Estimating classification errors under edit restrictions in composite survey-register data using Multiple Imputation Latent Class Modelling (MILC)

Boeschoten, Laura; Oberski, D.L.; de Waal, Ton

Published in:
Jounal of Official Statistics

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

DE GRUYTER OPEN

Journal of Official Statistics, Vol. 33, No. 4, 2017, pp. 921–962, http://dx.doi.org/10.1515/JOS-2017-0044

Estimating Classification Errors Under Edit Restrictions in Composite Survey-Register Data Using Multiple Imputation Latent Class Modelling (MILC)

Laura Boeschoten[1], *Daniel Oberski*[2], *and Ton de Waal*[3]

Both registers and surveys can contain classification errors. These errors can be estimated by making use of a composite data set. We propose a new method based on latent class modelling to estimate the number of classification errors across several sources while taking into account impossible combinations with scores on other variables. Furthermore, the latent class model, by multiply imputing a new variable, enhances the quality of statistics based on the composite data set. The performance of this method is investigated by a simulation study, which shows that whether or not the method can be applied depends on the entropy R^2 of the latent class model and the type of analysis a researcher is planning to do. Finally, the method is applied to public data from Statistics Netherlands.

1. Introduction

National Statistical Institutes (NSIs) often use large data sets to estimate population tables covering many different aspects of society. One way to create these rich data sets as efficiently and cost effectively as possible is to utilize already available register data. This has several advantages. First, known information is not collected again by means of a survey, saving collection and processing costs, as well as reducing the burden on the respondents. Second, registers often contain very specific information that could not have been collected by surveys (Zhang 2012). Third, statistical figures can be published more quickly, as conducting surveys can be time consuming. However, when more information is required than is already available, registers can be supplemented with survey data (De Waal 2016). Caution is then advised, as surveys likely contain classification errors. When a data set is constructed by integrating information at micro-level from both registers and surveys, we call this a composite data set. More information

[1] Tilburg University Tilburg School of Social and Behavioral Sciences – Methodology and Statistics, PO Box 90153, Tilburg 5000 LE, Netherlands and Centraal Bureau voor de Statistiek – Process development and methodology Henri Faasdreef 312, Den Haag 2492 JP, The Netherlands. Email: l.boeschoten@tilburguniversity.edu

[2] Universiteit Utrecht – Social and Behavioural Sciences, Utrecht, Utrecht, The Netherlands and Tilburg University Tilburg School of Social and Behavioral Sciences – Methodology and Statistics, Tilburg, The Netherlands. Email: d.l.oberski@uu.nl

[3] Centraal Bureau voor de Statistiek – Process development and methodology Den Haag, The Netherlands and Tilburg University Tilburg School of Social and Behavioral Sciences – Methodology and Statistics, Tilburg, The Netherlands. Email: T.deWaal@cbs.nl

Acknowledgments: The authors would like to thank the associate editor and the reviewers for their useful comments. Furthermore, the authors would like to thank Barry Schouten and Frank Bais for providing us with the application data.

Simulation code can be found on https://github.com/lauraboeschoten/MILC

on how to construct such a composite data set can be found in Zhang (2012) and Bakker (2010). Composite data sets are used by, among others, the Innovation Panel (Understanding Society 2016), the Millennium Cohort Study (UCL Institute of Education 2007), the Avon Longitudinal Study of Parents and Children (Ness 2004), the System of Social Statistical Databases of Statistics Netherlands, and the 2011 Dutch Census (Schulte Nordholt et al. 2014).

When using registers for research, we should be aware that they are collected for administrative purposes so they may not align conceptually with the target and can contain process delivered classification errors. These may be due to mistakes made when entering the data, delays in adding data to the register (Bakker 2009) or differences between the variables being measured in the register and the variable of interest (Groen 2012). This means that both registers and surveys may contain classification errors, although originating from different types of sources. This assumption is in contrast to what many researchers assume, namely that either registers or surveys are error-free. To illustrate, Schrijvers et al. (1994) used registers to validate a postal survey on cancer prevalence, Turner et al. (1997) used Medicare claims data to validate a survey on health status, and Van der Vaart and Glasner (2007) used optician database information to validate a telephone survey. In contrast, Jörgren et al. (2010) used a survey to validate the Swedish rectal cancer registry and Robertsson et al. (1999) used a postal survey to validate the Swedish knee arthroplasty register. Since neither surveys or registers are free of error, it is most realistic to approach them both as such. Therefore, we aim to develop a method which incorporates information from both to estimate the true value, without assuming that either one of them is error-free.

To distinguish between two types of classification errors, we classify them as either visibly or invisibly present. Both types can be estimated by making use of new information that is provided by the composite data set. Invisibly present errors in surveys or registers can be detected when responses on both are compared in the composite data set. Differences between the responses indicate that there is an error in one (or more) of the sources, although it is at this point unclear which score(s) exactly contain(s) error. The name 'invisibly present errors' is given because these errors could not have been seen in a single data set. They can be dealt with by estimating a new value using a latent variable model. To estimate these invisibly present errors using a latent variable model, multiple indicators from different sources within the composite data that measure the same attribute are used. This approach has previously been applied using structural equation models (Bakker 2012; Scholtus and Bakker 2013), latent class models (Biemer 2011; Guarnera and Varriale 2016; Oberski 2015) and latent markov models (Pavlopoulos and Vermunt 2015). Latent variable models are typically used in another context, namely as a tool for analysing multivariate response data (Vermunt and Magidson 2004).

Covariates (variables within the composite data set that measure something other than the attribute of interest) can help improve the latent variable model. Some errors can then be observed already when an impossible combination between a score on the attribute and a covariate is detected, which we define as a visibly present error. The name 'visibly present errors' is given here because (some of) these errors are visible in a single data set. An example of a combination which is not allowed is the score "own" on the variable *home ownership* and the score "yes" on the variable *rent benefit*. Such an, in practice,

impossible combination can be replaced by a combination that is deemed possible. Whether a combination of scores is possible and therefore "allowed" is commonly listed in a set of edit rules. An incorrect combination of values can be replaced by a combination that adheres to the edit rules. Different types of methods are used to find an optimal solution for different types of errors (De Waal et al. 2012). For errors caused by typing, signs or rounding, deductive methods have been developed by Scholtus (2009, 2011). For random errors, optimization solutions have been developed such as the Fellegi-Holt method for categorical data, the branch-and-bound algorithm, the adjusted branch-and-bound algorithm, nearest-neighbour imputation (De Waal et al. 2011, 115–156) and the minimum adjustment approach (Zhang and Pannekoek 2015). Furthermore, imputation solutions, such as nonparametric Bayesian multiple imputation (Si and Reiter 2013) and a series of imputation methods discussed by Tempelman (2007) can be used.

The solutions discussed two paragraphs above for invisibly present errors are not tailored to handle the invisibly and visibly present errors simultaneously, and they do not offer possibilities to take the errors into account in further statistical analyses; they only give an indication of the extent of the classification errors. In addition, uncertainty caused by both visibly and invisibly present errors is not taken into account when further statistical analyses are performed. An exception is the method developed by Kim et al. (2015), which simultaneously handles invisibly and visibly present errors using a mixture model in combination with edit rules for continuous data, and which has been extended by Manrique-Vallier and Reiter (2016) for categorical data. This method allows for an arbitrary number of invisible errors based on one file and one measurement, whereas we consider multiple linked files with multiple measurements of an attribute. Any method dealing with visibly or invisibly present classification errors should account for the uncertainty created by these errors. This can be done by making use of multiple imputations (Rubin 1987), and has previously been used in combination with solutions for invisibly present errors (Vermunt et al. 2008) and visibly present errors (Si and Reiter 2013; Manrique-Vallier and Reiter 2013).

We propose a new method that simultaneously handles the three issues discussed: it handles both visibly and invisibly present classification errors and it incorporates them both, as well as the uncertainty created by them, when performing further statistical analysis. By comparing responses on indicators measuring the same attribute in a composite data set we allow the estimation of the number of invisibly present errors using a Latent Class (LC) model. Visibly present errors are handled by making use of relevant covariate information and imposing restrictions on the LC model. In the hypothetical cross table between the attribute of interest and the restriction covariate, the cells containing a combination that is in practice impossible are restricted to contain zero observations. These restrictions are imposed directly when the LC model is specified. To also take uncertainty created by the invisibly and visibly present errors into account when performing further statistical analyses, we make use of Multiple Imputation (MI). Because MI and LC are combined in this new method, the method will be further denoted as MILC.

In the following section, we describe the MILC method in more detail. In the third section, a simulation study is performed to assess the novel method. In the fourth section, we apply the MILC method on a composite data set from Statistics Netherlands.

2. The MILC Method

The MILC method takes visibly and invisibly present errors into account by combining Multiple Imputation (MI) and Latent Class (LC) analysis. Figure 1 gives a graphical overview of this procedure. The method starts with the original composite data set comprising L measures of the same attribute of interest. In the first step, m bootstrap samples are taken from the original data set. In the second step, an LC model is estimated for every bootstrap sample. In the third step, m new empty variables are created in the original data set. The m empty variables are imputed using the corresponding m LC models. In the fourth step, estimates of interest are obtained from the m variables and in the last step, the estimates are pooled using Rubin's rules for pooling (Rubin 1987, 76). These five steps are now discussed in more detail.

The MILC method starts by taking m bootstrap samples from the original composite data set. These bootstrap samples are drawn because we want the imputations we create in a later step to take parameter uncertainty into account. Therefore, we do not use one LC model based on one data set, but we use m LC models based on m bootstrap samples of the original data set (Van der Palm et al. 2016).

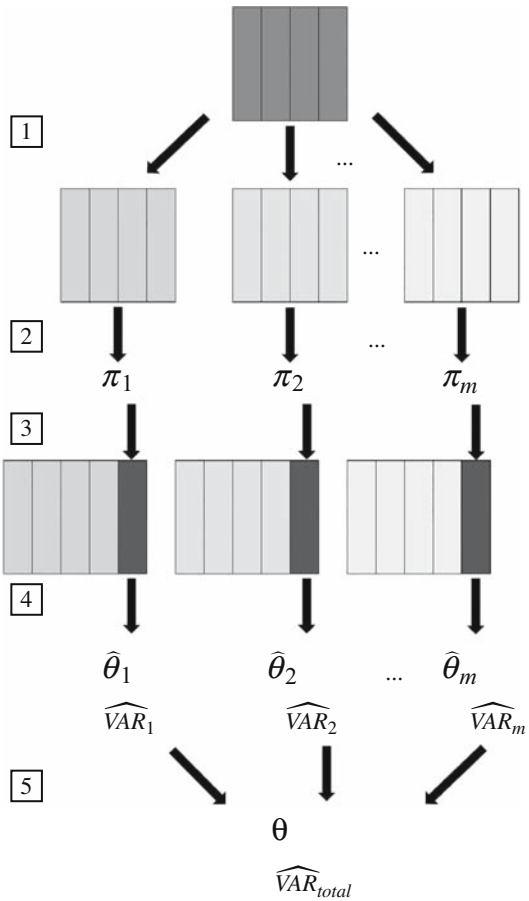

Fig. 1. Procedure of latent class multiple imputation for a multiply observed variable in a composite data set.

In the next step, we make use of LC analysis to estimate both visibly and invisibly present classification errors in categorical variables. We first link several data sets by unit identifiers, resulting in a composite data set matched on a common core set of identifiers (discarding all records where no match is obtained), and group variables measuring the same attribute present on more than one of the original source data sets. For each of the variable groups, we build a single latent variable (denoted by X) representing the underlining true measure, assuming discrepancies between different sourced measures.

For example, we have L dichotomous indicator variables ($Y_1, \ldots, Y_L$) measuring the same attribute *home ownership* (1 = "own", 2 = "rent") in multiple data sets linked on unit level. Differences between the responses of a unit are caused by what we described as invisibly present classification error in one (or more) of the indicators. Since the indicators all have an equal number of categories (C), we fix the number of categories of the latent variable X to C.

The LC model we then build using the indicator variables is based on five assumptions. The first assumption pertains to the marginal response pattern $\mathbf{y}$, which is a vector of the responses to the given indicators. For example, we have three indicators measuring home ownership, the response pattern $\mathbf{y}$ can be "own", "own", "rent". We assume here that the probability of obtaining this specific marginal response pattern $P(\mathbf{Y} = \mathbf{y})$ is a weighted average of the X class specific probabilities $P(\mathbf{Y} = \mathbf{y}|X = x)$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x)P(\mathbf{Y} = \mathbf{y}|X = x). \tag{1}$$

Here, $P(X = x)$ denotes the proportion of units belonging to category x in the underlying true measure, where x might be "own", the proportion of the population owning their own house.

The second assumption is that the observed indicators are independent of each other given a unit's score on the underlying true measure. This means that when a mistake is made when filling in a specific question in a survey, this is unrelated to what is filled in for the same question in another survey or register. This is called the assumption of local independence,

$$P(\mathbf{Y} = \mathbf{y}|X = x) = \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{2}$$

Combining Equation (1) and Equation (2) yields the following model for response pattern $P(\mathbf{Y} = \mathbf{y})$:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} P(X = x)\prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{3}$$

The model parameters ($P(X = x)$ and $P(Y_l = y_l|X = x)$) are estimated by Maximum Likelihood (ML). To find the ML estimates for the model parameters, Latent Gold uses both the Expectation-Maximization and the Newton-Raphson algorithm (Vermunt and Magidson 2013a).

In Equation (3), only the indicators are used to estimate the likelihood of being in a specific true category. However, it is also possible to make use of covariate information to estimate the LC model. The third assumption we then make is that the measurement errors are independent of the covariates. An example of a covariate which can help in identifying whether someone owns or rents a house is *marital status*, this covariate is denoted by Q and can be added to Equation (3):

$$P(\mathbf{Y} = \mathbf{y}|Q = q) = \sum_{x=1}^{C} P(X = x|Q = q) \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{4}$$

Covariate information can also be used to impose a restriction on the model, to make sure that the model does not create a combination of a category of the "true" variable and a score on a covariate that is in practice impossible. For example, when an LC model is estimated to measure the variable *home ownership* using three indicator variables and a covariate (denoted by Z) measuring *rent benefit*, the impossible combination of owning a house and receiving rent benefit should not be created.

