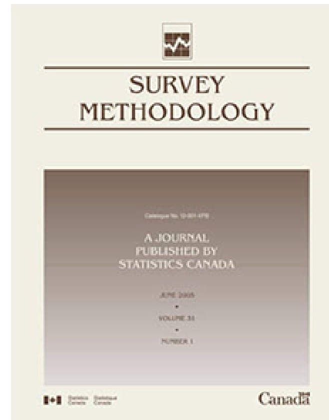
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology 41-1

Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach



by Alina Matei and M. Giovanna Ranalli

Release date: June 29, 2015

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 145-164
Statistics Canada, Catalogue No. 12-001-X

Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach

Alina Matei and M. Giovanna Ranalli[1]

Abstract

Nonresponse is present in almost all surveys and can severely bias estimates. It is usually distinguished between unit and item nonresponse. By noting that for a particular survey variable, we just have observed and unobserved values, in this work we exploit the connection between unit and item nonresponse. In particular, we assume that the factors that drive unit response are the same as those that drive item response on selected variables of interest. Response probabilities are then estimated using a latent covariate that measures the *will to respond to the survey* and that can explain a part of the unknown behavior of a unit to participate in the survey. This latent covariate is estimated using latent trait models. This approach is particularly relevant for sensitive items and, therefore, can handle non-ignorable nonresponse. Auxiliary information known for both respondents and nonrespondents can be included either in the latent variable model or in the response probability estimation process. The approach can also be used when auxiliary information is not available, and we focus here on this case. We propose an estimator using a reweighting system based on the previous latent covariate when no other observed auxiliary information is available. Results on its performance are encouraging from simulation studies on both real and simulated data.

Key Words: Unit nonresponse; Item nonresponse; Latent trait models; Response propensity; Rasch models.

1 Introduction

Nonresponse is an increasingly common problem in surveys. It is a problem because it causes missing data and, more importantly, because such gaps are a potential source of bias for survey estimates. In the presence of unit nonresponse, it is often assumed that each unit in the population has an associated probability to respond to the survey. Such a response probability is unknown and several methods are proposed to estimate it either explicitly, using response propensity modeling like logistic regression models (see e.g., Kim and Kim 2007), or implicitly, using response homogeneity groups or more generally calibration (see Särndal and Lundström 2005, for an overview). Once estimates are computed, a commonly used method to deal with unit nonresponse is reweighting: sampling weights of the respondents are adjusted by the inverse of the estimated response probability providing new weights. Estimation of response probabilities typically requires the availability of auxiliary information, either in the form of the value of some auxiliary variables for all units in the originally selected sample or of their population mean or total.

In this paper, we are particularly interested in the case where the missing data mechanism is non-ignorable, because nonresponse depends on characteristics of interest that are either observed only on the respondents or are completely unobserved, which leads to data that are Not Missing At Random (NMAR). This is typical of, but not limited to, surveys with sensitive questions (concerning drug abuse, sexual attitudes, politics, income, etc). Various approaches are proposed in the survey sampling literature to deal with non-ignorable nonresponse. These approaches can be roughly divided into likelihood based methods and reweighting methods. Note that all of these methods make use of observed auxiliary information. Survey problems with non-ignorable nonrespondents are discussed e.g., in Greenlees, Reece and

1. Alina Matei, Institute of Statistics, University of Neuchâtel, Pierre à Mazel 7, 2000, Neuchâtel, Switzerland. E-mail: alina.matei@unine.ch and Institute of Pedagogical Research and Documentation Neuchâtel, Switzerland; M. Giovanna Ranalli, Dept. of Political Sciences, University of Perugia, Italy. E-mail: giovanna.ranalli@stat.unipg.it.

Zieschang (1982), Little and Rubin (1987), Beaumont (2000), Qin, Leung and Shao (2002), Zhang (2002). Copas and Farewell (1998) introduce into the British National Survey of Sexual Attitudes and Lifestyles a variable called ‘enthusiasm-to-respond’ to the survey, which is expected to be related to probabilities of unit and item response. A method is proposed that estimates these probabilities using this variable to achieve unbiased estimates of population parameters. An approach based on the use of latent variables for modeling nonignorable nonresponse is given in Biemer and Link (2007), extending the ideas in Drew and Fuller (1980) and using a discrete latent variable based on call history data available for all sample units. The latent variable is computed using some indicators of level of effort based on call attempts.

We propose here a method of reweighting to reduce nonresponse bias in the case of non-ignorable nonresponse. The method does not require the availability of auxiliary information, on the sample or population level, but different assumptions are made. First, it is assumed that item nonresponse is present in the survey and that it affects m variables of particular interest. Thus a response indicator can be defined for each variable ℓ, for $\ell = 1, \ldots, m$, taking value 1 if item ℓ is observed on unit k and 0 otherwise. Next, the response indicators are assumed to be manifestations of an underlying continuous scale which determines a latent variable that is related to the response propensity of the units and to the variable of interest. It is possible to compute such a latent variable for all units in the sample, not only for the respondents, and thus to use it as an auxiliary variable in a response probability estimation procedure. The outcome of this estimation procedure can finally be used in a reweighting fashion.

The use of continuous latent variables to model item nonresponse is considered in Moustaki and Knott (2000). In this paper, we take a different perspective and use latent variable models to address non-ignorable unit nonresponse. We propose to use a latent variable called here ‘will to respond to the survey’, which is expected to be related to the probability of unit response, similar to the case of the ‘enthusiasm-to-respond’ variable as defined by Copas and Farewell (1998). Following Moustaki and Knott (2000), ‘*weighting through latent variable modeling is expected to perform well under non-ignorable nonresponse where conditioning on observed covariates only is not enough*.’ Moreover, in the absence of any covariate, we expect that an estimator based on the proposed weighting system using latent variables will perform better in terms of bias reduction than the naive estimator computed on the set of respondents. Moustaki and Knott (2000) propose a reweighting system for *item non-response* using covariates and one or more latent variables. Our major contribution over the existing literature is to construct a weighting system to deal with *unit and item non-response* based only on latent variables and that can also be used in the absence of any other covariate. On the other hand, our approach is different to that of Copas and Farewell (1998), because they survey their ‘enthusiasm-to-respond’ variable on the respondents to quantify the interest in answering the survey and a set of covariates, while we infer it from the data.

The paper is organized as follows. Section 2 introduces the survey framework and notation. Section 3 illustrates estimation of response probabilities. Section 4 describes the latent trait model used to this end. The proposed estimator and its variance estimation are shown in Section 5. In Section 6, the empirical properties of the proposed estimator are evaluated via simulation studies. In Section 7 we summarize our conclusions.

2 Framework

Let U be a finite population of size N, indexed by k from 1 to N. Let s denote the set of sample labels, so that $s \subset U$, drawn from the population using a probabilistic sampling design $p(s)$. The sample size is denoted by n. Let $\pi_k = \sum_{s; s \ni k} p(s)$ be the probability of including unit k in the sample. It is assumed that $\pi_k > 0, k = 1, \ldots, N$. Not all units selected in s respond to the survey. Denote by $r \subseteq s$ the set of respondents, and by $\bar{r} = s \setminus r$ the set of nonrespondents. The response mechanism is given by the distribution $q(r|s)$ such that for every fixed s we have

$$q(r|s) \geq 0, \text{ for all } r \in \mathcal{R}_s \text{ and } \sum_{s \in \mathcal{R}_s} q(r|s) = 1, \text{ where } \mathcal{R}_s = \{r | r \subseteq s\}.$$

Under unit nonresponse we define the response indicator $R_k = 1$ if unit $k \in r$ and 0 if $k \in \bar{r}$. Thus $r = \{k \in s | R_k = 1\}$. We assume that these random variables are independent of one another and of the sample selection mechanism (Oh and Scheuren 1983). Since only the units in r are observed, a response model is used to estimate the probability of responding to the survey of a unit $k \in U$, $p_k = P(k \in r | k \in s) = P(R_k = 1 | k \in s)$, which is a function of the sample and must be positive.

