Updating latent class imputations with external auxiliary variables

February 22, 2018

Abstract

Latent class models are often used to assign values to categorical variables that cannot be measured directly. This 'imputed' latent variable is then used in further analyses with auxiliary variables. The relationship between the imputed latent variable and auxiliary variables can only be correctly estimated if these auxiliary variables are included in the latent class model. Otherwise, point estimates will be biased. We develop a method that correctly estimates the relationship between an imputed latent variable and external auxiliary variables, by updating the latent variable imputations to be conditional on the external auxiliary variables using a combination of Multiple Imputation of Latent Classes (MILC) and the so-called three-step approach. In contrast with existing 'one-step' and 'three-step' approaches, our method allows the resulting imputations to be analyzed using the familiar methods favored by substantive researchers.

1 Introduction

In many different disciplines, multiple observed variables are used as indicators of one latent categorical variable that cannot be measured directly. For example in sociology, multiple indicators are used to distinguish latent classes of sexual morality and pro-life values (McCutcheon, 1987). In official statistics (the field of research concerned with the publishing of statistics for government or other official agencies), indicators from multiple sources are used to estimate the number of temporary and permanent employment contracts in the Netherlands (Pavlopoulos & Vermunt, 2015). In these settings, the latent variable of interest is estimated by including observed variables as indicators in a latent class (LC) model. This LC model is then used to assign values to the latent variable

itself. This 'imputed' latent variable (also known as a 'plausible value' (Mislevy, 1991; Mislevy et al., 1992)) is often used in further analyses with auxiliary variables. For example, to relate different levels of sexual morality and pro-life values to attitudes towards abortion (McCutcheon, 1987) or to relate type of employment contract to level of education (Pavlopoulos & Vermunt, 2015).

The relationship between the imputed latent variable and auxiliary variables can only be correctly estimated if the auxiliary variables of interest are included in the LC model. Otherwise, point estimates will be biased (Wu, 2005; Monseur & Adams, 2009). This bias is due to the estimates being conditional on the imputed latent variable, and not on the latent variable itself (Bolck et al., 2004). Therefore, all auxiliary variables potentially of interest should be included in the LC model. However, this may not be possible or desired. For example in cases where an auxiliary variable is considered a distal outcome of the latent variable (Bakk, 2015, p.2). Another example is when the constructors of the measurement model do not want to share the indicator variables with the analysts due to privacy concerns. A third example is when the auxiliary variables are unavailable when constructing the measurement model due to a longitudinal or composite nature of the dataset.

Bias in the point estimates, caused by the absence of the auxiliary variables in the LC model, can be seen as a form of misclassification in the imputed latent variable. Therefore, methods that correct for misclassification should be considered, and we distinguish between different groups of methods. The first group of methods focuses on correcting the imputations of the latent variable and include Multiple Imputation for Measurement Error (MIME) (Cole et al., 2006), Regression Calibration (RC) (Spiegelman et al., 1997) and the complete re-estimation of the LC model (Schofield et al., 2014). For the latter, Multiple Imputation of Latent Classes can be used (MILC, Boeschoten, Oberski, & de Waal, 2017). The advantage of these methods is that after correction, an adapted dataset is produced that can be used to perform any type of analysis. The main drawback is that every time that new external auxiliary variables are acquired, complete re-estimation of the LC model is required. The second group of methods correct the estimate describing the relationship that is prone to bias. This group includes methods as simulation extrapolation (SIMEX) (Cook & Stefanski, 1994) and the latent class three-step approach (Bolck et al., 2004). Their main advantage is that uncertainty due to misclassification is correctly incorporated into the estimates after new external variables are acquired. However, an important disadvantage is their inflexibility; a separate

procedure needs to be followed for every analysis and a likelihood needs to be available to obtain the estimates of interest. Such complications prevent these important corrections from gaining traction among substantive researchers.

We develop a general approach by combining a method based on model correction (implementation of the LC model using the MILC method) with a method that is based on correction for bias (the three-step approach). This new approach preserves the advantages of both methods while discarding their disadvantages due to its generic nature. More specifically, this combined method (from now on denoted as the three-step MILC method) uses an LC model to create multiple imputations of the latent variable, which includes both parameter uncertainty and latent variable uncertainty into the estimate of the variance. Next, information from the LC model is used to estimate the amount of misclassification in the imputed latent variable. The estimate of this misclassification is then used to correct the relationship between the imputed latent variable and external auxiliary variables. Finally, the latent variable imputations are updated to be conditional on the external variables.

In the second section, issues currently faced by researchers are discussed in more detail, for which we present the three-step MILC method as a solution in the third section. In the fourth section, a simulation study is conducted to investigate the performance of the three-step-MILC method. In the fifth section, the three-step-MILC method is applied on two empirical datasets, followed by a discussion in the sixth section.

2 Background

Researchers frequently summarize multiple observed variables $(Y_1, ..., Y_L)$ into one latent variable (X). A model $\mathrm{P}(\mathbf{Y}|X)$ is constructed to estimate the values of X. This model is used to assign estimated values to X, resulting in an imputed version of the latent variable, W. Different rules can be used to assign values to W using $\mathrm{P}(X|\mathbf{Y})$, such as modal (McLachlan, 1992), proportional (Dias & Vermunt, 2008) or random assignment. With the latter, individuals are assigned to classes by sampling from the posterior $\mathrm{P}(X|\mathbf{Y})$, so $W \sim \mathrm{P}(X|\mathbf{Y})$ (Bakk, 2015, p.11). Regardless of the method used for assigning values to W, W is never a perfect representation of X; some misclassification is always introduced (Bakk, 2015, p.12).

The imputed variable W is created so it can be used in further analyses with auxiliary variables ($\mathbf{Q}$). As addressed by Lanza et al. (2013) and implied by Blackwell et al. (2015), $\mathrm{P}(X|\mathbf{Q})$ can only be correctly estimated using $\mathrm{P}(W|\mathbf{Q})$ if $\mathbf{Q}$ is included in the model used to assign values to W. In other words, when the covariate-adjusted posterior $\mathrm{P}(X|\mathbf{Y},\mathbf{Q})$ is used to determine W. Otherwise, biased estimates for $\mathrm{P}(X|\mathbf{Q})$ are obtained (Bartlett et al., 2015; Bolck et al., 2004; Schofield, 2014; Tanner & Wong, 1987), unless the measurement is perfect, such that $\mathrm{P}(W|\mathbf{X}) = 1$ for exactly one value of X for each value of W (see also Marsman et al. (2016) for the same result in IRT).

