WORKING PAPER 3/2025 (STATISTICS)

Estimation with probability edited survey data under nonresponse

Maiki Ilves

ISSN 1403-0586

Örebro University School of Business
SE-701 82 Örebro, Sweden

Estimation with probability edited survey data under nonresponse

Maiki Ilves*

Abstract

Probabilistic editing has been introduced to enable valid inference using established survey sampling theory in situations when some of the collected data points may have measurement errors and are therefore submitted to an editing process. To reduce the editing effort and avoid over-editing, in current practice selective editing is most often used, which is a form of editing that limits the edit checks to those potential errors that, if indeed in error, are likely to have the biggest impact on estimates to be produced. However, selective editing is not grounded in probability theory associated with survey sampling, and cannot provide expressions for point and variance estimates that account for the uncertainties introduced by selective editing.

In the spirit of the total survey error paradigm, this paper extends the previous work on probabilistic editing by proposing an estimation procedure that provides valid inference when two kinds of nonsampling error are simultaneously present, in addition to the sampling error: the measurement error, requiring an editing step, and the practically unavoidable nonresponse error which also needs to be taken into account when producing unbiased estimates.

In a three-phase selection setup, bias due to measurement error is estimated through probabilistic editing while weight adjustment employing auxiliary information is used to deal with nonresponse. An estimator based on calibration for nonresponse and corrected for bias due to measurement error is introduced. Its theoretical variance and an estimator of the variance are derived. A simulation study illustrates the three-phase selection setup and the practical performance of the derived point and variance estimators.

Keywords: nonsampling errors, probabilistic editing, selective editing, calibration estimator, measurement bias estimation

JEL Classification: C13

*Örebro University School of Business, Sweden, maiki.ilves@gmail.com

1 Introduction

Probability sampling theory enables valid inference from a sample to the population that it was selected from by providing correct expressions to account for the sampling error - namely for the fact that just a subset of the population has been observed and measured.

To develop the theory of sampling, its pioneers started from abstract, idealised situations where all the data that theoretically needed to exist did exist, and correctness of the data was not in question: all the values were present and perfectly measured. While such an idealisation was needed in order to establish the field, in reality sampling frames are imperfect, not all units are responding and the responses given may not be correct. Nonresponse, frame errors and measurement errors belong to the group of survey error types jointly called nonsampling error (Groves, 1989;, Lessler and Kalsbeek, 1992; Biemer and Lyberg, 2003). Nonsampling errors can occur not only in survey data, but also in register data, administrative files, as well as in censuses. They are problematic because they arise outside of control of the survey statistician, their effects on the estimates and on the precision of the estimates are usually unknown, and it is difficult to evaluate the effects even when resources have been allocated for that purpose.

While insights about needing a more encompassing view originated much earlier (Platek and Särndal, 2001, section 8, give references starting from the 1950s), a concerted effort towards trying to build both theory and practice around embracing all error sources simultaneously started to happen early in the new millennium. The resulting total survey error (TSE) paradigm seeks optimal survey design by minimizing TSE subject to constraints (Biemer, 2010). Formally, TSE is composed of a bias and a variance component for each of the identified error sources, and their covariances. While regular workshops and occasional edited volumes and journal special issues mark clear steps in the outlined direction, the applications are still encompassing only a few error sources at the most, and measurement of the TSE is challenging. However, the insight of the interdependence of the error sources, and therefore of the need to treat them simultaneously, is important. In this paper, therefore, an earlier development of partial editing is here extended to include treatment of nonresponse as well.

Editing is an important process in statistics production. Its purpose is to address measurement errors but it also some errors occurring in data processing (Biemer and Lyberg, 2003). Granquist and Kovar (1997, referring to an earlier work of Granquist) define its goals as "to provide information about the quality of the data, to provide the basis for the (future) improvement of the survey vehicle, and to tidy up the data". It is also a rather resource-

demanding and costly operation, estimated to consume between 20% and 40% of the total survey budget (*ibid.*) and therefore producers of statistics have put efforts into finding ways to reduce the work on editing (partial editing) and to make it automatic (automatic editing) (de Waal et al., 2011, chapters 3-6).

An alternative description of editing, "[w]here data are considered incorrect, missing, unreliable or outdated, new values may be inserted or outdated data may be removed[...]" (UNECE, 2019), indicates that there is considerable liberty in choosing if, and what, to remove and what to insert. Randomness of this process is ordinarily not taken into account in point estimation and variance estimation. In other words, the inference - after editing - ordinarily continues as if all the resulting data, so in particular the edited data, have been provided in the original data collection stage (that is, as if no editing has been performed). But, in reality this doesn't hold as selection for editing introduces an additional source of variance, just like sampling, measurement, and so on, do. Therefore, in particular variance estimates of the point estimates not taking this source of error into account are likely downwards biased.

Probabilistic editing (Ilves and Laitila, 2009; Ilves, 2012) was developed to provide a probabilistic framework for inference on data that involve an editing step. While keeping the partial editing idea of selective editing, probabilistic editing expands the former by taking a probability sample from the observations, with or without a selective editing step performed prior to this (the cases C_1 and C_3 in Ilves and Laitila, 2009). In either case, probabilistic editing provides foundation for unbiased estimation, enabling expression of the total uncertainty (sampling and editing) in the point estimate, something that selective editing lacks.

In this paper an estimator is introduced that incorporates yet another error source - the nonresponse - into the previously developed probabilistic editing approach. It does so by combining weights adjustment for nonresponse and the probabilistic selection for editing in order to enable valid inference when both nonresponse and measurement errors are present in a sample survey. The main contribution of the paper is the derivation of the point estimator and its variance, as well as of the corresponding variance estimator, for the situation when both of these nonsampling errors need to be dealt with.

Section 2 gives an overview of literature regarding treatment of nonresponse, of measurement error, and editing, and of ways to deal with both error sources at the same time. Section 3 introduces a sampling design allowing a description of the sampling of units, the response set generation and of the setup for probabilistic editing, and presents an unbiased point esti-

mator accounting for nonresponse and measurement error. The variance of the estimator and an estimator of the variance are derived in Section 4, and the results from a simulation study are presented in Section 5. The paper concludes with a discussion in Section 6.

2 The research subject in the literature

This literature review section gives a brief summary of main issues and some developments regarding nonresponse error and measurement error. This is not, in either case, a comprehensive treatment of these areas, which are well covered in other sources; rather, it is a selective outline that primarily focuses on the approaches and techniques that have a bearing on the content developed further on in this paper.

2.1 Nonresponse

Nonresponse refers to the failure to obtain, from some of the sampled units (persons, households, business units), data measuring the variables that the survey aims to produce statistics about. This can be due to noncontact, refusal to participate, or inability to provide a response (Groves et al., 2004, chapter 6). The result can be a unit nonresponse or an item nonresponse (*ibid.*). Occurrence of nonresponse impacts the ability to produce correct estimates, both by potentially impacting the correct level (the point estimate) and by reducing the number of observations on which the variance of the point estimate is based, thus impacting the precision of the estimate.

Efforts to deal with nonresponse fall into two broad categories: reducing and adjusting (de Leeuw et al., 2008a; Lynn, 2008). The former focuses on designing surveys so as to minimise occurrence of nonresponse. This means paying particular attention to construction of the measurement instrument, choice of the data collection mode, training of interviewers, skilled use of incentives, and so on. The latter focuses on choosing best ways to statistically adjust the estimates so as to take account of the nonresponse that has occurred.

There are also some hybrid approaches, developed with the goal to improve the estimation (reduce the bias and variance of the estimates) but achieved through intervention at early stages of a survey, like its design or the data collection. For instance, an early way of taking nonresponse into account in inference, by Hansen and Hurwitz (1946), consisted of designing the sampling to take place in two phases, the latter of which was among the nonrespondents of the first sampling phase. Another, more recent set of

hybrid approaches, referred to as adaptive or responsive designs (Groves and Heeringa, 2006; Schouten et al., 2017; Chun et al., 2018), concentrates on activities conducted at the data collection stage of a survey. There the data collection is guided by the real-time observed nonresponse patterns and indicators of other nonsampling errors, by available existing information about all the sample members (yet surveyed or not), and by process data from the data collection. The aim is to prioritize data collection from the sampled units in such a way that will lead to a least (estimated) bias due to nonresponse and/or other nonsampling errors in the estimates that are to be produced.

Co-occurring with the increase in prominence of the total survey error paradigm, notable is a shift from the earlier focusing on a particular error source and its quality indicator to its impact on the final estimate, through the estimate's total bias and variance (i.e. the total mean squared error). For nonresponse, the shift means refocusing from the nonresponse rate in itself (de Heer, 1999; de Leeuw and de Heer, 2002; Willimack et al., 2002) to bias caused by the nonresponse (Groves, 2006; Groves and Peytcheva, 2008). Increasingly evident became an understanding that the relation between nonresponse and bias is not consistent or strong, and that the strength of the relation varies between the variables in the same survey but also—in meta-analyses—between surveys (Groves and Peytcheva, 2008; Brick and Tourangeau, 2017). Evidence of an additional weak relation comes from a further meta-analysis, by Peytcheva and Groves (2009). It concerns the relation between demographic variables and nonresponse. Because demographic variables may often be known for the whole sample (be it respondents or not) or even the whole population, they have standardly been used for adjusting for nonresponse (Bethlehem, 1988; Särndal and Lundström, 2005) and, later, as one of the data sources to guide data collection in adaptive and responsive designs (Schouten et al., 2009; Lundquist and Särndal, 2013). The prerequisite for their use is that there is a relation between them and the nonresponse. However, based on the meta-analysis, Peytcheva and Groves (2009), concluded that the difference between respondent and nonrespondent means for demographic variables is not predictive of the difference between respondent and nonrespondent means for study variables of the same survey. Schouten and colleagues (Schouten et al., 2009), Brick (2013), and Haziza and Lesage (2016) give distinct examples (own, of others, and based on simulation, respectively) of counterproductive effect of simplistic applications of nonresponse reducing efforts in data collection or during adjustment: the effect of such applications was an increase in bias of the point estimates.

In summary, relation between nonresponse, demographic variables, and bias is far from simple. Therefore, it invites caution, thought, and as good a

knowledge as possible about the relation of the study variable, nonresponse, and demographic and other auxiliary information for the particular survey and for each of its study variables, in order to apply any of these approaches for the purpose of successfully reducing bias due to nonresponse.

2.1.1 Calibration under nonresponse

Calibration (Deville and Särndal, 1992) is a class of point and variance estimators in the presence of auxiliary information (demographic, register, administrative or other) about units in the population. It adjusts the original sampling weights to those closest ones (under a distance function) that, when applied to the auxiliary information of the sample, result in known totals of these auxiliary variables for the population. Any particular instance of application of calibration assumes a specification of the distance function and the auxiliary information. The technique does not require a formal specification of a model of the relationship between the study variables and auxiliary variables, but - under certain specifications - generalised regression estimators (Särndal and Swensson, 1987; Särndal et al., 1992, chapters 6-8) and calibration estimators coincide (Deville and Särndal, 1992). The technique was developed with the assumption of full response.