Suppose that in the survey there are m variables of particular interest. Each respondent is exposed to these m questionnaire variables, labelled $\ell = 1, \ldots, m$. Suppose that the goal is to estimate the population total of some variables of interest and, in particular, of the variable of interest y_j, i.e., $Y_j = \sum_{k=1}^{N} y_{kj}$, with y_{kj} being the value taken by y_j on unit k. In the ideal case, if the response distribution $q(r|s)$ is known, then the p_k's would be known and available to estimate Y_j using a reweigthing approach. Suppose also that item nonresponse is present for variable y_j. Let $r_j = \{k \text{ answers } y_j | k \in r\}$ be the set of respondents for variable y_j. As in the case of unit nonresponse we assume that the units in r_j respond independently of each other. Let $q_{kj} = P(k \text{ answers } y_j | k \in r)$. The final set of weights to be used into a fully reweighting approach to handle unit and item nonresponse is given by $1/(\pi_k p_k q_{kj})$, for all $k \in r_j$, assuming $q_{kj} > 0$. These weights can be for example used in a three-phase fashion in the following Horvitz-Thompson (HT) estimator

$$\hat{Y}_{j, pq, \text{true}} = \sum_{k \in r_j} \frac{y_{kj}}{\pi_k p_k q_{kj}}, \tag{2.1}$$

(see Legg and Fuller 2009, for the properties of estimators under three-phase sampling).

Usually, p_k and q_{kj} are unknown and should be estimated. A nonresponse adjusted estimator is then constructed by replacing p_k and q_{kj} with estimates $\hat{p}_k$ and $\hat{q}_{kj}$ in (2.1). The following sections provide details with this regard.

3 Estimating response probabilities

3.1 Using logistic regression to estimate p_k

Different methods to estimate p_k are proposed in the literature. All of these methods are based on the use of auxiliary information known on the population or sample level. In the case of non-ignorable nonresponse, the variable of interest is itself the cause (or one of the causes) of the response behavior, and a covariance between the former and the response probability is produced through a direct causal relation (see Groves 2006). In such a case, the response probability p_k could be modeled for $k \in s$ using logistic regression as follows

$$p_k = P\left(R_k = 1 \middle| y_{kj}\right) = \frac{1}{1 + \exp\left(-\left(a_0 + a_1 y_{kj}\right)\right)}, \tag{3.1}$$

or as follows

$$p_k = P\left(R_k = 1 \middle| y_{kj}, \mathbf{z}_k\right) = \frac{1}{1 + \exp\left(-\left(a_0 + a_1 y_{kj} + \mathbf{z}'_k \boldsymbol{\alpha}\right)\right)}, \tag{3.2}$$

where $\mathbf{z}_k = (z_{k1}, \ldots, z_{kt})'$ is a vector with the values taken by $t \geq 1$ covariates on unit k, and a_0, a_1, and $\boldsymbol{\alpha}$ are parameters.

Nonresponse bias in the unadjusted respondent total of the variable of interest y_j depends on the covariance between the values y_{kj} and p_k (see Bethlehem 1988). An example of a covariate that reduces the covariance between y_{kj} and p_k is the interest in the survey topic, such as knowledge, attitudes, and behaviors related to the survey topic (see Groves, Couper, Presser, Singer, Tourangeau, Acosta and Nelson 2006). The set of covariates $\mathbf{z}_k$ could be also related to the variable of interest y_j to reduce sampling variance (Little and Vartivarian 2005).

Since y_{kj} is only observed on respondents, Models (3.1) and (3.2) cannot be estimated. Therefore, usually, the values of $\mathbf{z}_k$ that are known for both respondents and nonrespondents and are related to the y_{kj}'s by a 'hopefully strong regression' (Cassel, Särndal and Wretman 1983) are used in the following model

$$p_k = P(R_k = 1 | \mathbf{z}_k) = \frac{1}{1 + \exp(-(a_0 + \mathbf{z}'_k \boldsymbol{\alpha}))}. \tag{3.3}$$

Then, maximum likelihood can be used to fit Model (3.3) using the data $(R_k, \mathbf{z}_k)$ for $k \in s$. This leads to estimate $\hat{a}_0$ and $\hat{\boldsymbol{\alpha}}$ and to the estimated response probabilities $\hat{p}_k = 1/[1 + \exp(-(\hat{a}_0 + \mathbf{z}'_k \hat{\boldsymbol{\alpha}}))]$ to be used in (2.1). This procedure provides some protection against nonresponse bias if $\mathbf{z}_k$ is a powerful predictor of the response probability and/or of the variable of interest (Kim and Kim 2007).

In what follows, we propose a reweighting adjustment system based on an auxiliary variable that measures the propensity of each unit to participate to the survey. To this end, further assumptions on the response model are introduced in order to assume a dependence of the p_k's on one latent auxiliary variable that is connected to the propensity scores of Rosenbaum and Rubin (1983). The proposed approach can be used when no other auxiliary information is available on $k \in s$.

3.2 Latent variables as auxiliary information

To obtain a measure of response propensities, we consider the case in which item nonresponse on the variables of interest is also present. Then, following Chambers and Skinner (2003, page 278) '*from a* *theoretical perspective the difference between unit and item nonresponse is unnecessary. Unit nonresponse is just an extreme form of item nonresponse*', we assume that item response on the variables of interest is driven on respondents by the same attitude and factors that drive unit response. Latent variable models can be used to estimate such factors that, therefore, can be used as covariates in a logistic response model.

As we have already mentioned we assume that item nonresponse affects m survey variables of particular interest. A second response indicator is introduced for each item ℓ. For each item ℓ and each unit k, a binary variable $x_{k\ell}$ is defined that takes value 1 if unit k answers to item ℓ and 0 otherwise. Let $\mathbf{x}_k = (x_{k1}, \ldots, x_{k\ell}, \ldots, x_{km})'$ denote the vector of response indicators for unit k to the m items and let $\mathbf{y}_k = (y_{k1}, \ldots, y_{k\ell}, \ldots, y_{km})'$ be the study variable vector for unit k. Thus $y_{k\ell}$ is the response value of unit k to item ℓ and $x_{k\ell}$ is its response indicator.

Suppose the $x_{k\ell}$'s are related to an assumed underlying latent continuous scale; they are the indicators of a latent variable denoted by θ_k. De Menezes and Bartholomew (1996) call the variable θ_k the 'tendency to respond' to the survey. We call it here the 'will to respond to the survey' of unit k. A latent trait model with a single latent variable is used to compute θ_k for each $k \in s$ (we will see later how; see Section 4.4). Assume for the moment that θ_k is known on all sample units and, as with usual auxiliary information, can be used as a covariate. In the absence of other covariates, Model (3.3) is rewritten as

$$p_k = P(R_k = 1 | \theta_k) = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 \theta_k))}. \tag{3.4}$$

Covariate θ_k can be viewed as a variable explaining the behavior related to the survey topic, and thus having good properties to reduce the covariance between y_{kj} and p_k and, therefore, nonresponse bias. If other suitable auxiliary information is available, it can be inserted in the model as supplementary covariates. Now, to estimate the parameters of Model (3.4), the value of θ_k has to be available for all units in the sample. The following sections provide details on how to obtain estimated values of θ_k for both respondents and nonrespondents.