Throughout the article, we compare four approaches that researchers might administer when performing analyses using composite data sets containing classification errors and edit restrictions. In the first approach, researchers completely ignore the composite data structure and directly use one variable (which measures a construct that is measured by other variables in the composite data set as well) to obtain estimates of interest, for example a cross-table proportion or a logistic regression coefficient. In the second approach, researchers use an LC model to correct for classification errors, but are not aware of the edit restriction. The LC model used in this approach is equal to Equation (4); we call this the *unconditional model*. In the third approach, researchers are aware of the edit restriction, but they assume that including the restriction covariate (Z) in the LC model is enough to account for this; they do not explicitly mention the restriction itself. We call this the *conditional model*:

$$P(\mathbf{Y} = \mathbf{y}|Q = q, Z = z) = \sum_{x=1}^{C} P(X = x|Q = q, Z = z) \prod_{l=1}^{L} P(Y_l = y_l|X = x). \tag{5}$$

Only in the fourth approach, the restriction is imposed directly in the LC model to fix the cell proportion of the impossible combination to 0; we call this the *restricted conditional model*. In the example where Z measures *rent benefit*, and the latent "true" variable measures *home ownership*, the imposed restriction is:

$$P(X = \text{own}|Z = \text{rent benefit}) = 0. \tag{6}$$

By using such a restriction, we can take impossible combinations with other variables into account, while we estimate an LC model for the underlying true measure. The restriction is imposed by specifically denoting which cell in the cross-table between the covariate and the latent variable should contain zero observations and giving this cell a weight of 0, resulting in constrained estimation (Vermunt and Magidson 2013b).

By specifying a model as in Equation (4) or in Equation (5), we assume that the covariate measure is in fact error-free, which is the fourth assumption we make. A fifth

assumption is that the edit rules applied are hard edit rules,in contrast to soft edit rules where there is a small probability that the edit is in fact possible. These five assumptions (assumption that $P(\mathbf{Y} = \mathbf{y})$ is a weighted average of $P(\mathbf{Y} = \mathbf{y}|X = x)$; assumption of local independence; assumption that measurement errors are independent of covariates; assumption that the covariate is error-free; assumption of hard edits) are specific for the LC model we use.

However, in practice it is very likely that one of these assumptions is not met. For example, with the assumption of local independence, we assume that when a mistake is made in one indicator, this is unrelated to the answers on other indicators. This assumption is probably met when one indicator originates from a survey and another from a register. If two indicators both originate from surveys, it is much more likely that a respondent makes the same mistake in both surveys, this assumption would then not be met. We can also think of situations where the assumption that misclassification is independent of covariates is not met. For example with tax registration by businesses, the number of delays and mistakes tends to be related to company size, since appropriate administration is better institutionalized in larger companies. The assumption that a covariate is free of error is in practice almost never met, since all sources always contain some error. The last assumption made is that the edits applied are hard edits. In some cases soft edits might be more appropriate, for example when a combination of scores is highly unlikely but not impossible, such as the combination of being ten years old and having graduated from high school.

Luckily these assumptions can be relaxed by specifying more complex LC models. However, whether you are able to relax these assumptions depends on your specific data structure. More specifically, it depends on whether your model is still identifiable. Unfortunately, model identifiability is not straightforward. For example, a model with three dichotomous indicators is identifiable, while a model with two dichotomous indicators is not. Adding a covariate to this model would make it identifiable. Adding a restriction to a model can also help to make an unidentifiable model identifiable. Since it is not possible to present general recommendations here, we refer to Biemer (2011) for more information about model identifiability. Examples of complex latent variable models which incorporate the different assumptions discussed in official statistics data sets are Pavlopoulos and Vermunt (2015) and Scholtus and Bakker (2013). Model identification can be checked in Latent Gold by assessing whether the Jacobian of the likelihood is full rank at a larger number of random parameter values (Forcina 2008). All models in this article were confirmed to be identifiable.

How missing values in the indicators and covariates are handled is also dependent on model specification. We specified the model as such that the indicators are part of the estimation procedure. Missing values are therefore handled by Full Information Maximum Likelihood (FIML) (Vermunt and Magidson 2013b, 51–52). Covariates are treated as fixed and listwise deletion will be applied to missing values here.

By applying Bayes' rule to the LC models from Equation (4), Equation (5), or Equation (6), posterior membership probabilities can be obtained. These posterior membership probabilities represent the probability of being in an LC given a specific combination of scores on the indicators and covariates ($P(X = x|Y = y, Q = q, Z = z)$).

For example, the posterior membership probabilities for the *conditional model* are obtained by:

$$P(X=x|Y=y,Q=q,Z=z)=\frac{P(X=x|Q=q,Z=z)\prod_{l=1}^{L}P(Y_l=y_l|X=x)}{\sum_{x=1}^{C}P(X=x|Q=q,Z=z)\prod_{l=1}^{L}P(Y_l=y_l|X=x)}. \quad (7)$$

These posterior membership probabilities can be used to impute latent variable X. To distinguish between the unobserved latent variable X, described by the LC model, and the variable after imputation, we denote this imputed variable by W. Different methods exist to obtain W. An example is modal assignment, where each respondent is assigned to the class for which its posterior membership probability is the largest. To correctly incorporate uncertainty caused by the classification errors, we use multiple imputation to estimate W. We first create m empty variables $(W_1,\ldots,W_m)$ and we impute them by drawing one of the LCs by sampling from the posterior membership probabilities from the m LC models.

With the *restricted conditional model*, we want to make sure that cases are not assigned to categories on the latent "true" variable which would result in impossible combinations with scores on other variables, such as the combination "rent benefit" × "own". Therefore, the restriction set in Equation (6) is also used here.

After we created m variables by imputing them using the posterior membership probabilities obtained from each of the m LC models, the estimates of interest can be obtained. For example, we can be interested in a cross table between imputed "true" variable W and covariate Z, where our estimate of interest $\hat{\theta}$ can be the cell proportion $P(W=1,Z=1)$. The m estimates of $\hat{\theta}$ can now be pooled by making use of the rules defined by Rubin for pooling (Rubin 1987, 76). The pooled estimate is obtained by

$$\hat{\theta}=\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i. \quad (8)$$

The total variance is estimated as

$$\mathrm{VAR}_{\mathrm{total}}=\overline{\mathrm{VAR}}_{\mathrm{within}}+\mathrm{VAR}_{\mathrm{between}}+\frac{\mathrm{VAR}_{\mathrm{between}}}{m}, \quad (9)$$

where $\overline{\mathrm{VAR}}_{\mathrm{within}}$ is the within imputation variance calculated by

$$\overline{\mathrm{VAR}}_{\mathrm{within}}=\frac{1}{m}\sum_{i=1}^{m}\mathrm{VAR}_{\mathrm{within}_i}. \quad (10)$$

$\mathrm{VAR}_{\mathrm{within}_i}$ is estimated as the variance of the proportion of $\hat{\theta}_i$,

$$\frac{\hat{\theta}_i\times(1-\hat{\theta}_i)}{N}, \quad (11)$$

where N is the number of units in the composite data set, and $\mathrm{VAR}_{\mathrm{between}}$ is calculated by

$$\mathrm{VAR}_{\mathrm{between}}=\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i-\hat{\theta})(\hat{\theta}_i-\hat{\theta})'. \quad (12)$$

Besides the uncertainty caused by missing or conflicting data represented by the spread of parameter estimate values, $VAR_{between}$ also contains parameter uncertainty, which was introduced by the bootstrap performed in the first step of the MILC method.

3. Simulation

3.1. Simulation Approach

To empirically evaluate the performance of MILC, we conducted a simulation study using R (R Core Team 2014). We start by creating a theoretical population using Latent Gold (Vermunt and Magidson 2013a) containing five variables: three dichotomous indicators (Y_1, Y_2, Y_3) measuring the latent dichotomous variable (X); one dichotomous covariate (Z) which has an impossible combination with a score of the latent variable; and one other dichotomous covariate (Q). The theoretical population is generated using the restricted conditional model. When samples are drawn, it can happen that the LC model estimated from a sample assigns a non-zero probability to an impossible combination, so these errors are due to sampling. Furthermore, variations are made in the generated data sets according to scenarios described in the following sections.

When evaluating an imputation method, the relation between the imputed latent variable and other variables should be preserved since these relations might be the subject of research later on. When investigating the performance of MILC, there are two relations we are particularly interested in. We are interested in the relation between the imputed latent variable W and the covariate Z, which has an impossible combination with a score on the latent variable. The four cell proportions of the 2×2 table are denoted by: $W1 \times Z_1$, $W_2 \times Z_1$, $W_1 \times Z_2$ and $W_2 \times Z_2$. The cell $W_1 \times Z_2$ is the impossible combination, and should contain 0 observations. We compare the cell proportions of a 2×2 table of the population latent variable X and Z with the cell proportions of a table of the imputed latent variable W and Z from the samples. Furthermore, we are interested in the relation between W and covariate Q. To investigate this relation, we compare the coefficient of a logistic regression of the latent population variable X on Q with the logistic regression coefficient of the imputed W regressed on Q.

To investigate these relations, we look at three performance measures. First, we look at the bias of the estimates of interest. The bias is equal to the difference between the average estimate over all replications and the population value. Next, we look at the coverage of the 95% confidence interval. This is equal to the proportion of times that the population value falls within the 95% confidence interval constructed around the estimate over all replications. To confirm that the standard errors of the estimates were properly estimated, the ratio of the average standard error of the estimate over the standard deviation of the 1,000 estimates was also examined.

We expect the performance of MILC to be influenced by the measurement quality of the indicators, the marginal distribution of covariates Z and Q, the sample size, and the number of multiple imputations. The quality of the indicators is represented by classification probabilities. They represent the probability of a specific score on the indicator given the latent class. If the quality of the indicators is low, it will be more difficult for MILC to assign cases to the correct latent classes.

From Geerdinck et al. (2014) we know that classification probabilities of 0.95 and higher can be considered realistic for population registers. Pavlopoulos and Vermunt (2015) detected a classification probability of 0.83 in the Dutch Labour Force Survey. We investigate a range of classification probabilities around the values found, from 0.70 to 0.99. The marginal distribution of Z, $P(Z)$, is also expected to influence the performance of MILC. A higher value for $P(Z = 2)$ can give, for example, more information to the latent class model to assign scores to the correct latent class. Sample size may influence the standard errors and thereby the confidence intervals. The performance of MILC can also depend on the number of multiple imputations. Investigation of several multiple imputation methods have shown that five imputations are often sufficient (Rubin 1987). However, with complex data, it can be the case that more imputations are needed. As a result, the simulation conditions can be summarized as follows:

- Classification probabilities: 0.70; 0.80; 0.90; 0.95; 0.99.
- $P(Z = 2)$: 0.01; 0.05; 0.10; 0.20.
- Sample size: 1,000; 10,000.
- Logit coefficients of X regressed on Q of $\log(0.45/(1 - 0.45) = -0.2007$, $\log(0.55/(1 - 0.55) = 0.2007$ and $\log(0.65/(1 - 0.65) = 0.6190$ corresponding to estimated odds ratio of 0.81, 1.22 and 1.86. The intercept was fixed to 0
- Number of imputations: 5; 10; 20; 40.

To illustrate the measurement quality corresponding to different conditions, Figure 2 shows the entropy R^2 of the models under different values for $P(Z = 2)$ and classification probabilities. The entropy indicates how well one can predict class membership based on

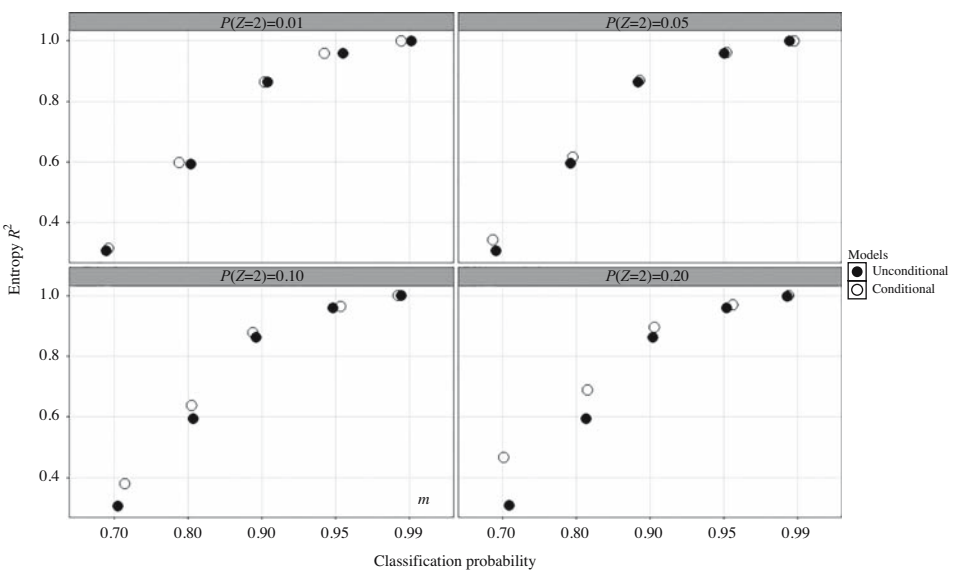

Fig. 2. Entropy R^2 of the unconditional and conditional model with different values for the classification probability and $P(Z = 2)$. The restricted conditional model has the same entropy R^2 as the conditional model because the models contain the same variables.

the observed variables, and is measured by:

$$EN(\alpha) = -\sum_{j=1}^{N}\sum_{x=1}^{X}\alpha_{jx}log\alpha_{jx}, \tag{13}$$

where α_{jx} is the probability that observation j is a member of class x, and N is the number of units in the composite data set. Rescaled with values between 0 and 1, entropy R^2 is measured by

$$R^2 = 1 - \frac{EN(\alpha)}{N\log X}, \tag{14}$$

where 1 means perfect prediction (Dias and Vermunt 2008). The *conditional* and the *restricted conditional model* have the same entropy R^2 because these models contain the same variables. All models with classification probabilities of 0.90 and above have a high entropy R^2 and are able to predict class membership well. When the classification probabilities are 0.70, the entropy R^2 is especially low. However, for the conditional and the restricted conditional model, the entropy R^2 under classification probability 0.70 increases as $P(Z = 2)$ increases. A larger $P(Z = 2)$ means that covariate Z contains more information for predicting class membership. Because covariate Z is not in the *unconditional model*, it makes sense that entropy R^2 remains stable for different values of $P(Z = 2)$ under this model. Furthermore, Figure 2 demonstrates that the performance of MILC is evaluated over an extreme range of entropy R^2 values and gives an indication of what we can expect from the MILC method under different simulation conditions.

3.2. Simulation Results

In this section we discuss our simulation results in terms of bias, coverage of the 95% confidence interval, and the ratio of the average standard error of the estimate over the standard deviation of the estimates. We do this in three sections. In the first section we discuss the 2 × 2 table of the imputed latent variable W and restriction covariate Z. In the second section, we investigate the relation between the imputed latent variable W and covariate Q. In the third section we investigate the influence of m, the number of bootstrap samples and multiple imputations. In the simulation results discussed in the first two sections, we used $m = 5$. When investigating the different simulation conditions, we focus on the performance of the four approaches discussed, using one indicator (Y_1), the *unconditional model*, the *conditional model* and the *restricted conditional model*. Interesting findings are illustrated with graphs containing results from situations when Y_1 is used and W is estimated using the restricted conditional model. For conditions that yielded approximately identical results, only one condition is shown in the figures. In Appendix A, tables with all results from the four approaches are given.