Applications of calibration to situations with nonresponse followed (Lundström and Särndal, 1999). In the already established tradition (e.g. Särndal and Swensson, 1987), the response set r is viewed as a result of a second selection, referred to as 'the response mechanism' because it is not a formal selection among the sampled units with known inclusion probabilities or under control of the survey statistician. However, with auxiliary information available, the calibration approach can be used to—presented intuitively—first calibrate on the basis of the auxiliary information for the sample s_a and then on the basis of the auxiliary information for the population U. This can be done in several ways, involving one of the two two-steps procedures or a single-step procedure; see Särndal and Lundström (2005, chapter 8) and a caution in for instance Haziza and Lesage (2016). A number of cases in which auxiliary information may be available and used are discussed by Estevao and Särndal (2002).

Here, a brief formal exposition based on that of Deville and Särndal (1992) and Särndal and Lundström (2005) is given.

The finite population U consists of N units, $U = \{1, ..., k, ..., N\}$, where k indexes the kth unit in the population under some fixed ordering. Of interest is a population function of the study variable y, for instance its total, t_y, where

$$t_y = \sum_{k \in U} y_k.$$

From U a sample s_a of size n_a is drawn using a probability mechanism with the known probability $p(s_a)$. The resulting inclusion probabilities $\pi_k = Pr(k \in s_a)$ are by design positive and known for each unit. The reciprocal of π_k is the design weight, d_k, thus $d_k = 1/\pi_k$.

A possible point estimator for a population total is

$$\hat{t}_{y\pi} = \sum_{k \in s_a} d_k y_k,$$

the Horvitz-Thompson (HT) estimator (Särndal et al., 1992).

With an auxiliary variable[1] x known for example for each unit k in the population and positively related to y, the estimation can be improved on the intuition similar to that for the ratio or regression estimator (Cochran, 1977, chapters 6-7). Calibration consists of finding a set of weights w_k that, when applied to x values x_k of the units k in s_a, will reproduce the known population total for x, and be as close as possible to the design weights d_k with respect to a distance function. Two constraints are involved. The first is the calibration equation,

$$\sum_{k \in s_a} w_k x_k = t_x,$$

where t_x can either be calculated on the basis of data available for the whole population, $t_x = \sum_{k \in U} x_k$, or is known from an external source (though some consistency problems may arise in the latter case). The second is a specified distance function between the original and the revised weights, for instance a chi-square (squared Euclidean) type of distance

$$\sum_{k \in s_a} (w_k - d_k)^2 / d_k$$

(other distance functions are possible, see Deville and Särndal, 1992).

Minimisation leads to the calibrated weights

$$w_k = d_k(1 + x_k \lambda) = d_k v_{sk},$$

where λ is the result of minimisation of the distance function over the sample respecting the constraints induced by the calibration equation,

[1]To simplify notation, a typographical distinction is not made between a single x variable and multiple x variables, which would formally be denoted as a matrix (if for the whole population) or as a column vector (if a specific unit k).

$$\lambda = (t_x - \hat{t}_{x\pi})(\sum_{k \in s_a} d_k x_k x_k')^{-1},$$

provided the inverse exists.

From this, the calibration estimator of the population total t_y is

$$\hat{t}_{yW} == \hat{t}_{y\pi} + (t_x - \hat{t}'_{x\pi})(\sum_{k \in s_a} d_k x_k x_k')^{-1} \sum_{k \in s_a} d_k x_k y_k,$$

that is, the expansion estimator corrected by the auxiliary information. This can also be motivated by a regression approach (Deville and Särndal, 1992).

With nonresponse, the study variable y values are collected from the response set r, $r \subset s_a$, of size n_r, but not from its complement, $s_a \setminus r$. Weight adjustment, in the case of nonresponse, proceeds by setting up either implicit or explicit models. The response mechanism is postulated to follow a distribution, $\Theta(r|s_a)$. Simpler assumptions about the distribution include that the unit k will respond with probability $Pr(k \in r|k \in s_a) = \theta_k > 0$, and that the events of the unit $k \in s_a$ responding and the unit $l \in s_a$ responding are mutually independent, $Pr(k, l \in r|k, l \in s_a) = \theta_k \theta_l$. For example, the response homogeneity group model (Särndal et al., 1992, chapter 15) relies on these, as well as some further assumptions.

However, calibration does not rely on explicit models of nonresponse. With nonresponse, the approach is essentially the same as without it, the main difference in that the y values for the estimation come from the response set r, $r \subset s_a$, of size n_r. In this case the known x information, with which calibration equations can be formulated, may be one of three kinds: that which relates to the sample (referred to as InfoS) and that which relates to the population (InfoU), as well as a join of both of them (InfoUS). In what follows, for brevity reasons the presentation is restricted to the InfoU case. Given the wide use of administrative data available to the researchers and statisticians, the use of InfoU approach is very common in practice. The other cases (InfoS and InfoUS) are given an account of in Särndal and Lundström (2005, chapter 8).

The calibration equation here is, similar to the full response case,

$$\sum_{k \in r} w_k x_k = t_x,$$

where t_x again can either be calculated on the basis of data available for the whole population, $t_x = \sum_{k \in U} x_k$, or is known from an external source.

With the squared Euclidean distance, λ is now found to be

$$\lambda_r = (t_x - \sum_{k \in r} d_k x_k)(\sum_{k \in r} d_k x_k x_k')^{-1},$$

provided the inverse exists (generalised inverse can be tried if needed, Särndal and Lundström, 2005, page 64).

Minimisation leads to the calibrated weights

$$w_k = d_k(1 + x_k \lambda_r) = d_k v_k.$$

From this, the calibration point estimator of the population total t_y with nonresponse is

$$\hat{t}_{yr} = \sum_{k \in r} w_k y_k. \tag{1}$$

Note that the calibration weight w_k is not dependent of the study variable y_k and thus is the same for all the study variables of a survey. It is assumed that the auxiliary vector satisfies the unity condition $\mu^T x_k = 1$ for all $k \in U$ where μ is a vector of constants.

As we have seen, the calibration estimator deals with nonresponse bias by adjusting weights through the use of auxiliary information x_k. Thus, while no explicit response model is entertained, the choice of suitable x vector is still important. In order to reduce nonresponse bias, one should include into the auxiliary vector those variables that explain the reciprocal of an assumed 'probability of responding', θ, associated with the response mechanism, and the study variable y (Särndal and Lundström, 2005). Särndal and Lundström (2010) introduce indicators that enable comparison of different sets of auxiliary vectors with the aim to choose the potentially best vector with regard to nonresponse bias.

The calibration estimator given above is a regression estimator. When other distance measures are used, then Deville and Särndal (1992) have shown that under mild conditions the calibration estimator can be approximated by a regression estimator. The regression estimator, and thus also the calibration estimator when applied on the full sample s_a, is a nearly unbiased estimator given the auxiliary information (Särndal and Lundström, 2005, section 4.3). When applied to the response set, the size of the bias cannot be evaluated. However, in the presence of good auxiliary information nearly unbiased estimates can be obtained (Särndal and Lundström, 2005). As the calibration estimator is a nonlinear estimator with respect to the response indicator $I_k = \{1, \text{ if } k \in r; 0 \text{ otherwise}\}$ in λ_r, only approximate variance can be derived.

2.2 Measurement error

An advantage of nonresponse error as the object of study in survey methodology, over measurement error, is the certainty about whether nonresponse has occurred or not. For measurement error, this certainty does not exist. This error type is therefore more difficult to conceptualise, to investigate, and to correct for. This section gives a brief overview of the causes of measurement error and of the approaches to modelling it.

2.2.1 Causes of measurement error

In direct data collections performed by producers of official statistics, information is generally collected from a respondent, who represents the sampled unit (herself, the household, the business, etc). Errors in measurement can occur due to the data collection instrument (wording of the questions and design of the questionnaire), the data collection mode, the interviewer, the respondent, or due to errors in automated or manual data processing after data collection (Groves, 1989; O'Muircheartaigh, 1997, and the volume that it is contained in, Lyberg et. al, 1997; Edwards et al., 2017).

In providing the data, the respondent is generally thought to go through four cognitive steps: question comprehension, data retrieval, judgement or evaluation of the retrieved data with respect to the requested information, and communication of the data (Tourangeau, 1984). These processes are prone to be under influence of the design choices made in designing the data collection instrument (Tourangeau et al., 2000). Further, some of the four steps may be skipped, replaced by less burdensome choices (Krosnick et al., 1996), leading to provision of less accurate data.

Impacts of modes of data collection were studied extensively (e.g. the contributions in Presser et al., 2004). Particularly well studied was influence of interviewer on responses collected, which demonstrated their substantial impact on the data submitted to the data collection instrument (e.g. Olson et al., 2020 and the references contained therein).

Predominantly in business surveys, data retrieval can also include consultation of computerised systems that store relevant information, which presumes competence of the data provider to work with such a system. In these surveys, the response process is also complicated by occasionally involving several respondents, by being influenced by already established practices of responding to the survey, as well as by there existing levels of authority in decisions on participation and on release of information (Willimack et al., 2002; Lorenc, 2007; Bavdaz, 2010). All these aspects can exert influence on, and introduce error into, the data collected by a survey.

2.2.2 Modelling the measurement error

Initial, and still predominant, contribution of methodological studies that demonstrated impacts on measurement of the factors mentioned above consisted of showing that there is a statistically significant difference in the variable of interest collected by (say) data collection mode A (e.g. paper questionnaire) versus mode B (e.g. web questionnaire), or (say) interviewer A versus interviewer B. The question of which of the alternatives gives a value that is the true one adds another level of complexity.

This question of philosophical origin has been scientifically studied within psychology, in particular psychometrics, and from there it has entered survey methodology. The key concepts are observed value, true score and true value. To illustrate this, consider the expression

$$y_i = \mu_i + \epsilon_i, \tag{2}$$

with i denoting a specific object of measurement and y its attribute (so, for instance, the turnover, y, of the business i.) Other characteristics of the measuring process are considered constant, like the measuring instrument, the external conditions of the measurement, and even time of the measurement).

The expression (2) can be interpreted as saying that an observed measure of the unit i consists of two components: a random error term and a *true score* defined as $\mu_i = E(y_i)$. Consequently, $E(\epsilon_i) = 0$ (Biemer, 2011). The process described by (2) can be seen as a hypothetical repeated random drawing from the distribution of ϵ_i under the same conditions resulting in different observed values y_i reflecting the same true score μ_i.

However, in addition to the random error component, there can be a systematic error component, which influences the true score itself and makes it distinct from the *true value* proper. Loosely following Scholtus (2018, chapter 1), we can express the true score as

$$\mu_i = a_i + b_i\tau_i. \tag{3}$$

The a_i is here seen as an intercept bias, that is, a linear shift for respondent i of the response on the scale of measurement; and b_i as a bias proportional to the true value for respondent i. And τ_i is the true value itself. When $a_i = 0$ and $b_i = 1$, the respondent i's true value and their true score on the concept in question are equal.

Joining (2) and (3), an observed value can be expressed as comprised of the mentioned components,

$$y_i = a_i + b\tau_i + \epsilon_i. \tag{4}$$

Thus, the quantity τ_i is the one sought for for the unit i, its true value on the concept of interest. But, as that quantity, as well as μ_i, are per definition not observable, they are referred to as latent variables or, in psychometrics, latent factors, analysed with methods such as latent class analysis or structural equation modelling (Biemer, 2011; Alwin, 2007).