4 Computing response propensities using latent trait models

The variable θ_k can be computed using a latent trait model. In general, latent variable models are multivariate regression models that link continuous or categorical responses to unobserved covariates. A latent trait model is essentially a factor analysis model for binary data (see Bartholomew, Steele, Moustaki and Galbraith 2002; Skrondal and Rabe-Hesketh 2007).

We start by creating the matrix with elements $\{x_{k\ell}\}_{k \in s; \ell = 1, \ldots, m}$. Figure 4.1 shows a schematic of the indicators $x_{k\ell}$ for respondents and nonrespondents. Then, we assume that the factors that drive unit response are the same as those that drive item response on selected variables of interest. In other words, item nonresponse is assumed nonignorable.

$$
\begin{array}{c}
\text{items} \\
\downarrow
\end{array}
$$

$$
\begin{array}{l}
\text{units} \rightarrow \left.\boxed{\begin{array}{cccc} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ & & \vdots & \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{array}}\right\} r \\
\text{units} \rightarrow \left.\boxed{\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}}\right\} \bar{r}
\end{array}
$$

Figure 4.1 Schematic representing variables $x_{k\ell}$ for the sets r and $\bar{r}$

Let $q_{k\ell}$ be the probability of response of unit k for item ℓ, for all $\ell = 1, \ldots, m$ and $k \in r$. As in the case of unit nonresponse, $q_{k\ell}$ is modelled as a function of the variable of interest using logistic regression as follows

$$q_{k\ell} = P(x_{k\ell} = 1 | y_{k\ell}, \theta_k, R_k = 1) = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k + \beta_{\ell 2} y_{k\ell}))}, \tag{4.1}$$

for $\ell = 1, \ldots, m$, and $k \in r$, where $\beta_{\ell 0}$, $\beta_{\ell 1}$ and $\beta_{\ell 2}$ are parameters. Since $y_{k\ell}$ is known only for units with $x_{k\ell} = 1$, $k \in r$, Model (4.1) cannot be estimated. As in the case of unit nonresponse, we propose to estimate $q_{k\ell}$ as a function of an auxiliary variable related to the variable of interest, that is θ_k. Model (4.1) is rewritten

$$q_{k\ell} = P(x_{k\ell} = 1 | \theta_k, R_k = 1) = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k))}, \tag{4.2}$$

for $\ell = 1, \ldots, m$, and $k \in r$. Model (4.2) is not an ordinary logistic regression model, because the θ_k's are unobservable values taken by a latent variable. Latent trait models can be used in this case to estimate $q_{k\ell}$, θ_k and the model parameters. Note that in the area of educational testing and psychological measurement, latent trait modelling is termed Item Response Theory.

The Rasch model (Rasch 1960) is a first simple latent trait model that is well known in the psychometrical literature and used to analyze data from assessments to measure variables such as abilities and attitudes. It takes the following form

$$q_{k\ell} = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_1\theta_k))} \text{ for } \ell = 1, \ldots, m \text{ and } k \in r. \tag{4.3}$$

The parameters $\beta_{\ell 0}$ are estimated for each item ℓ and reflect the extremeness (easiness) of item ℓ: larger values correspond to a larger probability of a positive response at all points in the latent space. The parameter β_1 is known as the 'discrimination' parameter and can be fixed to some arbitrary value without affecting the likelihood as long as the scale of the individuals' propensities is allowed to be free. In many situations the assumption that item discriminations are constant across items is too restrictive. The two-parameter logistic (2PL) model generalizes the Rasch model by allowing the slopes to vary. Specifically, the 2PL model assumes the form given in Equation (4.2). The parameters $\beta_{\ell 1}$ are now estimated for each item ℓ and provide a measure of how much information an item provides about the latent variable θ_k. To achieve identifiability of Model (4.2), we can fix the value of one or more parameters $\beta_{\ell 0}$ and $\beta_{\ell 1}$ in the estimation process. Moran (1986) showed that in the 2PL model, all the parameters are identifiable under wide conditions, provided the number of items exceeds two, and all the slopes are assumed to be strictly positive. A further generalization to Model (4.2) is considered in the literature - the 3PL model - that includes another parameter, the *guessing* parameter, to model the probability that a subject with a latent variable tending to $-\infty$ responds to an item. Such an extension does not seem necessary in the context at hand and will not be considered further.

4.1 Assumptions in latent trait models

Latent trait models typically rely on the following assumptions. The first one is the so-called *conditional independence* assumption, which postulates that item responses are independent given the latent variable (i.e., the latent variable accounts for all association among the observed variables $x_{k\ell}$). Consequently, given θ_k, the conditional probability of $\mathbf{x}_k$ is

$$P(\mathbf{x}_k | \theta_k) = \prod_{\ell=1}^{m} P(x_{k\ell} | \theta_k).$$

Following Bartholomew et al. (2002, page 181) 'the assumption of conditional independence can only be tested indirectly by checking whether the model fits the data. A latent variable model is accepted as a good fit when the latent variables account for most of the association among the observed responses.'

A second assumption of Models (4.2) and (4.3) is that of *monotonicity*: as the latent variable θ_k increases, the probability of response to an item increases or stays the same across intervals of θ_k. In other words, for two values of θ_k, say a and b, and arbitrarily assuming that $a < b$, monotonicity implies that $P(x_{k\ell} = 1 | \theta_k = a) < P(x_{k\ell} = 1 | \theta_k = b)$ for $\ell = 1, \ldots, m$. Larger values of θ_k are associated with a greater chance of a response to each item.

Finally, the third, and possibly strongest, assumption of Models (4.2) and (4.3) is that of *unidimensionality*, implying that a single latent variable fully explains the willingness of unit k to answer the questionnaire. All these basic assumptions imply that the dependence between the items $x_{k\ell}$ may be explained by the latent variable θ_k which represents the units' willingness and that the probability that a unit k responds to a given variable increases with θ_k.

4.2 Estimation of the model

In what follows we focus on the two-parameter logistic (2PL) model given in (4.2). Let $\boldsymbol{\beta}_\ell = (\beta_{\ell 0}, \beta_{\ell 1})'$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_\ell, \ell = 1, \ldots, m\}$. Model (4.2) can be fitted using maximum likelihood or bayesian methods. We focus here on the former. Under the maximum likelihood approach, three major methods - joint, conditional and marginal maximum likelihood - are developed. Here, we will concentrate on marginal maximum likelihood that can be applied to fit the 2PL model. This method is also used in the simulation studies of Section 6. It consists of maximizing the likelihood of the model after the θ_k are integrated out on the basis of a common distribution assumed on these parameters. In particular, it is assumed that θ_k is a random variable following a distribution with the density function $h(\cdot)$; typically $\theta_k \sim N(0,1)$. It is also assumed that the response vectors $\mathbf{x}_k$ are independent of one another and the conditional independence assumption holds.

