STRUCTURAL EQUATION MODELING
2025, VOL. 32, NO. 2
ISSN: 1070-5511 | E-ISSN: 1532-8007

Addressing Missing Data in Latent Class Analysis When Using a Three-Step Estimation Approach

Sarah Depaoli, Fan Jia, and Marieke Visser

University of California

ABSTRACT

This study specifically focuses on addressing the challenges related to employing missing data techniques when estimating a conditional Latent Class Analysis (LCA) model. In the context of a conditional LCA, where covariates are incorporated, the process of estimation becomes more intricate, introducing an additional layer of complexity linked to missing data. The simulation design is structured to examine the performance of different methods of estimation in the presence of covariates and missing data. Our primary focus revolves around the ML three-step approach, and we delve into alternative missing data techniques, including full information maximum likelihood (FIML), Bayesian estimation, and multiple imputation (MI), as viable alternatives to the default listwise deletion approach. By evaluating their performance under various covariate and missing data conditions, we aim to provide valuable recommendations for applied researchers who are navigating the implementation LCA with covariates when missing data are present.

KEYWORDS

Bayesian estimation; covariates; latent class analysis; missing data; three-step approach

Latent class analysis (LCA) provides a reliable approach to unveil underlying structures within observed data, enabling researchers to identify and characterize distinct subgroups in a population. LCA has been used in a variety of substantive settings, including the classification of adolescent smoking subgroups (Henry & Muthén, 2010), alcohol dependence subgroups (Moss et al., 2007), internet gambling subgroups (Lloyd et al., 2010), adolescent obesity subgroups (Huh et al., 2011), and peer victimization subgroups (Nylund, Bellmore, et al., 2007). Despite the popularity of using the LCA model within the applied literature, some methodological aspects of this modeling approach are still in a stage of development.

Conventionally, LCA models that include covariates (i.e., conditional LCA) would be estimated in a single step (i.e., *one-step approach*), where the latent class structure would be simultaneously estimated along with the inclusion of covariates. That model-building approach, however, has been criticized, mainly due to the influence that the inclusion of the covariates can have on the estimated latent class structure (see e.g., Asparouhov & Muthén, 2014). Specifically, the measurement model can be altered when covariates are included, putting into question what the measurement model (or latent class structure) should really look like. More recently, methodologists have recommended the use of a stepwise approach when building LCA models that include covariates. When using a stepwise approach to estimation, the LCA measurement model is established prior to the inclusion of covariates. The general procedure for implementing a typical stepwise approach is as follows:

1. The LCA measurement model is constructed. During this step, applied researchers must decide how many class indicators should be included in the LCA measurement model and how many classes should be specified. In addition, the local independence assumption should be evaluated during this step.
2. Using the parameter estimates from the LCA measurement model in Step 1, cases are assigned to the different latent classes based on their posterior membership probabilities.
3. The standard multinomial logistic regression is estimated, using the class membership assignment from Step 2 as an observed indicator of the latent class variable.

The described stepwise approach consistently underestimates the relationship between the covariate and latent class (Bolck et al., 2004). As the classification error in Step 2 increases, the relationship between the covariate and the latent class variable is attenuated. In response to these findings, several new methods have been proposed to address the measurement error issue in Step 2. One such method is the *maximum likelihood (ML) three-step approach*, which was developed by Vermunt (2010) with additional work by Bolck et al. (2004). The ML three-step approach is suitable for exploring the relationship between the LCA measurement model and covariate(s). Ample simulation research has

explored the performance of the ML three-step approach under different modeling conditions (e.g., Asparouhov & Muthén, 2014; Bakk et al., 2013; Nylund-Gibson et al., 2019; Vermunt, 2010). Results from these simulation studies suggest the three-step approach can produce unbiased parameter estimates if the latent classes are sufficiently distinct from one another in the population.

Statistical software capable of mixture modeling, such as M*plus* (Muthén & Muthén, 1998–2017) and Latent GOLD (Vermunt & Magidson, 2016) has significantly streamlined the implementation of the ML three-step approach, making it more accessible for applied researchers. However, challenges arise when dealing with missing data, particularly in covariates. The software's default setting is to exclude cases with incomplete covariate data during the analysis (i.e., *listwise deletion*), resulting in biased parameter estimates and diminished statistical power, as discussed by Collins and Lanza (2010). Questions remain regarding how missing data should be appropriately handled within the context of the three-step approach.

1. Goals of the Current Study and Organization of the Paper

When estimating a conditional LCA model, researchers must make decisions about many different facets, including the estimation strategy (one-step vs. three-step), specification of covariate relationships, and the handling of incomplete covariate variables. The current study combines many of these issues and aims to provide recommendations for how best to deal with incomplete covariates when using a three-step approach for LCA. Instead of the default listwise deletion approach, several alternative missing data techniques are available, including full information maximum likelihood (FIML), switching to Bayesian estimation in the third step, and multiple imputation (MI). The current study investigates the performance of these methods under different covariate missing conditions (e.g., proportion of missingness, missing data mechanism) and covariate distributions (e.g., standard normal, binomial). Results from this study will be valuable to applied researchers seeking to address incomplete covariates while using a three-step approach.