More than one observation is needed to make models such as that in (4)—or indeed most of the models resulting from the just mentioned methods—identifiable. In the given example of (4), as in Scholtus (2018, expression (1.3)), one can assume that the constants a and b are the same in the set of studied observations, thereby reducing the data need for estimating the model, but still more than one observation will be needed to estimate such a model. On the other hand, the model in (4) can be expanded with further parameters representing the contribution of the interviewer or of the data collection mode, thereby further increasing the amount of data to estimate the parameters.

Measurement error models are widely used to evaluate quality of produced statistics (Biemer, 2009) and to provide research insights into the measurement process in surveys (Saris and Gallhofer, 2014). However, as acknowledged by for instance Saris and Revilla (2016) and Scholtus (2018, chapter 1), their use for the purpose of reducing measurement errors in production of official statistics and other similar, routine statistical output are prevented by:

- the relative complexity of setting up these models,
- assuring the needed assumptions are fulfilled,
- the need for additional measurements, beyond those collected for the original purpose of producing statistics (commonly, repeated measures or reinterviews are needed),
- conformance of the results to the intended use (most of the methods not suitable for one-variable applications).

As Biemer (2009, page 281) notes, primary uses of measurement error analyses are: to evaluate or improve data collection methodology choices (e.g. the mode of data collection), to evaluate or improve the questionnaire, and to help understand limitations of the data. A direct use of the analyses for adjusting the statistical output is not among those uses.

Even with a somewhat less demanding set of analysis methods, that rely on existence of a gold standard (Biemer, 2011), adjusting the statistical output is still not sufficiently simple to enable its integration into a statistics

production process. Gold standard is a value obtained using a measurement process assumed to be error-free. For instance, a reinterview by an independent, experienced interviewer, the outcome of the reconciliation interview, or an administrative source assumed to be fully correct (ibid). However, it is worth noting that even such a relatively simple, considering the alternatives, approach does need that additional measurement, the gold standard. To do this collecting represents a further effort, a cost, and an addition to the total processing time. A possible exception to this would be an error free administrative data source, but on the other hand if such data existed then they could have been used for the original purpose of statistics production, without the need of survey data collection (that leads to data with measurement errors) by a survey measurement instrument. Such a data source can be useful for a retroactive evaluation of an already collected data set, but its existence is unrealistic as a source of gold standard for adjusting the measurement error in ongoing production of everyday official and similar statistics.

2.3 Editing

As indicated at the outset of this article, editing is a statistics processing step that addresses the measurement error in a wider sense. Thus it concerns any error that has been introduced in the data, not only in the narrower sense discussed in the preceding subsection but also errors that may occur in the course of, for instance, data recording, data coding, or data processing.

Because measurement error is not as easily recognisable as nonresponse is, efforts have been invested in developing indicators of potential measurement error, usable for producers of statistics. In general these are referred to as edit rules. Edit rules are expressed as conditions on (or limits for) the values of the collected data points of the variable to be edited that, when exceeded, signal a possible error in a data point or an inconsistency among some of the collected data points (see more in Luzi et al., 2007; de Waal et al., 2011; and Scholtus, 2018, for this and other topics of this subsection).

It is reasonable to assume that in any kind of data collection from which statistics are produced, the producer's goal is to reduce and possibly eliminate measurement errors. In particular this applies to producers of official statistics, as their results are publicly available and relied on when making various policy decisions on high government level.

2.3.1 Selective editing

To reduce costs and maintain quality, editing approaches have been introduced to only partially edit the inconsistencies. One is to carry out the

editing process starting from the most significant inconsistencies, and then continue in a descending order of inconsistency significance until a preset level of data quality has been reached.

To indicate influential inconsistencies, a score function is used (Latouche and Berthelot, 1992; Lawrence and McDavitt, 1994), which expresses significance of the data point (significance of the observation) with respect to the estimate to be produced. It is generally composed of an influence component - if the observation is indeed in error, how big influence on the produced estimate it may have; and a probability component - how likely it is that the observation is indeed in error. Again, the reader is referred to de Waal et al. (2011), and Scholtus (2018) for a more detailed exposition.

Gallagher et al. (2015) designed a sampling procedure and an estimator to review the selective editing thresholds in a particular survey. Data of a subsample of businesses that passed the selective editing in the survey (i.e. whose data were not edited) were subsequently traditionally edited, which enabled calculation of bias, which in turn was used to evaluate and possibly change the thresholds used in the survey's selective editing process.

Therefore, while successful at reaching its targeted goal to reduce editing costs, selective editing—if used with traditional estimators—results in estimates with unknown bias due to possible remaining errors; additionally, the variance cannot be estimated correctly due to it lacking the uncertainty introduced by the editing.

In that light, probabilistic editing has been introduced to enable correct inference with partially edited survey data.

2.3.2 Probabilistic editing

Probabilistic editing, introduced by Ilves and Laitila (2009) and developed further in Ilves (2011), combines the idea of selective editing with a probability sample of observations (including those not indicated for selective editing) to be checked and if necessary edited. For these units, if erroneous values have originally been recorded, the true values will be obtained in the editing process. Based on the sampling design and obtained true values, the effect of not editing all inconsistencies and possible erroneous values in the sample can be estimated and this bias estimate can be used to compute unbiased survey estimates and their variances.

So, like selective editing, this approach too does not verify all inconsistencies, but it gives tools for making inference in the probability sampling framework. An important aim of probabilistic editing approach is to minimize the work going into checking up correct units and this can be achieved by applying an efficient sampling design for selecting the units to be verified.

For efficient sampling, inclusion probabilities should be proportional to some measure reflecting the size of error, importance of the unit, importance of the variable etc. One possibility is to use a Poisson sampling as in Ilves (2011) or PoMix sampling as in Ilves and Laitila (2009) with measure being a global score, used in selective editing (Latouche and Berthelot, 1992) or some other auxiliary information related to the size of measurement error.

Laitila et al. (2017) conducted an experiment embedded in an ongoing monthly survey where routinely data were edited using selective editing. The authors evaluated the selective editing approach in two ways. Firstly, they subsampled records below the cutoff threshold for selective editing, edited those records and applied a bias corrected estimator from probabilistic editing approach (Ilves and Laitila, 2009). For the second approach, model-predicted values were calculated for subsampled records and for estimation design weights were applied. Application of selective editing was found successful by the authors. Regarding the remaining bias, estimates were larger, with larger standard errors, in the modelling approach.

Scholtus (2018) outlined an alternative to the standard editing process for economic statistics at Statistics Netherlands. The standard process was presented as a suite of deductive, selective, automatic and macro editing. The alternative would keep the deductive, automatic and macro editing steps but add a probability sample of records as outlined by Ilves and Laitila (2009) to evaluate the quality of the intended output. That process has not been tested, citing possible issues with application of the alternative process to surveys with a large number of variables as one reason.

2.4 Addressing both nonresponse and measurement error

As mentioned in the Introduction, the work presented in this study was set in a Total Survey Error context. Consequently, it aimed to integrate more than one source of error into a study or application (Groves and Lyberg, 2010). A review of literature on studies performed with the similar intent, to jointly treat measurement error and nonresponse, indicated a need to distinguish between retrospective studies and those that are concurrent with production of statistical output.

To clarify the difference, retrospective studies did not use their results to adjust estimates from the survey round on which they were conducted; rather, they aimed either to inform scientific discourse or to improve the survey's future rounds by providing an evaluation of these and possibly other quality components. Concurrent studies were designed so that their estimation of

nonresponse and measurement errors made possible adjustment of the current survey round's estimates. The data and time requirements for estimating measurement error models (Section 2.2.2) may have a direct bearing on this distinction and the fact that studies of the latter type are fewer than those of the former type.

2.4.1 Retrospective studies

In an early study aiming to simultaneously treat nonresponse and measurement error, Jackman (1999) built a model that included both of these error sources, and demonstrated its use by estimating voter turnout. However, some of the parameters that the model required were supplied as aggregate estimates taken from other surveys, which was the auxiliary information in this case.

In another early study, Biemer (2001) decomposed the differences associated with two data collection modes into measurement bias and nonresponse bias components in order to evaluate and compare the quality of survey data between the two modes. To assess the measurement bias, latent class analysis with interview-reinterview data was used. The author stressed importance of including the measurement error estimation and correction into statistics production in order to provide output of better quality.

Using administrative records to evaluate measurement error, Olson (2006) found nonresponse bias and measurement error bias to depend on type of nonresponse (noncontact or refusal) and on the variable studied. In this study, total bias decreased with increasing efforts to gain cooperation of sampled individuals.

Estimating nonresponse error and measurement error using administrative data on sampled units, Kreuter et al. (2010) found that increasing recruitment efforts led to significant reduction in nonresponse bias at the cost of only a minor increase in measurement error. However, they also found that the increased efforts did lead to an increase in total mean squared error.

Sakshaug et al. (2010), having administrative records as a gold standard, investigated the relative contributions of nonresponse and measurement error on total error under multiple modes of data collection and with two levels of sensitivity of the survey questions. They found that measurement error dominated nonresponse error in the case studied, but that this was mediated by sensitivity of questions. Similarly, Sakshaug et al. (2023), using linked administrative data to separate the effects of nonresponse error and measurement error in a panel survey that used a mixed mode data collection, found that mixing modes mostly reduced nonresponse bias without leading to a significant measurement bias, and also improved the total mean squared

error.

Buelens et al. (2012) conducted a complex study aiming to identify and estimate relative mode effect components arising from selection bias and measurement bias. Their design involved a second wave questionnaire, used to disentangling the relative mode effect into selection and measurement effects, but they did not consider it a reinterview, but as means to construct appropriate nonresponse adjustment weights for calibration estimator. Conducting their research on the Dutch Crime Victimization Survey and Labour Force Survey, they found that both the decomposition and the estimation strategy had relatively little impact on estimates.

In a setup similar to that of Buelens (2012), Klausch et al. (2015) evaluated measurement bias and selection bias in survey using sequential mixed-mode design. Again, special reinterview survey provided additional auxiliary data. Their evaluations had either a face-to-face reference survey as the benchmark, or a hybrid benchmark combining the measurements of a web survey with the selection bias of a face-to-face survey. Their conclusions on an optimal design depended on the benchmark used.

2.4.2 Concurrent studies

Calinescu and Schouten (2016) modelled, with separate models, both response propensity and response quality propensity (i.e tendency towards providing data without measurement error) in an adaptive survey design setting, to direct the data collection towards reducing both nonresponse bias and measurement bias. The report doesn't disclose whether the estimation procedure in the survey in which the research has been embedded has changed due to the interventions at the data collection stage or not.

Beyler and Beyler (2017), constructing a model for nonresponse and another for measurement error, demonstrated in a simulation study the importance of adjusting for both of this error types simultaneously in order to obtain unbiased point estimates. Estimating the measurement error model required two measures of the study variable per respondent in order to be identifiable. Any conducted work on estimation of variance of the point estimates was not reported, and the study seemed not to have been conducted within a finite population inference framework.

To estimate a measurement error model, either repeated measures or a gold standard need to exist. In a standard statistics production process these do not exist and require extra efforts and time to produce. This probably explains an apparent absence of studies where nonresponse and measurement error are addressed concurrently within an ongoing statistics production process.