3.2.1. The Relation of Imputed Latent Variable W with Restriction Covariate Z

When we investigate the results in terms of bias (Figure 3), the restricted conditional model produces bias when the classification probabilities of the indicators are below 0.80. The bias of the cells where $P(Z = 1)$ for the restricted conditional model decreases when the classification probabilities increase or when $P(Z = 2)$ increases. This trend coincides

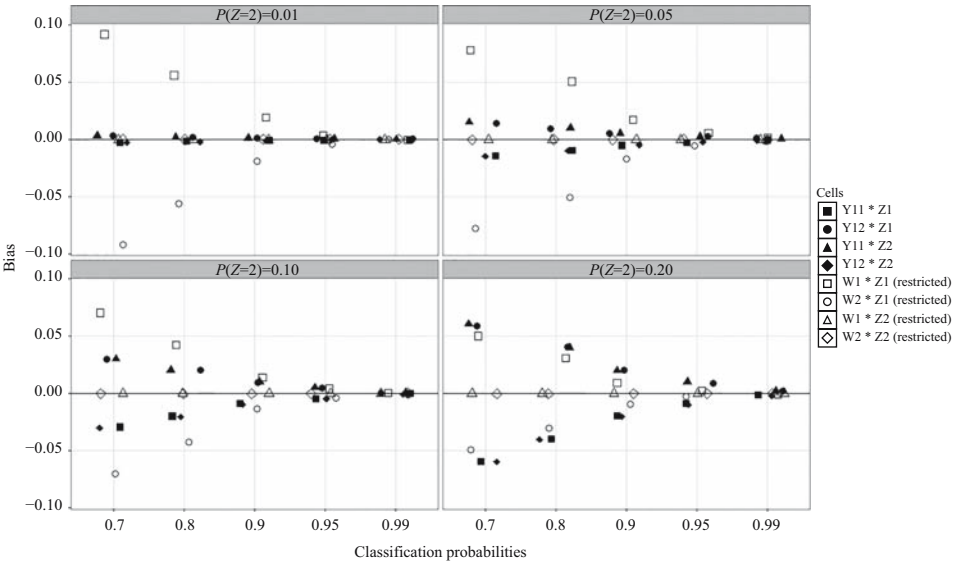

Fig. 3. Bias of the four cell proportions of the 2 × 2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $P(Z = 2)$. Sample size is 1,000 and $m = 5$.

with the trend we saw in Figure 2 for the entropy R^2, where a high entropy R^2 corresponds to a low bias. In contrast, when Y_1 is used, the bias of all cells is low when $P(Z = 2)$ is small, and increases as $P(Z = 2)$ increases. Furthermore, the restricted conditional model is the only model in which the cell representing the impossible combination $(W_1 \times Z_2)$ indeed contains 0 observations. $(Y1_1 \times Z_2)$ is never exactly 0.

When investigating the results for coverage of the 95% confidence intervals around the cell proportions (Figure 4), we see that the results differ over the different sample sizes. This is caused by the fact that even though the bias is not influenced by the sample size, the standard errors and therefore the confidence intervals are. Confidence intervals of biased estimates are therefore less likely to contain the population value. Furthermore, if the classification probabilities are larger,individuals are more likely to end up in the correct latent class, which also results in less variance, resulting in smaller confidence intervals. Confidence intervals cannot pe properly estimated for the impossible combination $Y1_1 \times Z_2$, since the proportions are very close to 0. This can be seen in Figure 4. Since $W_1 \times Z_2$ is not estimated with the restricted conditional model, confidence intervals cannot be estimated and coverage is therefore not shown.

The ratio of the average standard error of the estimate over the standard deviation of the simulated estimates tells us whether the standard errors of the estimates are properly estimated. In general, the values for both the situation of one indicator and the restricted conditional model, found in Figure 5, are both very close to 1. Only the standard errors for $W_1 \times Z_2$ are too small when one indicator is used. With the restricted conditional model, these are not estimated.

Overall, the small 2 × 2 cross tables investigated here containing a restriction covariate can be estimated when the LC model of the composite data set has an entropy R^2 of 0.90, or, when the sample size is large, an entropy R^2 of 0.95.

3.2.2. Relationship Between the Imputed Latent Variable W and Covariate Q

In the simulation results discussed in Subsection 3.2.1, the relation between the imputed latent variable W and covariate Z containing an impossible combination was investigated. Within the restricted conditional model, there was also another covariate, Q. We investigate the relation between W and Q with three different strengths of relations: intercepts are 0 and logit coefficients of W regressed on Q are -0.2007; 0.2007; 0.6190. Because the intercept is 0 in all conditions, we focus on the coefficients of Q when investigating the simulation results.

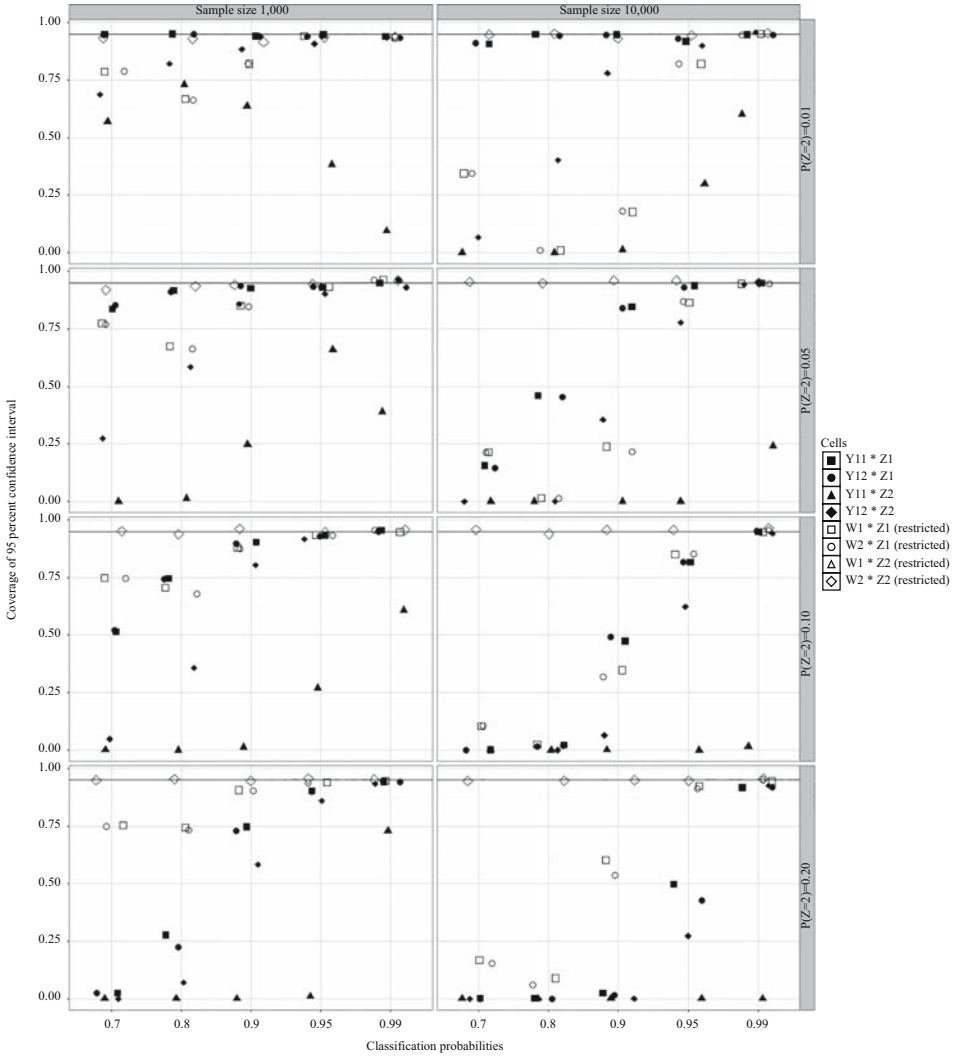

Fig. 4. Coverage of the 95% confidence interval of the four cell proportions of the 2 × 2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $P(Z = 2)$ and sample size, $m = 5$.

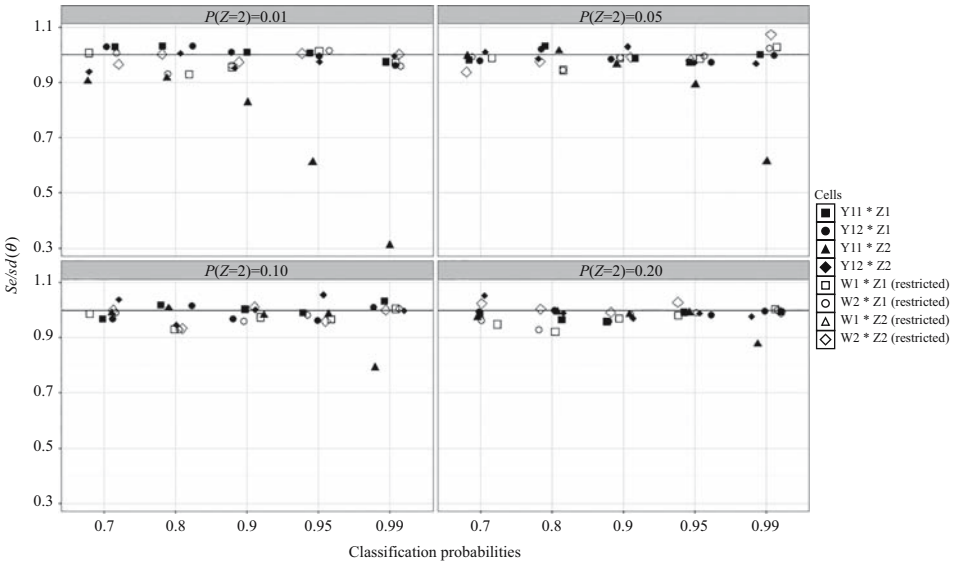

Fig. 5. $se/sd(\hat{\theta})$ of the four cell proportions of the 2 × 2 table of $Y_1 \times Z$ and $W \times Z$. W is estimated using the restricted conditional model. Results are shown for different values of the classification probabilities and $P(Z = 2)$. Sample size is 1,000 and $m = 5$.

In Figure 6 we see that for the restricted conditional model, the bias is very close to 0 in all conditions. When Y_1 is used, the bias is much larger and is related to the classification probabilities.

In Figure 7 we see the results in terms of coverage of the 95% confidence interval. The conclusions we can draw here are comparable to the conclusions we drew from the results in terms of bias. When W is used (estimated using the restricted conditional model), the

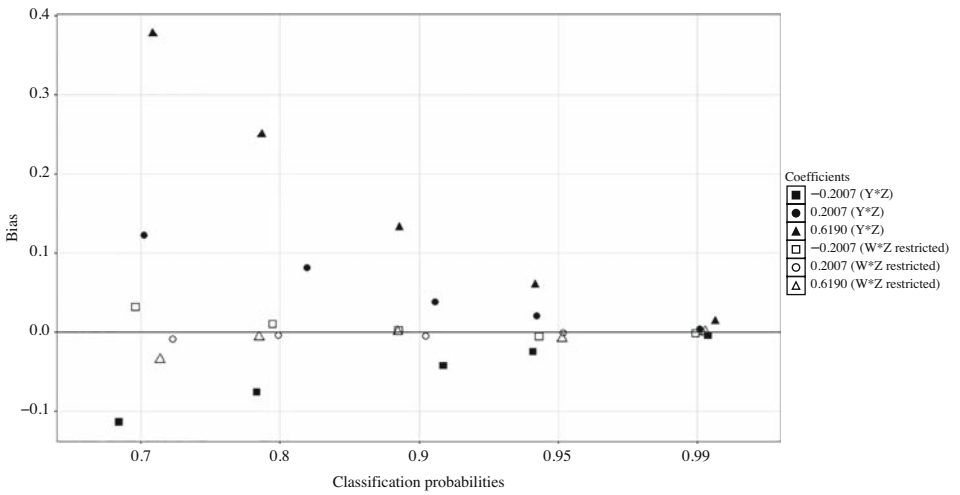

Fig. 6. Bias of the logistic regression coefficient of Y_1 regressed on covariate Q and of W regressed on Q. W is estimated using the restricted conditional model. Results are shown for different values of the logistic regression coefficient and the classification probabilities. $P(Z = 2) = 0.01$, sample size is 1,000 and $m = 5$.

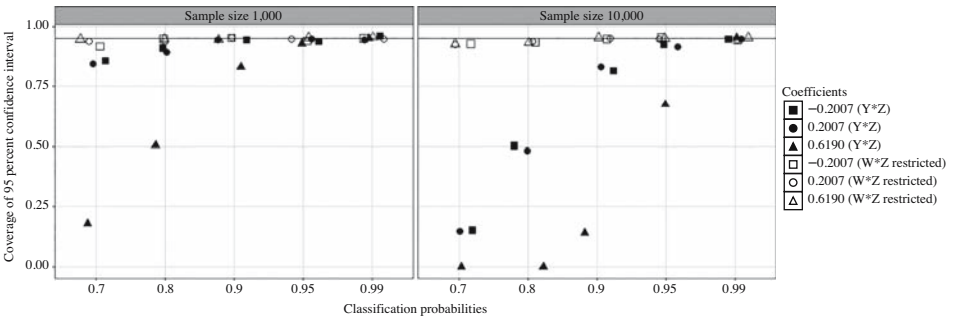

Fig. 7. Coverage of the 95% confidence interval of the logistic regression coefficient of Y_1 regressed on covariate Q and of W regressed on Q. W is estimated using the restricted conditional model. Results are shown for different values of the logistic regression coefficient, the classification probabilities and sample size. $P(Z = 2) = 0.01$ and $m = 5$.

coverage of the 95% confidence is approximately 95 in all discussed conditions. When only one indicator (Y_1) is used, we see undercoverage when the population value of the logistic regression coefficient is 0.6190. This undercoverage is related to the classification probabilities and increases when the sample size increases. Results in terms of the ratio of the average standard error of the estimate over the standard deviation of the simulated estimates are very close to the desired ratio of 1. This is the case for all investigated simulation conditions, both when Y_1 is used or when W is used. Results are reported in Appendix A.

Overall, for the investigated conditions, unbiased estimates can be obtained when the LC model of the composite data set has an entropy R^2 of 0.60 or larger.

