wave interval.

When intervals between waves are long, the burden of responding is lower (in any given time period), but the survey itself will become less salient. The longer the interval, the greater the investment that is needed in respondent tracking procedures. Frequent communication is necessary to make sure that the number of failures to locate or to make contact is limited. Change-of-address cards, or online forms, work well, especially in combination with incentives [15, 25, 56). When data collection is frequent, incentives can be kept low and communication between waves may not be necessary. For example, LISS uses incentives of €7,50 for every wave, and does not use between-wave mailings. Generally, incentives to limit nonresponse error in longitudinal surveys work slightly differently than in cross-sectional surveys [39]. They are as much an incentive for future participation, as a reward for past participation.

5. How should longitudinal instruments be designed?

Longitudinal studies aim to measure change, and questionnaires are constructed to enable the study of change. This implies that questions, or even entire sections of a questionnaire

are repeated over time and is a rationale for keeping questions the same over the entire course of the survey. However, keeping all questions unchanged over waves brings a risk for introducing specification error. There are two reasons why this can happen. First, concepts of interest may change over time (e.g. income sources, Internet use, attitudes towards gender roles), and not changing the questionnaire risks making the data irrelevant. Second, researchers may learn of specification errors in the initial questions. If sufficiently serious, it may be better to modify the questions, even if this means that estimates of change in a particular domain cannot then be produced for the first couple of waves. The more waves of the survey that remain, and the longer the time period that these will cover, the stronger the case for modifying questions to reduce specification error. In short-term rotating panels, it may not make sense to modify questions between waves of a panel. Rather, the modifications could be implemented starting from the first wave in the next panel.

In longitudinal surveys there is therefore a large onus on the design of the initial questionnaire: researchers need to be aware that they are not just designing a questionnaire for wave 1, but one that will continue to measure change appropriately after many waves. The implications of this should be understood and questions should be "future-proofed" as far as possible. The content, and to some extent the design, of the questionnaires for future waves should be developed at the same time as the first-wave questionnaire. It is the collective body of instruments administered across multiple waves that constitutes a well-designed survey, not the instruments for each wave considered separately.

There is also a potential trade-off between specification error and nonresponse error. As outlined earlier, one of the unique features of longitudinal surveys is that the respondent experience of taking part previously is likely to have a major impact on the decision to take part again. Thus, interview content that is interesting, engaging, and salient to respondents

may reduce nonresponse error, whereas the content that is in principle most relevant to the study objectives may not necessarily be so interesting, engaging or salient.

5. Examples of Trade-Offs in Three Longitudinal Surveys

Longitudinal surveys are well-suited to the study of different components of TSE, and possible trade-offs between these components. Because respondents are approached and interviewed repeatedly, there is often a lot of information available about each sample member, and also about the sample at the aggregate level. Here we present three examples from the LISS, SIPP and BHPS of how TSE components and interactions can be studied in detail. One of our later recommendations will be that trade-offs between survey errors require further study, and this section provides examples of how this can be done for longitudinal surveys.

Example 1: Trade-off between coverage, sampling and nonresponse error in LISS

The LISS study has included several additional samples since the start of the panel in 2007. Each additional sample was however recruited slightly differently, with different consequences for coverage, nonresponse and adjustment error. The initial recruitment of the sample was conducted by CentERdata in cooperation with Statistics Netherlands, who provided access to a rich address-based sampling frame from which a simple random sample of individuals was drawn [60]. All respondents were offered a computer and Internet, so that coverage error was minimized.

The recruitment rate, defined [9] as the percentage of eligible individuals becoming a panel member was 49%. 4,722 households were recruited, containing 8,148 individuals. For analytical purposes, the LISS study wanted to keep these two numbers around 5,000 and

8,000 respectively over the course of the panel. Due to attrition, additional samples were therefore added in 2009, 2011 and 2013.