Although this problem does not arise if $\mathbf{Q}$ is included in the LC model used when estimating X, we consider situations here where this is neither possible nor desired. For example, $\mathbf{Q}$ may not have been collected yet when $\mathrm{P}(\mathbf{Y}|X)$ was estimated, or researchers may be resistant to include $\mathbf{Q}$ in the initial measurement model. As a result, $\mathrm{P}(X|\mathbf{Y},\mathbf{Q})$ is not available, only $\mathrm{P}(X|\mathbf{Y})$ is. It is, however, possible to obtain information about the misclassification in W from the LC model, $\mathrm{P}(W|X)$, which is estimated as a byproduct of the parameters in $\mathrm{P}(\mathbf{Y}|X)$ and $\mathrm{P}(X)$ and the chosen assignment rule (Bakk et al., 2013). These two pieces of information, $\mathrm{P}(W|X)$ and $\mathrm{P}(W|\mathbf{Q})$, can be combined to obtain an estimate for $\mathrm{P}(X|\mathbf{Q})$ using maximum likelihood (Vermunt, 2010) or weighting (Bolck et al., 2004), which are both approaches of latent class three-step modelling. By specifying the log-linear model in its most general form, the newly imputed version of W can be used to estimate any type of relationship with $\mathbf{Q}$. Consequently, researchers do not have to think in advance about the kind of relationship to investigate at a later stage.

However, when a single imputation of W is created using this approach, uncertainty about X is not included in the estimate of the variance. Therefore, multiple imputations of W should be created so that the differences between the imputations reflect this uncertainty (Rubin, 1987, p.76). This approach is a combination of the MILC method used for model correction and the three-step approach used to correct for bias.

3 Methodology

In this section, we present a solution for the problem that biased estimates are obtained when an imputed latent variable is related to external auxiliary covariates. The methodology is discussed step by step, starting with the methodology of the MILC method (Boeschoten, Oberski, & de Waal,

2017) followed by its three-step (Vermunt, 2010) extension.

3.1 MILC

On the left hand side of Figure 1, a graphical overview of the MILC method is shown. The starting point of the method is a dataset comprising L indicator variables. In the **first** step, m bootstrap samples are drawn by sampling with replacement from the observed probability distribution of the original data, where m is equal to the number of multiple imputations created in a later stage. By using multiple imputations, we are able to include uncertainty due to measurement error in the indicators when estimating the variance. By using bootstrap samples, parameter uncertainty is also included. This is especially recommended when datasets with smaller sample sizes are used as parameter uncertainty can be substantial in such cases (Wisz et al., 2008).

In the **second** step, an LC model is estimated for every bootstrap sample using the L indicator variables $(Y_1, ..., Y_L)$ of latent variable X, which has C categories denoted by $x = 1, ..., C$. C is equal over the bootstrap samples. If MILC is used to correct for measurement error in combined datasets, C is equal to the number of categories in the indicators. Available auxiliary variables can also be incorporated in the LC model as covariates and are denoted by $\mathbf{Z}$. The LC model for the probability of response pattern $\mathrm{P}(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z})$ is then defined as:

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

In some applications, one may wish to account for combinations of scores between the covariate variables and the latent variable that are not possible in practice. An example of such an impossible combination of scores is a Dutch person having marital status 'married' and age 'below 16 years', as this is prohibited by law. Edit restrictions are used to account for such impossible combinations of scores (De Waal et al., 2012), and can be specified in the LC model:

$$\mathrm{P}(X = \text{'married'}|Z = \text{'age below 16'}) = 0. \tag{2}$$

This is especially relevant in cases where LC models are used to correct for misclassification (Biemer, 2011), because a violation of an edit restriction is by definition due to misclassification in one of the

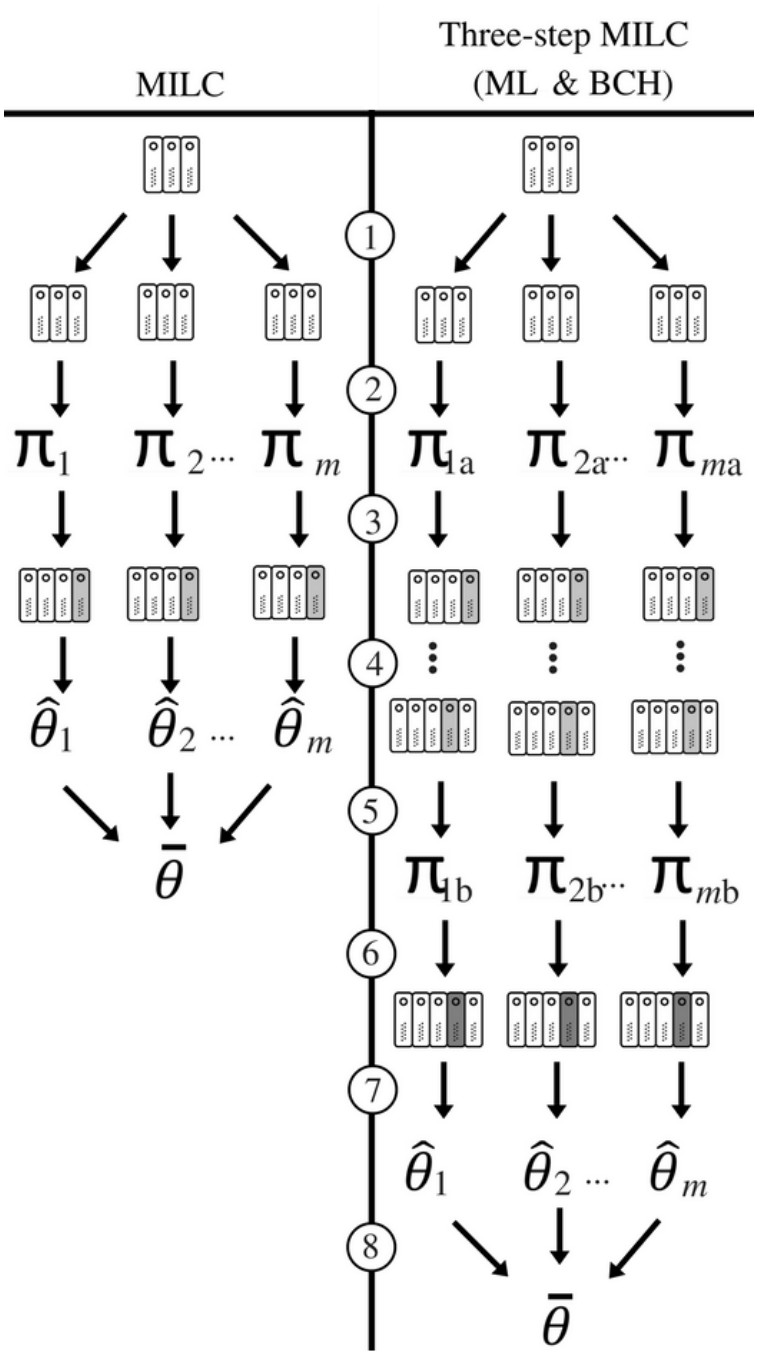

Figure 1: Graphical overview of the MILC method in the left pane and the Three-step MILC (using ML and BCH) in the right pane. All methods start with a dataset containing indicators and available covariate variables. At step 1, m bootstrap samples are drawn from the original dataset. At step 2, an LC model is built for each bootstrap sample (denoted by π). At step 3, m imputations for the latent variable are created. Estimates of interest are obtained from the m imputations, represented by $\hat{\theta}$ (step 4). Pooling these estimates to obtain $\bar{\theta}$ is the fifth step. If the imputed latent variable needs to be related to external variables, step 4 is obtaining the classification table. Step 5 is to apply the ML or BCH correction procedure. Posterior membership probabilities are used to update the imputations for the latent variable (step 7). From the imputations, estimates can then be obtained and pooled (step 8).

variables to which the edit restriction applies. By including the edit restriction in the LC model, the appearance of the impossible combination of scores is prevented by constraining the parameter estimates of the LC model.