3.2.3. Number of Imputations

To investigate the effect of the number of bootstrap samples and imputations (m), we performed 5, 10, 20, and 40 bootstrap samples and imputations. The results of $m = 5$ and $m = 40$ can be found in Figure 8, while more results can be found in Appendix A. Both in terms of bias and coverage the MILC method performs equally well over the different numbers of m. It is important to note that the fraction of missing information corresponds, in the worst case, to the amount of missing data (Rubin 1987, 114). In our case, it depends on the entropy R^2, which is dependent on the classification and the covariates. Although

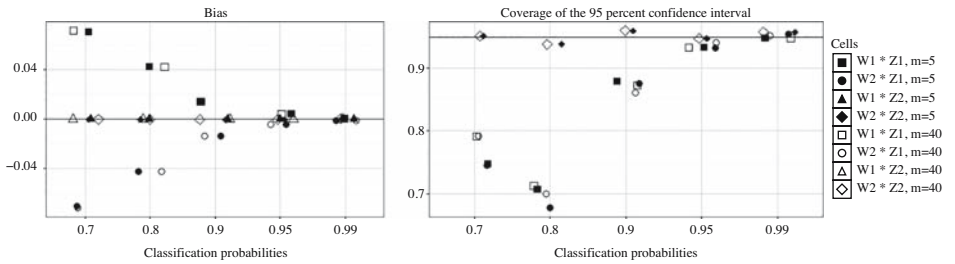

Fig. 8. Bias and coverage of the 95% confidence interval of four cells in the 2 × 2 table of covariate Z × W (estimated using the restricted conditional model). Number of bootstrap samples m = 5 and 40. The sample size is 1,000 and P(Z = 2) = 0.10.

the amount of missing values in W is 100%, the amount of missing information is much smaller when the entropy R^2 is larger than 0. This might explain why biased estimates with inappropriate coverage are obtained when the entropy R^2 is low, regardless of the size of m.

4. Application

4.1. Data

Home ownership is an interesting variable for social research. It has been related to a number of properties, such as inequality (Dewilde and Decker 2016), employment insecurity (Lersch and Dewilde 2015) and government redistribution (André and Dewilde 2016). Therefore, we apply the MILC method on a composite data set that brings together survey data from the LISS (Longitudinal Internet Studies for the Social sciences) panel from 2013 (Scherpenzeel 2011), which is administered by CentERdata (Tilburg University, The Netherlands) and a population register from Statistics Netherlands from 2013. Because samples for LISS were drawn by Statistics Netherlands, we were very well able to link these surveys and registers. From this composite data set, we use two variables indicating whether a person is either a home-owner or rents a house/other as indicators for the imputed "true" latent variable *home-owner/renter or other*. The composite data set also contains a variable measuring whether someone receives rent benefit from the government. A person can only receive rent benefit if this person rents a house. In a cross-table between the imputed latent variable *home-owner/renter* and *rent benefit*, there should be 0 persons in the cell "home-owner *s* receiving rent benefit". If people indeed receive rent benefit and own a house, this could be interesting for researchers and requires investigation. A more detailed LC model should then be specified, modelling local dependencies and allowing for error in the variable 'rent benefit'. However, this is outside the scope of the present study. We assume this to be measurement error, and therefore want this specific cell to contain 0 persons. Research has previously been done regarding the relation between home ownership and marital status (Mulder 2006). A research question here could be whether married individuals more often live in a house they own compared to non-married individuals. Therefore, a variable indicating whether a person is married or not is included in the latent class model as a covariate. The three data sets used to combine the data are discussed in more detail below:

- **Registration of addresses and buildings (BAG):** A register with data on addresses containing information about its buildings, owners and inhabitants originating from municipalities from 2013. Register information is obtained from persons who filled in the LISS studies and who declared that we are allowed to combine their survey information with registers. In total, this left us with 3,011 individuals. From the BAG we used a variable indicating whether a person "owns"/"rents"/"other" the house he or she lives in. Because our research questions mainly relate to home-owners, we recoded this variable into "owns"/"rents or other". This variable does not contain any missing values.
- **LISS background study:** A survey on general background variables from January 2013. From this survey we also have 3,011 individuals. We used the variable *marital*

status, indicating whether someone is "married"/"separated"/"divorced"/"widowed"/"never been married". As we are only interested in whether a person is married or not, we recoded this variable in such a way that "married" and "separated" individuals are in the recoded "married" category, and the "divorced", "widowed" and "never been married" individuals are in the "not married" category. It is difficult to handle a category as "separated" in such a situation. However, separated individuals are technically still married. Although they can in theory be more likely to live out of the registered address, it is difficult to make assumptions and therefore we decided to recode them into the category "married". This variable did not contain any missing values. We also used a variable indicating whether someone is a "tenant"/"sub-tenant"/"(co-) owner"/"other". We recoded this variable in such a way that we distinguish between "(co-) owner" and "(sub-) tenant or other". This variable had 14 missing values.

- **LISS housing study:** A survey on housing from June 2013. From this survey we used the variable *rent benefit*, indicating whether someone "receives rent benefit"/"the rent benefit is paid out to the lessor"/"does not receive rent benefit"/"prefers not to say". Because we are not interested in whether someone receives the rent benefit directly or indirectly, we recoded the first two categories into "receiving rent benefit". No one selected the option "prefers not to say". For this variable, we had 2,232 missing values resulting in 779 observations. The number of observations is small, because a selection variable (indicating whether someone rents their house) was used in the survey. Dependent interviewing has been used here. Only the individuals indicating that they rent their house in this variable were asked if they receive rent benefit. This selection variable could also have been used as an indicator in our LC model. However, because of the strong relation between this variable and the rent benefit variable we decided to leave it out of the model.

These data sets are linked at person level where matching is done on person identification numbers. In addition, matching could also have been done on date, since the surveys were conducted at different time points within 2013. However, mismatches on dates are a source of measurement error, and are therefore left in for illustration purposes. Although it is not necessarily the case in practice, the assumption is made that the covariate 'rent benefit' is measured without error, so we are able to apply the LC model investigated in the simulation study in practice. In Table 1, it can be seen that 48 individuals rent a home according to the BAG register, while stating to own a home in the LISS background survey. Furthermore, 155 individuals own a home according to the BAG register, while stating that they rent a home in the LISS background survey.

Table 1. Cross-table between the own/rent variable originating from the LISS background survey and the own/rent variable originating from the BAG register.

		Register	
		Rent	Own
Background survey	Rent	902	155
	Own	48	1,892

Not every individual is observed in every data set. This causes some missing values to be introduced when the different data sets are linked at a unit level. These records are not missing, but they are considered as non-sampled individuals. Full Information Maximum Likelihood was used to handle the missing values in the indicators (Vermunt and Magidson 2013b, 51–52).

The MILC method is applied to impute the latent variable *home owner/renter* by using two indicator variables and two covariates and the *restricted conditional model*. For results when the *unconditional* and the *conditional* model are applied we refer to Appendix B. In Table 2 classification statistics about the model is given, indicating how we can compare the results of this model to the information we obtained in the simulation study. Both the entropy R^2 and the classification probabilities are comparable to conditions we tested in the simulation study and in which the MILC method appeared to work very well. The classification probabilities for the LISS background survey and the BAG register indicate that theyboth have a high quality, but are error prone. Furthermore, P(married) and P(rent benefit) cannot be compared directly to the set up of the simulation study, but information provided by the covariates is taken into account in the entropy R^2.

For the two variables measuring home ownership, we can see from the cell totals in Table 3 whether individuals who say to own their home also receive rent benefit, which is not allowed. However, in practice these discrepancies can be caused by the fact that people make mistakes when filling in a survey, or for example because people were moving during the period the surveys took place. Furthermore, the total number of individuals who can be found in the table of the LISS background study are only 779, and for the BAG register 772. This is because only the people indicating that they rented a house in the LISS Housing study were asked the question whether they received rent benefit. For the LISS background study we see that eight individuals are in the cell representing the impossible combination of owning a house and receiving rent benefit, and for the BAG register 4. If we investigate the cell proportions estimated by the MILC method, we see that both the conditional and the unconditional model replicate the structure of the indicators very well, but that individuals are still assigned to the cell of the impossible combination (see Appendix B). To get this correctly estimated, we need the restricted conditional model. The marginals of the variable *own/rent* (in the upper block of Table 3) for the different models are all very close to each other, and closer to the estimates in the BAG register than to the estimates of the LISS background study. Also note that individuals with missing

Table 2. Entropy R^2 of the restricted conditional model; classification probabilities of the indicators and marginal probabilities of the covariates. The covariate rent benefit takes information of 779 individuals into account and marital status variable of 3,011 individuals.

			Restricted conditional model
Entropy R^2			0.9380
Classification probability	LISS background	P(rent|LC rent)	0.9344
		P(own|LC own)	0.9992
	BAG register	P(rent|LC rent)	0.9496
		P(own|LC own)	0.9525
P(rent benefit)			0.3004
P(married)			0.5284

Table 3. The first block represents the (pooled) marginal proportions of the variable own/rent. The second block represents the (pooled) proportions of the variable own/rent for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable own/rent for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last row represents the restricted conditional model used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.

	P(own)		P(rent)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.6450	[0.6448; 0.6451]	0.3550	[0.3549; 0.3511]
LISS background	0.6830	[0.6829; 0.6832]	0.3170	[0.3168; 0.3171]
Restricted conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
	P(own × rent benefit)		P(rent × rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
LISS background	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
Restricted conditional	0.0000	-	0.2978	[0.2649; 0.3307]
	P(own × no rent benefit)		P(rent × no rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
LISS background	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
Restricted conditional	0.0213	[−0.0116; 0.0542]	0.6773	[0.6444; 0.7102]

values on the variable *rent benefit* are not taken into account in the 2 × 2 table of *rent benefit* × *own/rent*.

After we investigated the cross table between home ownership and rent benefit, we were also interested in whether marriage can predict home ownership. When we consider the BAG register, we see that the estimated odds of owning a home when not married are $e^{-1.2331} = 0.29$ times the odds when married, while they are $e^{-1.3041} = 0.27$ when the LISS background survey is used. It is interesting to see that when the restricted conditional MILC model is used to obtain an estimate that also corrects for the impossible combination of owning a house and receive rent benefit, we see that this coefficient is even a little less strong, namely $e^{-1.3817} = 0.25$. Overall, these results show us that although non-married individuals are approximately equally likely to own or rent a house, married individuals are three times more likely to own a house than to rent one.

5. Discussion

In this article we introduced the MILC method, which combines latent class analysis with edit restrictions and multiple imputation to obtain estimates for variables of which we had multiple indicators in a composite data set. We distinguished between invisibly present and visibly present errors (commonly solved by edit restrictions), and argued the need for a method that takes them into account simultaneously. We evaluated the MILC method in terms of its ability to correctly take impossible combinations and relations with other

Table 4. The first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The third row represents the restricted conditional model used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval around the intercept and the logit coefficient of the variable owning/renting a house.

	Intercept		Marriage	
	Estimate	95% CI	Estimate	95% CI
BAG register	2.4661	[2.2090; 2.7233]	− 1.2331	[− 1.3901; − 1.0760]
LISS background	2.7620	[2.4896; 3.0343]	− 1.3041	[− 1.4678; − 1.1405]
Restricted conditional	2.7712	[2.5036; 3.0389]	− 1.3817	[− 1.6493; − 1.1140]

variables into account. We assessed these relations by investigating the bias of $\hat{\theta}$, coverage of the 95% confidence interval, and $se/sd(\hat{\theta})$ in different conditions in a simulation study. The performance of MILC appeared to be mainly dependent on the entropy R^2 value of the LC model. We conclude that a different quality of the composite data set is required to obtain unbiased estimates and standard errors for different types of estimates. In cases of 2×2 tables including an edit restriction, a higher quality of the composite data set was required (entropy R^2 of 0.90), while unbiased estimates and standard errors for logit coefficients can already be obtained with an entropy R^2 value of 0.60.

An example of a composite data set containing data from the LISS panel and the BAG register were shown to have adequate entropy R^2 and we investigated the MILC method using the unconditional model, the conditional model and the restricted conditional model. All models can potentially be used when using the MILC method in practice. However, if there are edit restrictions within the data that need to be taken into account, only the restricted conditional model is appropriate. In light of our main findings, the MILC method can be seen as an alternative for methods previously used for handling visibly and invisibly present errors. This was done either separately using latent variable models and edit rules, or simultaneously by Manrique-Vallier and Reiter (2016), by using one file and one measurement.

A number of limitations of the current study are related to the assumptions we made when specifying the LC model. We assumed that the observed indicators were independent of each other given a unit's score on the latent variable, which means that when a mistake is made on an indicator originating from one source, this is independent of mistakes made on indicators from other sources. For example, if multiple indicators originate from comparable surveys, there is a probability that a respondent makes the same mistake in both surveys; this assumption is then not met. There are ways to relax this assumption by extending the LC model, but we did not investigate the performance of the MILC method if this assumption is relaxed. We also assumed that the misclassification is independent of the covariates. This is also an assumption that in some cases should be relaxed, which we did not investigate as well. Furthermore, the assumption was made that the covariates are free of error. Since this assumption is often not met, ways to relax this assumption should be investigated as well as the performance of the MILC method in such cases. Finally, it was assumed that all edits applied were hard edits, while sometimes soft edits are better applicable. We applied the edits by specifying which cell in the cross table between the latent variable and a covariate should have a weight of 0, while it is also

possible to fix the relevant logit parameter to a very small number. In this way, it should be possible to apply hard or soft edit restrictions. However, we did not investigate the performance of the MILC method when edits are specified in such a manner. We also did not investigate the performance of the LC model used here when some of the previously discussed assumptions are not met.

If a researcher is interested in investigating the relationship between the imputed latent variable and many other variables, all these variables should be included in the LC model as covariates. With the LC three-step approach (Bakk et al. 2016), relationships between the imputed latent variable and other variables (not incorporated in the LC model) can be investigated as well. Edit restrictions could then be added later on as well. However, this three-step approach has not been incorporated in the MILC framework. More investigation can also be done on how the MILC framework handles missing values within covariates, linkage errors and selection errors. Furthermore, the current simulation study only considers dichotomous variables. The current simulation study shows how the method works and it gives some indications of when the method works. This simulation was also comprehensive enough to discover the relation between the quality of the results after imputation and the entropy R^2 value of the LC model. However, it should still be investigated if this relationship holds with larger numbers of indicators, covariates and larger numbers of edit restrictions, and what the exact limitations will be. Also situations when indicators have different numbers of categories are not yet investigated.

Another point of discussion is that we used three indicators in our LC model. In practice, it is more likely that researchers find only two indicators for an underlying true measure in their composite data set. However, a model with two indicators is not identifiable so an additional covariate is necessary. The fact that we used three indicators might seem like a disadvantage. However, a three indicator model and a two indicator plus covariate model are Markov equivalent, which means that they yield the same set of conditional inference assumptions and an identical likelihood.