The top-up sample of 2009 oversampled groups who had a lower propensity to become a panel member during the initial recruitment: elderly respondents, single-person households, and persons with a non-Western ethnic background [61]. For these variables, Table 1 shows that the additional sample of 2009 indeed successfully over-represented these groups. As a result, the 2011 panel composition is similar to that of the population, with respect to these variables, and coverage error reduced. At the same time, it however becomes hard to determine the selection probabilities of every sample member, , as one needs to take into account whether members of the 2009 additional sample were also eligible in 2007. Because of this, it becomes hard to compute design weights, and the variance of the weights is likely to increase. Further, the 2009 top-up sample is likely to include some new population members (i.e. immigrants who entered the population between 2007 and 2009), but as the top-up sample only represents about 15% of the total sample size, new immigrants will still be underrepresented in the panel.

Also, the disproportionate sampling of groups with low initial response propensities leads to larger variance in design weights and hence larger sampling error. Finally, by focusing on bias in the phase of the recruitment of the panel, any biases due to attrition between 2007 and 2009 were ignored, so the trade-off of sampling error for non-response error may not have been optimal.

In 2011, the additional sample followed proportionate sampling procedures very similar to 2007, and in 2013, a disproportionate stratified additional sample was used, similar to 2009. As a result of the additional samples, nonresponse error in the three variables which produced the largest errors in the recruitment in 2007, were reduced [61]. This reduction in nonresponse error comes at a price. Due to the focus on reducing nonresponse biases compared to the initial (2007) target population, coverage biases will increase when the LISS is used to study a cross-sectional population after 2008. Second, sampling error will

increase because of the disproportionate stratified design used in two of the additional samples. Design weights for LISS are not available, so errors due to undercoverage and sampling are not easily studied.

TABLE 14.1 Nonresponse and coverage error over the course of the LISS panel

Respondent percentages	2007 initial sample	2009 additional sample	2011 additional sample	2013 additional sample	Panel composition January 2011	Panel composition January 2013	Panel composition January 2015	Population (2008)
Aged over 65	10.3	29.3	13.1	9.9	13.0	16.8	17.6	15.3
Household composition								
- single person	12.7	24.0	9.6	19.2	13.8	14.5	17.8	14.9
- couple without children	31.7	38.3	27.5	18.0	27.6	28.6	26.4	27.0
- couple with children	49.4	31.7	54.5	50.2	50.9	48.9	46.1	50.4
- single with children	5.6	4.1	6.7	7.9	6.7	6.7	7.3	6.7
- other	0.6	1.9	1.7	4.8	1.0	1.3	2.4	1.0
Ethnicity								
- Non-western immigrants	3.2	9.3	2.5	7.2	4.0	3.4	5.1	6.3
- Western immigrants	2.0	7.0	2.5	6.2	3.2	2.9	3.8	5.0
Sample size	8148	1229	1678	2059	9166	8326	9319	

Notes: panel data statistics are available from www.lissdata.nl, population statistics from statline.cbs.nl [65].

Example 2: Trade-Off Between Nonresponse and Measurement Error in BHPS

'Difficult' to recruit respondents have been found to report with more measurement error than 'easy' respondents [12, 34]. Similar to this, respondents in longitudinal surveys at risk of dropping out may report with more measurement error. Investing resources to keep

such respondents in the panel may result in greater measurement error in estimates, but reduced non-response error. Loosveldt, Pickery and Billiet [41] for example, showed that item nonresponse in the first wave of a panel study is predictive of unit nonresponse in the second wave of the study. Zabel [78] uses item nonresponse in earlier waves as a covariate in a multivariate analysis of attrition, and finds a significant effect, as do Yan and Curtin [77].

In Figure 1 below we show for each reason for attrition (noncontact, refusal, becoming ineligible, or other (health) reasons) whether they are associated with higher levels of non-substantive responses to items (“don’t knows” and “refusals”) in the 5 waves before a respondent drops out from the BHPS. We contrast respondents who drop out with respondents always interviewed (no attrition). We find more item refusals to questions in the waves before attrition, but only for respondents who will later refuse to participate. This shows that refusals to give an answer to an income question may signal later dropout from the survey altogether.