In the **third step**, m new empty variables are created in the original dataset and imputed by sampling one the LC's using the posterior membership probabilities obtained from the corresponding m LC models:

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

These posterior membership probabilities represent the probability that a unit is a member of an LC given its combination of scores on the indicators and covariates used in the LC model. At this point, a dataset is obtained containing multiple imputations of the latent variable. From now on, the indicators themselves are no longer needed.

In the **fourth** step, estimates of interest are obtained from the m imputed variables. These can be logistic regression coefficients, tests for model fit, cell proportions in cross tables or any other estimate of interest to the researcher.

In the **fifth** step, the m estimates are pooled by using the rules defined by Rubin (Rubin, 1987, p.76). The pooled estimate is obtained by:

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i. \tag{4}$$

The total variance is estimated as

$$\mathrm{VAR}_{\mathrm{total}} = \overline{\mathrm{VAR}}_{\mathrm{within}} + \mathrm{VAR}_{\mathrm{between}} + \frac{\mathrm{VAR}_{\mathrm{between}}}{m}, \tag{5}$$

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

$$\overline{\mathrm{VAR}}_{\mathrm{within}} = \frac{1}{m} \sum_{i=1}^{m} \mathrm{VAR}_{\mathrm{within}_i}, \tag{6}$$

and $\text{VAR}_{\text{between}}$ is calculated by

$$\text{VAR}_{\text{between}} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta})(\hat{\theta}_i - \hat{\theta})'. \tag{7}$$

$\text{VAR}_{\text{between}}$ contains both uncertainty caused by missing or conflicting data and parameter uncertainty (Van der Palm et al., 2016).

3.2 Three-step MILC

The MILC method can be expanded to incorporate the three-step approach, enabling the investigation of relationships between latent variable X and auxiliary variables not included in the initial LC model ($\mathbf{Q}$). This procedure is shown on the right hand side of figure 1.

The first three steps of the MILC method are applied to create imputations for W: bootstrap samples are created (step **one**), LC models are estimated (step **two**) and m empty variables are imputed (step **three**). An extra step is now required to estimate the classification error of the imputed variables $W_1, .., W_m$ (step **four**):

$$\text{P}(W = w|X = x) = \frac{\sum_{\mathbf{y}}\sum_{\mathbf{z}}\text{P}(\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})\text{P}(X = x|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})\text{P}(W = w|\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z})}{\text{P}(X = x)}. \tag{8}$$

Note that $\text{P}(W = w|X = x)$ can be estimated from the imputed dataset directly and that a separate estimate for $\text{P}(W = w|X = x)$ is obtained for every imputation of W.

In step **five**, the relationship between external auxiliary variables $\mathbf{Q}$ and latent variable X ($\text{P}(X = x|\mathbf{Q} = \mathbf{q})$) is estimated by using the m imputations obtained in step **three** and the corresponding classification errors obtained in step **four**. $\text{P}(X = x|\mathbf{Q} = \mathbf{q})$ is estimated either by using the ML approach (Vermunt, 2010) or the Bolck-Croon-Hagenaars (BCH) approach (Bolck et al., 2004). For both approaches, an LC model is estimated. With the ML approach this is done using a procedure that is comparable to estimating a regular LC model, while for the BCH approach this is done by using a weighting procedure.

With the ML approach, an LC model is specified where W is used as the only indicator of X and this relationship is fixed to the classification error $\text{P}(W = w|X = x)$. The form of the

$\mathrm{P}(X = x|\mathbf{Q} = \mathbf{q})$ distribution is specified in its most general form, and is estimated as:

$$\mathrm{P}(W = w|\mathbf{Q} = \mathbf{q}) = \sum_{x=1}^{C} \mathrm{P}(X = x|\mathbf{Q} = \mathbf{q})\mathrm{P}(W = w|X = x). \tag{9}$$

Based on equation 9, posterior membership probabilities can be obtained for every combination of scores on W and $\mathbf{Q}$:

$$\mathrm{P}(X = x|\mathbf{Q} = \mathbf{q}, W = w) = \frac{\mathrm{P}(X = x|\mathbf{Q} = \mathbf{q})\mathrm{P}(W = w|X = x)}{\sum_{x'=1}^{C} \mathrm{P}(X = x'|\mathbf{Q} = \mathbf{q})\mathrm{P}(W = w|X = x')}. \tag{10}$$

With the BCH correction method, $\mathrm{P}(X = x|\mathbf{Q} = \mathbf{q})$ is estimated by weighting $\mathrm{P}(W = w|\mathbf{Q} = \mathbf{q})$ by the inverse of $\mathrm{P}(W = w|X = x)$:

$$\mathrm{P}(X = x|\mathbf{Q} = \mathbf{q}) = \sum_{w=1}^{C} P(W = w|\mathbf{Q} = \mathbf{q})d^*_{wx}, \tag{11}$$

where d^*_{wx} represents an element of the inverted $C \times C$ matrix $\mathbf{D}$ with elements $\mathrm{P}(W = w|X = x)$. The obtained result can be plugged into equation 9 to obtain posterior membership probabilities for every combination of scores on W and $\mathbf{Q}$, as shown in equation 10.

The original BCH method has two major drawbacks. First, it can create negative values for the elements $\mathrm{P}(X = x|\mathbf{Q} = \mathbf{q})$, resulting in inadmissible solutions. Second, edit restrictions (to prevent the appearance of impossible combinations of scores on X and $\mathbf{Q}$) cannot be incorporated. To circumvent these issues, Boeschoten, Croon, & Oberski (2017) placed the BCH approach in a framework of quadratic loss functions and linear equality and inequality constraints. This approach is used throughout the remainder of this paper.

Both the ML and the BCH correction procedure result in a set of posterior membership probabilities for every combination of scores on W and $\mathbf{Q}$ (and for each of the m bootstrap samples), which can be used to create m new imputations for W. The same procedure as followed in step three is used here, although the posteriors are now also conditional on $\mathbf{Q}$. Performing these new imputations is the **sixth** step of the procedure. Step **seven** is then to obtain estimates of interest for each bootstrap sample, which are likely to be parameter estimates describing the relationship between the imputed latent variable W and the external auxiliary variables $\mathbf{Q}$. In step **eight**, these

estimates are pooled using Rubin's rules.

4 Simulation

4.1 Simulation setup

To empirically evaluate the performance of the three-step MILC method, we conducted a simulation study using R (R Core Team, 2014). We started by creating a theoretical population using Latent GOLD (Vermunt & Magidson, 2013) containing five variables: three dichotomous indicators (Y_1, Y_2, Y_3) of the property of interest (X) and two dichotomous variables Q_1 and Q_2 that we consider as external auxiliary variables, so they are not included in the initial LC model. Variations are made according to scenarios described in the subsequent subsections. A theoretical population is used to draw $1,000$ samples and these are used to evaluate the performance of the three-step MILC approach, following the steps described in section 3.2. In the initial LC model (step two), only the three indicators are included in the LC model. At step five, the three-step procedure is applied for external auxiliary variables Q_1 and Q_2 simultaneously.