It should also be noted that MILC can be applied to indicators coming from both population registers and sample surveys. When the indicators only come from sample surveys, we can use the standard rules for pooling as defined by Rubin (1987). However, when at least one of the indicators is sourced from a complete population register, we can choose to either only impute the survey variables, and weigh them to appropriately represent the population variables, or we can choose to impute both the survey and population variables, and use adjusted rules for pooling (Vink and van Buuren 2014). We use these adjusted rules because in the case of register indicators all sampling variability is captured by the between imputation variance, so the within variance should be left out of the equation. In this article, we consider the situation where samples and population registers are linked at a unit level, resulting in a composite data set consisting of only the individuals that were also in the survey sample. However, it is important to be aware of necessary adjustments when population registers are used.

Appendix A

- **Table 1** *Y*1 × *Z***:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of *Y*1 and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N)*, number of bootstrap samples, $m = 5$.
- **Table 2** *W* × *Z* **unconditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of imputed 'true' variable *W* (imputed using the unconditional latent class model) and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N)*, number of bootstrap samples, $m = 5$.
- **Table 3** *W* × *Z* **conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of imputed 'true' variable *W* (imputed using the conditional latent class model) and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N)*, number of bootstrap samples, $m = 5$.
- **Table 4** *W* × *Z* **restricted conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of imputed 'true' variable *W* (imputed using the restricted conditional latent class model) and covariate *Z* with different values for classification probabilities, different values for *P* ($Z = 2$) and different values for sample size (*N)*, number of bootstrap samples, $m = .$
- **Table 5** *Y*1 × *Q***:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of Y_1 on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N)*, $m = 5$.
- **Table 6** W × Q **unconditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of *W* (imputed using the unconditional latent class model) on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N*), $m = 5$.
- **Table 7** *W* × *Q* **conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of *W* (imputed using the conditional latent class model) on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N)*, $m = 5$.
- **Table 8** W × Q **restricted conditional:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of *W* (imputed using the restricted conditional latent class model) on covariate *Q* with different values for the population values of the logit coefficient, classification probabilities, *P* ($Z = 2$) and sample size (*N*), $m = 5$.

- **Table 9** W × Z **restricted conditional** *m***:** This table shows the results in terms of bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W (imputed using the restricted conditional model) and covariate Z with classification probabilities 0.90, $P(Z = 2) = 0.1$, sample size = 1000, 00 and different values for m.

Table 1. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of Y_1 and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and different values for sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		1	−.0030	.9490	1.0282	−.0020	.9490	1.0315	−.0011	.9400	1.0072	−.0009	.9460	1.0053	−.0006	.9370	0.9730
		2	.0031	.9480	1.0301	.0020	.9480	1.0312	.0011	.9390	1.0089	.0009	.9410	0.9946	.0005	.9350	0.9611
	.01	3	.0030	.5690	0.9032	.0019	.7290	0.9144	.0010	.6370	0.8257	.0005	.3820	0.6093	.0001	.0950	0.3097
		4	−.0031	.6880	0.9381	−.0020	.8210	1.0064	−.0010	.8840	0.9526	−.0005	.9070	0.9748	−.0000	.9330	0.9967
		1	−.0146	.8340	0.9803	−.0096	.9150	1.0300	−.0055	.9250	0.9860	−.0030	.9280	0.9716	−.0008	.9470	0.9994
		2	.0144	.8530	0.9799	.0097	.9120	1.0207	.0052	.9370	0.9843	.0027	.9330	0.9723	.0010	.9630	0.9980
	.05	3	.0150	.0000	0.9944	.0099	.0120	1.0129	.0050	.2480	0.9638	.0025	.6600	0.8902	.0005	.3880	0.6134
		4	−.0147	.2740	1.0104	−.0100	.5850	0.9858	−.0047	.8560	1.0292	−.0022	.9020	0.9713	−.0007	.9300	0.9680
1,000		1	−.0297	.5130	0.9659	−.0202	.7450	1.0181	−.0093	.9010	1.0017	−.0051	.9320	0.9894	−.0004	.9540	1.0310
		2	.0297	.5210	0.9686	.0204	.7410	1.0152	.0093	.8970	0.9680	.0047	.9300	0.9626	.0002	.9520	1.0109
	.10	3	.0300	.0000	0.9885	.0202	.0000	1.0051	.0099	.0110	0.9797	.0050	.2690	0.9833	.0010	.6070	0.7905
		4	−.0300	.0480	1.0377	−.0204	.3570	0.9469	−.0099	.8030	1.0010	−.0046	.9150	1.0553	−.0008	.9460	0.9981
		1	−.0600	.0240	0.9841	−.0400	.2760	0.9638	−.0198	.7460	0.9571	−.0092	.9020	0.9908	−.0018	.9420	0.9911
		2	.0593	.0270	0.9931	.0405	.2260	0.9983	.0203	.7310	0.9611	.0090	.8990	0.9822	.0020	.9430	0.9949
	.20	3	.0605	.0000	0.9737	.0399	.0000	0.9931	.0199	.0000	0.9830	.0100	.0100	0.9891	.0020	.7300	0.8763
		4	−.0597	.0000	1.0523	−.0404	.0710	0.9887	−.0204	.5840	0.9690	−.0098	.8610	0.9887	−.0022	.9350	0.9769

Table 1. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		1	−.0031	.9060	1.0085	−.0022	.9470	1.0526	−.0009	.9460	0.9997	−.0006	.9170	0.9558	−.0000	.9460	0.9592
		2	.0031	.9090	1.0003	.0021	.9440	1.0559	.0009	.9450	1.0010	.0006	.9310	0.9685	.0000	.9470	0.9599
	.01	3	.0030	.0000	1.0166	.0020	.0000	0.9763	.0010	.0130	0.9765	.0005	.2980	0.9257	.0001	.6010	0.7449
		4	−.0030	.0670	1.0061	−.0020	.4020	1.0053	−.0010	.7800	0.9837	−.0005	.8990	1.0001	−.0001	.9570	1.0448
		1	−.0149	.1550	0.9830	−.0102	.4580	0.9722	−.0049	.8440	1.0069	−.0024	.9360	1.0488	−.0005	.9460	0.9902
		2	.0148	.1460	0.9879	.0103	.4550	0.9905	.0048	.8380	0.9870	.0024	.9290	1.0282	.0005	.9530	1.0145
	.05	3	.0150	.0000	0.9912	.0100	.0000	0.9905	.0050	.0000	0.9780	.0025	.0000	1.0130	.0005	.2400	1.0204
		4	−.0150	.0000	1.0053	−.0101	.0000	0.9859	−.0049	.3550	1.0221	−.0025	.7780	0.9873	−.0005	.9410	0.9776
		1	−.0298	.0000	1.0235	−.0200	.0210	0.9755	−.0101	.4720	1.0048	−.0052	.8140	0.9984	−.0010	.9460	1.0379
10,000		2	.0298	.0000	1.0135	.0202	.0160	0.9866	.0100	.4910	1.0333	.0050	.8180	0.9866	.0010	.9510	1.0240
	.10	3	.0300	.0000	1.0127	.0199	.0000	1.0195	.0100	.0000	0.9927	.0050	.0000	0.9951	.0010	.0140	1.0053
		4	−.0300	.0000	1.0421	−.0201	.0000	1.0141	−.0100	.0640	1.0627	−.0048	.6220	1.0058	−.0010	.9400	1.0022
		1	−.0600	.0000	1.0166	−.0401	.0000	1.0019	−.0202	.0230	1.0247	−.0098	.4960	1.0247	−.0023	.9160	0.9981
		2	.0599	.0000	1.0085	.0400	.0000	1.0031	.0201	.0140	0.9854	.0099	.4280	1.0534	.0021	.9180	0.9838
	.20	3	.0600	.0000	1.0068	.0401	.0000	0.9837	.0199	.0000	1.0067	.0100	.0000	1.0267	.0020	.0000	0.9471
		4	−.0599	.0000	0.9514	−.0399	.0000	1.0110	−.0199	.0000	0.9832	−.0100	.2730	1.0208	−0.0018	.9260	0.9964

Table 2. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W estimated using the unconditional model and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		1	.0366	.8470	1.0021	.0394	.7770	0.9261	.0173	.8320	0.9552	.0037	.9420	1.0147	.0006	.9480	1.0053
		2	−.0368	.8520	1.0019	−.0393	.7780	0.9289	−.0173	.8370	0.9612	−.0037	.9430	1.0136	−.0005	.9510	1.0140
	.01	3	.0027	.9370	1.2328	.0010	.8400	1.2466	.0002	.3170	0.8594	.0000	.0810	0.4324	.0000	.0040	0.0974
		4	−.0025	.8160	1.1160	−.0010	.9080	1.0406	−.0002	.9130	0.9804	−.0000	.9340	1.0045	−.0000	.9240	0.9772
		1	.0249	.8540	1.0331	.0333	.8040	0.9736	.0153	.8770	0.9936	.0054	.9350	0.9894	.0016	.9620	1.0258
		2	−.0248	.8540	1.0347	−.0332	.8070	0.9743	−.0152	.8630	0.9943	−.0055	.9360	0.9994	−.0012	.9600	1.0235
	.05	3	.0131	.6080	1.1421	.0049	.8320	1.3231	.0009	.7990	1.1260	.0002	.3140	0.7553	.0000	.0160	0.1506
		4	−.0132	.6620	1.0697	−.0050	.8740	1.0218	−.0010	.9320	0.9923	−.0001	.9430	0.9812	−.0005	.9590	1.0732
1,000		1	.0102	.8460	0.9645	.0253	.8430	0.9270	.0124	.8870	0.9782	.0041	.9340	0.9633	.0006	.9460	1.0048
		2	−.0102	.8450	0.9640	−.0250	.8380	0.9310	−.0120	.8820	0.9706	−.0039	.9330	0.9830	−.0009	.9520	1.0054
	.10	3	.0260	.4920	1.0207	.0098	.5010	1.1634	.0017	.9490	1.3165	.0004	.4960	0.9755	.0000	.0270	0.2549
		4	−.0260	.5490	1.0473	−.0100	.7800	0.9656	−.0021	.9390	1.0089	−.0005	.9350	0.9589	.0003	.9580	1.0001
		1	−.0215	.8540	0.9901	.0112	.8720	0.8983	.0076	.9240	0.9714	.0022	.9440	0.9852	−.0007	.9450	1.0061
		2	.0220	.8540	0.9917	−.0107	.8700	0.9106	−.0075	.9300	0.9971	−.0022	.9430	0.9964	.0007	.9420	1.0025
	.20	3	.0501	.4130	1.0317	.0189	.2250	1.0623	.0031	.9330	1.3470	.0006	.7070	1.1819	.0000	.0620	0.3982
		4	−.0506	.4990	1.0398	−.0194	.7040	0.9939	−.0032	.9290	0.9898	−.0005	.9590	1.0322	.0000	.9530	0.9921

Table 2. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		1	.0439	.7410	0.9406	.0420	.1180	0.9371	.0163	.2460	0.9242	.0049	.8280	0.9624	.0005	.9490	1.0112
		2	−.0440	.7390	0.9405	−.0420	.1160	0.9362	−.0163	.2560	0.9124	−.0049	.8260	0.9596	−.0005	.9480	1.0085
	.01	3	.0025	.0170	1.2140	.0010	.2510	1.3434	.0002	.9530	1.2920	.0000	.4960	1.0528	.0000	.0200	0.2565
		4	−.0024	.3630	1.0981	−.0009	.8280	1.0447	−.0002	.9300	1.0030	−.0000	.9440	0.9849	−.0000	.9530	1.0248
		1	.0314	.7980	0.9143	.0347	.1910	0.9367	.0143	.3370	0.9716	.0043	.8760	1.0095	.0001	.9450	0.9906
		2	−.0314	.7990	0.9117	−.0346	.1910	0.9476	−.0144	.3100	0.9818	−.0042	.8740	0.9970	−.0001	.9460	0.9851
	.05	3	.0122	.0000	0.9927	.0047	.0000	1.2805	.0008	.3410	1.3722	.0001	.9330	1.3488	.0000	.1130	0.6631
		4	−.0123	.0170	1.0485	−.0047	.4330	1.0148	−.0007	.9470	1.0245	−.0002	.9540	1.0424	.0001	.9520	0.9968
		1	.0172	.8690	0.9719	.0269	.3530	0.9559	.0123	.4490	0.9795	.0042	.8530	0.9596	.0001	.9440	0.9698
10,000		2	−.0172	.8660	0.9749	−.0267	.3540	0.9753	−.0123	.4190	0.9746	−.0042	.8640	0.9899	−.0002	.9560	1.0132
	.10	3	.0243	.0000	1.0485	.0092	.0000	1.2298	.0016	.0250	1.3530	.0003	.9630	1.4935	.0000	.1750	0.8093
		4	−.0243	.0030	1.0305	−.0094	.1400	1.0139	−.0016	.9100	1.0152	−.0003	.9580	1.0118	.0001	.9630	1.0389
		1	−.0135	.8600	0.9420	.0111	.7430	0.9338	.0078	.7000	0.9701	.0025	.9220	0.9899	.0004	.9470	0.9896
		2	.0134	.8600	0.9490	−.0113	.7220	0.9505	−.0078	.6810	0.9894	−.0025	.9090	0.9983	−.0003	.9500	0.9825
	.20	3	.0474	.0000	0.9731	.0177	.0000	1.0894	.0030	.0000	1.3102	.0005	.7840	1.4592	.0000	.2640	1.0452
		4	−.0473	.0010	0.9562	−.0175	.0180	0.9876	−.0030	.8760	0.9922	−.0005	.9470	0.9829	−.0001	.9550	1.0227

Table 3. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W estimated using the conditional model and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		1	−.2642	.4510	0.9192	.0544	.6940	0.9307	.0188	.8180	0.9471	.0037	.9420	0.9971	.0102	.9050	1.0478
		2	.2640	.4530	0.9210	−.0544	.6980	0.9304	−.0188	.8150	0.9519	−.0037	.9430	0.9962	−.0102	.9170	1.0569
	.01	3	−.0025	.8190	1.1097	.0005	.5420	0.9409	.0001	.2030	0.7171	.0000	.0610	0.4060	−.0100	.0000	0.0746
		4	.0027	.9640	1.3693	−.0005	.9190	1.0043	−.0001	.9140	0.9741	−.0000	.9350	1.0036	.0100	.0100	1.0072
		1	.1163	.6880	1.0093	.0543	.6570	0.9572	.0173	.8460	0.9963	.0055	.9310	0.9908	.0016	.9640	1.0268
		2	−.1162	.6890	1.0128	−.0543	.6400	0.9564	−.0171	.8420	0.9956	−.0056	.9330	0.9989	−.0011	.9620	1.0251
	.05	3	.0055	.9170	1.0301	.0011	.7930	0.9647	.0002	.3500	0.8031	.0000	.0960	0.5438	.0000	.0150	0.1936
		4	−.0056	.8770	0.9839	−.0012	.9370	0.9854	−.0003	.9410	0.9926	.0000	.9430	0.9793	−.0005	.9590	1.0746
1,000		1	.1040	.6550	1.0122	.0474	.6600	0.9330	.0144	.8740	0.9752	.0042	.9320	0.9683	.0006	.9480	1.0060
		2	−.1040	.6480	1.0132	−.0472	.6510	0.9297	−.0140	.8610	0.9594	−.0041	.9360	0.9862	−.0009	.9560	1.0074
	.10	3	.0084	.9190	0.9661	.0018	.8420	0.9172	.0002	.3530	0.7467	.0000	.0650	0.4029	.0000	.0160	0.2591
		4	−.0084	.8970	1.0497	−.0021	.9250	0.9437	−.0006	.9590	1.0080	−.0002	.9470	0.9595	.0002	.9570	1.0004
		1	.0723	.6580	0.9574	.0354	.7010	0.9143	.0100	.9090	0.9647	.0024	.9450	0.9822	−.0009	.9440	1.0038
		2	−.0718	.6520	0.9722	−.0349	.6840	0.9238	−.0099	.9050	0.9785	−.0024	.9420	0.9975	.0008	.9420	1.0013
	.20	3	.0118	.9290	0.9710	.0026	.9150	0.9981	.0003	.3880	0.7857	.0000	.0390	0.2919	.0000	.0250	0.3337
		4	−.0123	.9050	1.0121	−.0031	.9480	1.0107	−.0004	.9460	0.9904	.0000	.9580	1.0282	−.0000	.9550	0.9929