2.5 Synthesis

Current state of art in the reviewed areas allows these conclusions:

- treatment of nonresponse error benefits from beginning early, at the data collection stage, prioritising those cases that—would their data not be collected—are likely to have the largest impact on nonresponse bias; for this, various approaches to case prioritisation (balance, distance, or representativity indicators under adaptive or responsive designs) are available;

- for unbiased estimation in the case of nonresponse, availability of variables that are good predictors of the nonresponse and of variables that are good predictors of the study variable (not necessarily the same as the preceding ones) are essential in reducing the bias; with the use of such variables one can proceed with methods that either explicitly model these relations (model-based and model-assisted methods of inference) or those that demonstrate advantages of using such variables but without formally stating any model (e.g. the calibration approach as reflected in Särndal and Lundström (2005));

- irrespective of which approach to inference is used, evidence has emerged that the procedure that applies a so-called two-step approach—inference first from the response set to the sample and then from the sample to the target population—is to be preferred over one-step approaches;

- while it is likely that an approach to correcting nonresponse bias along the lines suggested in the preceding item will indeed improve the point estimate, there is no guarantee that such an outcome will indeed happen; the methodologist for any particular survey needs to be alert to possible shifts in explanatory power of the used auxiliary variables and verify these occasionally, because otherwise it may happen that the adjustment just increases the variance while not improving the point estimate;

- selective editing is a method that reduces editing efforts and costs as well as reduces measurement error in the data and thereby improves the point estimate; however, by itself—as described in the editing literature—this method does not take into account (a) possible remaining bias after the selective editing, and (b) the extra variance introduced by the editing process, resulting in possibly biased point estimates with underestimated variances;

– impact of nonresponse and measurement error varies between the studies and the data collection modes they use; therefore, including them both into a study's design and estimation provides increased assurance that their effects, if they exist, will be included into estimation of the statistics produced, and thereby improve on the statistics produced without accounting for these two nonsampling errors.

Next, an estimator is introduced that, while keeping to the probabilistic approach to editing, in addition uses calibration for nonresponse adjustment in order to deal with two types of nonsampling error simultaneously.

3 Point estimation for partially edited data under nonresponse

This section introduces an unbiased estimator of the population total in the presence of nonresponse and measurement error.

3.1 Quasi-randomization setup

Oh and Scheuren (1983) framed nonresponse in sample surveys as the outcome of a second phase of sampling. However, this was not a pure design-based framework due to the implicit modelling involved concerning the nonresponse part, and therefore the authors called this a quasi-randomization approach for dealing with nonresponse. In this paper, the approach also contains a third phase of sampling to accommodate the editing of measurement errors. This leads to a three-phase selection design where consecutive phases are realized given the outcome of the previous stage. The setup is described by the following three phases:

1. Randomness caused by random sampling of units from the population (sampling error);

2. Randomness caused by an unknown selection mechanism of the sampled units into the response set (nonresponse error);

3. Randomness caused by subsampling units among respondents in order to edit their responses and evaluate the measurement bias (measurement error).

Note that the randomness mechanisms in the first and third phases are known due to the chosen probability sampling designs, while the randomness

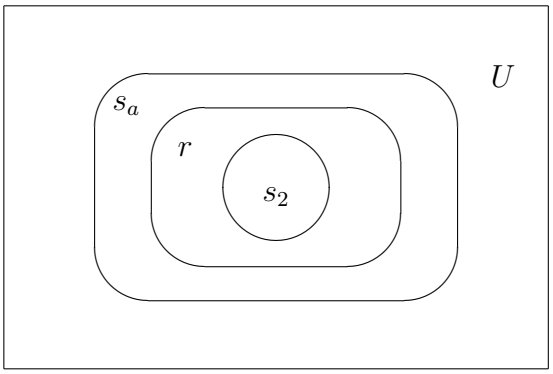

Figure 1: Sampling scheme

mechanism in the second phase is unknown. In order to perform statistical inference, assumptions must be placed on how the response set is obtained given the first phase sample.

Let us consider population $U = \{1, \ldots, N\}$ of size N from which a sample s_a of size n_a is drawn according to some probability sampling design, $p(s_a)$. The first-order and second-order inclusion probabilities are known and denoted by π_k and π_{kl}, with $k, l \in U$, respectively. Drawing a random sample corresponds to the first phase of the three phases.

The second phase of the design makes assumptions about how the response set was obtained. Due to nonresponse, only a subset of the sample is observed, the response set, r, where $r \subseteq s_a$, and the size of the subset is n_r. The unknown response distribution that creates the response set r is denoted by $\Theta(r|s_a)$.

In the third phase of the design, information is collected about the size of the measurement error and its effect on the variability of the estimates. For this purpose, a subset is drawn from the response set, denoted by s_2, of size n_2 according to some sampling design $p(s_2|r)$. The first-order and second-order inclusion probabilities of the design $p(s_2|r)$ are denoted by $\pi_{k|r}$ and $\pi_{kl|r}$, where $k, l \in r$, respectively. The three phases of sampling are illustrated in Figure 1.

Let us denote by $d_k = 1/\pi_k$ the design weight assigned to the unit k in drawing the first phase sample, by v_k the weight adjustment due to observing study variable values in the set r and not the whole sample s_a, and by w_{2k} the weight assigned for subsampling of units of the response set r into the third phase sample s_2. Weights d_k, v_k and w_{2k} are, for phases one and three, functions of their corresponding sampling designs, and for phase two functions of the unknown nonresponse-generating mechanism. In all the

cases, the weights may also be functions of some auxiliary information (if available). The product of the three weights is denoted by w_k,

$$w_k = d_k v_k w_{2k}. \tag{5}$$

3.2 An estimator corrected for nonresponse and measurement biases

The goal is to estimate the population total of the variable of interest z, $t_z = \sum_{k=1}^{N} z_k$. Due to nonresponse, responses are obtained not from the whole intended sample s_a but from a reduced set r, $r \subset s_a$. Further, due to measurement error, the variable observed is y which, for some of the population units, is different from the variable z. After taking a sample, s_2, from the response set r, true values z_k are obtained for the units in s_2. Both z_k and y_k are treated as fixed, nonrandom variables.

Let $\hat{t}_y = \sum_{k \in r} d_k v_k y_k$ be an estimator of the population total of y_k calculated using the observed y values in r. The bias of the estimator due to measurement error – that is, due to observing y instead of z – can be estimated by using the pairs of observations y_k and z_k, $k \in s_2$. Different estimators of that bias can be considered. In the case of the HT-estimator, the bias estimator is given by

$$\hat{t}_{qs_2} = \sum_{k \in s_2} \frac{d_k v_k (y_k - z_k)}{\pi_{k|r}}. \tag{6}$$

where $\pi_{k|r}$ is the first-order inclusion probability of the sampling design $p(s_2|r)$.

And the estimator of the population total of the variable of interest z, corrected for the measurement error bias, is:

$$\hat{t}_z = \sum_{k \in r} d_k v_k y_k - \sum_{k \in s_2} \frac{d_k v_k (y_k - z_k)}{\pi_{k|r}}. \tag{7}$$

The estimator (7) is unbiased with respect to measurement errors as $E(\hat{t}_z|r) = \sum_r d_k v_k z_k$.

Replacing the inclusion probabilities $1/\pi_{k|r}$ by w_{2k}, the estimator (7) can be generalized so that it applies to any linear approximately unbiased estimator used for estimation of the population total and the measurement error bias:

$$\hat{t}_z = \hat{t}_{yr} - \hat{t}_{qs_2} = \sum_{k \in r} d_k v_k y_k - \sum_{k \in s_2} d_k v_k w_{2k} q_k. \tag{8}$$

Here, $\hat{t}_{yr}$ is an estimator of the population total disregarding the effect of measurement errors, $\hat{t}_{qs_2}$ is the estimate of the size of the measurement errors, the subindices r and s_2 indicate the set used for computing the estimates, and $q_k = y_k - z_k, \quad k \in s_2$, denotes the size of the measurement error for unit k and variable z.

It may also be of interest to consider the ratio estimator for estimating the measurement bias. Then, the weight is $w_{2k} = \frac{t_x}{\hat{t}_x}\frac{1}{\pi_{k|r}}$ in (8) yielding a ratio estimator

$$\hat{t}_{qs_2}^{ratio} = \frac{t_x}{\hat{t}_x}\sum_{k\in s_2}\frac{d_k v_k q_k}{\pi_{k|r}}, \tag{9}$$

where $\hat{t}_x = \sum_{k\in s_2} w_k x_k$. Using ratio type estimator will help to reduce the variance of the measurement bias estimator if there exists auxiliary information x roughly proportional to measurement error q_k. The ratio estimator is nearly unbiased given the sample.

The formula given by (8) is simple but computationally not very practical as one needs to have data in the form of two separate sets, response set r and subsample s_2. Alternatively, the estimator (8) can be written as an estimate over the response set only:

$$\hat{t}_z = \sum_{k\in r} d_k v_k y_k^\star, \qquad y_k^\star = y_k - w_{2k} I_{k|r}(y_k - z_k), \tag{10}$$

$$= \sum_{k\in r}(w_k^\star y_k + (d_k v_k - w_k^\star) z_k), \quad w_k^\star = d_k v_k (1 - w_{2k} I_{k|r}), \tag{11}$$

where

$$I_{k|r} = \begin{cases} 1 & k \in s_2 \\ 0 & k \notin s_2 \end{cases}$$

indicates whether unit k is selected into subsample s_2 or not.

The expression given by (10) is of a familiar linear form and shows that the total estimate can be computed as any other weighted total once the transformed variable $y_k^\star$ is formed. In a typical survey, numerous variables are collected and not all of them will be checked for measurement errors. Thus, it is important that the same estimation procedure can be applied to all the variables and the same set of weights can be used when computing different totals.

The expression (11) gives the computational formula when, instead of the transformed variable, original and edited values are used. Note that in this latter case two different sets of weights need to be computed.

4 Variance expressions for the point estimate

The previous section introduced an approximately unbiased estimator that can be used in a practical survey situation where both nonresponse and measurement errors occur and where measurement errors are corrected among a subsample of responding units. In this section, the accuracy of the estimator is studied. For this purpose, expressions of the variance and the variance estimators are derived for the point estimator (8).

4.1 Variance formulae

The variance of $\hat{t}_z$, $V(\hat{t}_z)$, can be expressed as sum of three terms:

$$V(\hat{t}_z) = V(\hat{t}_{yr}) + V(\hat{t}_{qs_2}) - 2Cov(\hat{t}_{yr}, \hat{t}_{qs_2}) \quad (12)$$

the terms being variance of the estimator of the total of the study variable, $\hat{t}_{yr} = \sum_{k \in r} d_k v_k y_k$, variance of the estimator of the size of measurement error, $\hat{t}_{qs_2} = \sum_{k \in s_2} d_k v_k w_{2k} q_k$, and the covariance of the two estimators.

Each term of the (12) can be developed further by taking into account described quasi-randomization setup and applying the laws of iterated expectations. The first term can be expressed as:

$$\begin{aligned} V(\hat{t}_{yr}) &= V_1 E_2(\hat{t}_{yr}|s_a) + E_1 V_2(\hat{t}_{yr}|s_a) \\ &= V_1(\hat{t}_{y_{s_a}} + B(\hat{t}_{yr}|s_a)) + E_1 V_2(\hat{t}_{yr}|s_a) \end{aligned}$$

where $\hat{t}_{ys_a} = \sum_{s_a} d_k y_k$. Here, E and V refer to the expectation and variance of the estimators, respectively, and B refers to the bias of the estimator. Sub-indexes 1 and 2 indicate the randomness mechanism in steps 1 and 2, respectively (see Section 3.1).