For a set of n_r respondents having the response vectors $\mathbf{x}_k, k = 1, \ldots, n_r,$ the marginal likelihood can be expressed as

$$L\left(\boldsymbol{\beta}; \mathbf{x}_1, \ldots, \mathbf{x}_{n_r}\right) = \prod_{k=1}^{n_r} f(\mathbf{x}_k | \boldsymbol{\beta}),$$

where $f(\mathbf{x}_k | \boldsymbol{\beta}) = \int_{-\infty}^{\infty} g(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}) h(\theta_k) \, d\theta_k,$

$$g(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}) = \prod_{\ell=1}^{m} q_{k\ell}^{x_{k\ell}} (1 - q_{k\ell})^{1-x_{k\ell}} = \prod_{\ell=1}^{m} \frac{\exp(x_{k\ell}(\beta_{\ell 0} + \beta_{\ell 1}\theta_k))}{1 + \exp(\beta_{\ell 0} + \beta_{\ell 1}\theta_k)},$$

and h now denotes the density of the $N(0,1)$ distribution. The method consists in maximizing the corresponding log-likelihood, given by

$$\log L\left(\boldsymbol{\beta}; \mathbf{x}_1, \ldots, \mathbf{x}_{n_r}\right) = \sum_{k=1}^{n_r} \log(f(\mathbf{x}_k | \boldsymbol{\beta})),$$

with respect to $\boldsymbol{\beta}$ using, for example, the EM algorithm. Estimates of $\beta_{\ell 0}$ and $\beta_{\ell 1}, \ell = 1, \ldots, m$ are thus provided. Afterwards, θ_k is estimated using the empirical Bayes method by maximizing the posterior density

$$h(\theta_k | \mathbf{x}_k) = \frac{g(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}) h(\theta_k)}{g(\mathbf{x}_k)} \propto g(\mathbf{x}_k | \theta_k, \boldsymbol{\beta}) h(\theta_k),$$

with respect to θ_k and keeping item parameters and observations fixed. Estimates of $q_{k\ell}$ are obtained using Expression (4.2), where $\beta_{\ell 0}, \beta_{\ell 1}$ and θ_k are replaced with their estimates.

4.3 Goodness-of-fit measures of the model

Different goodness-of-fit measures are proposed in the literature to test whether the model given in (4.2) adequately fits the data (see e.g., Bartholomew et al. 2002). One uses two-way and three-way margins of the response items. Discrepancies between the expected (E) and observed (O) counts in these tables are measured using the statistic $R = (O - E)^2 / E$. Large values of R for the second-order or third-order margins will identify sets of items for which the model does not fit well. Note that the residuals $(O - E)^2 / E$ are not independent and they cannot be summed to give an overall test statistics distributed as a chi-squared (see Bartholomew et al. 2002, page 186). Item fit indexes (Bond and Fox 2007) can be used to this end as well. On the basis of estimated latent variables and item parameters, the expected response of a unit to an item can be computed. The similarity between the observed and expected responses to any item can be assessed through two fit mean-square statistics: the outlier-sensitive fit statistic (item outfit) and the information-weighted fit statistic (item infit). The estimate produced by the item outfit is relatively more affected by unexpected responses different from a person's measure, i.e., it is more sensitive to unexpected observations by units on items that are relatively very easy or very hard for them to answer. The item infit has each observation weighted by the information and, on the other side, is relatively more affected by unexpected responses closer to a person's measure, i.e., it is more sensitive to unexpected patterns of observations by units on items that are roughly targeted on them according to their latent variable value. The expected value for both statistics is one. For infit and outfit values greater/less than one indicate more/less variation between the observed and the predicted response patterns, a range of 0.5 to 1.5 is generally acceptable (Bond and Fox 2007).

In addition, point-measure correlations (Olsson, Drasgow and Dorans 1982) can be used to estimate the correlation between the latent variable and the single item response. Items for which such measures take negative or zero values should be removed from the analysis or may be evidence that the latent construct is not unidimensional. Unidimensionality can be tested by running a Principal Components Analysis (PCA) of the standardized residuals for the items (Wright 1996). In this way the first component (dimension) has already been removed, and it is possible to look at secondary dimensions, components or contrasts. Unidimensionality is supported by observing that the eigenvalue of the first PCA component in the correlation matrix of the residuals is small (usually less than 2.0). If not, the loadings on the first contrast indicate that there are contrasting patterns in the residuals.

Finally, when items are used to form a scale, they need to have internal consistency. Cronbach alpha can be used to test whether items have the reliability property, i.e., if they all measure the same thing, then they should be correlated with one another.

4.4 Estimation of p_k

Two solutions are shown here to estimate p_k using information from the latent trait model. The first solution uses logistic regression to estimate p_k for all $k \in s,$ and a two-stage approach.

Stage 1: First, an estimate $\hat{\theta}_k$ of θ_k is provided. To compute a value $\hat{\theta}_k$ for $k \in \bar{r},$ we assume again that unit nonresponse is just an extreme form of item nonresponse. Thus, a nonrespondent does not answer any item ℓ and thus $x_{k\ell} = 0,$ for all $\ell = 1, \ldots, m.$ The computation of $\hat{\theta}_k$ for $k \in \bar{r}$ is handled as follows: we add to the set r a phantom respondent unit $\tilde{k}$ having $x_{\tilde{k}\ell}$ equal to 0, for all $\ell = 1, \ldots, m.$ We denote this new set by $\tilde{r} = r \cup \{\tilde{k}\}.$ We estimate the parameters of Model (4.2) using all units $k \in \tilde{r},$ and compute the values $\hat{\theta}_k, k \in \tilde{r}.$ Model (4.2) allows the computation of $\hat{\theta}_k$ for all $k \in \tilde{r}.$ Unit $\tilde{k}$ has an estimated value $\hat{\theta}_{\tilde{k}}.$ We assign to all units $k \in \bar{r}$ an estimate $\hat{\theta}_k$ equal to $\hat{\theta}_{\tilde{k}}.$ Thus, the same value of $\hat{\theta}_k$ is provided for all $k \in \bar{r}.$ Using this method, each unit $k \in s$ has associated an estimate $\hat{\theta}_k.$ This is the key feature for the estimation of the response probabilities p_k provided in the next stage.

Stage 2: We use the estimate $\hat{\theta}_k$, for $k \in s$, provided in the first stage as a covariate in Model (3.4) instead of the unknown value of θ_k; in particular

$$p_k = P\left(R_k = 1 \middle| \hat{\theta}_k\right) = \frac{1}{1 + \exp\left(-\left(\alpha_0 + \alpha_1 \hat{\theta}_k\right)\right)}, \text{for all } k \in s. \tag{4.4}$$

Model (4.4) provides estimates $\hat{p}_k$ of p_k, for all $k \in s$.