Table 3. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
10,000	.01	1	.1083	.3180	0.9304	.0580	.0140	0.9355	.0177	.1860	0.9253	.0050	.8180	0.9618	.0005	.9480	1.0122
		2	−.1084	.3150	0.9299	−.0581	.0140	0.9341	−.0178	.1750	0.9133	−.0050	.8140	0.9592	−.0005	.9480	1.0094
		3	.0006	.9300	0.9719	.0002	.8490	0.9536	.0000	.3510	0.7248	.0000	.1330	0.5662	.0000	.0180	0.2357
		4	−.0006	.9140	1.0145	−.0001	.9510	1.0016	−.0000	.9330	0.9943	−.0000	.9440	0.9827	−.0000	.9530	1.0248
	.05	1	.1036	.1730	0.9255	.0524	.0130	0.9195	.0159	.2330	0.9716	.0044	.8530	1.0045	.0001	.9450	0.9891
		2	−.1036	.1710	0.9222	−.0524	.0110	0.9285	−.0160	.2270	0.9780	−.0044	.8620	0.9924	−.0001	.9470	0.9840
		3	.0018	.8800	0.9726	.0004	.9500	0.9899	.0001	.5210	0.8260	.0000	.1170	0.4172	.0000	.0230	0.3223
		4	−.0018	.9020	1.0385	−.0005	.9430	0.9846	.0001	.9600	1.0233	−.0001	.9590	1.0413	.0001	.9520	0.9969
	.10	1	.0896	.0740	0.9529	.0455	.0180	0.9337	.0140	.3380	0.9850	.0043	.8500	0.9594	.0001	.9440	0.9693
		2	−.0897	.0730	0.9515	−.0453	.0170	0.9555	−.0140	.3200	0.9809	−.0043	.8570	0.9918	−.0002	.9560	1.0124
		3	.0026	.8500	0.9642	.0007	.9210	0.9950	.0001	.6890	0.8811	.0000	.1460	0.4519	.0000	.0080	0.1825
		4	−.0025	.9170	1.0370	−.0008	.9310	0.9986	−.0001	.9560	1.0123	−.0000	.9560	1.0116	.0001	.9640	1.0389
	.20	1	.0601	.0970	0.9126	.0312	.0760	0.9378	.0098	.5890	0.9694	.0027	.9240	0.9922	.0004	.9470	0.9894
		2	−.0602	.0910	0.9402	−.0313	.0570	0.9654	−.0097	.5310	0.9889	−.0027	.9080	1.0019	−.0003	.9510	0.9822
		3	.0036	.8420	0.9642	.0009	.9090	0.9216	.0002	.8180	1.0068	.0000	.2220	0.5693	.0000	.0020	0.0715
		4	−.0035	.8960	1.0040	−.0007	.9330	0.9912	−.0002	.9500	0.9984	−.0000	.9480	0.9811	−.0001	.9550	1.0229

Table 4. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2 × 2 table of W estimated using the restricted conditional model and covariate Z with different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		1	.0920	.7870	1.0070	.0561	.6670	0.9301	.0190	.8210	0.9555	.0038	.9410	1.0149	−.0003	.9360	0.9726
		2	−.0922	.7880	1.0075	−.0561	.6630	0.9322	−.0190	.8220	0.9608	−.0038	.9420	1.0149	.0002	.9400	0.9595
	.01	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0002	.9330	0.9656	−.0001	.9290	1.0017	.0000	.9140	0.9738	.0000	.9360	1.0045	.0001	.9380	1.0022
		1	.0780	.7740	0.9893	.0507	.6730	0.9450	.0171	.8520	0.9882	.0056	.9330	0.9864	.0016	.9630	1.0270
		2	−.0779	.7690	0.9927	−.0507	.6630	0.9456	−.0170	.8460	0.9896	−.0056	.9320	0.9962	−.0011	.9610	1.0249
	.05	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0001	.9200	0.9375	−.0001	.9360	0.9757	−.0001	.9410	0.9919	.0001	.9440	0.9800	−.0005	.9590	1.0741
1,000		1	.0705	.7470	0.9859	.0425	.7060	0.9299	.0141	.8790	0.9741	.0042	.9330	0.9663	.0006	.9480	1.0044
		2	−.0705	.7450	0.9894	−.0423	.6770	0.9283	−.0137	.8760	0.9607	−.0041	.9320	0.9822	−.0009	.9540	1.0052
	.10	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
		1	.0501	.7530	0.9482	.0309	.7440	0.9212	.0095	.9060	0.9686	.0024	.9410	0.9801	−.0007	.9470	1.0041
		2	−.0496	.7500	0.9611	−.0304	.7330	0.9279	−.0094	.9040	0.9845	−.0024	.9390	0.9907	.0007	.9430	1.0000
	.20	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0005	.9490	1.0238	−.0005	.9530	1.0044	−.0001	.9460	0.9915	.0001	.9560	1.0284	.0000	.9530	0.9923

Table 4. Continued.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	$P(Z=2)$	$\hat{\theta}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		1	.1021	.3450	0.9331	.0581	.0110	0.9398	.0178	.1780	0.9292	.0050	.8200	0.9664	.0005	.9500	1.0132
		2	−.1021	.3440	0.9330	−.0582	.0120	0.9391	−.0178	.1820	0.9168	−.0050	.8220	0.9633	−.0005	.9480	1.0101
	.01	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0001	.9470	0.9863	.0001	.9510	1.0032	.0000	.9310	0.9913	−.0000	.9440	0.9825	−.0000	.9530	1.0248
		1	.0916	.2140	0.9175	.0513	.0140	0.9214	.0158	.2380	0.9697	.0044	.8630	1.0099	.0001	.9460	0.9894
		2	−.0916	.2120	0.9145	−.0512	.0130	0.9309	−.0159	.2170	0.9757	−.0044	.8690	0.9979	−.0001	.9470	0.9840
	.05	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	−.0000	.9530	0.9966	−.0001	.9490	0.9824	.0001	.9600	1.0234	−.0001	.9590	1.0416	.0001	.9520	0.9971
		1	.0794	.1050	0.9382	.0441	.0240	0.9314	.0139	.3470	0.9666	.0043	.8510	0.9603	.0001	.9460	0.9697
10,000		2	−.0794	.1050	0.9361	−.0439	.0180	0.9537	−.0139	.3200	0.9653	−.0043	.8520	0.9925	−.0002	.9560	1.0129
	.10	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0000	.9580	1.0304	−.0002	.9390	0.9953	−.0000	.9560	1.0115	−.0000	.9560	1.0118	.0001	.9640	1.0389
		1	.0531	.1680	0.9064	.0300	.0890	0.9311	.0096	.6020	0.9619	.0027	.9250	0.9907	.0004	.9450	0.9898
		2	−.0532	.1530	0.9333	−.0301	.0620	0.9565	−.0096	.5350	0.9772	−.0026	.9130	0.9982	−.0003	.9510	0.9831
	.20	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
		4	.0001	.9460	0.9888	.0002	.9470	0.9933	−.0001	.9490	0.9993	−.0000	.9480	0.9803	−.0001	.9550	1.0229

Table 5. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of Y_1 regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		.01	−.1139	.8550	1.0087	−.0763	.9080	1.0065	−.0432	.9430	0.9924	−.0253	.9360	0.9852	−.0047	.9590	1.0096
		.05	−.1254	.8360	1.0070	−.0756	.9220	1.0013	−.0503	.9230	0.9830	−.0170	.9290	0.9661	−.0025	.9500	0.9876
	.45	.10	−.1156	.8620	1.0103	−.0823	.8970	0.9963	−.0335	.9550	1.0339	−.0228	.9420	0.9940	−.0066	.9460	0.9528
		.20	−.1144	.8610	1.0121	−.0762	.9030	0.9795	−.0378	.9430	1.0232	−.0157	.9500	1.0179	−.0063	.9480	0.9803
		.01	.1221	.8440	1.0323	.0808	.8930	0.9760	.0378	.9440	1.0077	.0204	.9460	0.9819	.0040	.9450	0.9799
1,000		.05	.1156	.8590	1.0206	.0795	.9000	0.9768	.0329	.9400	0.9950	.0217	.9320	0.9681	.0027	.9560	1.0533
	.55	.10	.1250	.8440	1.0136	.0753	.9050	1.0200	.0376	.9310	0.9820	.0214	.9400	1.0049	.0080	.9450	0.9828
		.20	.1164	.8530	1.0150	.0759	.9000	0.9940	.0403	.9350	0.9989	.0178	.9410	0.9894	.0052	.9580	1.0157
		.01	.3768	.1790	0.9792	.2498	.5060	1.0233	.1322	.8310	0.9955	.0596	.9270	0.9876	.0135	.9480	1.0014
		.05	.3677	.1700	1.0074	.2448	.5270	0.9782	.1249	.8260	0.9749	.0604	.9210	1.0037	.0021	.9440	0.9610
	.65	.10	.3700	.1660	1.0439	.2456	.5260	1.0248	.1248	.8500	1.0019	.0659	.9230	1.0028	.0018	.9450	0.9874
		.20	.3799	.1570	1.0223	.2557	.4880	0.9879	.1258	.8490	1.0087	.0648	.9290	1.0449	.0173	.9440	0.9946
		.01	−.1211	.1510	0.9838	−.0794	.5020	0.9382	−.0417	.8140	1.0021	−.0222	.9240	1.0272	−.0014	.9460	0.9907
		.05	−.1201	.1510	0.9765	−.0810	.4700	1.0168	−.0399	.8320	1.0117	−.0183	.9190	0.9928	−.0018	.9440	0.9919
	.45	.10	−.1200	.1490	0.9682	−.0813	.4550	1.0213	−.0398	.8330	1.0032	−.0199	.9100	0.9696	−.0028	.9550	1.0037
		.20	−.1179	.1730	0.9627	−.0793	.4890	0.9764	−.0411	.8190	0.9954	−.0199	.9330	1.0578	−.0029	.9550	1.0297
		.01	.1195	.1470	1.0344	.0802	.4830	0.9700	.0383	.8320	1.0036	.0200	.9150	0.9717	.0058	.9470	0.9691
10,000		.05	.1211	.1400	0.9898	.0800	.4870	0.9555	.0407	.8260	1.0384	.0193	.9200	1.0074	.0060	.9370	0.9844
	.55	.10	.1221	.1420	0.9849	.0814	.4810	0.9800	.0373	.8600	1.0222	.0180	.9310	1.0053	.0010	.9610	1.0464
		.20	.1213	.1360	1.0191	.0786	.5080	1.0016	.0413	.8110	0.9906	.0214	.9160	0.9963	.0027	.9530	1.0121
		.01	.3727	.0000	0.9855	.2499	.0000	0.9948	.1251	.1420	1.0078	.0617	.6740	0.9620	.0104	.9520	1.0347
		.05	.3735	.0000	1.0096	.2484	.0000	1.0109	.1270	.1300	0.9911	.0649	.6380	0.9782	.0145	.9310	1.0156
	.65	.10	.3746	.0000	0.9988	.2493	.0000	1.0022	.1241	.1580	0.9729	.0624	.6740	1.0216	.0115	.9360	0.9955
		.20	.3706	.0000	1.0016	.2510	.0000	0.9942	.1266	.1250	1.0222	.0642	.6440	0.9656	.0115	.9450	1.0145

Table 6. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the unconditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		.01	.0157	.9350	0.9918	.0090	.9530	1.0346	−.0021	.9520	1.0093	.0015	.9460	0.9795	−.0053	.9500	1.0045
		.05	.0081	.9410	1.0813	−.0010	.9320	0.9798	.0024	.9490	1.0050	−.0047	.9530	1.0270	.0029	.9510	0.9889
	.45	.10	.0167	.9310	0.9873	−.0028	.9410	0.9611	.0050	.9450	0.9953	.0011	.9580	1.0166	.0016	.9460	0.9968
		.20	.0112	.9210	0.9989	.0079	.9400	0.9890	.0006	.9520	0.9672	−.0046	.9580	1.0020	−.0005	.9520	1.0285
		.01	−.0308	.9310	0.9980	−.0031	.9350	0.9906	−.0006	.9440	0.9985	−.0093	.9470	0.9968	.0084	.9520	1.0232
1,000		.05	−.0090	.9230	0.9695	−.0047	.9350	0.9788	−.0006	.9540	0.9978	−.0043	.9520	1.0365	−.0002	.9610	1.0240
	.55	.10	−.0214	.9310	0.9734	−.0039	.9440	1.0123	−.0013	.9420	0.9833	.0019	.9450	0.9809	−.0045	.9600	1.0210
		.20	−.0278	.9360	0.9902	−.0112	.9460	1.0053	.0018	.9510	1.0056	.0049	.9560	1.0109	.0006	.9560	1.0235
		.01	−.0345	.9310	1.0286	−.0049	.9470	1.0130	−.0013	.9360	0.9722	−.0071	.9500	1.0091	−.0009	.9490	0.9949
		.05	−.0444	.9320	1.0059	−.0098	.9400	0.9911	−.0001	.9510	1.0101	.0024	.9580	1.0171	.0028	.9420	0.9774
	.65	.10	−.0578	.9190	0.9891	−.0149	.9470	1.0139	−.0062	.9590	0.9939	−.0091	.9430	0.9583	−.0054	.9530	0.9910
		.20	−.0361	.9460	1.0222	−.0051	.9430	1.0044	−.0044	.9600	1.0294	−.0080	.9510	1.0019	−.0048	.9410	0.9982
		.01	−.0046	.9280	0.9998	.0031	.9410	0.9762	.0024	.9530	1.0250	−.0022	.9550	1.0359	.0006	.9510	0.9851
		.05	−.0020	.9210	0.9806	.0005	.9500	1.0033	.0016	.9410	1.0022	.0021	.9380	0.9844	.0014	.9500	0.9966
	.45	.10	.0044	.9200	0.9365	−.0017	.9520	1.0266	.0018	.9670	1.0462	−.0010	.9510	1.0170	.0018	.9510	1.0040
		.20	.0038	.9110	0.9531	.0012	.9510	0.9738	.0002	.9560	1.0014	−.0003	.9680	1.0423	−.0019	.9400	0.9699
		.01	−.0043	.9150	0.9728	−.0009	.9350	0.9699	−.0019	.9360	0.9678	−.0000	.9510	0.9889	−.0002	.9410	0.9808
10,000		.05	−.0034	.9320	0.9775	−.0007	.9320	0.9671	−.0003	.9480	1.0049	−.0004	.9550	1.0133	.0019	.9500	0.9753
	.55	.10	−.0080	.9310	0.9924	−.0009	.9530	1.0127	−.0033	.9430	0.9760	.0003	.9510	0.9899	−.0005	.9450	0.9955
		.20	−.0069	.9210	0.9813	−.0028	.9460	1.0044	−.0010	.9340	0.9476	.0008	.9550	1.0331	.0001	.9490	1.0081
		.01	−.0083	.9250	0.9643	−.0014	.9300	0.9458	−.0037	.9530	1.0423	−.0033	.9500	0.9952	−.0001	.9630	1.0809
		.05	−.0062	.9260	1.0011	−.0024	.9390	0.9908	−.0029	.9520	0.9845	.0002	.9410	0.9877	.0013	.9620	1.0461
	.65	.10	−.0153	.9300	0.9918	−.0053	.9440	0.9832	−.0028	.9570	1.0140	−.0013	.9530	1.0042	.0008	.9570	1.0047
		.20	−.0134	.9030	0.9355	−.0032	.9370	0.9849	.0001	.9470	0.9967	−.0005	.9490	1.0010	−.0001	.9610	1.0555