Using the assumption of unbiasedness of the estimator $\hat{t}_{yr}$ given the sample, $B(\hat{t}_{yr}|s_a) \approx 0$, an assumption made by for example Särndal and Lundström (2005) for the calibration estimator, then

$$V(\hat{t}_{yr}) = V_1(\hat{t}_{y_{s_a}}) + E_1 V_2(\hat{t}_{yr}|s_a). \quad (13)$$

Similarly approaching the second term in (12), one gets

$$\begin{aligned} V(\hat{t}_{qs_2}) &= V_1 E_{23}(\hat{t}_{qs_2}) + E_1 V_{23}(\hat{t}_{qs_2}) = \\ &= V_1 E_2 E_3(\hat{t}_{qs_2}|s_a, r) + E_1 V_2 E_3(\hat{t}_{qs_2}|s_a, r) + \\ &+ E_1 E_2 V_3(\hat{t}_{qs_2}|s_a, r) \end{aligned}$$

Here, sub-index 3 refers to the randomness mechanism in 3 as explained in Section 3.1).

Again, assuming that such estimator $\hat{t}_{qs_a}$ is used that is unbiased given the previous phases of sampling, $B(\hat{t}_{qr}|r, s_a) \approx 0$ and $B(\hat{t}_{qr}|s_a) \approx 0$, then $V(\hat{t}_{qs_2})$ can be further derived,

$$\begin{aligned} V(\hat{t}_{qs_2}) &= V_1E_2(\hat{t}_{qr}|s_a) + E_1V_2(\hat{t}_{qr}|s_a) + E_1E_2V_3(\hat{t}_{qs_2}|s_a, r) \\ &= V_1(\hat{t}_{qs_a}) + E_1V_2(\hat{t}_{qr}|s_a) + E_1E_2V_3(\hat{t}_{qs_2}|s_a, r) \end{aligned} \quad (14)$$

Using the same technique and assumptions made above, the covariance term in (12) takes the form:

$$\begin{aligned} Cov(\hat{t}_{yr}, \hat{t}_{qs_2}) &= Cov_1(E_2E_3(\hat{t}_{yr}|s_a, r), E_2E_3(\hat{t}_{qs_2}|s_a, r)) \\ &+ E_1Cov_2(E_3(\hat{t}_{yr}|s_a, r), E_3(\hat{t}_{qs_2}|s_a, r)) \\ &+ E_1E_2Cov_3((\hat{t}_{yr}|s_a, r), (\hat{t}_{qs_2}|s_a, r)) \\ &= Cov_1(\hat{t}_{y_{s_a}}, \hat{t}_{q_{s_a}}) + E_1Cov_2(\hat{t}_{yr}|s_a, \hat{t}_{qr}|s_a) + 0 \end{aligned} \quad (15)$$

Here, Cov refers to the covariance of the estimator.

Combining (13), (14), and (15) will give a variance formula:

$$\begin{aligned} V(\hat{t}_z) &= V_1(\hat{t}_{ys_a}) + E_1V_2(\hat{t}_{yr}|s_a) \\ &+ V_1(\hat{t}_{qs_a}) + E_1V_2(\hat{t}_{qr}|s_a) + E_1E_2V_3(\hat{t}_{qs_2}|s_a, r) \\ &- 2Cov_1(\hat{t}_{ys_a}, \hat{t}_{qs_a}) - 2E_1Cov_2(\hat{t}_{yr}|s_a, \hat{t}_{qr}|s_a) \end{aligned} \quad (16)$$

Variance (16) is general and can be applied with any estimators $\hat{t}_{yr}$ and $\hat{t}_{qs_2}$ satisfying the unbiasedness assumptions given above. The four first terms and the two last terms on the right hand side represent the variances of the differences between the two estimators — between the total of the observed study variable y and the total of the measurement error q — calculated on each of the first two phases, $V_1(\hat{t}_{ys_a} - \hat{t}_{qs_a})$ and $E_1V_2((\hat{t}_{yr} - \hat{t}_{qr})|s_a)$, as can be seen by rewriting (16) as

$$\begin{aligned} V(\hat{t}_z) &= V_1(\hat{t}_{ys_a}) + V_1(\hat{t}_{qs_a}) - 2Cov_1(\hat{t}_{ys_a}, \hat{t}_{qs_a}) \\ &+ E_1V_2(\hat{t}_{yr}|s_a) + E_1V_2(\hat{t}_{qr}|s_a) - 2E_1Cov_2(\hat{t}_{yr}|s_a, \hat{t}_{qr}|s_a) \\ &+ E_1E_2V_3(\hat{t}_{qs_2}|s_a, r). \end{aligned} \quad (17)$$

The same estimator, that is, the same weights, are used for both estimators in the same phase. The two sets of terms in (17), excluding the closing term $E_1E_2V_3(\hat{t}_{qs_2}|s_a, r)$, are not dependent on the third phase sampling $p(s_2 \mid s_a, r)$.

Let us consider the calibration estimator (1) in place of $\hat{t}_{yr}$, a Poisson design for the probabilistic editing, and the HT-estimator for estimating the

measurement error bias, $\hat{t}_{qs_2}$. Then the terms in (16) can be approximated as:

$$
\begin{aligned}
V_1(\hat{t}_{y_{s_a}}) &\approx \sum_{k,l\in U}\sum \Delta_{kl} d_k U_k d_l U_l && (18)\\
E_1 V_2(\hat{t}_{yr}|s_a) &\approx E_1\left(\sum_{k,l\in s_a}\sum \Delta_{kl|s_a} d_k v_k U_k d_l v_l U_l\right) && (19)\\
V_1(\hat{t}_{qs_a}) &\approx \sum_{k,l\in U}\sum \Delta_{kl} d_k q_k d_l q_l && (20)\\
E_1 V_2(\hat{t}_{qr}|s_a) &\approx E_1\left(\sum_{k,l\in r}\sum \Delta_{kl|s_a}(d_k v_k q_k)(d_l v_l q_l)\right) && (21)\\
E_1 E_2 V_3(\hat{t}_{qs_2}|s_a, r) &\approx E_1 E_2\left(\sum_{k,l\in r}\sum \Delta_{kl|r}(w_k q_k)(w_l q_l)|s_a\right) && (22)\\
Cov_1(\hat{t}_{y_{s_a}}, \hat{t}_{qs_a})) &\approx \sum_{k,l\in U}\sum \Delta_{kl} d_k U_k d_l q_l && (23)\\
E_1 Cov_2(\hat{t}_{yr}, \hat{t}_{qr}|s_a) &\approx E_1\left(\sum_{k\in s_a}\sum \Delta_{kl|s_a}(d_k v_k U_k)(d_l v_l q_l)\right) && (24)
\end{aligned}
$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, $\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}$, and $\Delta_{kl|r} = \pi_{kl|r} - \pi_{k|r}\pi_{l|r}$, are the covariances of the first, second and third phase sampling indicators, respectively, and $U_k = y_k - x_k^T(\sum_{k\in U} x_k x_k^T)^{-1}(\sum_{k\in U} x_k y_k)$ are regression residuals.

Above, (18) is the variance of a GREG estimator and (20) is the variance of the HT-estimator, and the (23) the covariance between these two estimators. Other terms do not have explicit form as they may depend on the actually realized samples s_a and r. In many situations it is reasonable to assume that the probability of unit responding or the probability that unit has measurement errors is independent of the other units response probability and probability to have measurement errors in the records. This means that $\Delta_{kl|s_a} = 0$ and $\Delta_{kl|r} = 0$ if $k \neq l$ and the variance terms not having explicit form above can be written as single sums.

$$E_1 V_2(\hat{t}_{yr}|s_a) \approx E_1\left(\sum_{k\in s_a}\frac{v_k-1}{{v_k}^2}(d_k v_k U_k)^2\right) \quad (25)$$

$$E_1 V_2(\hat{t}_{qr}|s_a) \approx E_1\left(\sum_{k\in s_a}\frac{v_k-1}{{v_k}^2})(d_k v_k q_k)^2\right) \quad (26)$$

$$E_1 E_2 V_3(\hat{t}_{qs_2}|s_a,r) \approx E_1 E_2\left[\sum_{k\in r}\left(\frac{w_{2k}-1}{w_{2k}^2}\right)(w_k q_k)^2|s_a\right] \quad (27)$$

$$E_1 Cov_2(\hat{t}_{yr},\hat{t}_{qr}|s_a) \approx E_1\left(\sum_{k\in s_a}\frac{v_k-1}{{v_k}^2}(d_k v_k U_k q_k)^2\right) \quad (28)$$

4.2 Variance estimators

Estimation of $V(\hat{t}_z)$ can be based on derivation of expressions for the terms in equation (16) similar to those derived in equations (18)-(24). In that case, the variance estimator proposed by Särndal and Lundström (2005, chapter 11) for the calibration estimator $\hat{t}_{yr}$ is obtained using expansion estimators of (18) and (19) based on the response set r.

Similarly, the terms (20)-(24) can be estimated using expansion estimators based on the third phase sample s_2. For this, define

$$u_k = y_k - x_k^T\left(\sum_{k\in r} d_k v_k x_k x_k^T\right)^{-1}\left(\sum_{k\in r} d_k v_k x_k y_k\right) \quad (29)$$

Under the three phase quasi-randomization sampling scheme (Section 3.1) with simple random sampling in the first phase, some unspecified selection process with independent inclusions in the second phase and Poisson sampling in the third phase, the variance estimate,$\hat{V}(\hat{t}_z)$, has three terms:

$$\hat{V}(\hat{t}_z) = \hat{V}(\hat{t}_{yr}) + \hat{V}(\hat{t}_{qs_2}) - 2\hat{Cov}(\hat{t}_{yr},\hat{t}_{qs_2}), \quad (30)$$

Each term is expressed below.

$$\hat{V}(\hat{t}_{yr}) = \sum_{k,l\in r}\sum(d_k d_l - d_{kl})v_k u_k v_l u_l$$
$$-\sum_{k\in r} d_k(d_k-1)v_k(v_k-1){u_k}^2 + \sum_{k\in r} v_k(v_k-1)(d_k u_k)^2 \quad (31)$$

Inserting SI-design weights for the first phase into (31), we get

$$\hat{V}(\hat{t}_{yr}) = \frac{N^2(1-f)}{n(n-1)}\left[\sum_r (v_k u_k)^2 - \frac{1}{n}(\sum_r v_k u_k)^2\right]$$
$$-\frac{1}{f}\left(\frac{1}{f}-1\right)\sum_r v_k(v_k-1)u_k^2 + \frac{1}{f^2}\sum_r v_k(v_k-1)u_k^2 \tag{32}$$

where $f = n/N$ is the first phase sampling fraction, v_k is the calibration weight, $w_{2k} = 1/\pi_{k|r}$ is the weight in the Poisson sampling and u_k is given by (29).