The remaining sections of the paper are organized as follows. Next, we present the details of the LCA model, with extensions to allow for covariates. This is followed by a section detailing the background of how missing data interfaces with LCA models. Then we present on three of the main methods that can be used for addressing incomplete data for covariates in LCA models. We then present an extensive simulation study examining the performance of several methods under a variety of missing data and modeling situations. We conclude with a discussion of how these findings can inform methods implemented in the applied LCA modeling context.

2. The Latent Class Analysis Model

In the LCA model, there are two types of parameters of interest: measurement and structural parameters. The measurement parameters describe the relationship between the observed latent class indicators and the latent class variables (i.e., the class-specific item endorsements probabilities, which are the distribution of the binary class indicators conditional on the latent class variable). In contrast, the structural parameters describe the multinomial distribution of the latent class variable (i.e., the proportion of cases in each latent class). The parameterization of the LCA model with binary indicators is detailed below, borrowing notation presented in Nylund-Gibson and Masyn (2016) and Masyn (2017).

Figure 1 provides a visual representation of the unconditional LCA model. In Figure 1, each latent class indicator is observed on n individuals with u_{mi} representing individual i's response to class indicator m. The latent class variable has K classes where $c_i = k$ when individual i belongs to Class k. The latent classes are mutually exclusive; therefore, individual i can only be assigned to one of K classes. The relationship between the observed class indicator variables and the latent class variable can be formulated with:

$$\Pr(u_{1i}, u_{2i}, \ldots, u_{Mi}) = \sum_{k=1}^{K} [\pi_k \cdot \Pr(u_{1i}, u_{2i}, \ldots, u_{Mi} | c_i = k)], \tag{1}$$

where π_k is a structural parameter representing the prevalence of individuals in Class k (i.e., class proportions). Considering the latent classes are mutually exclusive, $\sum \pi_k = 1$.

The measurement model for the latent class variable can be parameterized as the relationship between the observed class indicators $u_1, u_2, \ldots, u_M$ and the latent class variable c, which can be formulated with:

$$\Pr(u_{mi} | c_i = k) = \frac{1}{1 + \exp(\tau_{mki})}, \tag{2}$$

where τ_{mki} is the negative log odds of endorsing class indicator u_m given membership to latent class k. In other words, τ_{mki} is equal to -logit($\mathrm{E}[u_{mi} | c_i = k]$). $\Pr(u_{mi} | c_i = k)$ is known as the class-specific item response probability, which suggests how likely an individual is to endorse a particular item given latent class membership.

For the structural model, the unconditional distribution of the multinomial latent class variable, c, can be parameterized with a multinomial logistic regression formulation. Specifically, the π_k parameters can be defined as intercepts on the inverse multinomial logit scale, such that:

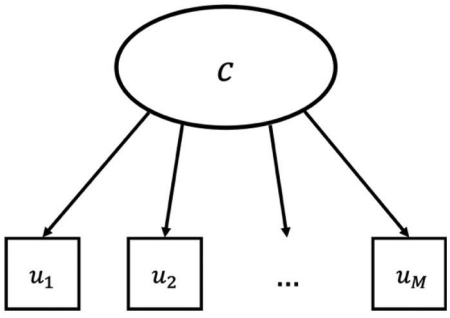

Figure 1. The unconditional LCA model with M binary latent class indicators. The latent class indicators are represented with $u_1, u_2, \ldots, u_M$, and c represents the underlying multinomial latent class variable.

$$\pi_k = \Pr(c = k) = \frac{\exp(\gamma_{0k})}{\sum_{j=1}^{K}\exp(\gamma_{0j})}. \quad (3)$$

The γ_{0k} represents the log odds of membership in class k, given membership in either class k or K. For identification purposes, γ_{0K} is constrained to 0.

The LCA model assumes local independence, which suggests the M binary latent class indicators are uncorrelated conditional on class membership. In other words, latent class membership fully explains any correlations between observed class indicators. Software capable of LCA imposes the local independence assumption by default. By making the local independence assumption, Equation (1) is further simplified to:

$$\Pr(u_{1i}, u_{2i}, \ldots, u_{Mi}) = \sum_{k=1}^{K}\left[\pi_k \cdot \left(\prod_{m=1}^{M}\Pr(u_{mi}|c_i = k)\right)\right]. \quad (4)$$

2.1. The One-Step Approach to Include Covariates

In many practical applications of LCA, the LCA measurement model is used as part of a larger structural equation model (SEM). These models often include observed explanatory variables (i.e., covariates, predictors, independent variables, external variables, or concomitant variables[1]) that predict the latent class variable. For example, Quirk et al., (2013) extended their LCA model for kindergarten readiness to include several predictors (e.g., student's prior preschool experiences, age, language skills, and gender). The addition of these variables allows researchers to explore research questions about why an individual was assigned to a particular latent class. A visual example of the latent class model with a covariate (i.e., conditional LCA model) can be seen in Figure 2. The covariate, x_1, can be categorical (Clogg & Goodman, 1985; Goodman, 1974; Haberman, 1979; Hagenaars, 1990; 1993; Vermunt, 1997) or continuous (Bandeen-Roche et al., 1997; Dayton & Macready, 1988; Kamakura et al., 1994; Yamaguchi, 2000).