In this simulation study, the performance of two different approaches to the three-step MILC method are evaluated: ML and BCH. As a reference, we also include estimates obtained when no correction method was applied, so where W is imputed using an LC model containing only indicators, and its relationship with Q_1 and Q_2 is investigated directly.

When evaluating the correction methods, the relationship between X and Q_1 and Q_2 should be preserved. There are two types of relationships we are specifically interested in. For the first, we compare the logit coefficient of latent variable X regressed on Q_1 in the theoretical population with the logit coefficient of imputed W regressed on Q_1. This relationship is investigated using four performance measures:

- The bias of the logit coefficient, which is equal to the difference between the average estimate over all replications and the value found in the theoretical population.
- The coverage of the 95% confidence interval.
- The ratio of the average standard error of the estimate over the standard deviation of the $1,000$ replication estimates is examined to confirm that the standard errors of the estimates

are properly estimated.

- The root mean squared error, which is the root of the average of the squares of the errors.

Second, we are interested in a restricted relationship as described in equation 2. In the theoretical population, $\mathrm{P}(X = 1|Q_2 = 2) = 0$. When an impossible combination of scores between X and an external auxiliary variable exists but is not accounted for in the LC model, it can appear in the imputed dataset because LC models do not assign probabilities of exactly 0 by default. Therefore, we investigate whether the restricted cell in the imputed dataset $\mathrm{P}(W = 1|Q_2 = 2)$ indeed contains zero observations. This is done by investigating the observed frequency of this specific cell (the observed cell proportion is multiplied by the sample size to obtain the observed frequency).

Previous research has shown that the performance of the MILC method is strongly related to the entropy R^2 value of the LC model (Boeschoten, Oberski, & de Waal, 2017). The entropy R^2 indicates how well one can predict class membership based on the observed variables, and is influenced by the measurement quality of the indicators. Therefore, we investigate a range of realistic values for the measurement quality of the indicators in the simulation study. The conclusion was also made that 5 imputations are sufficient to obtain unbiased estimates and appropriate coverage of the 95% confidence interval (Boeschoten, Oberski, & de Waal, 2017), so $m = 5$ is used in this simulation study as well. Furthermore, different sample sizes are investigated, since they influence the standard errors and thereby the confidence intervals. The main properties of this simulation study are summarized as follows:

- Class-specific response probabilities of the three dichotomous indicators of dichotomous X (Y_1, Y_2, Y_3): 0.70; 0.80; 0.90; 0.95; 0.99 (corresponding entropy R^2 values respectively: 0.31; 0.59; 0.86; 0.96; 0.99).

- Logit coefficients of X regressed on Q_1: 0.00; 0.50; 1.00; 2.00.

- Different proportions for $\mathrm{P}(Q_2 = 2)$, where $\mathrm{P}(W = 1|Q_2 = 2)$ should contain zero observations: $\mathrm{P}(Q_2)$= 0.01; 0.05; 0.10; 0.20.

- Sample size: 200; 500; 1,000.

- Number of bootstrap samples and multiple imputations $m = 5$.

- Correction methods: No correction method; ML; BCH.

4.2 Simulation results

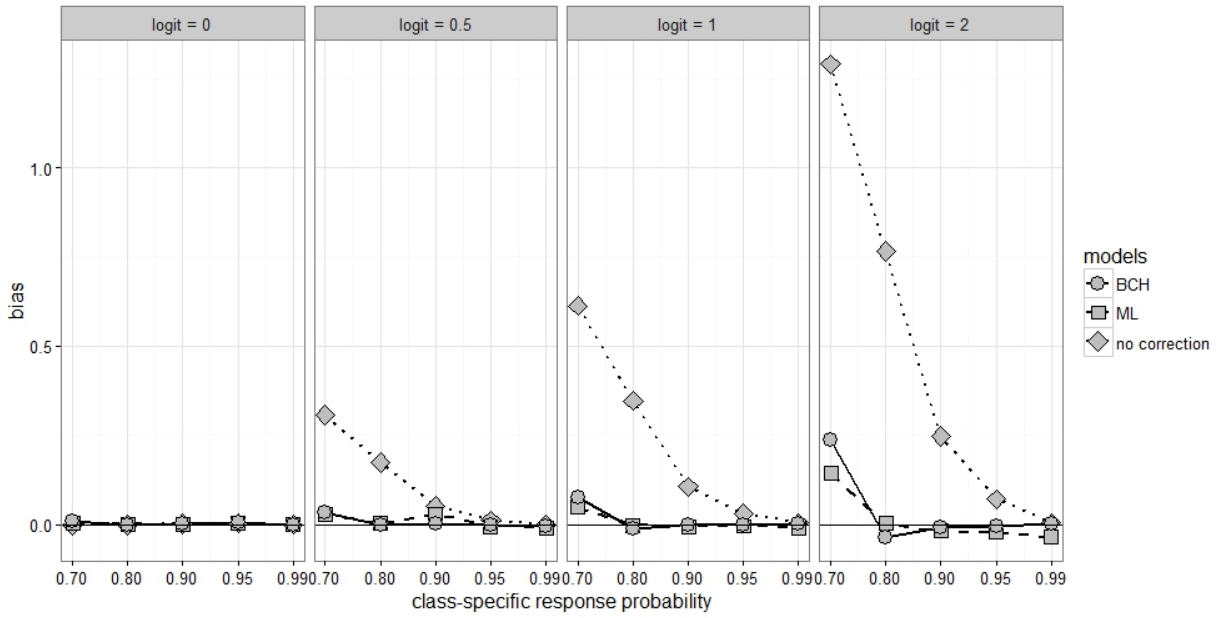

Figure 2: Displayed is the bias of the logit coefficient of imputed variable W regressed on covariate Q_1. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variable. Sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.2$.

Figure 2 shows the bias of the logit coefficient of latent variable X regressed on covariate Q_1 when estimated by using imputed variable W regressed on covariate Q_1. When comparing the results over different strengths of the logit coefficients, in general there is more bias when the logit coefficient increases. When the logit coefficient is 0 (i.e. there is no effect), there is approximately no bias in all conditions for both correction methods, and when no correction method is applied. When the logit coefficient increases, bias increases as well if no correction method is applied, while it remains low when correction methods are applied. The only exception is when the class-specific response probabilities are low (0.70).

Figure 3 shows the coverage of the 95% confidence interval of the logit coefficient of imputed variable W regressed on covariate Q_1. If the population logit coefficient is 0, the correction methods perform approximately equally well, and not using a correction method also leads to desired results.