The second term of (30) can be expressed as

$$\hat{V}(\hat{t}_{yr}) = \sum_{k,l\in s_2}\sum \frac{\Delta_{kl}}{\pi_{kl}\pi_{kl|s_a}\pi_{kl|r}} d_k q_k d_l q_l$$
$$+\sum_{k,l\in s_2}\sum \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}\pi_{kl|r}} d_k v_k q_k d_l v_l q_l + \sum_{k,l\in s_2}\sum \frac{\Delta_{kl|r}}{\pi_{kl|r}} d_k v_k w_2 k q_k d_l v_l w_2 l q_l \tag{33}$$

Applying our specific sampling designs on (33), again with SI design in the first phase, we get

$$\hat{V}(\hat{t}_{qs_2}) = \frac{1-f}{f^2}\sum_{s_2} v_k w_{2k} {q_k}^2$$
$$-\frac{1-f}{f^2(n-1)}\left[(\sum_{s_2} v_k w_{2k} q_k)^2 - \sum_{s_2}(v_k w_{2k} q_k)^2\right]$$
$$+\frac{1}{f^2}\left[\sum_{s_2} v_k^2 w_{2k}(w_{2k}-1){q_k}^2 + \sum_{s_2} v_k(v_k-1)w_{2k}q_k^2\right] \tag{34}$$

And the general form of the estimator of the covariance term is

$$\hat{Cov}(\hat{t}_{yr}, \hat{t}_{qs2}) = \sum_{k\in r}\sum_{l\in s_2} \frac{\Delta_{kl}}{\pi_{kl}\pi_{kl|s_a}\pi_{l|r}} d_k u_k d_l q_l$$
$$+\sum_{k\in r}\sum_{l\in s_2} \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}\pi_{l|r}} d_k v_k u_k d_l v_l q_l \tag{35}$$

which yields the following expression in our specific case

$$\hat{Cov}(\hat{t}_{yr}, \hat{t}_{qs2}) = \frac{1}{f^2} \sum_{s_2} v_k(v_k - f) w_{2k} u_k q_k$$
$$- \frac{1-f}{f^2(n-1)} \sum_{\substack{k \in r, l \in s_2 \\ k \neq l}} \sum (v_k u_k)(v_l w_{2l} q_l) \tag{36}$$

The estimator (31) is obtained by applying variance estimator given in Särndal and Lundström (2005, expressions (11.4), (11.5)) for a special case where only population level auxiliary information is used and an instrument vector $z_k = x_k$. Derivation of the equations (34) and (36) is given in the Appendix.

Now, if the ratio estimator is used instead of the HT-estimator for the bias estimation then approximate variance can be computed using estimated Taylor deviates. Estimated Taylor deviates in the case of the ratio estimator have the following form: $\hat{g}_k = \hat{t}_x^{-1}(y_k - \hat{R}x_k)$, where $\hat{R} = \hat{t}_x^{-1}\hat{t}_y$. The variance of the ratio estimator can be approximated by the variance of the HT-estimator, $\hat{V}(\hat{R}) \approx \hat{V}(\hat{t_g})$, where $\hat{t}_g$ is the HT-estimator of the estimated Taylor deviates. This is a large-sample property, derived in Fuller (2009, pages 58-60).

Applying the notation and framework used in this paper to the general formula, the deviates are: $\hat{g}_k = \hat{t}_x^{-1}(q_k - \hat{R}x_k)$ where $\hat{R} = \hat{t}_x^{-1}\hat{t}_{qs_2}$. Approximating the variance of the ratio estimator by $\hat{V}(t_x\hat{R}) \approx {t_x}^2\hat{V}(\hat{t}_g)$, where $\hat{t}_g = \sum_{s_2} w_k\hat{g}_k$ and w_k is defined by (11), the total variance can be computed as

$$\hat{V}(\hat{t}_z) = \hat{V}(\hat{t}_{yr}) + {t_x}^2\hat{V}(\hat{t}_g) - 2t_x\hat{Cov}(\hat{t}_{yr}, \hat{t}_g), \tag{37}$$

where the terms are given by (32), (34), and (36) by replacing q_k with $\hat{g}_k$. For certain populations ratio estimator can be highly efficient (Cassel et al., 1977, chapter 7)

5 Simulation study

To investigate the performance of developed estimators, I conducted a Monte Carlo study on synthetic data with a known population-generating mechanism. It means that variables were generated from known gamma distributions and the parameters of the distribution were chosen so that the resulting variables were correlated.

A population in this setup was characterised by five variables, conceived to be as follows: a target variable (the variable of interest), z; an observed

variable related to the target variable, y; and three auxiliary variables, x_1, x_2, and x_3. The variable z can be seen as for instance actual turnover of a business, and y as the turnover as reported by the business in a specific statistical survey conducted by a national statistical institute; the $x_i, i \in \{1, 2, 3\}$ were auxiliary variables, in different ways related to z, y, and to each other, as follows. Obtained from a very large population (100 Million units [2]), some numerical properties of these variables, generated as a set of correlated gamma distributions, are presented in Table 1 and Table 2.

Table 1: Univariate properties of the variables z, y, x_1, x_2, and x_3, as well as of z/x_3. Note: Measurement error is not conceived of as a population characteristics, and is therefore introduced at a later stage; therefore, z and y are the same at the time of population generation.

	z	y	x_1	x_2	x_3	z/x_3
Min.	4.50	4.50	3.21	0.73	7.25	0.02
1st Qu.	168.44	168.44	41.69	117.58	150.75	0.76
Median	233.56	233.56	53.68	180.48	202.04	1.16
Mean	250.01	250.01	56.00	201.68	213.58	1.40
3rd Qu.	313.76	313.76	67.79	262.89	263.86	1.75
Max.	1351.38	1351.38	223.15	1469.23	1008.56	45.66

Table 2: Pearson's correlation r between the variables in the population.

	z	y	x_1	x_2	x_3	z/x_3
z	1.000	1.000	.300	.788	.000	.637
y	1.000	1.000	.300	.788	.000	.637
x_1	.300	.300	1.000	.000	.873	-.313
x_2	.788	.788	.000	1.000	.000	.502
x_3	.000	.000	.873	.000	1.000	-.577
z/x_3	.637	.637	-.313	.502	-.577	1.000

For the purpose of the simulation study, each generated population created in the described way consisted of $N = 10\,000$ units.

[2]Population of 100 Million units was used only to illustrate the numerical properties of the population. Later on in the study, smaller population is used.

Subsets from the population, according to the three-phase sampling process, were created as follows:

1. First phase: sample s_a of size $n_a = 1\,000$ was taken from the population using simple random sampling (SRS).

2. Second phase: among the units in s_a, a subset of respondents r of the random size n_r was generated using a proportional to size selection (as Poisson sampling but – for speed – without separately treating the cases of $\pi_k > 1$), with x_1 as the size variable and with 700 units as the targeted size of the subset. This resulted in an actual 'response rate' of around 68%.

3. For simulation purposes, an indicator of measurement error was assigned to the responding units, again using a proportional to size selection (again, as Poisson sampling but without separately treating the cases of $\pi_k > 1$), with z/x_3 as the size variable, conceived roughly as 'turnover per (a correlate of) employee'. Of those, to reduce the number of erroneous records, only every fourth record was randomly picked and set to indicate measurement error. In this way, a measurement error was indicated for around 20% of the response set r units.

4. For each of the units i where a measurement error was indicated, the error's magnitude was determined using a linear function, $e_i = z_i * u_i$, where u_i was assigned from values in the range $(.30, .55)$ using a uniform random number generator. From this, the observed target variable y_i was created, $y_i = z_i + e_i$.

5. Third phase: In order to correct for the measurement error, a subset s_2 of a random size n_2 from the response set was drawn using a Poisson sampling design with x_2 as the size variable[3] and with an expected sample size of 90. The value without measurement error z_i was observed for these units. For the purpose of this application, it was assumed that all the units sampled into s_2 responded.

Summary of the distributions of sizes of the subsets at the three phases, over 10 000 draws of both the population and subsets, is presented in Table 3.

The generated simulation variable values, sampling indicators and weights were recorded in a data table, enabling application of the developed point

[3]The variable x_2 can here be seen as an illustration of a global score, identifying the most important records to be edited. How such a global score would be derived is beyond the scope of this paper, but some relevant references were given in Section 2.3.

Table 3: Achieved sizes of the subgroups, over a set of 10 000 independently created populations and drawn samples.

	n_a	n_r	n_2
Min.	1 000	629.0	60.0
1st Qu.	1 000	671.0	84.0
Median	1 000	680.0	90.0
Mean	1 000	680.2	90.0
3rd Qu.	1 000	689.0	96.0
Max.	1 000	734.0	123.0

estimator (7) and the ratio estimator (9). Variance estimates were computed as given in Section 4.2. For the case of the ratio estimator, the parameters estimated on the level of the sample s_2 are those in equations (9) and (37), based on the quantity $\hat{g}_k$.

In the simulation study the auxiliary information was used in two ways. First, the auxiliary vector in the calibration estimator consisted of two variables known on population level, $x_k = (1, x_{1k}), k \in U$. And the other use of additional information was in the ratio estimator (9) where $\hat{t}_x = \sum_{s_2} d_k v_k w_{2k} x_{3k}$ and $t_x = \sum_U x_{3k}$.

The simulation study was performed as a set of B repetitions ($b = 1, ..., B$) of the procedure above, where the population was held constant but the three-phase selection and the assignment of measurement errors was independently repeated, and the estimators recorded ("inner"repetitions). A set of B repetitions was nested within a set of C repetitions ($c = 1, ..., C$) where a new population with the characteristics presented above was drawn at the start of an outer repetition, independently from the previous drawings ("outer" repetitions).

In each of the b repetitions, the calculated point estimates $\hat{t}_z$, $\hat{t}_{yr}$, and $\hat{t}_{qs_2}$ were recorded, as well as the difference $\hat{t}_z - t_z$ and also the estimated variance $\hat{V}(\hat{t}_z)$ and its components $\hat{V}(\hat{t}_{yr})$, $\hat{V}(\hat{t}_{qs_2})$, and $\hat{Cov}(\hat{t}_{yr}, \hat{t}_{qs_2})$. The empirical, true value of t_z for the population (the same in each of the b "inner" repetitions drawn from the same population in a c "outer" repetition) was also recorded, as well as a coverage indicator set to mark when the absolute difference $|\hat{t}_z - t_z|$ was smaller than $1.96 \times (\hat{V}(\hat{t}_z))^{1/2}$. Also calculated was the relative bias (RB) for the different estimated values of an estimand, θ, defined as $RB = (\hat{\theta} - \theta)/\theta$, expressed in the results table as a percentage (i.e. $\%RB = RB(\hat{\theta}) \times 100$).

After a set of B repetitions has been completed, the empirical means and

Table 4: Simulation results for the two estimators used in the third phase, the HT and ratio estimators, with $B = 1\,000$, $C = 1\,000$: means over the C repetitions.

	HT			Ratio		
Parameter	Estimated	Empirical	$\%RB$	Estimate	Actual	$\%RB$
$\bar{t}_z/10^3$	2 499.7	2 499.8	-0.0	2 498.3	2 499.8	-0.1
$\bar{t}_{yr}/10^3$	2 720.3			2 720.3		
$\bar{t}_{qs_2}/10^3$	220.6			222.0		
$\bar{V}(\hat{t}_z)/10^6$	4 410.2	4 394.9	0.3	4 226.9	4 225.1	0.0
$\bar{V}(\hat{t}_{yr})/10^6$	2 399.8	2 426.0	-1.1	2 399.8	2 426.0	-1.1
$\bar{V}(\hat{t}_{qs_2})/10^6$	3 171.5	3 173.4	-0.1	3 050.8	3 088.2	-1.2
$-2\overline{Cov}(\hat{t}_{yr}, \hat{t}_{qs_2})/10^6$	-1 161.1	-1 204.5	-3.6	-1 223.8	-1 289.1	-5.1
$\bar{I}_{95\%Coverage} \times 10^2$		94.8			94.8	

variances of the produced point estimates were calculated, and thus their empirical biases and MSEs; also calculated were means of the estimated variances, and the means of the relative bias and the coverage indicator. These all were recorded as a row of the C data table. Finally, after completing the set of C repetitions, means of the recorded empirical means and variances over the C repetitions were calculated.