To include a covariate, the latent class model is combined with the latent class regression model into a joint model, which is typically estimated with the maximum-likelihood (ML) estimator. ML estimation involves the repeated audition of different combinations of population parameter values until a specific combination of values obtains the highest log-likelihood, which then represents the best fit of the model to data (Enders, 2010). This approach is often referred to as the *one-step approach* in the methodological literature because the measurement model and the structural model (i.e., the logistic regression in which the latent classes are related to the covariates) are simultaneously estimated in a single step (Asparouhov & Muthén, 2014; Bandeen-Roche et al., 1997; Dayton & Macready, 1988; Vermunt, 2010). More specifically, the latent class variable is regressed on the covariate using multinomial logistic regression parametrization (Nylund-Gibson & Masyn, 2016). Using notation first presented in Nylund-Gibson and Masyn (2016), the relationship between the LCA model and covariate x_i can be expressed as a multinomial logistic regression model:

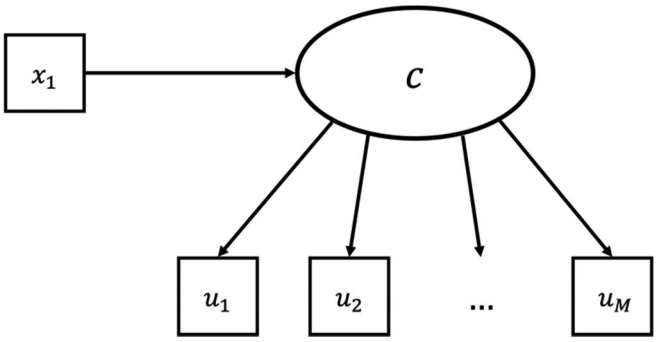

Figure 2. The conditional LCA model with M binary latent class indicators and a single covariate, x_1. The latent class indicators are represented with $u_1, u_2, \ldots, u_M$, and c represents the underlying multinomial latent class variable.

$$\Pr(c_i = k|x_i) = \frac{\exp(\gamma_{0k} + \gamma_{1k}\ x_i)}{\sum_{j=1}^{K}\exp(\gamma_{0j} + \gamma_{1j}\ x_i)}, \quad (5)$$

where $\gamma_{0K} = \gamma_{1K} = 0$ for model identification. In Equation (5), the latent class indicator variables are considered independent of the covariate conditional on class membership. Therefore, Equation (4) can be adapted to include covariate x_i such that

$$\Pr(u_{1i}, u_{2i}, \ldots, u_{Mi}|x_i) = \sum_{k=1}^{K}\left[\Pr(c_i = k|x_i)\cdot\left(\prod_{m=1}^{M}\Pr(u_{mi}|c_i = k)\right)\right]. \quad (6)$$

When K number of classes have correctly been identified, the exclusion of x_i from the model has no impact on the point estimates (i.e., τ_{mk}) for each class indicator (i.e., $u_1, u_2, \ldots, u_M$). In other words, latent class membership will depend on x_i, but the class indicator responses should only depend on class membership. Thus, the covariate only has an indirect effect on the latent class indicators via the latent class variable.

The one-step approach may appear straightforward enough, but several drawbacks have been noted in the methodological literature. Specifically, Vermunt (2010) notes that the one-step approach is impractical when using many covariates, as is typical in exploratory studies. With each additional covariate, the LCA measurement model and the structural model must be estimated again. In addition, Vermunt (2010) highlights the model-building issues surrounding the inclusion of covariates. Users must decide whether to pick the number of latent classes before or after including covariates. Although covariates can aid class enumeration when properly specified (Li & Hser, 2011; Lubke & Muthén, 2007; Muthén, 2002), it is always challenging to properly specify the covariates in the model without prior knowledge. Misspecifying the covariate relationships with the LCA measurement model can impact the class enumeration procedure, resulting in an over-extraction of the number of classes (Nylund-Gibson & Masyn, 2016). Therefore,

[1]A concomitant variable is a variable that is not the focus of the study, but the variable may influence variables of interest to the study (e.g., the dependent variable).

several methodological studies suggest the number of latent classes should be established prior to including covariates (Collins & Lanza, 2010; Masyn, 2013; Petras & Masyn, 2010). Vermunt (2010) also notes that applied researchers do not find the joint model to be intuitive because they often wish to introduce covariates after classifying individuals. In addition, the applied researcher who establishes the latent class measurement model may not be the same person who is building the structural model.

2.2. The ML Three-Step Approach

To address some of the drawbacks of the one-step approach, methodologists have proposed a stepwise approach to LCA estimation (Vermunt, 2010; Asparouhov & Muthén, 2014), the maximum likelihood (ML) three-step approach. In this three-step method, latent class model and the relationship between the latent class variable and covariate are independently evaluated, which can resolve many of the issues with the one-step approach.

The ML three-step approach uses the following general steps:

1. The best-fitting unconditional LCA model is identified.
2. A most likely class variable, *M*, is created using the latent class posterior distribution from Step 1. The conditional probabilities for the class assignment given true latent class membership are also computed. These computed quantities will be used as the estimated classification errors in class assignment.
3. The structural model is estimated, allowing for the inclusion of covariates of the latent class variable. *M* is used as a single, nominal indicator of the latent class variable. The computed quantities form Step 2 are used as fixed parameter values that describe the relationship between the latent class variable and *M*.