One of the Referees suggested the following solution to estimate p_k. Let $S_k = \sum_{\ell=1}^{m} x_{k\ell}$ be the raw score for unit k, i.e., the number of items unit k has responded to: if $k \in \bar{r}$, then $S_k = 0$; if $k \in r$, then $S_k > 0$. Then p_k can be estimated by modelling $P(S_k > 0 | \theta_k)$. By the conditional independence assumption we have

$$\begin{aligned} p_k &= P(S_k > 0|\theta_k) = 1 - P(S_k = 0|\theta_k) = 1 - P\left(\bigcap_{\ell=1}^{m} (x_{k\ell} = 0|\theta_k)\right) \\ &= 1 - \prod_{\ell=1}^{m} (1 - P(x_{k\ell} = 1|\theta_k)). \end{aligned}$$

We have $P(x_{k\ell} = 1|\theta_k) = P(R_k = 1|\theta_k) P(x_{k\ell} = 1|\theta_k, R_k = 1) + P(R_k = 0|\theta_k) P(x_{k\ell} = 1|\theta_k, R_k = 0) = p_k q_{k\ell}$, because $P(x_{k\ell} = 1|\theta_k, R_k = 0) = 0$. As a result, we obtain

$$p_k = 1 - \prod_{\ell=1}^{m} (1 - p_k q_{k\ell}), k \in r.$$

The estimated response probability $\hat{p}_k$, $k \in r$ is obtained as a solution to the polynomial equation

$$\hat{p}_k = 1 - \prod_{\ell=1}^{m} (1 - \hat{p}_k \hat{q}_{k\ell}).$$

This solution, although very elegant, has two drawbacks. If m is large, the above polynomial equation is difficult or even impossible to solve. If it possible to solve the polynomial equation for moderate m, the real solutions are not necessarily in (0, 1). This solution has not been considered here further.

5 The proposed estimator and its variance estimation

Recall that we have a variable of particular interest y_j and that item nonresponse is present for it. If we wish to estimate the population total Y_j of y_j, then a naive estimator that does not correct neither for unit nor for item nonresponse is given by

$$\hat{Y}_{j,\text{naive}} = N \sum_{k \in r_j} \frac{y_{kj}}{\pi_k} \bigg/ \sum_{k \in r_j} \frac{1}{\pi_k}. \tag{5.1}$$

Reweighting item responders is also an approach to handle item nonresponse. Moustaki and Knott (2000) propose to weight item responders by the inverse of the fitted probability of item response $\hat{q}_{k\ell}$, assuming $\hat{q}_{k\ell} > 0$. Therefore, a possible adjustment weight for item and unit nonresponse associated with unit $k \in r_j$ is given by $1/\left(\hat{p}_k \hat{q}_{kj}\right)$. We propose using the three-phase estimator adjusted for item and unit nonresponse via reweighting given by

$$\hat{Y}_{j,pq} = \sum_{k \in r_j} \frac{y_{kj}}{\pi_k \hat{p}_k \hat{q}_{kj}}, \tag{5.2}$$

where $\hat{p}_k$ is provided by Model (4.4), and $\hat{q}_{kj}$ by Model (4.2). Proposals that use imputation of y_{kj} values for $k \in r \setminus r_j$ to deal with item nonresponse are also considered but not reported for reasons of space. They are available from the Authors upon request.

The properties of the proposed estimator (5.2) depend on the assumptions made about the unit and the item nonresponse mechanisms. In particular, Estimator (5.2) assumes a second phase of sampling with unknown response probabilities. If we ignore estimation of θ_k in Model (4.4), the results in Kim and Kim (2007) on design consistency of the two-phase estimator that uses estimated response probabilities hold here as well when considering maximum likelihood estimates for the parameters α_0 and α_1. Again, ignoring estimation of the latent variable θ_k and using marginal maximum likelihood estimates for the parameters $\beta_{\ell 0}$ and $\beta_{\ell 1}$ in Model (4.2), estimator $\hat{Y}_{j,pq}$ will be consistent if the models for unit and item nonresponse probabilities are correctly specified.

We can consider replication methods for variance estimation of the proposed estimator and combine proposals for two-phase sampling (Kim, Navarro and Fuller 2006) and for generalized calibration in the presence of nonresponse (Kott 2006). In particular, the replicate variance estimator can be written as

$$\hat{V}_r = \sum_{l=1}^{L} c_l \left(\hat{Y}_{j,pq}^{(l)} - \hat{Y}_{j,pq}\right)^2,$$

where $\hat{Y}_{j,pq}^{(l)}$ is the l^{th} version of $\hat{Y}_{j,pq}$ based on the observations included in the l^{th} replicate, L is the number of replications, c_l is a factor associated with replicate l determined by the replication method. The l^{th} replicate of $\hat{Y}_{j,pq}$ can be written as $\hat{Y}_{j,pq}^{(l)} = \sum_{k \in r_j} w_{3k}^{(l)} y_{kj}$, where $w_{3k}^{(l)}$ denotes the replicate weight for the k^{th} unit in the l^{th} replication. These replicate weights are computed using a two-step procedure.

First, note that, if we ignore for the moment the presence of item nonresponse, the two-phase estimator $\hat{Y}_{j,p} = \sum_{k \in r} w_{2k} y_{kj}$, has weights

$$w_{2k} = 1/(\pi_k p_k) = w_{1k} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right),$$

with, $w_{1k} = 1/\pi_k$, $F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right) = 1 + \exp\left(-\left(\alpha_0 + \alpha_1 \hat{\theta}_k\right)\right)$ (see Equation (4.4)). Let $\hat{\mathbf{z}}_1 = \sum_{k \in s} w_{1k} \mathbf{z}_{1k}$ be the first phase estimate of the total of variable $\mathbf{z}_1$ defined as $\mathbf{z}_{1k} = \pi_k p_k \left(1, \hat{\theta}_k\right)'$. Then, parameters α_0 and α_1 are such that

$$\sum_{k \in r} w_{1k} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right) \mathbf{z}_{1k} = \hat{\mathbf{z}}_1. \tag{5.3}$$

This procedure is equivalent to obtaining unweighted maximum likelihood estimates, but is convenient to set it as a non-linear generalized calibration problem. In this way, it is possible to use the approach in Kott (2006), combined with that in Kim et al. (2006), to obtain replicate weights using the following steps.

Step 1: Compute the first phase estimate of the total of $\mathbf{z}_{1k}$ with l^{th} observation deleted, i.e., $\hat{\mathbf{z}}_1^{(l)} = \sum_{k\in s} w_{1k}^{(l)} \mathbf{z}_{1k}$, where $w_{1k}^{(l)}$ is the classical jackknife replication weight for unit k in replication l. Compute the jackknife weights for the second phase sampling using $\hat{\mathbf{z}}_1^{(l)}$ as a benchmark. In particular, $w_{2k}^{(l)}$ are chosen to be $w_{2k}^{(l)} = w_{2k} w_{1k}^{(l)} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right) / w_{1k}$ with α_0 and α_1 such that

$$\sum_{k\in r} w_{2k}^{(l)} \mathbf{z}_{1k} = \hat{\mathbf{z}}_1^{(l)}.$$

This procedure provides weights that are very similar to those considered in Kott (2006) and can be computed using existing software that handles generalized calibration.

Item nonresponse is handled similarly by considering $w_{3k} = 1/\left(\pi_k p_k q_{kj}\right) = w_{2k} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right)$ (compare Equation (4.3)). A major approximation here is to assume that, given $\hat{\theta}_k$, parameters β_{j0} and β_{j1} are estimated using a classical logistic model (instead of a 2PL model) and are such that

$$\sum_{k\in r_j} w_{2k} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right) \mathbf{z}_{2k} = \hat{\mathbf{z}}_2,$$

where $\hat{\mathbf{z}}_2 = \sum_{k\in r} w_{2k} \mathbf{z}_{2k}$ and $\mathbf{z}_{2k} = \pi_k p_k q_{kj} \left(1, \hat{\theta}_k\right)^T$. Another drawback is that auxiliary variables $\mathbf{z}_{2k}$ depend on j and, therefore, different sets of weights have to be produced for the different variables of interest.