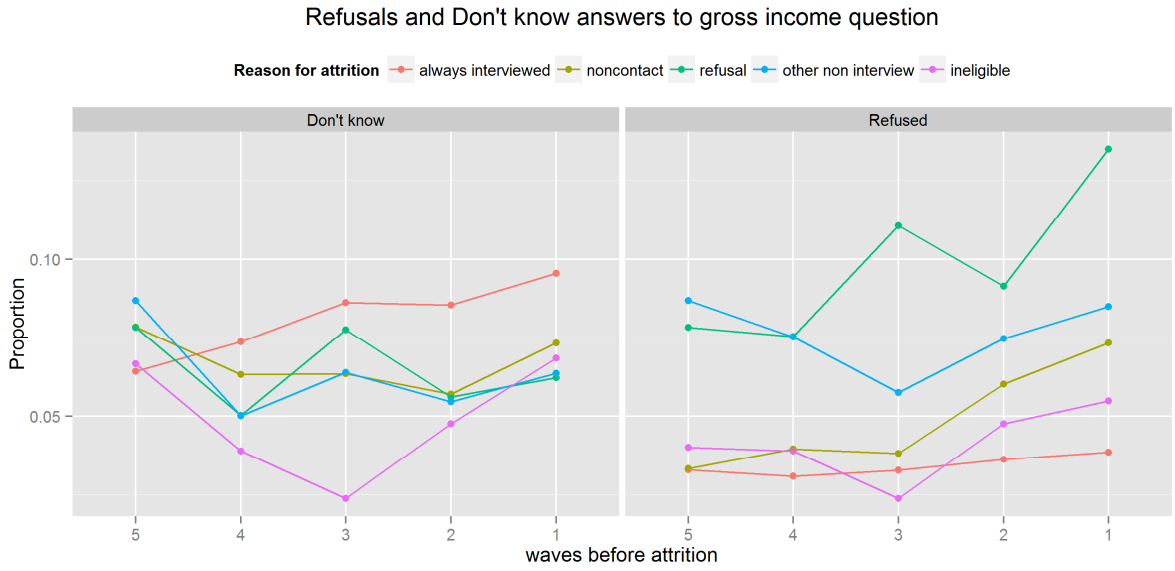

FIGURE 1. Proportion of " don't know" and "refusal" answers to personal gross income question in last five waves before attrition in BHPS

Example 3: Trade-off between specification and measurement error in SIPP

The SIPP has gone through several phases of redesign in its history [14]. The SIPP redesign of 1996 changed the sample size and number of months respondents were followed from 36 to 48 months. SIPP also largely moved from Paper and Pencil Interviewing (PAPI) to Computer Assisted Personal Interviewing (CAPI). The mode change also led to a general redesign of the questionnaire with the goal to reduce measurement and specification error [37].

The change from PAPI to CAPI enabled the use of both extensive routing and longitudinal edit checks in the questionnaire. Routing allowed for more detailed questions and the edit checks were used to achieve more consistent longitudinal survey estimates on for example employment status.

In order to evaluate the effects on data quality and survey estimates, the introduction of the CAPI was carried out experimentally. Half the survey respondents received the old PAPI questionnaire, while half received the new CAPI questionnaire.

Table 2 shows that the CAPI instrument in the employment section is more elaborate, and also takes a different approach to reconstructing weekly employment status in the reference period. In the PAPI instrument, the full employment history is constructed, before details about each job are collected. In the CAPI instrument, employment status is determined first at the start of the reference period, along with details on the job, after which it is determined whether a respondent has been a worker in the reference period. If so, details about each job, as well as the current employment status of the respondent are asked. By asking more questions, and putting the questions in a temporal order, the CAPI instrument aims to improve on the old PAPI instrument in two ways:

1. It aims to reduce measurement error by clearly bounding the reference period, and asking respondents specifically about details of each employer, instead of for example asking the sum of all earnings with all employers.