In the third step, the latent class variable is regressed on *x* while taking into account the measurement error. Specifically, the most likely class variable *M* is used as a single, nominal class indicator of the latent class variable *c*. The logits are used as fixed parameter values that describe the relationship between the latent class variable and the most likely class variable. The multinomial regression of *c* on predictor *x* is freely estimated. All estimations are maximum likelihood based. A visual representation of the three-step approach can be seen in Figure 3, which was adapted from Asparouhov and Muthén (2014). In the following section we link the ML three-step approach to another

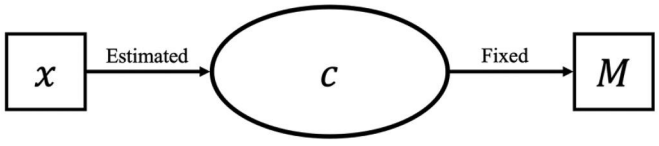

Figure 3. A visual of the ML three-step approach. The latent class variable *c* is regressed on covariate *x*. The most likely class variable, *M*, is used as a single class indicator of the latent class variable *c*. The relationship between *M* and *c* is fixed, and the relationship between *x* and *c* is freely estimated.

important facet of applied LCA modeling—namely, the presence of missing data.

3. Background to Missing Data in LCA Models

3.1. Missing Data Mechanisms

Missing data occurs in most empirical datasets, regardless of efforts by substantive researchers to minimize missingness. Ignoring missing values can be problematic for two key reasons. First, partially complete cases could be substantively different from complete cases, and these differences need to be adjusted to prevent bias in parameter estimates. Second, partially complete cases are often removed from statistical analyses, a practice known as listwise deletion. The reduction in sample size can lead to estimation problems and poor statistical power (Collins & Lanza, 2010). Therefore, methodologists strongly recommend applied researchers address missing data in their analyses.

The methodological literature has identified three missing-data mechanisms (Little & Rubin, 2002; Rubin, 1976; Schafer, 1997; Schafer & Graham, 2002). Data can be missing not at random (MNAR), missing completely at random (MCAR), or missing at random (MAR).

Data are considered MNAR when the probability of having incomplete data on variable Y is related to the values of Y itself, despite controlling for observed variables (Enders, 2010). When the probability of incomplete data on *Y* depends on another observed variable but not the values of *Y* itself, the data are considered MAR. Data are considered MCAR when the probability of incomplete *Y* data is not related to any observed variable or the values of *Y* itself. MNAR is the most problematic because there is a systematic reason for the missing cases, making it particularly challenging to account for that reason. Therefore, the parameter estimates that cannot be adjusted for the unknown reason will be biased. On the other hand, MCAR and MAR are considered ignorable missingness. This is because modern missing data techniques are available to handle these types of missing data, and the reason of missingness can be ignored (Collins & Lanza, 2010).

3.2. Missing Data in Class Indicators

In latent class analysis models, missing cases in the class indicator variables can be easily addressed if the missing data mechanism is MCAR or MAR (Collins & Lanza, 2010; Kolb & Dayton, 1996). Software packages capable of fitting LCA models are typically equipped to address missing latent class indicators rigorously. SEM software (e.g., M*plus*) often defaults to model-based missing data procedures without requiring additional syntax. More specifically, M*plus* computes the parameter estimates with all the available information (i.e., using the MAR assumption). These software defaults allow missing class indicators to be addressed without much consideration from researchers. Unfortunately, including a grouping variable (i.e., covariate, predictor, independent variable) can complicate the situation. When

estimating an LCA model that includes covariates, the user must decide on not only the estimation strategy (one-step vs. stepwise), but also the plan for addressing incomplete covariates.

3.3. Limitation of Automatic ML Three-Step Approach with Incomplete Covariates

We focus on the ML three-step approach for model estimation, as the stepwise strategy becomes more favorable in methodological studies. Without missing data on the covariates, statistical software can help automate these estimation steps. Unfortunately, this automation can make for an inflexible modeling experience in which users cannot adapt the model specification. The inflexibility of the automatic method can be especially problematic when addressing incomplete covariates because the software defaults to listwise deletion, which deletes cases with any incomplete covariate data in the analysis.

Statistical software defaults to listwise deletion for incomplete covariates (and not for incomplete latent class indicators) because covariates are exogenous variables in the larger SEM. The outcome (i.e., the latent class variable) is conditional on the covariate, which has no distributional assumptions. Therefore, the covariate is assumed to be fixed and fully observed in a conditional mixture model (Sterba, 2014). The listwise deletion of exogenous covariates is not unique to conditional mixture models; conditional non-mixture models often specify covariates as exogenous variables (Sterba, 2014). The unique issue here is the common practice of using an automated stepwise approach that does not allow users to adapt model specifications to address the incomplete covariates. When users rely on the automatic ML three-step approach, all the available data are used to estimate the unconditional LCA model in the first step, but individuals with incomplete covariates will be removed from the analysis in the third step. Applied users may be caught off-guard by the reduced sample size caused by the listwise deletion of cases with incomplete covariates. A different method is preferred because listwise deletion decreases statistical power and biases parameter estimates under the missing-at-random (MAR) assumption (Little, 1992; Little & Zhang, 2011).