Step 2: Third phase jackknife weights are obtained by first computing the second phase estimate of the total of $\mathbf{z}_{2k}$ with unit l removed by using weights coming from Step 1, i.e., $\hat{\mathbf{z}}_2^{(l)} = \sum_{k\in r} w_{2k}^{(l)} \mathbf{z}_{2k}$. Then, using $\hat{\mathbf{z}}_2^{(l)}$ as a benchmark, $w_{3k}^{(l)}$ are chosen to be $w_{3k}^{(l)} = w_{3k} w_{2k}^{(l)} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right) / w_{2k}$ with β_{j0} and β_{j1} computed via

$$\sum_{k\in r_j} w_{3k}^{(l)} \mathbf{z}_{2k} = \hat{\mathbf{z}}_2^{(l)}.$$

6 Simulation studies

We evaluate the performance of the estimator presented in Section 5 by means of a Monte Carlo simulation under two different settings. The first one uses a real data set as the population and considers variables of interest that are all binary, while the second one uses simulated population data with variables of interest that are continuous. Results from the first setting are presented in Section 6.1, while those from the second setting are presented in Section 6.2.

In both settings, simple random sampling without replacement is employed and the following estimators are considered:

- $\mathrm{HT} = \sum_{k \in s} y_{kj} / \pi_k$: the Horvitz-Thompson estimator in the case of full response is computed as a benchmark in the absence of nonresponse.

- $\hat{Y}_{j,\text{naive}}$: the naive estimator given in (5.1); no explicit action is taken to adjust for unit and item nonresponse. Note that for simple random sampling without replacement, it reduces to $\hat{Y}_{j,\text{naive}} = N \sum_{k \in r_j} y_{kj} / n_{r_j}$, where n_{r_j} is the size of the set r_j, and it is the same as the Horvitz-Thompson estimator adjusted for unit nonresponse that assumes uniform response probabilities estimated by n_{r_j} / n.

- $\hat{Y}_{j,pq}$: the three-phase estimator proposed in Section 5, Equation (5.2).

- $\hat{Y}_{j,pq,\text{true}}$: the three-phase estimator that uses the true values for the response probabilities p_k and q_{kj} is also computed for comparison with $\hat{Y}_{j,pq}$ to understand the effect of estimating the response probabilities.

The simulations are carried out in R version 2.15, using the R package '`ltm`' (Rizopoulos 2006) to fit the latent trait models. The following performance measures are computed for each estimator, generically denoted below by $\hat{Y}$ where suffix j is dropped for ease of notation (Y denotes the population total):

- the Monte Carlo Bias

$$\mathrm{B} = E_{\text{sim}}\left(\hat{Y}\right) - Y,$$

where $E_{\text{sim}}\left(\hat{Y}\right) = \sum_{i=1}^{M} \hat{Y}_i / M$, $\hat{Y}_i$ is the value of the estimator $\hat{Y}$ at the i^{th} simulation run and M is total number of simulation runs;

- the Relative Bias

$$\mathrm{RB} = \frac{\mathrm{B}}{Y};$$

- the Monte Carlo Standard Deviation

$$\sqrt{\mathrm{VAR}} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{Y}_i - E_{\text{sim}}\left(\hat{Y}\right)\right)^2};$$

- the Monte Carlo Mean Squared Error

$$\mathrm{MSE} = \mathrm{B}^2 + \mathrm{VAR}.$$

6.1 Simulation setting 1

We consider the Abortion data set formed by four binary variables extracted from the 1986 British Social Attitudes Survey and concerning the attitude towards abortion. The data is available in the R package ‘`ltm`’ (Rizopoulos 2006). $N = 379$ individuals answered the following questions after being asked if the law should allow abortion under the circumstances presented under each item:

1. The woman decides on her own that she does not wish to keep the baby.
2. The couple agrees that they do not wish to have a child.
3. The woman is not married and does not wish to marry the man.
4. The couple cannot afford any more children.

The variable of interest y_j is selected to be the second one $(j = 2)$ with a total $Y_j = 225$ in the population.

The data is analyzed by Bartholomew et al. (2002) as an example in which a latent variable can be found that measures the attitude towards abortion. At the population level, we compute the latent variable (denoted here by θ_k^a) using Model (4.2) on the $\{y_{k\ell}\}_{k=1,\ldots,N;\ell=1,\ldots,4}$ data. The correlation between the values $y_{k\ell}$ and θ_k^a is approximately equal to 0.85, for $\ell = 1,\ldots,4$. Afterwards, we have set $\theta_k = \hat{\theta}_k^a$, for all $k = 1,\ldots,N$.

At the population level, the unit response probabilities are generated using the following response model

$$p_k = 1/(1 + \exp(-(0.7 + y_{k2} + \theta_k + 0.2\varepsilon_k))), \tag{6.1}$$

with $\varepsilon_k \sim U(0,1)$, to simulate nonignorable nonresponse. The population mean of p_k is approximately 0.74.

To generate item response probabilities at the population level, the following model is used

$$q_{k\ell} = 1/(1 + \exp(-(b_\ell\theta_k + a_\ell + y_{k\ell}))), \quad \text{for } \ell = 1,\ldots,4, \tag{6.2}$$

where $b_\ell = 3$, for $\ell = 1,\ldots,4$, while a_ℓ takes different values according to ℓ; in particular, $a_1 = 1, a_2 = 0, a_3 = -0.5$ and $a_4 = 1$. The nominal item nonresponse rate for the four items in the population is 35%, 42%, 47%, 31%, respectively.

We draw $M = 10{,}000$ simple random samples without replacement from the population using two sample sizes: $n = 50$ and $n = 100$. In each sample s, the units are classified as respondents according to Poisson sampling, using the probabilities p_k computed as in Equation (6.1) and resulting in the set r. Then, given r, the matrix $\{x_{k\ell}\}_{k\in r;\ell=1,\ldots,4}$ is constructed where the values $x_{k\ell}$ are drawn using Poisson sampling with probabilities $q_{k\ell}$ defined in (6.2). In each simulation run, Model (4.2) and the respondents set r are used to compute the variable $\hat{\theta}_k$ for all $k \in s$ as described in Section 4.4. Model (4.4) is fitted to obtain $\hat{p}_k$. The average item nonresponse rate over simulations for the four items is found to be 26%, 33%, 38% and 23%. The jackknife variance estimator was computed as described in Section 5 using the `gencalib()` function in R package ‘sampling’ (Tillé and Matei 2012) and the logistic distance (Deville, Särndal and Sautory 1993).

Table 6.1 reports the results for $n = 50$ and $n = 100$. As expected, HT and $\hat{Y}_{j,pq,\text{true}}$ have almost zero bias, with the second one showing a relatively larger MSE that is due uniquely to the smaller sample size. The naive estimator shows a very large negative bias. This is due to the fact that units with a zero value of y_j are less likely to respond and the total is clearly underestimated. The estimator $\hat{Y}_{j,pq}$ shows a much smaller bias than the naive estimator. Note that the performance of the proposed estimator is mostly driven by absolute bias, so that the performance is not particularly different when increasing the sample size, apart from a decrease in variance. If we compare $\hat{Y}_{j,pq,\text{true}}$ and $\hat{Y}_{j,pq}$, we note that $\hat{Y}_{j,pq}$ still suffers from some bias that comes from response model misspecification (we are not accounting for the variables of interest values).