This simulation study was run in two variants, the first with the HT-estimator in the third phase, and the second with the ratio estimator (9). In both cases, the calibration estimator $\hat{t}_{yr}$ was the same. Both were run with the same random generator seed value, meaning that the created populations and samples for them were identical. The results are presented combined in Table 4

Results in Table 4 confirm that the proposed bias corrected estimator is indeed unbiased, with relative bias measure being near zero. Without correcting for the measurement error bias the point estimator would overestimate the total by 8% as all measurement errors were generated upwards per the simulation setup, meaning all corrections were done downwards during editing. One should though note that due to the simulation setup the bias due to the nonreponse was likely to be small or near zero as the same auxiliary information was used in the calibration estimator as was used in the response generating mechanism.

To get some insight about the nonresponse bias, the same simulation setup was carried out by using expansion estimator instead of calibration

Table 5: Simulation results with expansion estimator in the second phase instead of the calibration estimator, and ratio estimator in the third phase, with $B = 1\,000$, $C = 1\,000$: means over the C repetitions.

Parameter	Estimated	Empirical	$\%RB$
$\bar{t}_z/10^3$	2 617.3	2 499.8	4.7
$\bar{t}_{yr}/10^3$	2 821.5		
$\bar{t}_{qs_2}/10^3$	204.2		
$\bar{V}(\hat{t}_z)/10^6$	4 110.6	4 222.1	-2.6
$\bar{V}(\hat{t}_{yr})/10^6$	2 641.1	2 661.7	-0.8
$\bar{V}(\hat{t}_{qs_2})/10^6$	2 611.7	2 708.6	-3.6
$-2\overline{Cov}(\hat{t}_{yr}, \hat{t}_{qs_2})/10^6$	-1 142.2	-1 148.2	-0.5
$\bar{I}_{95\%Coverage} \times 10^2$		52.8	

estimator to deal with nonresponse at the estimation stage. The results in Table 5 show, as expected, biased point estimate and loss in the coverage rate. What is interesting though is that the nonresponse bias of about 5% leads to corresponding coverage rate of 53%. In comparison, unbiased estimate had coverage rate of 95%.

Variance estimates in Table 4 are compared to the empirical variances of the estimators as population level estimates cannot be mathematically explicitly expressed. In addition to the total variance, each term of the total variance is also presented. Results show good fit between estimates and empirical results, the largest difference being in the covariance estimate. Still, the coverage rate is near the designated 95%, indicating that the variance estimate quite precisely estimates the variation of the point estimate.

Both estimators, HT and ratio type estimator, yield nearly unbiased estimates. However, there is slight efficiency gain in variance when using ratio type estimator instead of the HT-estimator.

As we see from the numerical results, it is important to use variance estimates that take into account the variability in all three phases. If we would use only readily available variance estimate of the calibration estimator and ignore the variability stemming from the editing phase, we would considerably underestimate the variance—almost by half—and overestimate the coverage rate.

6 Summary and discussion

In this paper, an estimation procedure was proposed when two kind of non-sampling error are present in a sample survey: nonresponse error and measurement error. Estimation relies on quasi-randomization setup consisting of three phases of random selections, where calibration approach utilizing auxiliary information is applied in the second phase to primarily treat nonresponse error, and probabilistic editing carried out in the last phase addresses measurement error. Unbiased estimator of the total, its variance estimator and corresponding variance estimate were derived and applied on synthetic data in a simulation study.

Proposed approach relies, firstly, on good understanding of the response mechanism and on information available to adjust for nonresponse. Secondly, one needs to carry out editing of all records in a subsample to verify responses and correct them, if necessary. One can increase the efficiency of the editing procedure by using global scores in the third phase of sampling, which more likely highlights suspicious records.

The main attention was on the measurement error and its effect on estimates. The method described in this paper gives a tool to estimate bias due to measurement error and ultimately, correct the total estimate of interest. Simulation results confirm the correctness of the theoretical derivations as both the point estimator of the study variable and its variance have near zero relative bias. Thus, the proposed estimator removes bias due to measurement error and should be preferred, in general, since an uncorrected estimator leaves bias of unknown size and may invalidate the decisions made based on it.

As concerns the nonresponse error, it was assumed to be present but was not explicitly estimated, however it was removed through applying the calibration method. Some insight into the nonresponse bias was offered by comparing expansion estimator and calibration estimator. Simulation results showed that expansion estimator led to the biased estimate of the study variable with relative bias about 5%, the variance estimate of the study variable was relatively similar in both cases, however, coverage rate dropped from 95% to 53% when expansion estimator was used in the estimation setup. Quantifying nonresponse bias is very difficult and thus, in practice, it is rarely possible to check how well the calibration method works. Following the guidelines for how to choose the best set of auxiliary variables (e.g. Särndal and Lundström 2010) should give some confidence that nonresponse bias is reduced and with that the effect on coverage rate.

In this paper, in addition to the HT-estimator also ratio estimator was considered for estimating the size of measurement error because it is known

that the ratio estimator can be more efficient compared to the HT-estimator for sampling designs with variable sample size. In the simulation study, the ratio estimator did not behave very differently from the HT-estimator, most likely due to the rather small sample size used in the third phase. A larger subsample size can increase the relative efficiency of the ratio estimator.

Further, it should be noted that probabilistic editing approach relies on the assumption that all sampled units get true values during the editing process. This is unlikely to be so in practice, but keeping the sample size relatively small the effort put into editing can be manageable. In the simulation study only 9% of the initial sample size was used to estimate the size of the measurement bias. So compromise needs to be found between the effort and resources put into editing versus decreasing the precision of the estimates.

Implementation of the introduced method is not difficult as Poisson sampling is available in standard statistical packages. Computing the bias corrected estimates is not difficult as both terms can be obtained separately while variance estimation needs a bit more effort as the covariance term is not computed by standard software.

In the simulation study, a Poisson design was used in the third phase. Different sampling designs can be implemented in practice. Combination of selective editing, and probabilistic editing can be used where units very likely in error are treated by selective editing and probabilistic editing is carried out on the rest of the units. Also, Bernoulli design can be used when no information is available for indicating the likely errors and it is probably also more efficient in case of systematic errors in data.

Employing probabilistic editing provides information about the size of the measurement bias, however removing measurement bias comes at the price of an increased variance as simulation study indicated. Further research how to balance the size of measurement bias and precision of an estimator can be of interest. For example, the choice of point or variance estimator may be done differently depending on whether the size of measurement bias is considerable or negligible.

The proposed estimator does not take into account frame errors. However, the calibration estimator can be used also to deal with frame errors given that auxiliary information about the actual target population is available. The estimator dealing with nonresponse and frame errors is suggested by Särndal and Lundström (2005). Thus, the calibration estimator and probabilistic editing procedure proposed here have potential to simultaneously address these three nonsampling error types.

References

Alwin D.F. (2007). *Margins of Error*. New York: John Wiley and Sons.

Bavdaz, M. (2010). The multidimensional integral business survey response model. *Survey Methodology*, **36**, 81-93.

Beyler N. and Beyler A. (2017). Adjusting for measurement error and nonresponse in physical activity surveys: a simulation study. *Journal of Official Statistics*, **33**, 533-550.

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

Biemer P.P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, **17**, 295-320.

Biemer, P. (2009). Measurement errors in sample surveys. In D. Pfeffermann and C. R Rao (eds.), *Sample Surveys: Design, Methods and Applications, Vol. 29A*. Amsterdam: Elsevier, pp. 281-315

Biemer, P. (2010). Total survey error: design, implementation and evaluation. *Public Opinion Quarterly*, **74**, 817-848.

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. New York: John Wiley and Sons.

Biemer, P., de Leeuw E., Eckman S., Edwards B., Kreuter F., Lyberg L.E., Tucker N.C., West B.T. (eds.) (2017). *Total Survey Error in Practice*. New York: John Wiley and Sons.

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S. (eds.) (1991). *Measurement Errors in Surveys*. New York: John Wiley and Sons.

Biemer, P. and Lyberg, L. (2003). *Introduction to Survey Quality*. New York: John Wiley and Sons.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: a critical review. *Journal of Official Statistics*, **29**, 329-353.

Brick, J. M. and Tourangeau R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, **33**, 735-752.

Buelens, B., van der Laan, J., Schouten, B., van den Brakel, J., and Klausch, T. (2012). *Disentangling Mode-specific Selection and Measurement Bias in Social Surveys*. Discussion paper No. 201211. The Hague, The Netherlands: Statistics Netherlands. Available at: *https://www.cbs.nl/-/media/imported/documents/2012/32/2012-11-x10-pub.pdf*, accessed January 2024.

Calinescu, M. and Schouten, B. (2016). Adaptive survey designs for nonresponse and measurement error in multi-purpose surveys. *Survey Research Methods*, **10**, 35-47.

Cassel, C. M., Särndal, C.E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley and Sons.

Cochran G.W. (1977). *Sampling Techniques*, 3rd edition. New York: John Wiley and Sons.

Chun A.Y., Heeringa S.G. and Schouten B. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, **34**, 581-597.

de Heer W. (1999). International response trends: results of an international survey. *Journal of Official Statistics*, **15**, 129-142.

de Leeuw, E. D. and de Heer W.(2002). Trends in household survey nonresponse: a longitudinal and international comparison. In R. M. Groves, D.A. Dillman, J. L. Eltinge and R. J. A. Little (eds.), *Survey Nonresponse*. New York:John Wiley and Sons, pp. 41-54.

de Leeuw, E. D., Hox, J. J., and Dillman, D. A. (2008a). The cornerstones of survey research. In E.D. de Leeuw, J.J. Hox and D.A. Dillman (eds.), *International Handbook of Survey Methodology*. UK: Taylor & Francis Group, pp. 1-17.

de Leeuw, E. D., Hox, J. J., and Dillman, D. A. (eds.) (2008b). *International Handbook of Survey Methodology*. UK: Taylor & Francis Group.

de Waal, T., Pannekoek J. and Scholtus S. (2011). *Handbook of Statistical Data Editing and Imputation*. New York: John Wiley and Sons.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Edwards B., Maitland A., and Connor S. (2017). Measurement error in survey operations management: detection, quantification, visualization, and reduction. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West (eds.), *Total Survey Error in Practice*. New York: John Wiley and Sons, pp. 253-277.

Estevao V.M. and Särndal C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, **18**, 233-255.

Fuller, W. A. (2009). *Sampling Statistics*. New York: John Wiley and Sons.

Gallagher S., Graham B. and Gaughan C. (2015). Method for reviewing selective editing thresholds at ONS, RSI pilot study. Paper presented at *UN-ECE Work Session on Statistical Data Editing*, Budapest, Hungary, 14-16 September 2015.

Granquist, L. (1997). The New View on Editing. *International Statistical Review*, **65**, 381-387.