3.4. Addressing Incomplete Covariates in the Manual ML Three-Step Approach

Three alternative methods have been identified to address incomplete covariates when using a manual ML three-step approach. The first method analyzes the data from individuals with complete and partially complete data together, and the model estimates are adjusted iteratively based on this information provided. The second method uses Bayesian estimation to account for the incomplete covariate values using all available information. The third method imputes plausible values in place of missing values to generate multiple datasets, analyzes each dataset, and pools the results from each analysis. Each of these methods will be described in detail next.

3.5. FIML

One state-of-the-art method for addressing missing data is *full information maximum likelihood* (FIML), which utilizes ML estimation to handle missing values. FIML addresses missing data by maximizing the likelihood function of the observed data while accounting for individual missing data patterns. These computations often require the use of an iterative optimization algorithm such as the expectation-maximization (EM) algorithm. In the methodological literature, FIML has earned state-of-the-art status because it yields unbiased parameter estimates under MCAR and MAR data (Schafer & Graham, 2002).

When using the ML estimator, statistical software often automatically addresses missing values in endogenous variables. For example, the ML three-step approach automatically handles missing latent class indicators via FIML during the first estimation step. To address incomplete covariates with FIML, a relatively straightforward programming trick is required. In the third step, the user must specify model parameters specific to the covariates (e.g., means, variances, and covariances) in addition to the conditional LCA model. By estimating covariate parameters, the software treats the covariates as endogenous variables and applies a normality assumption. The EM algorithm will then address the missing values by maximizing the joint likelihood (Sterba, 2014). Significantly, this programming trick does not change the interpretation of model parameters (e.g., the regression coefficient for "c on x" maintains the same meaning). Past simulation research suggests the EM algorithm effectively addresses incomplete covariates when using a one-step approach to estimation (Sterba, 2014). One drawback of this approach is that the EM algorithm can be computationally intensive when addressing missing data for several covariates in a single model. For each additional covariate with missing data, increasingly heavy numerical integration computations are required, which increases how long the analysis takes (Asparouhov & Muthén, 2021). Although the heavy numerical integration issue has primarily been discussed anecdotally (Asparouhov & Muthén, 2021), it does suggest there is reason to explore alternative strategies that may yield unbiased parameter estimates in a shorter time.

3.6. Bayesian Third Step

An alternative strategy for addressing missing data is to use Bayesian estimation. In contrast to the ML estimator, the Bayesian estimator handles missing values with an internal imputation process (Asparouhov & Muthén, 2021). In M*plus*, Bayesian estimation is implemented with a MCMC estimation algorithm. There are multiple samplers available for use with MCMC methods, but we focus here on the Gibbs sampler (Gelman et al., 2013). The Gibbs sampler iteratively generates a sequence of model parameters, latent variables, and missing observations, which can be used to construct the posterior distribution upon convergence (Asparouhov & Muthén, 2010). The Bayesian estimator is considered a full-information estimator and typically produces similar results to the ML estimator with missing data

(i.e., FIML; Asparouhov & Muthén, 2021). When using the ML three-step approach, the user can switch the estimator in the third step (i.e., *Bayesian third step*) and estimate parameters related to the incomplete covariate (e.g., means, variances, and covariances). The missing values are then modeled and imputed internally using an unrestricted model (Asparouhov & Muthén, 2021). Correlating the covariates is helpful because an observed covariate may help impute the missing values in the other covariate.

Perhaps the most significant advantage of using the Bayesian third step is that the method allows for the specification of priors on parameters of substantive interest (e.g., regression coefficients). Past methodological research suggests the stepwise approaches yield biased structural parameter estimates when the latent classes are not very distinct (Bakk & Vermunt, 2016; Nylund-Gibson et al., 2019; Vermunt, 2010). Depending on the prior specification for the covariate effect, Bayesian estimation may improve the accuracy of the parameter estimate, especially under poor class separation. Although no previous simulation studies have explored the potential benefits of the Bayesian third step, previous mixture modeling research suggests Bayesian estimation can aid estimation for latent class models (Depaoli, 2013, 2014; Lu et al., 2011); therefore, it follows that Bayesian estimation may be helpful in this situation. Another advantage of the Bayesian estimator is computation speed, which can sometimes be much quicker than the numerical integration required by the EM algorithm (Asparouhov & Muthén, 2021). This is most evident when the analysis has many covariates with missing values, which requires heavier numerical integration. Despite these potential advantages, the Bayesian third step does have disadvantages in certain modeling situations. One critical limitation of the Bayesian third step in M*plus* is that the covariates are assumed to be normally distributed (Asparouhov & Muthén, 2021). In other words, the Bayesian third step may produce biased results when applied to an incomplete nonnormal or categorical covariate.