For the proposed estimator, the jackknife variance estimator was also tested by looking at the empirical coverage of a 95% confidence interval computed for each replicate as $\hat{Y}_{j,pq} \pm 1.96\sqrt{\hat{V}_r}$. For $n = 50$, the mean value of $\sqrt{\hat{V}_r}$ over simulations was 54.8, while for $n = 100$, 53.3, with a 95% coverage rate of 94.6% and 96.3%, respectively. The replicate estimator overestimates the Monte Carlo standard deviation reported for $\hat{Y}_{j,pq}$ in Table 6.1 in both cases, but shows good coverage rates.

Table 6.1
Simulation results for setting 1 - Abortion data set

Estimator	B	$\sqrt{\text{VAR}}$	MSE	% RB
$n = 50$				
HT	0.05	24.5	600.5	$<$ 0.1
$\hat{Y}_{j,\text{naive}}$	-126.5	19.4	16,378.6	-56.2
$\hat{Y}_{j,pq}$	20.6	32.4	1,474.1	9.1
$\hat{Y}_{j,pq,\text{true}}$	0.02	35.0	1,225.0	$<$ 0.1
$n = 100$				
HT	-0.06	16.0	255.5	$<$ 0.1
$\hat{Y}_{j,\text{naive}}$	-126.9	13.5	16,284.1	-56.4
$\hat{Y}_{j,pq}$	17.9	21.9	802.2	8.0
$\hat{Y}_{j,pq,\text{true}}$	-0.1	23.7	559.9	$<$ 0.1

To study the performance of the latent model on the population level and the correlation between the variable of interest and the estimated latent variable, we apply the procedure described earlier using $q_{k\ell}$ defined in (6.2) to construct the matrix $\{x_{k\ell}\}_{k=1,\ldots,N;\ell=1,\ldots,4}$ for all population units. We fit Model (4.2) on the population level and compute the variable θ_k for all $k = 1,\ldots,N$. The Cronbach's alpha measure takes value 0.83 showing a good internal consistency of the items. The correlation coefficient between the variable of interest and the estimated latent variable takes value 0.76, indicating that the latent auxiliary information has a strong power of predicting y_{k2}, as advocated in the model of Cassel et al. (1983). Inspection of the two-way margins for the matrix $\{x_{k\ell}\}$ gives the residuals $(O - E)^2/E$ between 0.03 and 0.23. Similarly, the three-way margins for the matrix $\{x_{k\ell}\}$ give residuals between 0 and 1.19. This indicates that we have no reason to reject here the one-factor latent Model (4.2) (see Bartholomew et al. 2002, page 186).

6.2 Simulation setting 2

We generate $\{y_{k1}, \ldots, y_{k6}, \theta_k\}$ for $k = 1, \ldots, N = 2{,}000$ using a multivariate normal distribution with mean 1. The degree of correlation between y_ℓ and $y_{\ell'}$ is 0.8, with $\ell, \ell' = 1, \ldots, 6, \ell \neq \ell'$. We set the variable of interest to be y_6 and consider different degrees of correlation between its values and those taken by θ_k, namely 0.3, 0.5, 0.8. The values of θ_k are afterwards standardized to have mean 0 and variance 1.

The response probabilities are obtained by first computing

$$p_k^\circ = 1/[1 + \exp(-(0.5 + y_{k1} + \theta_k))], \quad \text{for } k = 1, \ldots, N, \tag{6.3}$$

and then rescaling them to take values between 0.1 and 0.9, with a population mean approximatively equal to 0.7.

The item response probabilities are generated by first computing

$$q_{k\ell}^\circ = 1/(1 + \exp(-(b_\ell \theta_k + a_\ell + y_{k\ell}))), \quad \text{for } k = 1, \ldots, N \text{ and } \ell = 1, \ldots, 6, \tag{6.4}$$

where $\{a_\ell\}_{\ell=1,\ldots,6} = \{1, 0, -0.5, 1, 0, -0.5\}$ and $\{b_\ell\}_{\ell=1,\ldots,6} = \{1, 1, 1, 1.5, 1.5, 1.5\}$, and then rescaling the values to be between 0.1 and 0.95.

We draw $M = 10{,}000$ samples by simple random sampling without replacement of size $n = 200$. For each sample s, a response set r is created by carrying out Poisson sampling with parameter p_k defined in (6.3). Each element of the matrix $\{x_{k\ell}\}_{k \in r, \ell=1,\ldots,6}$ is generated using Poisson sampling with parameter $q_{k\ell}$ defined in (6.4). Item nonresponse rates over simulations take approximately value 18%, 28%, 35%, 19%, 29%, 34%, for $\ell = 1, \ldots, 6$, respectively. For each simulation run, Model (4.2) is used to compute the variable $\hat{\theta}_k$ for all $k \in s$. Model (4.4) is then fitted to obtain $\hat{p}_k$.

Table 6.2
Simulation results for setting 2 - Simulated continuous data

Estimator	**B**	$\sqrt{\textbf{VAR}}$	**MSE**	**RB%**
correlation coefficient 0.3				
HT	-0.7	131.6	17,331.2	≈ -0.0
$\hat{Y}_{j,\text{naive}}$	825.6	177.1	713,039.3	41.0
$\hat{Y}_{j,pq}$	-227.4	188.0	87,033.0	-11.3
$\hat{Y}_{j,pq,\text{true}}$	48.4	231.8	56,073.2	2.4
correlation coefficient 0.5				
HT	0.1	135.0	18,220.5	≈ 0.0
$\hat{Y}_{j,\text{naive}}$	972.6	176.2	977,009.5	50.7
$\hat{Y}_{j,pq}$	-180.0	175.5	63,552.0	-9.4
$\hat{Y}_{j,pq,\text{true}}$	74.8	212.7	50,844.0	3.9
correlation coefficient 0.8				
HT	-0.1	134.1	17,992.0	≈ -0.0
$\hat{Y}_{j,\text{naive}}$	1,154.6	168.1	1,361,388.1	57.7
$\hat{Y}_{j,pq}$	-184.8	164.4	61,173.0	-9.2
$\hat{Y}_{j,pq,\text{true}}$	100.6	196.2	48,597.9	5.0

Table 6.2 reports on the performance of the estimators for the three values taken by the nominal correlation coefficient between y_{k1} and θ_k : 0.3, 0.5 and 0.8. The proposed estimator is always able to reduce bias over the naive estimator, even when the correlation between the variable of interest and the latent variable gets smaller. The relative bias takes acceptable values in most cases. Bias deserves a closer look. The naive estimator in all cases largely overestimates the total. This is expected, because the values p_k, q_{k6}, θ_k and y_{k6} all go in the same direction. Therefore, in our respondents sample, we are more likely to find relative larger values for y_6 by this providing overestimation for the naive estimator. On the other hand, $\hat{Y}_{j,pq}$ underestimates the total because it is based only on the observed units of r_j that do have relatively large values for y_6, but also relatively large values for p_k and q_{k6} and, therefore, end up having a small weight.

The matrix of population values $\{x_{k\ell}\}_{k=1;\ldots;2{,}000;\ell=1,\ldots,6}$ is constructed in the same way as in Section 6.1 to validate the assumptions behind the 2PL model. The Cronbach's alpha takes approximately value 0.5 for the correlation coefficient equal to 0.3, 0.6 for 0.5, and 0.7 for 0.8; the pairwise association between the six items reveals p-values smaller than 0.01. Inspection of the two-way and three-way margins of the matrix $\{x_{k\ell}\}$ gives residuals $(O-E)^2/E$ that all take values smaller than 4. Therefore, the one factor latent model can be accepted and items all seem to be measuring the same latent trait.