Table 7. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the conditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
		.01	−.4303	.7100	1.3597	.0097	.9550	1.0372	−.0036	.9580	1.0172	.0022	.9450	0.9789	−.0053	.9470	1.0040
		.05	.0075	.9490	1.0825	−.0027	.9360	0.9812	.0025	.9520	0.9975	−.0043	.9550	1.0289	.0032	.9500	0.9899
	.45	.10	.0135	.9310	0.9973	−.0021	.9410	0.9624	.0065	.9440	0.9964	.0014	.9570	1.0134	.0016	.9480	0.9982
		.20	.0027	.9290	0.9731	.0027	.9410	1.0048	.0003	.9540	0.9760	−.0048	.9470	0.9990	−.0006	.9540	1.0283
		.01	.4050	.7500	1.2376	−.0039	.9360	0.9990	−.0015	.9470	1.0016	−.0096	.9490	0.9959	.0084	.9520	1.0231
1,000		.05	−.0028	.9220	1.0031	−.0041	.9450	0.9927	−.0008	.9490	0.9984	−.0037	.9540	1.0433	−.0001	.9590	1.0239
	.55	.10	−.0113	.9490	0.9894	−.0018	.9470	1.0198	−.0020	.9470	0.9814	.0025	.9460	0.9805	−.0044	.9600	1.0211
		.20	−.0087	.9450	1.0109	−.0080	.9490	1.0205	.0015	.9490	0.9990	.0056	.9590	1.0146	.0006	.9560	1.0236
		.01	1.3336	.0800	1.3089	−.0060	.9480	1.0146	−.0027	.9440	0.9774	−.0076	.9520	1.0107	−.0010	.9540	0.9959
		.05	−.0381	.9430	1.6091	−.0101	.9370	0.9926	−.0012	.9540	1.0142	.0018	.9560	1.0148	.0028	.9420	0.9773
	.65	.10	−.0316	.9310	1.0027	−.0114	.9460	1.0251	−.0070	.9590	0.9969	−.0089	.9340	0.9574	−.0052	.9520	0.9904
		.20	−.0004	.9480	1.0298	.0040	.9450	0.9902	−.0037	.9570	1.0298	−.0082	.9560	1.0068	−.0048	.9430	0.9975
		.01	.0027	.9240	0.9560	.0031	.9390	0.9819	−.0003	.9450	0.9913	−.0009	.9510	1.0202	.0030	.9410	0.9946
		.05	.0011	.9400	1.0213	.0005	.9490	1.0116	.0021	.9560	1.0079	.0020	.9420	0.9893	.0024	.9450	0.9938
	.45	.10	.0004	.9220	0.9820	−.0026	.9540	1.0325	.0004	.9530	1.0206	.0008	.9430	0.9837	.0015	.9600	0.9878
		.20	.0006	.9380	0.9926	−.0005	.9510	0.9893	−.0019	.9550	1.0067	.0009	.9590	1.0582	.0005	.9640	1.0294
		.01	−.0070	.9220	0.9893	−.0012	.9350	0.9709	−.0018	.9440	0.9918	−.0009	.9470	0.9815	.0018	.9470	0.9713
10,000		.05	.0012	.9200	0.9515	−.0003	.9320	0.9704	.0005	.9520	0.9998	−.0012	.9570	1.0107	.0018	.9420	0.9900
	.55	.10	−.0001	.9260	0.9837	−.0001	.9570	1.0173	−.0044	.9530	1.0407	−.0028	.9480	1.0010	−.0025	.9580	1.0482
		.20	.0029	.9290	0.9886	−.0011	.9430	1.0107	−.0001	.9410	0.9692	.0013	.9520	0.9884	−.0009	.9550	1.0200
		.01	−.0105	.9220	1.1451	−.0026	.9300	0.9572	−.0044	.9580	1.0195	−.0029	.9530	0.9930	−.0022	.9540	1.0358
		.05	−.0074	.9280	0.9838	−.0025	.9390	0.9986	−.0011	.9500	0.9978	.0004	.9370	0.9773	.0014	.9540	1.0102
	.65	.10	−.0014	.9400	0.9904	−.0034	.9410	0.9913	−.0043	.9500	1.0043	−.0017	.9490	1.0400	−.0009	.9530	0.9892
		.20	.0017	.9330	1.0127	.0023	.9470	0.9892	−.0024	.9480	0.9917	.0002	.9450	0.9932	−.0011	.9600	1.0095

Table 8. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the logit coefficients of W estimated using the restricted conditional model regressed on covariate Q with different values for the population values of the logit coefficient, different values for the classification probabilities, different values for P (Z = 2) and sample size (N), number of bootstrap samples m = 5.

			class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
N	coef	$P(Z=2)$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
1,000	.45	.01	.0318	.9170	0.9730	.0099	.9470	1.0215	.0018	.9530	0.9893	−.0054	.9400	0.9693	−.0013	.9500	1.0134
		.05	.0099	.9390	1.0035	−.0011	.9380	0.9586	−.0096	.9410	0.9932	.0077	.9490	0.9912	.0024	.9500	0.9816
		.10	.0057	.9370	1.0108	−.0027	.9420	0.9728	.0047	.9580	1.0362	−.0029	.9520	1.0184	−.0033	.9430	0.9480
		.20	.0060	.9350	1.0079	.0050	.9530	1.0086	.0063	.9550	1.0681	.0051	.9530	1.0049	−.0008	.9450	0.9851
	.55	.01	−.0088	.9370	1.0423	−.0040	.9380	0.9916	−.0055	.9540	1.0004	−.0011	.9470	0.9818	−.0002	.9480	0.9880
		.05	−.0048	.9400	1.0348	−.0037	.9460	0.9922	−.0094	.9480	1.0012	.0016	.9430	0.9808	−.0014	.9550	1.0565
		.10	.0058	.9340	1.0123	−.0021	.9350	0.9904	−.0011	.9460	0.9986	.0031	.9510	1.0202	.0023	.9530	0.9840
		.20	−.0010	.9320	0.9809	−.0065	.9470	1.0168	−.0019	.9510	0.9913	−.0015	.9490	0.9833	.0009	.9550	1.0090
	.65	.01	−.0351	.9450	1.0301	−.0061	.9480	1.0107	.0008	.9450	0.9898	−.0086	.9530	1.0075	.0004	.9540	1.0022
		.05	−.0192	.9410	0.9847	−.0044	.9460	1.0003	−.0059	.9430	0.9849	−.0033	.9540	1.0015	−.0108	.9460	0.9629
		.10	−.0080	.9340	1.0162	−.0139	.9480	1.0160	−.0038	.9580	1.0093	−.0006	.9580	0.9895	−.0109	.9430	0.9965
		.20	.0016	.9460	0.9996	.0044	.9490	0.9849	.0044	.9550	1.0192	−.0029	.9600	1.0578	.0045	.9480	0.9885
10,000	.45	.01	−.0039	.9280	1.0269	.0033	.9340	0.9739	−.0005	.9470	0.9897	−.0005	.9540	1.0205	.0030	.9410	0.9939
		.05	−.0019	.9260	0.9895	.0006	.9510	1.0142	.0020	.9500	1.0154	.0019	.9400	0.9886	.0024	.9430	0.9934
		.10	−.0006	.9190	0.9415	−.0024	.9510	1.0333	.0003	.9520	1.0188	.0006	.9400	0.9867	.0015	.9580	0.9889
		.20	−.0004	.9370	0.9883	−.0005	.9550	1.0033	−.0019	.9510	1.0037	.0011	.9610	1.0530	.0006	.9620	1.0290
	.55	.01	−.0065	.9230	0.9828	−.0015	.9380	0.9802	−.0021	.9500	0.9876	−.0007	.9490	0.9789	.0019	.9450	0.9721
		.05	.0032	.9130	0.9552	.0003	.9380	0.9804	−.0000	.9490	1.0015	−.0011	.9590	1.0104	.0018	.9420	0.9893
		.10	.0008	.9370	0.9855	.0006	.9520	1.0206	−.0046	.9540	1.0307	−.0030	.9490	1.0022	−.0024	.9580	1.0489
		.20	.0041	.9360	0.9792	−.0007	.9490	1.0056	.0002	.9380	0.9725	.0012	.9510	0.9893	−.0009	.9560	1.0200
	.65	.01	−.0136	.9270	0.9887	−.0031	.9310	0.9522	−.0040	.9530	1.0141	−.0028	.9500	0.9954	−.0022	.9540	1.0352
		.05	−.0047	.9320	0.9980	−.0024	.9440	1.0046	−.0009	.9550	0.9956	.0006	.9360	0.9753	.0014	.9520	1.0110
		.10	.0016	.9370	0.9886	−.0028	.9460	0.9957	−.0038	.9460	0.9928	−.0018	.9510	1.0382	−.0009	.9510	0.9892
		.20	.0058	.9370	1.0049	.0029	.9460	0.9893	−.0022	.9510	0.9946	.0001	.9460	0.9941	−.0011	.9600	1.0092

Table 9. Bias, coverage of the 95% confidence interval and $se/sd(\hat{\theta})$ of the 4 cell proportions of the 2×2 table of W estimated using the restricted conditional model and covariate Z with classification probabilities 0.90, P (Z = 2) = 0.1, sample size = 1000,00 and different values for the number of bootstrap samples m.

		class. prob. 0.70			class. prob. 0.80			class. prob. 0.90			class. prob. 0.95			class. prob. 0.99		
m	θ	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$	bias	cov	$\frac{se}{sd(\hat{\theta})}$
	1	.0705	.7470	0.9859	.0425	.7060	0.9299	.0141	.8790	0.9741	.0042	.9330	0.9663	.0006	.9480	1.0044
	2	−.0705	.7450	0.9894	−.0423	.6770	0.9283	−.0137	.8760	0.9607	−.0041	.9320	0.9822	−.0009	.9540	1.0052
5	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	−.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
	1	.07146	.7750	0.9870	.0427	.7090	0.9399	.014	.8780	0.9829	.0042	.9340	0.9690	.0006	.9460	1.0042
	2	−.0715	.7660	0.9929	−.0424	.6970	0.9406	−.0137	.8600	0.9685	−.0040	.9390	0.9861	−.0009	.9540	1.0056
10	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
	1	.0710	.8010	1.0161	.0425	.7090	0.9360	.0138	.8880	0.9753	.0041	.9300	0.9646	.0006	.9480	1.0052
	2	−.0710	.8060	1.0222	−.0423	.6930	0.9399	−.0135	.8680	0.9626	−.0040	.9340	0.9828	−.0009	.9540	1.0061
20	3	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994
	1	0.0720	.7910	1.0152	.0426	.7120	0.9485	.0141	.8720	0.9826	.0042	.9330	0.9661	.0007	.9480	1.0045
	2	−0.0720	.7920	1.0208	−.0424	.7000	0.9512	−.0137	.8610	0.9692	−.0040	.9400	0.9835	−.0009	.9520	1.0055
40	3	0.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-	.0000	-	-
	4	−0.0000	.9510	0.9994	−.0003	.9380	0.9340	−.0004	.9600	1.0104	−.0002	.9470	0.9589	.0003	.9580	0.9994

Appendix B

- **Table 10 (Application) entropy R^2, classification probabilities, marginal probabilities:** This table shows the entropy $R2$, classification probabilities for the indicators and marginal probabilities for the covariates for the unconditional, the conditional and the restricted conditional model. Note that the *rent benefit* variable takes information of 779 individuals into account and *marital status* variable of 3, 011.
- **Table 11 (Application) proportions and marginal proportions:** The first block of tihs table represents the (pooled) marginal proportions of the variable *own/rent*. The second block represents the (pooled) proportions of the variable *own/rent* for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable *own/rent* for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.
- **Table 12 (Application) estimates of intercept and logit coefficient:** In this table, first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval (total) standard error of the intercept and the logit coefficient of the variable *owning/renting* a house.

Table 10. Entropy R^2, classification probabilities for the indicators and marginal probabilities for the covariates for the unconditional, the conditional and the restricted conditional model. Note that the rent benefit variable takes information of 779 individuals into account and marital status variable of 3,011.

			Unconditional model	Conditional model	Restricted conditional model
Entropy R^2			0.9334	0.9377	0.9380
	LISS	P(rent| LC rent)	0.8937	0.8938	0.9344
Classification	background	P(own| LC own)	0.9997	0.9997	0.9992
probability	BAG	P(rent| LC rent)	0.9501	0.9500	0.9496
	register	P(own| LC own)	0.9749	0.9749	0.9525
P(rent benefit)				0.3004	0.3004
P(married)			0.5284	0.5284	0.5284

Table 11. The first block represents the (pooled) marginal proportions of the variable own/rent. The second block represents the (pooled) proportions of the variable own/rent for persons receiving rent benefit. The third block represents the (pooled) proportions of the variable own/rent for persons not receiving rent benefit. Within each block, the first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. For each proportion a (pooled) estimate and a (pooled) 95% confidence interval is given.