3.7. Multiple Imputation

A final option for addressing missing data is another state-of-art method called *multiple imputation* (MI). MI consists of three distinct steps: the imputation phase, the analysis phase, and the pooling phase. During the imputation phase, *W* copies of the original dataset are created. Each copy contains plausible values for the missing values in the original dataset. During the analysis step, each of the *W* datasets is analyzed using the same statistical procedure that would have been performed had the original dataset been complete. Considering the analysis phase results in *W* parameter estimates and SEs, the results need to be pooled together into a single set of results during the pooling phase (Enders, 2010). In the context of the ML three-step approach, MI can either be performed before or after the first estimation step (Asaprouhov & Muthén, 2021). Using MI prior to the first estimation step may be useful if the researchers wish to include direct effects between the covariate and latent class indicators. In contrast, using MI after the first estimation step presents a more simplified strategy for addressing incomplete covariate data.

One factor that can further improve the accuracy of MI parameter estimates is the addition of auxiliary variables (AVs). AVs are ancillary variables that are not part of the primary analysis but are possibly correlated with missingness or the incomplete analysis model variables (Enders, 2010; Schafer, 1997). Although AVs are not of substantive interest, they can increase power and reduce bias in parameter estimates (Enders, 2010). The omission of an important AV has the potential to transform the nature of missing data from MAR to MNAR because no variable is explaining the missingness in the analysis. Even in modeling situations that are firmly MCAR or MAR, the addition of AVs can improve power and reduce bias with almost no downside (Collins et al., 2001; Graham, 2003; Schafer & Graham, 2002); therefore, methodologists strongly recommend an inclusive analysis strategy. MI can be implemented following the first analysis step when using the ML three-step approach. After estimating the unconditional LCA model, the user can impute the missing covariate values by specifying an imputation model for the incomplete covariates, which should include variables that are possibly correlated with the missingness (e.g., covariates, latent class indicators, AVs). After the imputation phase, each complete dataset is analyzed with a conditional LCA model. The parameter estimates are then pooled. Statistical software packages have automated much of the MI process, but the user still needs to make important decisions concerning the number of imputed datasets and how convergence will be assessed. In addition, users should only select the AVs that can help predict missingness during the imputation phase (Enders, 2010).

Perhaps the biggest advantage of MI is the ability to specify the distribution of the incomplete covariate during the imputation phase. Both FIML and the Bayesian third step impose normality assumptions on the covariate, whereas MI allows the user to specify the covariate distribution. In practical settings with more trivial the missing data, the assumption violation introduced by FIML and the Bayesian third step may not be as problematic (Asparouhov & Muthén, 2021). One issue that should be considered is the past methodological work, which strongly recommends against the MI of covariates in mixture models because it can lead to substantial parameter bias (Enders & Gottschall, 2011). MI produces biased parameter estimates because covariate relations can vary across classes, making it highly unlikely that the imputation model will be specified correctly in applied settings (Enders & Gottschall, 2011; Sterba, 2014, 2017). Although there is good reason to be wary of MI in mixture modeling, there are situations with non-normal covariates (e.g., binary covariate) that may benefit.

4. Design

There is some evidence suggesting each of the described alternative methods could effectively address incomplete

covariates. The aim of the present simulation study is to investigate the performance of four methods (listwise deletion, FIML, Bayesian third step, and MI) for handling incomplete covariates in LCA models. To achieve this aim, we varied the missing data mechanism (MAR and MCAR), percentage of missing data (15%, 35%, and 55%), covariate distribution (Bernoulli and standard normal), and the strength of the covariate effect (weak, moderate, and strong). The simulation study consisted of 288 cells, and each cell had 500 replications. Previous studies using conditional LCA models have found 500 replications to be sufficient (Di Mari & Bakk, 2018; Janssen et al., 2019; Nylund-Gibson & Masyn, 2016).

To investigate the performance of each of the missing data methods (i.e., listwise deletion, FIML, Bayesian third step, and MI), data were generated in *Mplus* using a conditional LCA model specification. The population model for this study was divided into two parts: the measurement part of the model and the structural part of the model. The measurement part of the model consisted of the unconditional latent class model. The structural part of the model related the covariates to the latent class variable. A visual representation of the population model is displayed in Figure 4, where $u_1 - u_5$ are the binary observed class indicators, c is the categorical latent class variable, and x_1 and x_2 are the observed covariates predicting the latent class variable. The regression coefficient for the effect of x_1 on c is labeled, γ_{11}, and the regression coefficient for the effect of x_2 on c is labeled, γ_{21}.

4.1. Measurement Model

The measurement model consisted of two classes ($K = 2$) of equal size ($\pi_1 = 0.5, \pi_2 = 0.5$) with five binary class indicator variables. All population models had moderate class separation, which was achieved by setting the item thresholds in Class 1 to $\tau = -1.25$ and the item thresholds in Class 2 to $\tau = 1.25$. Similar population values have been used in other simulation studies with conditional LCA models (Masyn, 2013; Nylund-Gibson & Masyn, 2016). These item threshold specifications corresponded to a conditional item-response probability (for all items) of 0.78 for Class 1 and 0.22 for Class 2. The sample size was fixed at $n = 500$ across conditions.[2]

4.2. Structural Model

The structural model of interest was a binomial logistic regression in which the latent class variable, c, was regressed on two covariates, x_1 and x_2. The x_1 variable was fully observed (i.e., no missing data) and followed a standard normal distribution. The effect of x_1 on c was held constant at $\gamma_{11} = -1$ across conditions. In contrast, x_2 varied across conditions. The structural model varied on the following factors: the strength of the effect of x_2 on c, the distribution of x_2, percentage of missing data on x_2, and the missing data mechanism.