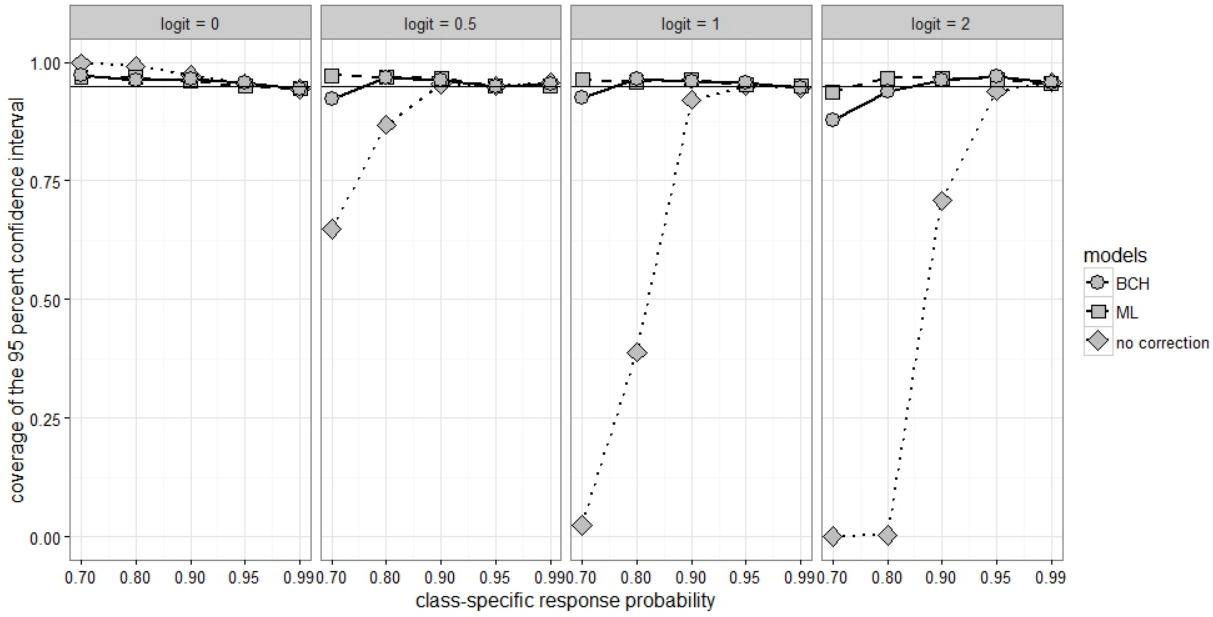

Figure 3: Displayed is the coverage of the 95% confidence interval of the logit coefficient of imputed variable W regressed on covariate Q_1. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variable. Sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.2$.

As the population logit coefficient becomes larger, undercoverage becomes more of a problem when no correction method is applied, especially when the class-specific response probabilities are low. The results obtained for the ML and BCH method are very similar. The coverage rates are generally somewhat higher for ML when the class-specific response probabilities are lower, while the coverage rates are almost identical when the class-specific response probabilities are higher. This is unrelated to the strength of the population logit coefficient.

Figure 4 shows the ratio of the average standard error of the estimated logit coefficients over the standard deviation of the logit coefficient of imputed variable W regressed on covariate Q_1. Here we investigate whether the estimated standard errors are indeed equal to the standard deviation of the estimates. When no correction is applied, the standard errors are too large when the class-specific response probabilities are low, and the ratio comes closer to 1 as the class-specific response probabilities increase. This is unrelated to the size of the population logit coefficient. A comparable trend is seen for the ML and BCH correction methods. The ratio is however much closer to the desired value of 1 for both methods compared to when no correction method is applied. The trend

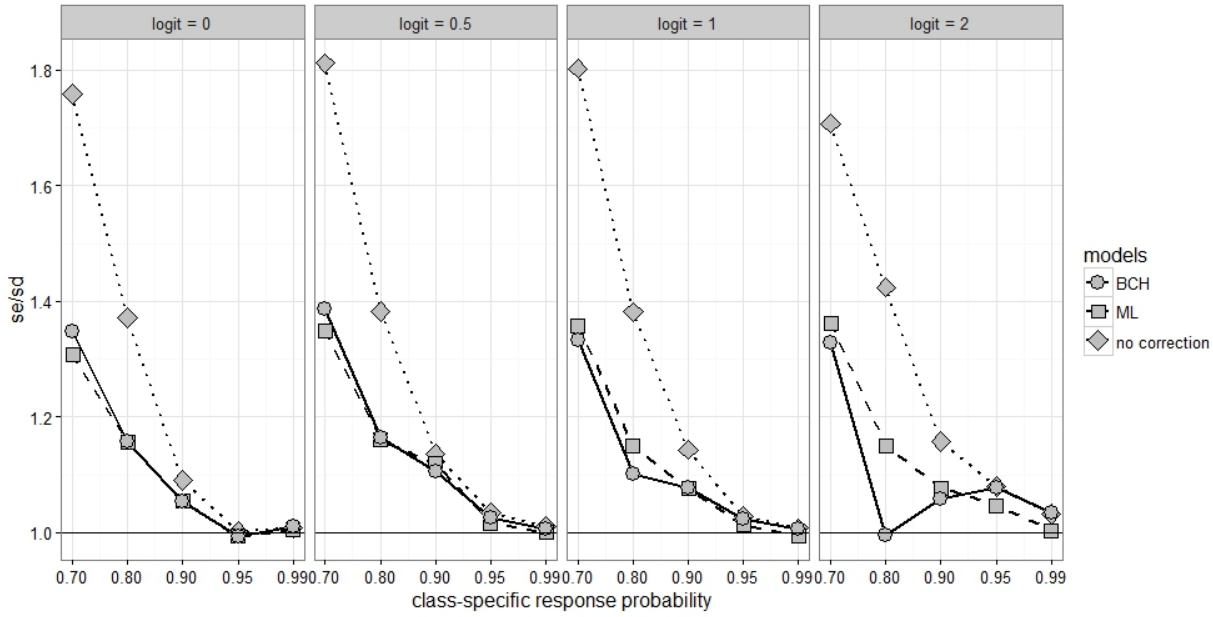

Figure 4: Displayed is the ratio of the average standard error of the logit coefficients over the standard deviation of the logit coefficients of imputed variable W regressed on covariate Q_1. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variables. Sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.2$.

for BCH method becomes a bit more unstable as the size of the logit coefficient increases, while the trend for the ML method seems more stable.

Figure 5 shows us the root mean square error, where the errors are represented by the difference between the logit coefficient of imputed W and Q_1 and its value in the theoretical population. When the logit coefficient is 0, using no correction is the best option in terms of RMSE. However, as soon as the logit coefficient increases, the RMSE of using no correction becomes larger compared to both correction methods. This effect becomes stronger as the logit coefficient increases. The correction methods perform approximately equally well, where the RMSE is generally a bit lower for the ML method.