2. It aims to reduce measurement error by facilitating recall for respondents. This should it reduce underreports, which were found to be problematic in earlier rounds of SIPP (Census Bureau 1998).

The CAPI instrument can potentially reduce measurement error. However, there is a risk it can also *increase* measurement error. Respondents who are self-employed, work irregularly or do informal work may have trouble recalling details, such as start and end dates of all their jobs in the CAPI instrument

TABLE 2 Flow through employment section of the 1996 SIPP questionnaire

	PAPI instrument	**CAPI instrument**
1.	General employment status – weekly labor status collected using a calendar instrument	Initial general employment status – worker vs. nonworker status in reference period determined
2.	Job/business characteristics – industry and occupation codes, hours worked, and job specific information	Job/business characteristics – employer name, dates of employment, industry, occupation, and job specific information
3.	Earnings – hourly wage rate and monthly earnings	Final general employment status – information on time spent looking or layoff
4.		Current situation – employment status and employer name at data of interview
5.		Job/business earnings – up to 2 jobs and 2 businesses.

Note: adapted from [37].

The results from the 1996 SIPP mode-experiment showed that survey estimates on employment rates were largely similar. The largest difference is in the people reporting to be looking for work or in the process of layoff. This amounts 0.9% of respondents in CAPI

and 3.5% in PAPI on average per month. There are no register data to validate these two statistics, but it appears that in the CAPI instrument, respondents underreport being in the process of layoff and instead report no labour force activity (CAPI 27.8% vs. 25.7% in PAPI) [37]. So, a reduction of specific types of measurement errors in this case was counteracted by an increase in other types of measurement error.

6. Discussion

In this article we have highlighted ways in which longitudinal surveys are distinctive in terms of TSE components and interactions between them. We hope that these insights may give pause for thought to researchers involved in designing and planning longitudinal surveys. We have mentioned in a couple of places that control of these errors often involves a trade-off with survey costs. Thus, to make rational decisions about error control survey researchers would ideally be equipped with knowledge about the value of error reduction. In principle, an error reduction technique is only worth implementing if the value of doing so exceeds the cost. However, examples of studies that have attempted to put a value on survey errors are few and far between. And examples of attempts to develop generalisable principles about the value of survey errors are, to our knowledge, non-existent. We therefore believe that survey design decisions would benefit from the development of methods to estimate the value of error reduction. This is particularly challenging in the longitudinal survey context, as the value of survey estimates is typically realized only as the survey matures and more waves of data become available. However, the volume of well-documented data now available from well-established longitudinal surveys with a track record of research records and impacts should make this task possible.

It is also important to be aware that controlling total survey error often involves a trade-off with other dimensions of survey quality, such as timeliness and relevance. For example,

methods to improve the trade-off between measurement error and non-response error may involve call-backs or extended fieldwork efforts which are likely to delay the completion of fieldwork and consequently also the release of data. In seeking the optimum solution, researchers should also consider that the value of quality dimensions such as timeliness may also be different in the longitudinal survey context. We have argued that survey error sources have distinct characteristics in the case of longitudinal survey and we would suggest that the same is likely to be true also regarding other survey quality dimensions.

We believe that the value and quality of longitudinal surveys could in many cases be improved with better decision-making regarding error trade-offs, and regarding the trade-offs between these and costs and other quality dimensions. However, further research is needed to better understand those trade-offs. A good starting point would be to extend existing research which has examined error trade-offs in cross-sectional estimates to longitudinal estimates such as measures of change or predictors of change. Beyond that, research into the error and quality implications of some of the design features that are unique to longitudinal surveys would be highly valuable. For example, the effect of between-wave interval on non-response error and measurement error in the context of different measurement approaches (e.g. dependent vs. independent interviewing) and different data collection modes (interviewer-administered vs. self-administered) is as yet poorly understood, as are several other error trade-offs mentioned in this chapter. We hope that these methodological issues will receive greater attention in future.