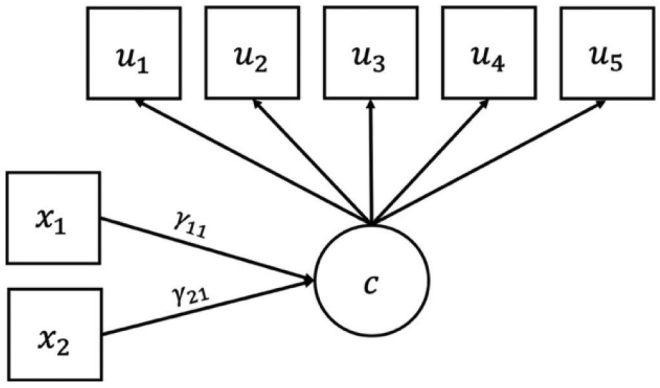

Figure 4. The population model.

4.2.1. Strength of the Covariate Effect

The strength of the covariate effect of x_2 on c was set to be either weak ($\gamma_{21} = 0.5$), moderate ($\gamma_{21} = 1$), or strong ($\gamma_{21} = 1.5$). These regression coefficient specifications correspond to an odds ratio of 1.65, 2.72, and 4.48, respectively. Previous simulation studies on conditional LCA models have used similar regression coefficient values (Bakk et al., 2014; Nylund-Gibson & Masyn, 2016; Vermunt & Magidson, 2021). Manipulating the strength of the regression coefficient may be an important factor for illustrating the pitfalls of the automatic ML three-step approach, which results in listwise deletion. Even strong effects can be lost to reduced statistical power from listwise deletion.

4.2.2. Distribution of the Covariate

One factor that is likely to impact the performance of the methods for addressing incomplete covariates is the distribution of covariate x_2. Therefore, x_2 will either follow a standard normal distribution (like x_1) or a Bernoulli distribution. For conditions with a binary x_2, the logit threshold was set to 0, which equates to a response probability of 0.5. To our knowledge, no previous simulation study has explored missing binary covariates in conditional LCA models, but it is likely missing binary covariates will pose a greater estimation challenge (Asparouhov & Muthén, 2021).

4.3. Missing Data Conditions

Missing data were generated on x_2 according to six different missing conditions. Specifically, missing data were generated under two different missing mechanisms: MAR and MCAR. In addition, the percentage of missing data on x_2 was generated as 15%, 35%, or 55%. These missing data conditions are in line with the limited methodological research on incomplete covariates in conditional mixture models (Sterba, 2014). Past research suggests listwise deletion can produce unbiased estimates under MCAR assumptions, but

[2]Sterba (2014) found $n = 500$ to be a common sample size in mixture model studies in the social sciences. In addition, Nylund-Gibson and Masyn (2016) uses $n = 500$ for a smaller sample size condition in their simulation study with an LCA model with a continuous covariate.

estimation efficiency and statistical power are reduced (Little, 1992; Little & Zhang, 2011; Skrondal & Rabe-Hesketh, 2014). In contrast, listwise deletion produces biased estimates under MAR assumptions. One way to illustrate the consequences of using an automatic ML three-step approach (i.e., listwise deletion) is to include both MAR and MCAR missing mechanisms. The missing percentages selected (15% and 35%) represent realistic missing data situations in practice (Enders & Bandalos, 2001; Merkle, 2011; Wothke, 2000), and the 55% condition represents a worst-case scenario.

4.3.1. Missing Data Generation

Missing data were generated in M*plus* using the MISSING option in the MONTECARLO command, which allows the user to specify a logistic regression model to generate the missing data for one or more variables (i.e., x_2). More specifically, the logistic regression model was used to derive the intercept (α) and slope (β) parameters from the regression of a missing data indicator (R) on one of the latent class indicators (u_1):

$$p(R = 1|u_1) = \frac{\exp(\alpha + \beta u_1)}{1 + \exp(\alpha + \beta u_1)}. \quad (7)$$

In Equation (7), the missing data indicator (R) is a binary dependent variable that is scored as 0 for not missing and scored as 1 for missing on the dependent variables in the data generation model (i.e., x_2). The latent class indicator u_1 is considered a predictor of missingness in the model. Depending on the values selected for the intercept and slope of the logistic regression model, the probability of missingness, the expected value of $p(R = 1|u_1)$, changes.

For conditions with MCAR missing data on x_2, the slope of u_1 was set to 0, suggesting the missingness was unrelated to any other variables in the data. Depending on the desired percent missing on x_2, different values for the logistic regression intercept were used (i.e., −1.734, −0.619, and 0.201) to achieve the desired probability of missingness (i.e., 15%, 35%, and 55%). For conditions with MAR missing data, the slope of u_1 was set to 1.48. The slope selected produced a squared correlation of 0.40, which is in line with Enders and Mansolf (2018). The selected slope value indicates a moderately strong relationship between the cause of missingness (u_1) and the underlying latent probability of missing data. In addition to setting the slope to 1.48, different values for the logistic regression intercept were used (i.e., −2.66, −1.445, and −0.51) to achieve the desired probability of missingness (i.e., 15%, 35%, and 55%). Table 1 summarizes the population values used to generate missing data conditions.