Figure 6 shows the number of times that the combination of scores that is in practice impossible, $\mathrm{P}(W = 1|Q_2 = 2)$, is observed in the imputed dataset averaged over $1,000$ replicates. We see the results when no correction is applied, and for the ML and BCH correction methods. Furthermore, we see the results for different class-specific response probabilities. When no correction is applied, the observed frequency (the cell proportion multiplied with the sample size) is strongly related to

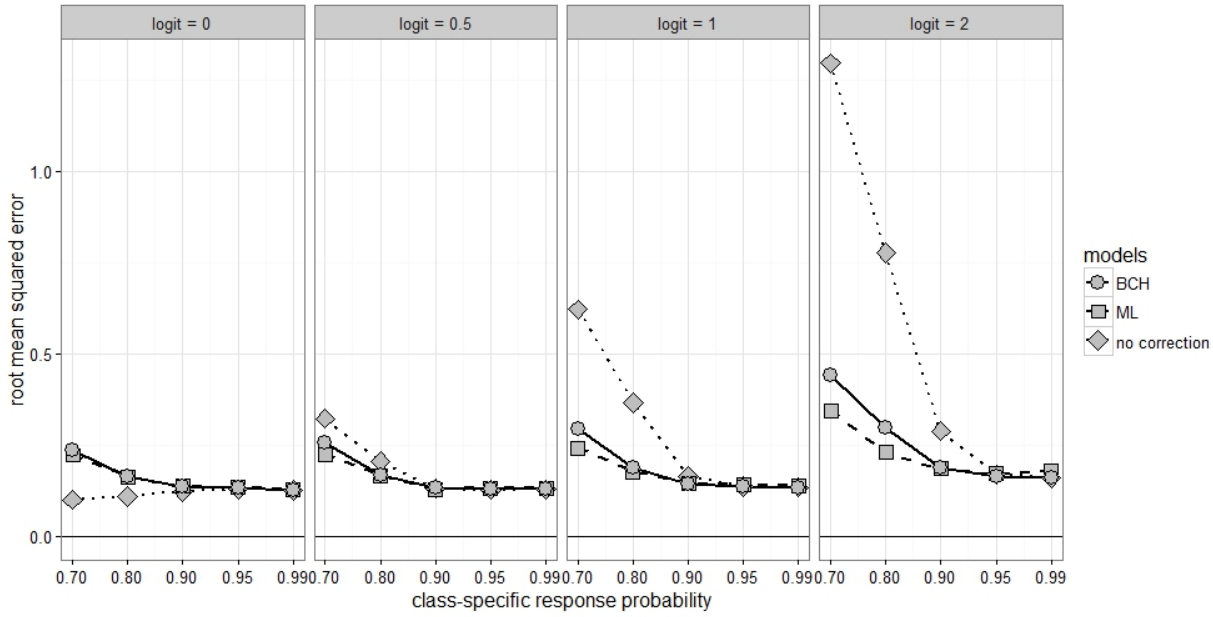

Figure 5: Displayed is the root mean squared error, where the errors are represented by the difference between the logit coefficient of imputed W on Q_1 and its value in the theoretical population. The different shapes represent the different correction methods (ML; BCH) and when no correction method is used, they are connected by lines of different types. Results are shown for different population values of the logit coefficient and for different class-specific response probabilities of the indicators of the latent variables. Sample size is $1,000$ and $\mathrm{P}(Z = 2) = 0.2$.

these class-specific response probabilities. We see that when the class-specific response probabilities are low (0.70), the observed frequency of $\mathrm{P}(W = 1|Q_2 = 2)$ in this condition is around 65. This number decreases as the class-specific response increases, but even when the class-specific response probabilities are 0.90, there are still impossible combinations of scores created. When the ML correction method is applied, 0 impossible combinations of scores are created under all conditions investigated, only when the class-specific response probabilities are 0.70, the number of impossible combinations created is not exactly 0, but still below 1. With the BCH correction method, impossible combinations of scores are only created when the class-specific response probabilities are 0.70. This makes sense, the entropy R^2 is very low in these conditions so we did not expect the correction methods to perform well here. In all other conditions, no impossible combinations of scores are created.

Overall it can be said that problems in terms of bias and coverage of estimates can be severe if no correction is applied when performing a latent class three-step method. Both correction methods (ML and BCH) have shown to improve these results and the differences in results between

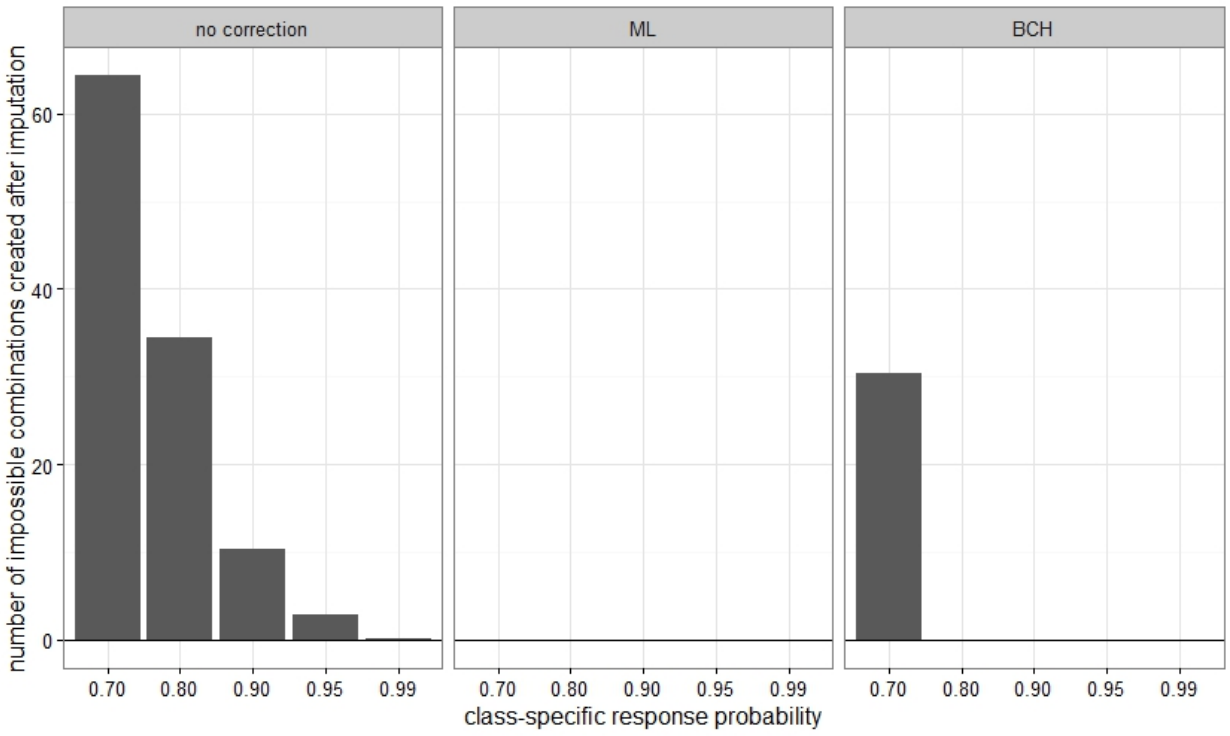

Figure 6: The bars in this histogram display the number of times that the impossible combination $P(W = 1|Q_2 = 2)$ is observed when no correction, ML and BCH are applied. The sample size is $1,000$ and the marginal of Q_2 is 0.2. Results are displayed for different class-specific response probabilities of the indicator variables.

the methods are minimal. Even with low class-specific response probabilities, improvements were detected for both methods, although not all problems are solved in these low-quality cases. For example, BCH was not able to handle edit restrictions in combination with low class-specific response probabilities and the coverage rates for BCH were also somewhat lower in these cases.