References

1. K. Becker, S.H. Berry, N. Orr, and J. Perlman, “Finding the hard to reach and keeping them engaged in research,” *in: Hard-to-Survey Populations (Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M., Bates, N., Eds.)*, Cambridge University Press, Cambridge (2014) pp. 619-641.
2. R. Berthoud and J. Burton (eds.), *In Praise of Panel Surveys*. ESRC Longitudinal Studies Centre, Institute for Social and Economic Research, Colchester (2008).
3. P. P. Biemer, “Total survey error: design, implementation and evaluation,” *Public Opinion Quarterly*, 74(5), pp. 817-848 (2010).
4. D. Binder, “Longitudinal surveys: why are these surveys different from all other surveys?” *Survey Methodology*, 24, pp. 101-110 (1998).
5. A. Blom, M. Bosnjak, A. Cornilleau, A.s.. Cousteaux, M. Das, S. Douhou, and U. Krieger, “A comparison of four probability-based online and mixed mode panels in Europe,” *Social Science Computer Review* (2015). Published online 31 March 2015.
6. M. Bosnjak, M. Das, and P. Lynn, “Methods for probability-based online and mixed-mode panels: selected recent trends and future perspectives,” *Social Science Computer Review*, (2015). Published online 7 April 2015.
7. N. Buck and S. McFall, “Understanding Society: design overview,” *Longitudinal and Life Course Studies*, 3, pp. 5-17 (2012).
8. M. Callegaro, “Seam effects in longitudinal surveys,” *Journal of Official Statistics*, 24(3), pp. 387-409 (2008).
9. M. Callegaro and C. DiSogra, “Computing response metrics for online panels,” *Public Opinion* Quarterly, 72(5), pp. 1008-1032 (2008).
10. P. Campanelli and C. O’Muircheartaigh, “Interviewers, interviewer continuity, and panel survey nonresponse,” *Quality & Quantity*, 33, pp. 59-76 (1999).
11. P. Campanelli and C. O’Muircheartaigh, “The importance of experimental control in testing the impact of interviewer continuity on panel survey nonresponse,” *Quality & Quantity*, 36, pp. 129-144 (2002).
12. C.F. Cannell and F.J. Fowler, “Comparison of a self-enumerative procedure and a personal interview: A validity study,” *Public Opinion Quarterly*, *27*(2), pp. 250-264 (1963).
13. Census Bureau, “Survey of Income and Program Participation quality profile. 3rd edition,” US Census Bureau: *SIPP working paper 230*, (1998). Retrieved from https://www.census.gov/sipp/workpapr/wp230.pdf