	P(own)		P(rent)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.6450	[0.6448; 0.6451]	0.3550	[0.3549; 0.3511]
LISS background	0.6830	[0.6829; 0.6832]	0.3170	[0.3168; 0.3171]
Unconditional	0.6405	[0.6404; 0.6407]	0.3595	[0.3593; 0.3596]
Conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
Restricted conditional	0.6597	[0.6595; 0.6598]	0.3403	[0.3402; 0.3405]
	P(own × rent benefit)		P(rent × rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0051	[0.0001; 0.0102]	0.2953	[0.2632; 0.3273]
LISS background	0.0104	[0.0032; 0.0175]	0.2889	[0.2568; 0.3209]
Unconditional	0.0028	[0.0023; 0.0034]	0.2950	[0.2944; 0.2955]
Conditional	0.0064	[−0.0263; 0.0392]	0.2914	[0.2587; 0.3241]
Restricted conditional	0.0000	-	0.2978	[0.2649; 0.3307]
	P(own × no rent benefit)		P(rent × no rent benefit)	
	Estimate	95% CI	Estimate	95% CI
BAG register	0.0552	[0.0391; 0.0713]	0.6444	[0.6107; 0.6781]
LISS background	0.0285	[0.0167; 0.0403]	0.6723	[0.6391; 0.7054]
Unconditional	0.0157	[0.0151; 0.0162]	0.6829	[0.6824; 0.6835]
Conditional	0.0159	[−0.0168; 0.0487]	0.6827	[0.6499; 0.7154]
Restricted conditional	0.0213	[−0.0116; 0.0542]	0.6773	[0.6444; 0.7102]

Table 12. The first two rows represent the BAG register and the LISS background survey, used as the indicators for the MILC method. The last three rows represent the three different models used to apply the MILC method. The columns represent the (pooled) estimate and 95% confidence interval (total) standard error of the intercept and the logit coefficient of the variable owning/renting a house.

	Intercept		Marriage	
	Estimate	95% CI	Estimate	95% CI
BAG register	2.4661	[2.2090; 2.7233]	−1.2331	[−1.3901; −1.0760]
LISS background	2.7620	[2.4896; 3.0343]	−1.3041	[−1.4678; −1.1405]
Unconditional	2.6869	[2.4251; 2.9487]	−1.3875	[−1.6493; −1.1257]
Conditional	2.7698	[2.5034; 3.0363]	−1.3982	[−1.6646; −1.1317]
Restricted conditional	2.7712	[2.5036; 3.0389]	−1.3817	[−1.6493; −1.1140]

6. References

André, S. and C. Dewilde. 2016. "Home Ownership and Support for Government Redistribution." *Comparative European Politics* 14: 319–348. Doi: http://dx.doi.org/10.1057/cep.2014.31.

Bakk, Z., D.L. Oberski, and J.K. Vermunt. 2016. "Relating Latent Class Membership to Continuous Distal Outcomes: Improving the LTB Approach and a Modified Three-Step Implementation." *Structural Equation Modeling: A Multidisciplinary Journal* 23: 278–289. Doi: http://dx.doi.org/10.1080/10705511.2015.1049698.

Bakker, B.F.M. 2009. Trek alle registers open! *Rede in verkorte vorm uitgesproken bij de aanvaarding van het ambt van bijzonder hoogleraar Methodologie van registers voor sociaalwetenschappelijk onderzoek bij de Faculteit der Sociale Wetenschappen van de Vrije Universiteit Amsterdam op 26 november 2009*. Available at: http://dare.ubvu.vu.nl/bitstream/handle/1871/15588/Oratie%20Bakker.pdf (accessed April 24, 2017).

Bakker, B.F.M. 2010. "Micro-Integration, State of the Art." *Paper presented at the joint UNECE-Eurostat expert group meeting on registered based censuses in The Hague, May 11, 2010*. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2010/wp.10.e.pdf (accessed April 24, 2017).

Bakker, B.F.M. 2012. "Estimating the Validity of Administrative Variables." *Statistica Neerlandica* 66: 8–17. Doi: http://dx.doi.org/10.1111/j.14679574.2011.00504.x.

Biemer, P.P. 2011. *Latent Class Analysis of Survey Error* (Vol. 571). Hoboken, New Jersey: John Wiley & Sons.

De Waal, T. 2016. "Obtaining Numerically Consistent Estimates from a Mix of Administrative Data and Surveys." *Statistical Journal of the IAOS* 32: 231–243. Doi: http://dx.doi.org//10.3233/SJI-150950.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation* (Vol. 563). John Wiley & Sons.

De Waal, T., J. Pannekoek, and S. Scholtus. 2012. "The Editing of Statistical Data: Methods and Techniques for the Efficient Detection and Correction of Errors and Missing Values." *Wiley Interdisciplinary Reviews: Computational Statistics* 4: 204–210. Doi: http://dx.doi.org/10.1002/wics.1194.

Dewilde, C. and P.D. Decker. 2016. "Changing Inequalities in Housing Outcomes Across Western Europe." *Housing, Theory and Society* 33: 121–161. Doi: http://dx.doi.org/10.1080/14036096.2015.1109545.

Dias, J.G. and J.K. Vermunt. 2008. "A Bootstrap-Based Aggregate Classifier for Model-Based Clustering." *Computational Statistics* 23: 643–659. Doi: http://dx.doi.org/10.1007/s00180-007-0103-7.

Forcina, A. 2008. "Identifiability of Extended Latent Class Models with Individual Covariates." *Computational Statistics & Data Analysis* 52: 5263–5268. Doi: http://dx.doi.org/10.1016/j.csda.2008.04.030.

Geerdinck, M., M. Goedhuys-van der Linden, E. Hoogbruin, A. De Rijk, N. Sluiter, and C. Verkleij. 2014. *Monitor Kwaliteit Stelsel van Basisregistraties: Nulmeting van de Kwaliteit van Basisregistraties in Samenhang, 2014* (13114th ed.). Henri Faas-dreef 312, 2492 JP Den Haag: Centraal Bureau voor de Statistiek. Available at:

https://www.cbs.nl/-/media/pdf/2016/50/monitor-kwaliteit-stelsel-van-basisregistraties.pdf (accessed April 25, 2017).

Groen, J.A. 2012. "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." *Journal of Official Statistics* 28: 173–198.

Guarnera, U. and R. Varriale. 2016. "Estimation from Contaminated Multi-Source Data Based on Latent Class Models." *Statistical Journal of the IAOS* 32: 537–544. Doi: dx.doi.org//10.3233/SJI-150951.

Jörgren, F., R. Johansson, L. Damber, and G. Lindmark. 2010. "Risk Factors of Rectal Cancer Local Recurrence: Population-Based Survey and Validation of the Swedish Rectal Cancer Registry." *Colorectal Disease* 12: 977–986. Doi: http://dx.doi.org/10.1111/j.1463-1318.2009.01930.x.

Kim, H.J., L.H. Cox, A.F. Karr, J.P. Reiter, and Q. Wang. 2015. "Simultaneous Edit-Imputation for Continuous Microdata." *Journal of the American Statistical Association* 110: 987–999. Doi: http://dx.doi.org/10.1080/01621459.2015.1040881.

Lersch, P.M. and C. Dewilde. 2015. "Employment Insecurity and First-Time Homeownership: Evidence from Twenty-Two European Countries." *Environment and Planning A* 47: 607–624. Doi: http://dx.doi.org//10.1068/a130358p.

Manrique-Vallier, D. and J.P. Reiter. 2013. "Bayesian Multiple Imputation for Large-Scale Categorical Data with Structural Zeros." *Survey Methodology* 40: 125–134. Available at: https://ecommons.cornell.edu/handle/1813/34889 (accessed April 25, 2017).

Manrique-Vallier, D. and J.P. Reiter. 2016. "Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data." *Journal of the American Statistical Association*. Doi: http://dx.doi.org/10.1080/01621459.2016.1231612.

Mulder, C.H. 2006. "Home-Ownership and Family Formation." *Journal of Housing and the Built Environment* 21: 281–298. Doi: http://dx.doi.org/10.1007/s10901-006-9050-9.

Ness, A.R. 2004. "The Avon Longitudinal Study of Parents and Children (ALSPAC)- a Resource for the Study of the Environmental Determinants of Childhood Obesity." *European Journal of Endocrinology 151*(Suppl 3): U141–U149. Doi: http://dx.doi.org//10.1530/eje.0.151U141.

Oberski, D.L. 2015. "Total Survey Error in Practice." In *Total Survey Error*, edited by P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. Tucker, and B. West. New York: Wiley.

Pavlopoulos, D. and J. Vermunt. 2015. "Measuring Temporary Employment. Do Survey or Register Tell the Truth?" *Survey Methodology* 41: 197–214. Available at: http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14151-eng.pdf (accessed April 25, 2017).

R Core Team. 2014. "R: A Language and Environment for Statistical Computing [Computer software manual]." Vienna, Austria. Available at: http://www.R-project.org/ (accessed October 13, 2017).

Robertsson, O., M. Dunbar, K. Knutson, S. Lewold, and L. Lidgren. 1999. "Validation of the Swedish Knee Arthroplasty Register: A Postal Survey Regarding 30,376 Knees Operated on Between 1975 and 1995." *Acta Orthopaedica Scandinavica* 70: 467–472. Doi: http://dx.doi.org/10.3109/17453679909000982.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys* (Vol. 81). John Wiley & Sons. Doi: http://dx.doi.org//10.1002/9780470316696.

Scherpenzeel, A. 2011. "Data Collection in a Probability-Based Internet Panel: How the LISS Panel was Built and How it can be Used." *Bulletin of Sociological Methodology/Bulletin de Methodologie Sociologique* 109: 56–61. Doi: http://dx.doi.org//10.1177/0759106310387713.

Scholtus, S. 2009. "Automatic Detection of Simple Typing Errors in Numerical Data with Balance Edits." *Statistics Netherlands Discussion Paper* (09046). Available at: https://www.cbs.nl/-/media/imported/documents/2009/48/2009-46-x10-pub.pdf (accessed April 25, 2017).

Scholtus, S. 2011. "Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data." *Journal of Official Statistics* 27: 467–490.

Scholtus, S. and B.F.M. Bakker. 2013. "Estimating the Validity of Administrative and Survey Variables through Structural Equation Modeling: A Simulation Study on Robustness." *Statistics Netherlands Discussion Paper*. Available at: https://www.cbs.nl/-/media/imported/documents/2013/12/2013-02-x10-pub.pdf (accessed April 25, 2017).

Schrijvers, C.T.M., K. Stronks, D.H. van de Mheen, J.-W. W. Coebergh, and J.P. Mackenbach. 1994. "Validation of Cancer Prevalence Data from a Postal Survey by Comparison with Cancer Registry Records." *American Journal of Epidemiology* 139: 408–414. Doi: https://doi.org/10.1093/oxfordjournals.aje.a117013.

Schulte Nordholt, E., J. Van Zeijl, and L. Hoeksma. 2014. "Dutch Census 2011, Analysis and Methodology." *Statistics Netherlands*. Available at: https://www.cbs.nl/-/media/imported/documents/2014/44/2014-b57-pub.pdf (accessed April 25, 2017).

Si, Y. and J.P. Reiter. 2013. "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys." *Journal of Educational and Behavioral Statistics* 38: 499–521. Doi: dx.doi.org//10.3102/1076998613480394.

Tempelman, C. 2007. *Imputation of Restricted Data: Applications to Business surveys* (Doctoral dissertation, Rijksuniversiteit Groningen). Available at: https://www.cbs.nl/-/media/imported/documents/2007/05/2007-i76-pub.pdf (accessed April 25, 2017).

Turner, C.F., T.K. Smith, L.K. Fitterman, T. Reilly, K. Pate, M.B. Witt, and B.H. Forsyth. 1997. "The Quality of Health Data Obtained in a New Survey of Elderly Americans: A Validation Study of the Proposed Medicare Beneficiary Health Status Registry (mbhsr)." *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 52B: S49–S58. Doi: http://dx.doi.org//10.1093/geronb/52B.1.S49.

Understanding Society. 2016. "Understanding Society: Innovation Panel, Waves 1–7, 2008–2014 [data collection]. 6th edition [Computer software manual]. UK Data Service. Doi: 10.5255/UKDA-SN-6849-7.

University of London. Institute of Education. Centre for Longitudinal Studies, Millennium Cohort Study: First Survey, 2001–2003 [computer file]. 6th edition. Colchester, Essex: UK Data Archive [distributor], SN: 4683. (2007, March). Available at: http://dx.doi.org/10.5255/UKDA-SN-4683-1.

Van der Palm, D.W., L.A. Van der Ark, and J.K. Vermunt. 2016. "Divisive Latent Class Modeling as a Density Estimation Method for Categorical Data." *Journal of Classification* 1–21. Doi: http://dx.doi.org/10.1007/s00357-016-9195-5.

Van der Vaart, W. and T. Glasner. 2007. "Applying a Timeline as a Recall Aid in a Telephone Survey: a Record Check Study." *Applied Cognitive Psychology* 21: 227–238. Doi: http://dx.doi.org/10.1002/acp.1338.

Vermunt, J.K. and J. Magidson. 2004. "Latent Class Analysis." *The Sage Encyclopedia of Social Sciences Research Methods* 549–553. Available at: http://members.home.nl/jeroenvermunt/ermss2004a.pdf (accessed April 25, 2017).

Vermunt, J.K. and J. Magidson. 2013a. Latent GOLD 5.0 Up-grade Manual [Computer software manual]. Belmont, MA. Available at: https://www.statisticalinnovations.com/wp-content/uploads/LG5manual.pdf (accessed April 25, 2017).

Vermunt, J.K. and J. Magidson. 2013b. "Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax." *Statistical Innovations Inc., Belmont, MA*. Available at: https://www.statisticalinnovations.com/wp-content/uploads/LGtecnical.pdf (accessed April 25, 2017).

Vermunt, J.K., J.R. Van Ginkel, L.A. Van Der Ark, and K. Sijtsma. 2008. "Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis." *Sociological Methodology* 38: 369–397. Doi: http://dx.doi.org/10.1111/j.1467-9531.2008.00202.x.

Vink, G. and S. van Buuren. 2014. "Pooling Multiple Imputations When the sample Happens to be the Population." *arXiv preprint arXiv:1409.8542*. Available at: https://arxiv.org/abs/1409.8542.

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Available at: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Zhang, L.-C. and J. Pannekoek. 2015. "Optimal Adjustments for Inconsistency in Imputed Data." *Survey Methodology* 41: 127–144. Available at: http://www.statcan.gc.ca/pub/12-001-x/12-001-x2015001-eng.pdf (accessed April 25, 2017).

Received July 2016
Revised April 2017
Accepted May 2017