5 Applications

In this section, the flexibility of the three-step MILC method is illustrated by applying the method to two empirical datasets. First, the method is applied to a composite dataset used in official statistics, where researchers use LC models to correct for measurement error. Second, the method is applied to a dataset containing scores on math items, where discretized IRT models can be used to measure mathematical ability.

In each application, the relationship between an imputed latent variable and an external auxiliary variable is investigated. First, an LC model is applied without including the auxiliary variable, this will be denoted as uncorrected MILC. The relationship between the imputed latent variable and the auxiliary variable is then investigated. Next, the estimate of this relationship is corrected using the ML and BCH method. For comparison, the estimate of this relationship is also investigated when the external variable was included in the initial LC model.

5.1 A latent class model to correct for measurement error in official statistics

We investigate the relationship between home ownership and marital status. To estimate this relationship, a composite dataset is used that consists of two surveys carried out by LISS (Longitudinal Internet Studies for the Social sciences) from 2013 (Scherpenzeel, 2011), which is administered by CentERdata (Tilburg University, The Netherlands) and a population register from Statistics Netherlands from 2013. Since this composite dataset contains two variables indicating whether a person is either a 'home-owner' or 'home-renter or other', we use these as indicators to measure the 'true' variable 'home-owner'/'home-renter or other', which can correct for misclassification in the indicator variables. Since an LC model with only two indicators is not identifiable, we also included a covariate measuring whether someone receives rent benefit from the government. Since a person can only receive rent benefit if this person rents a house, we included an edit restriction here. For a

Table 1: The columns represent the (pooled) estimate and 95% confidence interval around the intercept and the logit coefficient of the variable owning/renting a house. The first row represents the results obtained when no correction method is applied. The second and third row represent the ML and BCH correction methods. The last row represents the results obtained when the auxiliary variable 'marital status' is included in the initial LC model.

	intercept		marriage	
	estimate	95% CI	estimate	95% CI
No correction	-2.6829	[-2.9533; -2.4126]	1.2524	[0.9820; 1.5227]
ML	-2.7221	[-2.9898; -2.4544]	1.3247	[1.0570; 1.5924]
BCH	-2.6097	[-2.8669; -2.3526]	1.3866	red[1.1294; 1.6437]
Included	-2.7712	[-3.0389; -2.5036]	1.3817	[1.1140; 1.6493]

detailed description of the composite dataset and the processing it, we refer to (Boeschoten, Oberski, & de Waal, 2017). Next, we impute the true variable measuring 'home-owner'/'home-renter or other' using the LC model, and we investigate the relationship between this variable and a covariate measuring marital status.

More specifically, we investigate whether marriage can predict home ownership. First, uncorrected MILC was applied (marriage was not included in the LC model). Second, MILC was applied while the auxiliary variable 'marriage' was included as a covariate in the LC model. Results of both these models can be found in Table 1. Here we see that both the intercept and the logit coefficient are closer to 0 when the auxiliary variable was not included in the initial model, compared to when it is included.

Furthermore, if we apply either the ML or BCH method to correct the imputations made using a model without covariate, we see that the differences between these corrected estimates and the estimates when including the auxiliary variable into the model are much smaller, although they are not exactly equal.

In general we can conclude that non-married individuals are approximately equally likely to own than rent a house (non-married individuals are approximately $e^{-2.7} = 0.07$ times more likely to own than rent a house). Married individuals are more likely to own a house than to rent it. However, if we would not have included this auxiliary variable in the LC model, we would conclude that they are approximately $e^{1.2524} = 3.4987$ times more likely to own than to rent a house. If we would have either included the auxiliary variable in the model or used a correction method, we would conclude that this relation is actually somewhat stronger, either $e^{1.3247} = 3.7611$ (ML), $e^{1.3866} = 4.0012$ (BCH) or $e^{1.3817} = 3.9817$ (included) times more likely. In general, the results from

Table 2: The columns represent the (pooled) estimate and 95% confidence interval around the intercept and the logit coefficient of the outcome variable 'math ability'. The first row represents the results obtained when no correction method is applied. The second and the third row represents the results of the ML and BCH correction methods, the last row represents the results obtained when the auxiliary variable is included in the initial LC model.

	intercept		high ability	
	estimate	95% CI	estimate	95% CI
No correction	0.6837	[0.1917; 1.1757]	-1.0867	[-1.8377; -0.3358]
ML	0.7259	[0.2223; 1.2294]	-1.1329	[-1.8794; -0.3864]
BCH	0.7351	[0.2176; 1.2527]	-1.0726	[-1.8128; -0.3325]
Included	0.7421	[0.1767; 1.3075]	-0.9293	[-1.6718; -0.1869]

the included model and from both the correction methods are quite close.

5.2 A discretized IRT model in psychometrics

Here we investigate the relationship between mathematical ability of a child and the level of education of its mother. To investigate this relationship, we make use of the 2015 PISA data. More specifically, we use a subset of the data containing 141 Dutch 15-year old pupils who conducted booklet number 43 of the mathematical ability test. This booklet contains 20 mathematical ability questions which can be graded with either 'correct'/'incorrect' or 'correct'/'partly correct'/'incorrect'. The scores obtained by answering these questions are used as indicators of the latent variable 'mathematical ability', which we measure using a discretized IRT model in Latent GOLD with 2 classes ('low level'/'high level'). We impute the latent variable of mathematical ability using the discretized IRT model, and we investigate the relationship between mathematical ability of the child and its mothers' level of education. To measure the level of education of the mother, we used a dichotomous variable indicating whether she finished the International Standard Classification Level 4 (ISCL 4: post-secondary non-tertiary education).

We investigate whether mothers' level of education can predict math ability of her child. First, uncorrected MILC was applied (where mothers' level of education was not included in the LC model as a covariate). Here, the intercept is $e^{0.6837} = 1.9812$, which can be interpreted as the odds that a mother has education level ISCL 4 when a child hass mathematical ability 'above level'. The logit coefficient is $e^{-1.0867} = 0.3373$. The odds for a mother to have education level ISCL 4 when her child does not have mathematical ability 'above level' is 0.3373 times the odds when her child has mathematical ability 'below level'.

In this example, it can be particularly undesirable that mothers' level of education is included in the LC model as a covariate, since it then contributes to the assignment of children to classes measuring their mathematical ability, and researchers can find it undesirable that children are assigned to a math ability class based on their mothers' ability. It can be seen in the results that the relationship between the two variables is stronger when the mothers' level of education is included in the model. When MILC is applied with mothers' level of education included in the LC model as covariate, this logit coefficient is $e^{-0.9293} = 0.3948$, so the estimated relationship between mothers' level of education and math ability of her child is stronger when this variable is included in the model compared to when it's not included in the model.

However, when the model without covariate is used, there is no correction for the fact that the relationship of interest is estimated using an imputed version of 'math ability' and not the true values of 'math ability'. Therefore, correction methods are applied and they both result in small adjustments compared the uncorrected results. For both ML and BCH, the intercept is a bit larger compared to the uncorrected results, while the logit coefficient is a bit smaller.