14. C.F. Citro and J.K. Scholz, *Reengineering the Survey of Income and Program Participation*, The National Academies Press, Washington, D.C. (2009).
15. M.P. Couper and M.B. Ofstedal, “Keeping in contact with mobile sample members,” *in: Methodology of Longitudinal Surveys (Lynn, P., Ed.)*, Wiley, Chichester (2009) pp. 183-204.
16. R.E. Davis, M.P. Couper, N.K. Janz, C.H. Caldwell, and K. Resnicow, “Interviewer effects in public health surveys,” *Health Education Research*, 25(1), pp. 14-26 (2010).
17. E.D. de Leeuw, “To mix or not to mix data collection modes in surveys,” *Journal of Official Statistics*, 21(2), pp. 233-255 (2005).
18. D.A. Dillman, “Some consequences of survey mode changes in longitudinal surveys,” *in: Methodology of Longitudinal Surveys (Lynn, P., Ed.)*, Wiley, Chichester (2009) pp. 127-139.
19. W. Dijkstra, “How interviewer variance can bias the results of research on interviewer effects,” *Quality and Quantity*, 17, pp. 179-87 (1983).
20. J. Dykema, J.M. Lepkowski, and S. Blixt, “The effect of interviewer and respondent behavior on data quality: analysis of interaction coding in a validation study,” *in: Survey measurement and process quality (Lyberg, L., 3, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. et al., Eds.)*, Wiley, New York (1997) pp. 287-310.
21. S. Eckman, F. Kreuter, A. Kirchner, A. Jäckle, R. Tourangeau, and S. Presser, “Assessing the mechanisms of misreporting to filter questions in surveys,” *Public Opinion Quarterly*, 78, pp. 721-733 (2014).
22. J. Edgar, D.V. Nelson, L. Paszkiewicz, and A. Safir, *The Gemini Project to redesign the Consumer Expenditure Survey: redesign proposal*, Bureau of Labor Statistics, Washington D.C. (2013). http://www.bls.gov/cex/geminiproject.htm
23. Eurostat, “Labour Force Survey in the EU, candidate and EFTA countries: main characteristics of national surveys 2011 (2012 edition),” *Methodologies and Working Papers Series*, Eurostat, Luxembourg (2012).
24. Eurostat, “Quality report of the European Union Labour Force Survey 2013 (2014 edition),” *Methodologies and Working Papers Series*, Eurostat, Luxembourg (2014).
25. L. Fumagalli, H. Laurie, and P. Lynn, “Experiments with methods to reduce attrition in longitudinal surveys,” *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 176(2), pp. 499-519 (2013).
26. T. Glasner and W. van der Vaart, “Applications of calendar instruments in social surveys: a review,” *Quality & Quantity,* 43, pp. 333-349 (2009).

27. R. M. Groves and L. Lyberg, “Total survey error: past, present and future,” *Public Opinion Quarterly*, 74(5), pp. 849-879 (2010)
28. M. Iacovou and P. Lynn, “Design and implementation issues to improve the research value of the longitudinal component of EU-SILC,” *in: Monitoring Social Europe (Atkinson, A., Guio, A.-C. and Marlier, E., Eds.)*, EU Publications (2016) chapter 27.
29. A. Jäckle, “Dependent interviewing: A framework and application to current research,” *in: Methodology of Longitudinal Surveys (Lynn, P., Ed.)*, Wiley, Chichester (2009) pp. 93-112.
30. A. Jäckle and P. Lynn, “Dependent interviewing and seam effects in work history data,” *Journal of Official Statistics*, 23, pp. 529-552 (2007).
31. A. Jäckle, P. Lynn, and J. Burton, “Going online with a face-to-face household panel: effects of a mixed mode design on item and unit non-response,” *Survey Research Methods*, 9(1), pp. 57-70 (2015).
32. G. Kalton and C. F. Citro, “Panel surveys: adding the fourth dimension,” *Survey Methodology*, 19, pp. 205-215 (1993).
33. O. Kaminska and P. Lynn, “Taking into account unknown eligibility in nonresponse correction,” *Paper presented to the International Workshop on Household Survey Nonresponse, Ottawa* (2012).
34. O. Kaminska, A. McCutcheon, and J. Billiet, “Satisficing among reluctant respondents in a cross-national context,” *Public Opinion Quarterly*, 74(5), pp. 956-984 (2010).
35. T. Klausch, J. Hox, and B. Schouten, “Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions,” *Sociological Methods and Research, 42,* pp. 227-263 (2013).
36. F. Kreuter, S. Presser, and R. Tourangeau, “Social desirability in CATI, IVR, and Web surveys. The effect of mode and question sensitivity,” *Public Opinion Quarterly*, 72(5), pp. 847-865 (2008).
37. E. Lamas, T. Palumbo, and J. Eargle, “The effect of the SIPP redesign on employment and earnings data,” *US Census Bureau: SIPP Working Paper* 96-06, (1996). http://www.census.gov/sipp/workpapr/wp9606.pdf
38. H. Laurie, “From PAPI to CAPI: consequences for data quality on the British Household Panel Study,” *ISER Working Paper 2003-14*, Institute for Social and Economic Research, Colchester (2003).
39. H. Laurie and P. Lynn, “The use of respondent incentives in longitudinal surveys,” *in: Methodology of Longitudinal Surveys (Lynn, P., Ed.)*, Wiley, Chichester (2009) pp. 205-233.