7 Discussion and conclusions

We have proposed a reweighting system to compensate for non-ignorable nonresponse based on a latent auxiliary variable. This variable is computed for each unit in the sample using a latent model assuming the existence of item nonresponse and that the same latent structure is hidden behind item and unit nonresponse. Unit response probabilities are then estimated by a logistic model that uses as a covariate the latent trait extracted by the response patterns using a latent trait model. The proposed reweighting system is then used in a three-phase estimator to handle nonresponse, together with a replication method to estimate its uncertainty. The main goal is to reduce nonresponse bias in the estimation of the population total. The proposed estimator performs well in our simulation studies compared with the naive estimator, and the gain in efficiency is substantial in certain cases. Reductions in bias are also seen when the correlation between the latent trait and the variable of interest is modest.

By design, the estimated latent variable $\hat{\theta}_k$ is related to the response indicators x_{kj} for the variable of interest y_j; since nonresponse is assumed to be non-ignorable, y_{kj} and x_{kj} are related as well. If the following condition holds,

$$\rho^2_{y_j,x_j} + \rho^2_{\hat{\theta},x_j} > 1,$$

where the correlation coefficients $\rho_{y_j,x_j}, \rho_{\hat{\theta},x_j} > 0$, then y_j and $\hat{\theta}$ are positively correlated (see Langford, Schwertman and Owens 2001). Note that the minimum degree of correlation between the variable of interest and the latent variable capable of reducing the nonresponse bias was found to be 0.3 in simulation setting 2 (Section 6.2). Of course, bias reduction depends on model assumptions. If response indicators are not good predictors of unit response behavior, then model misspecification is present and, of course, reduction in bias may not be present and variance could be introduced in estimation. Nonetheless, diagnostic tools from item response theory can be used to assess the goodness of fit of the latent trait model employed to estimate values for θ_k.

We have considered the case in which no auxiliary information is available at the sample or population level to reduce nonresponse bias. Observed covariates (if available) and the latent variable can be, however, used together in the estimation of response probabilities. Moreover, latent trait models can, themselves, be fitted with covariates. The introduction of covariates in these models should be carried out with increasing prudence on variance.

The proposed estimator is a three-phase estimator using a reweighting system based on $\hat{p}_k$ and $\hat{q}_{kj}$. It is known that small values of $\hat{p}_k$ and $\hat{q}_{kj}$ may lead to unstable reweighted estimators because of large nonresponse weights. To overcome this problem, the propensity score method (e.g., Eltinge and Yansaneh 1997) is often used in practice, providing a good solution against extreme weights adjustments. In order to apply this method in our framework, the respondents to y_j should be grouped in different classes given by the quantiles of $1/(\hat{p}_k \hat{q}_{kj})$. The final step is the calculation of a weight for each class.

Final remarks concern the conditional independence assumption in latent trait models. In nonresponse literature, it is usual to use Poisson sampling to model unit response behavior by assuming that units in the set r are selected with unknown response probabilities and that response is independent from unit to unit. The conditional independence assumption in the latent trait models is a similar condition applied to items. Both assumptions are strong, sometimes they are in doubt, yet they are necessary in the statistical inferential process.

Different methods were developed in psychometric literature to relax the conditional independence assumption. We cite here the *partial independence* approach by Reardon and Raudenbush (2006), developed for the case where responses to earlier questions determine whether later questions are asked or not, and where the usual conditional independence assumption of standard models fails. This approach could be used in our framework for the case where $q_{k\ell}$ is defined as $P(x_{k\ell} = 1 | x_{kj}$, for some $j \in \{1, \ldots, m\}, \ell \neq j, \theta_k)$ instead of $P(x_{k\ell} = 1 | \theta_k)$, $k \in r$. Another useful approach for cases where items are clustered is the latent trait hierarchical modeling. A random effect is introduced into a latent trait model to account for potential residual dependence due to the common sources of variation shared by clusters of items (see e.g., Scott and Ip 2002). Further research should be done to accommodate these approaches in the survey sampling framework.

Acknowledgements

The work of M. Giovanna Ranalli has been developed partially under the support of the project PRIN-SURWEY (grant 2012F42NS8, Italy).

References

Bartholomew, D.J., Steele, F., Moustaki, I. and Galbraith, J.I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Chapman and Hall/CRC.

Beaumont, J.-F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology,* 26, 2, 131-136.

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

Biemer, P.P., and Link, M.W. (2007). *Evaluating and Modeling early Cooperator Effects in RDD Surveys*. New York: John Wiley & Sons, Inc.

Bond, T., and Fox, C. (2007). *Applying the Rasch model: Fundamental Measurement in the Human Sciences* (2nd Ed.). Lawrence Erlbaum Associates, Inc, Mahwah, N.J.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys,* (Eds., W.G. Madow and I. Olkin), New York: Academic Press. 3, 143-160.

Chambers, R.L., and Skinner, C. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.

Copas, A.J., and Farewell, V.T. (1998). Dealing with non-ignorable non-response by using an ‘enthusiasm-to-respond’ variable. *Journal of the Royal Statistical Society, Series A,* 161, 385-396.

De Menezes, L.M., and Bartholomew, D.J. (1996). New developments in latent structure analysis applied to social attitudes. *Journal of Royal Statistical Society, Series A,* 159, 213-224.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Drew, J.H., and Fuller, W.A. (1980). Modeling nonresponse in surveys with callbacks. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 1, 33-40.

Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association,* 77, 251-261.

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

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

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics,* 35, 501-514.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association,* 101, 473, 312-320.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology,* 32, 2, 133-142.

Langford, E., Schwertman, N. and Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician*, 55, 4, 322-325.

Legg, J.C., and Fuller, W.A. (2009). Two-phase sampling. *Handbook of Statistics,* 29, 55-70.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology,* 31, 2, 161-168.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Moran, P.A.P. (1986). Identification problems in latent trait models. *British Journal of Mathematical and Statistical Psychology,* 39, 2, 208-212.

Moustaki, I., and Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of Royal Statistical Society, Series A,* 163, 445-459.

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

Olsson, U., Drasgow, F. and Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika,* 47, 337-347.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association,* 97, 193-200.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. *The Danish Institute of Educational Research,* Copenhagen.

Reardon, S.F., and Raudenbush, S.W. (2006). A partial independence item response model for surveys with filter questions. *Sociological Methodology,* 36, 1, 257-300.

Rizopoulos, D. (2006). `ltm`: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17, 5, 1-25.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score inobservational studies for causal effects. *Biometrika*, 70, 41-55.

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

Scott, S.L., and Ip, E.H. (2002). Empirical bayes and item-clustering effects in a latent variable hierarchical model: A case study from the national assessment of educational progress. *Journal of American Statistical Association,* 97, 459, 1-11.

Skrondal, A., and Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics,* 34, 712-745.

Tillé, Y., and Matei, A. (2012). *Sampling: Survey Sampling*. R package version 2.5.

Wright, B. (1996). Local dependency, correlations and principal components. *Rasch Meas Trans,* 10-3, 509-511.

Zhang, L.C. (2002). A method of weighting adjustment for survey data subject to nonignorable nonresponse. DACSEIS research paper no. 2, http://w210.ub.unituebingen.de/dbt/volltexte/2002/451.