Granquist and L., Kovar, J. G. (1997). Editing of survey data: how much is enough? In L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo,

N. Schwartz, D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: John Wiley and Sons, pp. 415-435.

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley and Sons.

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

Groves, R. M., Dillman, D. A., Eltinge, J. L., Little, R. J. A. (eds.) (2002). *Survey Nonresponse*. New York: John Wiley and Sons.

Groves, R. M., Fowler F. J. Jr., Couper M. P., Lepkowski J. M., Singer E., Tourangeau R. (2004). *Survey Methodology*. New York: John Wiley and Sons.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **169**, 439-457.

Groves R.M. and Lyberg L. (2010). Total survey error: past, present, and future. *Public Opinion Quarterly*, **74**, 849-879.

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

Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, **41**, 517-529.

Haziza, D. and Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, **32**, 129-145

Ilves, M. (2011). GREG estimation and probabilistic editing. *Metron*, **70**, 1-12.

Ilves, M. and Laitila, T. (2009). Probability-sampling approach to editing. *Austrian Journal of Statistics*, **38**, 171-183.

Jackman S. (1999). Correcting surveys for non-response and measurement error using auxiliary information. *Electoral Studies*, **18**, 7-27.

Klausch T., Schouten B., and Hox J.J. (2017). Evaluating bias of sequential mixed-mode designs against benchmark surveys. *Sociological Methods and Research*, **46**, 456-489.

Kreuter F., Müller G., and Trappmann M. (2010). Nonresponse and measurement error in employment research: making use of administrative data. *Public Opinion Quarterly*, **74**, 880-906.

Krosnick J.A., Narayan S., and Smith W.R. (1996). Satisficing in surveys: initial evidence. *New Directions for Evaluation*, **1996**, 29-44.

Laitila, T., Lindgren K., Norberg A. and Tongur C. (2017). Quantifying measurement errors in partially edited business survey data. In P. P. Biemer, E. de

Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, and B. T. West (eds.), *Total Survey Error in Practice*. New York: John Wiley and Sons, pp. 319-337.

Latouche, M. and Berthelot, J-M. (1992). Use of score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, **8**, 38-400.

Lawrence, D. and McDavitt, C. (1994). Significance editing in the australian survey of average weekly earnings. *Journal of Official Statistics*, **10**, 437-447.

Lessler, J. T., Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley and Sons.

Little, R. J. A. (1986). Survey nonresponse adjustments. *International Statistical Review*, **54**, 139-157.

Lorenc, B. (2007). Using the theory of socially distributed cognition to study the establishment survey response process. *Proceedings of the Third International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, pp. 881-891.

Lundquist, P. and Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, **29**, 557-582.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, **15**, 305-327.

Luzi et al. (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Recommended Practice Manual. Luxembourg, LU: Eurostat. Available at: *https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf/6e51b229-8628-422d-8c4c-7ede411e107f*, accessed January 2024.

Lyberg L., Biemer P., Collins , de Leeuw E., Dippo, Schwarz N., and Trewin D. (eds.) (1997). *Survey Measurement and Process Quality*. New York: John Wiley and Sons.

Lynn P. (2008). The problem of nonresponse. In E.D. de Leeuw, J.J. Hox and D.A. Dillman (eds.), *International Handbook of Survey Methodology*. UK: Taylor & Francis Group, pp. 35-55.

Oh, H. L. and Scheuren, F. J. (1983). Weighting adjustments for unit nonresponse. In Madow, W.G, Olkin, I. and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys*, Vol.2. New York: Academic Press, pp. 143-184.

Olson K. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly*, **70**, 737-758.

Olson K., Smyth J .D., Dykema J., Holbrook A. L., Kreuter F., and West B. T. (eds.) (2020). *Interviewer Effects from a Total Survey Error Perspective*. New York, NY: Chapman and Hall/CRC.

O'Muircheartaigh (1997). Measurement error in surveys: a historical perspective. In L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwartz, D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: John Wiley and Sons, pp. 1-25.

Peytcheva E. and Groves R. M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics*, **25**, 193-201.

Platek, R., Särndal, C-E. (2001). Can a statistician deliver? *Journal of Official Statistics*, **17**, 1-20.

Presser S., Rothgeb J. M., Couper M. P., Lessler J. T., Martin E., Martin J., and Singer E. (eds.) (2004). *Methods for Testing and Evaluating Survey Questionnaires*. New York: John Wiley and Sons.

Sakshaug J. W., Beste J., and Trappmann M. (2023). Effects of mixing modes on nonresponse and measurement error in an economic panel survey. *Journal for Labour Market Research*, **57**, article number: 2. https://doi.org/10.1186/s12651-022-00328-1

Sakshaug J. W., Yan T., and 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**, 907-933.

Saris W. E. and Gallhofer I. N. (2014). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, 2nd edition. New York: John Wiley and Sons.

Saris W. E. and Revilla M. (2016). Correction for measurement errors in survey research: necessary and possible. *Social Indicators Research*, **127**, 1005-1020.

Särndal, C-E., Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley and Sons.

Särndal, C-E., Lundström, S. (2010). Design for estimation: identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, **36**, 131-144.

Särndal, C-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, **55**, 279-294.

Särndal, C-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scholtus, S. (2018). *Editing and Estimation of Measurement Errors in Administrative and Survey Data*, PhD thesis. Research and graduation internal, Vrije Universiteit Amsterdam. Available at: *https://research.vu.nl/en/publications/editing-and-estimation-of-measurement-errors-in-administrative-an*, accessed January 2024.

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

Schouten, B., Peytchev A., and Wagner J. (2017). *Adaptive Survey Design.* Boca Raton: CRC Press.

Statistics Sweden (2023). Quality declaration: Foreign trade - exports and imports of goods. Available at: *https://www.scb.se/contentassets/4584f3f1df19400a885428aaa914d06e/ha0201_kd_2023_eng_230228.pdf*, accessed January 2024.

Tourangeau R., Rips L.J., and Rasinski K. (2000). *The Psychology of Survey Response.* Cambridge, UK: Cambridge University Press.

UNECE (2019). *Generic Statistical Business Process Model*, Version 5.1. Geneva, Switzerland. Available at *https://unece.org/sites/default/files/2023-11/GSBPM%20v5_1.pdf*, accessed January 2024.

Willimack, D.K. and Nichols, E. (2001). Building an alternative response process model for business surveys. *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Willimack D.K., Nichols, E., and Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, R. J. A. Little (eds.), *Survey Nonresponse.* New York: John Wiley and Sons, pp. 213-227.

Appendix

A. Derivation of (34)

The variance estimator of the measurement bias estimator, $\hat{V}(\hat{t}_{qs_2})$, is the sum of three terms:

$$\begin{aligned}
\hat{V}_1(\hat{t}_{qs}) &= \sum_{k,l\in s_2}\sum \frac{\Delta_{kl}}{\pi_{kl}\pi_{kl|s_a}\pi_{kl|r}}\frac{q_k}{\pi_k}\frac{q_l}{\pi_l} = \\
&= \sum_{k\in s_2} d_k(d_k-1)v_k w_{2k} q_k^2 + \sum_{k\neq l\in s_2}\sum (d_k d_l - d_{kl})v_k v_l w_{2k} w_{2l} q_k q_l = \\
&= \frac{1-f}{f^2}\left[\sum_{k\in s_2} v_k w_{2k} q_k^2 - \frac{1}{n-1}\left(\left(\sum_{k\in s_2} v_k w_{2k} q_k\right)^2 - \sum_{k\in s_2}\left(v_k w_{2k} q_k\right)^2\right)\right] \\
\hat{V}_2(\hat{t}_{qr|s_a}) &= \sum_{k,l\in s_2}\sum \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}\pi_{kl|r}}\frac{q_k}{\pi_k\pi_{k|s_a}}\frac{q_l}{\pi_l\pi_{l|s_a}} = \\
&= \sum_{k\in s_2} \frac{\pi_{k|s_a}(1-\pi_{k|s_a})}{\pi_{k|s_a}\pi_k^2\pi_{k|r}\pi_{k|s_a}^2} q_k^2 = \frac{1}{f^2}\sum_{k\in s_2} v_k(v_k-1)w_{2k}q_k^2 \qquad (38) \\
\hat{V}_3(\hat{t}_{qs_2|r,s_a}) &= \sum_{k,l\in s_2}\sum \frac{\Delta_{kl|r}}{\pi_{kl|r}}\frac{q_k}{\pi_k\pi_{k|s_a}\pi_{k|r}}\frac{q_l}{\pi_l\pi_{l|s_a}\pi_{l|r}} = \\
&= \sum_{k\in s_2} \frac{(1-\pi_{k|r})}{\pi_{k|r}\pi_{k|r}}(d_k v_k q_k)^2 = \frac{1}{f^2}\sum_{k\in s_2} v_k^2 w_{2k}(w_{2k}-1)q_k^2 \qquad (39)
\end{aligned}$$

The expressions presented in (38) and (39) simplify from double sums to single sums due to $\Delta_{kl|s_a} = 0$ and $\Delta_{kl|r} = 0$ when $k \neq l$.

B. Derivation of (36)

The covariance estimator, $\hat{Cov}(\hat{y}_{yr}, \hat{t}_{qs_2})$, is the sum of two terms:

$$\begin{aligned}
\hat{Cov}(\hat{t}_{yr}, \hat{t}_{qs2}) &= \sum_{k\in r}\sum_{l\in s_2} \frac{\Delta_{kl}}{\pi_{kl}\pi_{kl|s_a}\pi_{l|r}} d_k u_k d_l q_l \\
&+ \sum_{k\in r}\sum_{l\in s_2} \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}\pi_{l|r}} d_k v_k u_k d_l v_l q_l
\end{aligned}$$

Having $\Delta_{kl} = (d_k - 1){d_k}^{-2}$, if $k = l$ and $\Delta_{kl} = (d_k d_l - d_{kl})/(d_{kl} d_k d_l)$, if $k \neq l$, also $\Delta_{kl|s_a} = (v_k - 1){v_k}^{-2}$, if $k = l$ and $\Delta_{kl|s_a} = 0$ if $k \neq l$, we get

$$\begin{aligned}
\hat{Cov}(\hat{t}_{yr}, \hat{t}_{qs2}) &= \sum_{k \in s_2} (d_k - 1) d_k v_k w_{2k} u_k q_k \\
&+ \sum_{\substack{k \in r, l \in s_2 \\ k \neq l}} \sum (d_k d_l - d_{kl})(v_k u_k)(v_l w_{2l} q_l) + \sum_{k \in s_2} {d_k}^2 (v_k - 1) v_k w_{2k} u_k q_k \\
&= \frac{1}{f}\left(\frac{1}{f} - 1\right) \sum_{k \in s_2} v_k w_{2k} u_k q_k - \frac{1 - f}{f^2 (n-1)} \sum_{\substack{k \in r, l \in s_2 \\ k \neq l}} \sum (v_k u_k)(v_l w_{2l} q_l) \\
&+ \frac{1}{f^2} \sum_{k \in s_2} v_k (v_k - 1) w_{2k} u_k q_k \\
&= \frac{1}{f^2} \sum_{s_2} v_k (v_k - f) w_{2k} u_k q_k - \frac{1 - f}{f^2 (n-1)} \sum_{\substack{k \in r, l \in s_2 \\ k \neq l}} \sum (v_k u_k)(v_l w_{2l} q_l)
\end{aligned}$$