40. P. Lavallée, *Indirect Sampling*. Springer, Berlin (2007).
41. G. Loosveldt, J. Pickery, and J. Billiet, “Item nonresponse as a predictor of unit nonresponse in a panel survey,” *Journal of Official Statistics*, 18, pp. 545-557 (2002).
42. P. Lugtig, “Panel attrition: separating stayers, fast attriters, gradual attriters and lurkers,” *Sociological Methods and Research*, 43, pp. 699-723 (2014).
43. P. Lugtig, T. Glasner, and A.J. Boevé, “Reducing underreports of behaviors in retrospective surveys: the effects of three different strategies,” *International Journal of Public Opinion Research,* (2015). published online 2 September 2015
44. P. Lugtig, and A. Jäckle, “Can I just check...? Effects of edit check questions on measurement error and survey estimates,” *Journal of Official Statistics*, 30(1), pp. 1-19 (2014).
45. P. Lynn, “Methods for longitudinal surveys,” *in: Methodology of Longitudinal Surveys (Lynn, P., Ed.)*, Wiley, Chichester (2009) pp. 1-19.
46. P. Lynn, “Maintaining cross-sectional representativeness in a longitudinal general population survey,” *ISER Working Paper 2011-04,* Institute for Social and Economic Research, Colchester (2011).
47. P. Lynn, “Alternative sequential mixed mode designs: effects on attrition rates, attrition bias and costs,” *Journal of Survey Statistics and Methodology*, 1(2), pp. 183-205 (2013).
48. P. Lynn, “Longer interviews may not affect subsequent survey participation propensity,” *Public Opinion Quarterly,* 78(2), pp. 500-509 (2014).
49. P. Lynn, “Design of the immigrant and ethnic minority boost sample,” *Understanding Society Working Paper*, Institute for Social and Economic Research, Colchester (forthcoming).
50. P. Lynn and O. Kaminska, “Criteria for developing non-response weight adjustments for secondary users of complex longitudinal surveys,” *Paper presented to the International Workshop on Household Survey Nonresponse, Nürnberg* (2010).
51. P. Lynn and O. Kaminska, “Combining refreshment or boost samples with an existing panel sample: challenges and solutions,” *Paper presented to the International Workshop on Panel Survey Methods, Melbourne* (2012).
52. P. Lynn and E. Sala, “Measuring change in employment characteristics: the effects of dependent interviewing,” International Journal of Public Opinion Research, 18(4), pp. 500-509 (2006).