6 Discussion

In this paper we introduced the three-step MILC method, which updates latent variable imputations to be conditional on external variables. If the latent variable imputation is not corrected to be conditional on external variables, the point estimates of the relationship between the imputed latent variable and the external variables are biased. This bias is caused by the fact that an imputation of a latent variable is generally not a perfect representation of that latent variable, it contains some measurement error. While a method that corrects for measurement error is required, current methods lack either general applicability or flexibility. Therefore, the MILC method and the three-step approach of the latent class model are combined into one generic procedure.

We incorporated two alternative correction procedures in the three-step MILC method, and evaluated them in terms of their ability to correct for bias in point estimates due to measurement error in the imputed latent variable. We assessed the different procedures in terms of bias of the estimates, coverage of their 95% confidence interval, standard error of the estimate over the standard deviation over the estimates and root mean squared error. Furthermore, we investigated

whether the different correction procedures were able to successfully incorporate edit restrictions. This all was investigated under a number of different conditions in a simulation study.

From the simulation study, it can be concluded that the necessity of applying the three-step MILC method (or probably any other correction procedure) was strongly related to the strength of the relationship under investigation. If the true logit coefficient was 0, i.e. there was no relationship between the imputed latent variable and the external covariate, then there was also no bias if no correction procedure was used. In other words, there was nothing to correct. Furthermore, the necessity of applying the three-step MILC was also related to class separation.

When class separation was higher, results of better quality were obtained when no correction was applied. However, regardless of the strength of the relationship under investigation or the strength of the class separation, results always improved when a correction method was applied compared to when no correction method was applied. Furthermore, the BCH and ML correction methods performed in a very comparable way. ML can be recommended over BCH in cases where the class-specific response probabilities of the indicators are low, since the simulation results showed that the coverage rates where somewhat lower for BCH here, and they showed that BCH was not able to succesfully incorporate edit restrictions in these situations.

It should also be noted that estimates of interest can also be obtained directly after applying the BCH or ML correction procedure. These results are then obtained from the output generated by these correction procedures and not by creating new imputations and investigating these. To be able to directly obtain these corrected estimates, the researcher needs to think about the type of relationship that he or she wants to investigate and specify the correction procedure correspondingly. However, using an imputed variable allows for much more flexibility, because when investigating this variable in relationship with other variables, the researcher is not limited to how these relationships are specified in the correction procedure.

The two applications show the great flexibility of the three-step MILC method. In the first application, a composite dataset from official statistics is used, where LC models are applied to correct for measurement error. In the second application, a dataset containing childrens' scores on math items is used, where a discretized IRT model is used to investigate childrens' math ability. In both applications, both ML and BCH perform approximately equally well. Furthermore, the second application also shows that directly including an external variable into the LC model can

have an undesirable influence on the class assignments.

The results in this paper invite further investigation of the applicability of the three-step MILC method to continuous data or other types of models that can be fit into the latent class framework, such as hidden Markov or multilevel latent class models.

Further research into the three-step MILC method is also required due to the limited scope of the current simulation study. Only a small latent class model is investigated, where various model assumptions were made. For example, the auxiliary variables were assumed to be free of measurement error. This can be strange in an official statistics setting where the LC model itself is used to correct for measurement error in the indicator variables. In addition, the indicator variables are assumed to be conditionally independent and the misclassification error in these indicators is assumed to be unrelated to the auxiliary variables. Unfortunately, these assumptions will not always be met in practice and thus how robust this method is to violations of these assumptions should be looked into.

In summary, the three-step extension of the MILC method presented in this paper allows correct estimation of relationships between an imputed latent variable and external auxiliary variables. This method is a promising solution to correct for misclassification, due to its general applicability and flexibility.

References

Bakk, Z. (2015). *Contributions to bias adjusted stepwise latent class modeling* (Doctoral dissertation). Retrieved from `https://pure.uvt.nl/portal/files/8521154/Bakk_Contributions_16_10_2015.pdf`

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*(1), 272–311.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487.

Biemer, P. P. (2011). *Latent class analysis of survey error* (Vol. 571). Hoboken, New Yersey: John Wiley & Sons. (ISBN: 978-0-470-28907-5)

Blackwell, M., Honaker, J., & King, G. (2015). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research*, 0049124115585360.

Boeschoten, L., Croon, M., & Oberski, D. (2017). A note on applying the bch method under linear equality and inequality constraints. *unpublished*. Retrieved from `http://daob.nl/wp-content/papercite-data/pdf/boeschoten2017note.pdf`

Boeschoten, L., Oberski, D., & de Waal, T. (2017). Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (milc). *Journal of Official Statistics*, *33*(4), 921–962.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27.

Cole, S. R., Chu, H., & Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International journal of epidemiology*, *35*(4), 1074–1081.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association*, *89*(428), 1314–1328.

De Waal, T., Pannekoek, J., & Scholtus, S. (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*(2), 204–210. Retrieved from `http://dx.doi.org/10.1002/wics.1194` doi: 10.1002/wics.1194

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*(4), 643–659.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural equation modeling: a multidisciplinary journal*, *20*(1), 1–26.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *psychometrika*, *81*(2), 274–289.

McCutcheon, A. L. (1987). Sexual morality, pro-life values, and attitudes toward abortion: A simultaneous latent structure analysis for 1978-1983. *Sociological Methods & Research*, *16*(2), 256-275. Retrieved from https://doi.org/10.1177/0049124187016002003 (PMID: 11655913) doi: 10.1177/0049124187016002003

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition* (Vol. 544). John Wiley & Sons.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. Retrieved from https://doi.org/10.1007/BF02294457 doi: 10.1007/BF02294457

Mislevy, R. J., Beaton, A. E., & Kaplan, a., B. (1992, June). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133-161. Retrieved from dx.doi.org//10.1111/j.1745-3984.1992.tb00371.x doi: 10.1111/j.1745-3984.1992.tb00371.x

Monseur, C., & Adams, R. (2009, January). Plausible values: How to deal with their limitations. *Journal Of Applied Measurement*, *10*(3), 320-334. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19671992

Pavlopoulos, D., & Vermunt, J. (2015). Measuring temporary employment. do survey or register tell the truth? *Survey Methodology*, *41*(1), 197-214. Retrieved from http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14151-eng.pdf (date visited 2017.04.25)

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys (wiley series in probability and statistics).

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 Méthodologie*

Sociologique, *109*(1), 56–61. Retrieved from dx.doi.org//10.1177/0759106310387713 doi: 10.1177/0759106310387713

Schofield, L. S., Junker, B., Taylor, L. J., & Black, D. A. (2014). Predictive inference using latent variables with covariates. *Psychometrika*, 283–314.

Spiegelman, D., McDermott, A., & Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *The American journal of clinical nutrition*, *65*(4), 1179S–1186S.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification*, 1–21. Retrieved from http://dx.doi.org/10.1007/s00357-016-9195-5 doi: 10.1007/s00357-016-9195-5

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political analysis*, *18*(4), 450–469.

Vermunt, J. K., & Magidson, J. (2013). Latent GOLD 5.0 Upgrade Manual [Computer software manual]. Belmont, MA.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & Group, N. P. S. D. W. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. Retrieved from http://dx.doi.org/10.1111/j.1472-4642.2008.00482.x doi: 10.1111/j.1472-4642.2008.00482.x

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*, 114-128. doi: 10.1016/j.stueduc.2005.05.005