53. P. Lynn, A. Jäckle, S.P. Jenkins, and E. Sala, “The effects of dependent interviewing on responses to questions on income sources,” Journal of Official Statistics, 22(3), pp. 357-384 (2006).
54. P. Lynn, A. Jäckle, S.P. Jenkins, and E. Sala, “The impact of questioning method on measurement error in panel survey measures of benefit receipt: evidence from a validation study,” *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 175(1), pp. 289-308 (2012).
55. P. Lynn, O. Kaminska, and H. Goldstein, “Panel attrition: how important is it to keep the same interviewer?” *Journal of Official Statistics*, 30, pp. 443-457 (2014).
56. K. McGonagle, M.P. Couper, and R. Schoeni, “Keeping track of panel members: an experimental test of a between-wave contact strategy,” *Journal of Official Statistics*, 27, pp. 319-338 (2011).
57. J. Neter and J. Waksberg, “A study of response errors in expenditures data from household interview,” *Journal of the American Statistical Association*, 59, pp. 18-55 (1964).
58. H. Sadig. “Unknown Eligibility whilst Weighting for Non-response: The Puzzle of who has Died and who is still Alive?” ISER working Paper series 2014-35. Institute for Social and Economic Research, Colchester (2014).
59. E. Sala, S.C.N. Uhrig, and P. Lynn, P. “It is time computers do clever things!: the impact of dependent interviewing on interviewer burden” *Field Methods*, 23(1), pp. 3-23 (2011).
60. A.C. Scherpenzeel, “Data collection in a probability-based Internet Panel: How the LISS panel was built and how it can be used,” *Bulletin of Sociological Methodology*, 109(1), pp. 56-61 (2011).
61. A.C. Scherpenzeel, “Survey participation in a probability-based online panel in the Netherlands,” *in: Improving Survey Methods: Lessons from Recent Research (Engels, U., Jann, B., Lynn, P, Scherpenzeel, AC., and Sturgis, P., Eds.)*, Taylor & Francis, Boca Raton (2014).
62. A.C. Scherpenzeel and M. Das, “True longitudinal and probability-based internet panels: evidence from the Netherlands,” *in: Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies (Das, M., P. Ester, and L. Kaczmirek, Eds.),* Taylor & Francis, Boca Raton (2010) pp. 77-104.
63. M. Schonlau, N. Watson, and M. Kroh, “Household panel surveys: how much do following rules affect sample size?” *Survey Research Methods*, 5(2), 53-61 (2011).

64. T. Smith, “Refining the total survey error perspective,” *International Journal of Public Opinion Research*, 23, pp. 464-484 (2011).
65. Statistics Netherlands, Statline database (2008). Retrieved August 4, 2015, from: http://statline.cbs.nl/statweb/?LA=en.
66. P.R. Stopher, K. Kockelman, S.P. Greaves, and E. Clifford, “Reducing burden and sample sizes in multiday household travel surveys,” *Transportation Research Record*, 2064, pp. 12-18 (2008).
67. R. Tourangeau, *Recurring Surveys: Issues and Opportunities. A Report to the National Science Foundation*, National Science Foundation, Arlington VA (2003).
68. R. Tourangeau and T. Yan, Sensitive questions in surveys. *Psychological Bulletin*, 133(5), pp. 859-883 (2007).
69. S.C.N. Uhrig, “Understanding panel conditioning: an examination of social desirability bias in self-reported height and weight in panel surveys using experimental data,” *Longitudinal and Life Course Studies*, 3(1), 120-136 (2012).
70. US Census Bureau, *Survey of Income and Program Participation users’ Guide. Revised third edition,* U.S. Census Bureau, Washington (2009).
71. S. van Buuren, *Flexible imputation of missing data*, Chapman & Hall/CRC, Boca Raton (2012).
72. G. Wagner, J. Frick, and J. Schupp, “The German Socio-Economic Panel Study (SOEP) – Scope, evolution and enhancements,” *Schmollers Jahrbuch*, 127(1), pp. 139-169 (2007).
73. J.R. Warren and A. Halpern-Manners, “Panel conditioning effects in longitudinal social science surveys,” *Sociological Methods and Research*, 41, pp. 491-534 (2012).
74. H.R. Weisberg, “The total survey error approach. A guide to the new science of survey research. University of Chicago Press, Chicago (2009)
75. M. Wood, S. Kunz, “CAWI in a mixed mode longitudinal design,” *Understanding Society Working paper series 2014-07*, Institute for Social and Economic Research, Colchester (2014). https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2014-07.pdf
76. M. Wooden and N. Watson, “The HILDA survey and its contribution to economic and social research (so far),” *The Economic Record* 83, pp. 208-231 (2007).
77. T. Yan and R. Curtin, “The relation between unit nonresponse and item nonresponse: A response continuum perspective,” *International Journal of Public Opinion Research, 22*(4), pp. 535 (2010).

78. J. E. Zabel, “An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor market behavior,” *Journal of Human Resources,* pp. 479-506 (1998).