Table 1. The population values for generating missing data conditions.

	α		β	
Missing assumption	MAR	MCAR	MAR	MCAR
Missing %				
15	−2.66	−1.734	1.48	0
35	−1.445	−0.619	1.48	0
55	−0.51	0.201	1.48	0

4.4. Analysis Models

In accordance with the manual ML three-step approach procedure in M*plus*, each generated dataset was first analyzed with a 2-class unconditional LCA model. Model constraints were applied to prevent label-switching across replications. After completing the first estimation step, the most likely class variable, M, was saved, and the conditional probabilities for the class assignment given true latent class membership was also computed.[3] Then, the measurement part of the model needed to be combined with the structural part of the model as part of a larger SEM while addressing the missingness on the covariate x_2. Four methods were implemented to address the incomplete covariates: listwise deletion, FIML, Bayesian third step, and MI. The first three methods were implemented during the third estimation step, whereas MI was implemented prior to the third estimation step. All analyses were performed in M*plus*, and the process was automated using the M*plus* Automation package (Hallquist & Wiley, 2018) in R (R Core Team, 2019). In the following sections, the implementation of each of the methods will be described.

4.4.1. Listwise Deletion

The listwise deletion method only required the specification of a conditional LCA model with a single latent class indicator, M, and two covariates, x_1 and x_2. With this model specification, covariates were considered exogenous variables in the SEM and were listwise deleted from the analysis during the third estimation step. Notably, this model specification is equivalent to using the automatic ML three-step approach.

4.4.2. FIML

To address missing data in the third step with FIML, a conditional LCA model with a single latent class indicator, M, and two covariates, x_1 and x_2 was specified. In addition, the means, variances, and covariance of x_1 and x_2 were specified. By also estimating parameters specific to the covariates, x_1 and x_2 were treated as dependent variables by the M*plus* software. All missingness on x_2 was automatically addressed during model estimation by maximizing the joint likelihood, and an assumption is made that x_2 is normally distributed. Numerical integration was required to estimate the joint likelihood; therefore, the Monte Carlo integration algorithm was applied.

4.4.3. Bayesian Third Step

To implement Bayesian estimation in the third step, the estimator was switched from ML to Bayesian. In addition, a

[3]The most likely class variable, M, is saved using the SAVEDATA option and including the statements "FILE = Step3.dat" and "SAVE = CPROB". In addition, x_1 and x_2 are saved in the new dataset by including these variables in the auxiliary statement. The new dataset contains $u_1 - u_5$, x_1, x_2, the individual posterior probabilities for each latent class, and M. During the third estimation step, the new dataset is used to specify the structural model "c on x_1 x_2", and M is fixed in each class using the logits from the "Logits for the Classification Probabilities the Most Likely Latent Class Membership (Column) by Latent Class (Row)" section of the output from the first step. These logits are used to account for measurement error in M.

conditional LCA model with a single latent class indicator, M, and two covariates, x_1 and x_2 was specified. As was the case for FIML, the means, variances, and covariance of x_1 and x_2 were also specified in the model. By switching estimators and estimating parameters specific to the covariates (i.e., means, variances, covariances), the missing values on x_2 were imputed internally, and an assumption was made that x_2 was normally distributed.

To investigate the impact of prior specifications (or misspecifications), five prior conditions were considered for coefficients, γ_{11} and γ_{21}. The five prior specifications included the default prior in M*plus* for regression coefficients, $N(0,5)$, which is considered diffuse. In addition, priors correctly centered on the population value (i.e., correct priors) and priors that are centered on a wrong value (i.e., wrong priors) were also considered. For both the correct and wrong prior conditions, the degree of informativeness was varied (i.e., informative vs. weakly informative). The combination of these prior means and variances produced four additional prior conditions: correct-informative, correct-weakly informative, wrong-informative, and wrong-weakly informative. Figure 5 provides a visual of the five prior conditions utilized for both γ_{11} and γ_{21}. Notably, the prior specification changed for γ_{21} depending on the strength of the covariate effect (e.g., 0.5, 1.0, or 1.5) with correct priors centered on 0.5, 1.0, or 1.5 and wrong priors centered on 1.0, 1.5, or 2.0, respectively. The variance hyperparameter was set to 0.04 in informative conditions and 0.16 in weakly-informative conditions. Figure 5 Panels A, B, and C provide the prior specifications for each γ_{21} condition, respectively. In contrast, γ_{11} was set at -1.0 across conditions, and all five prior conditions can be seen in Figure 5 Panel D. A single MCMC chain was utilized for parameter estimation to prevent between-chain label switching. The number of iterations was set to 30,000 for all analyses, and the first 15,000 iterations were discarded as burn-in. Convergence was assessed via careful examination of trace plots and autocorrelation plots and monitoring the potential scale reduction factor (PSRF).

4.4.4. Multiple Imputation

The procedure for implementing MI was quite different from the previous methods discussed. A manual ML three-step approach was still required, but the second step was to impute missing data instead of generating a single variable M and proceeding directly to the final estimation step. The missing values on x_2 were imputed using the "Data Imputation" option in M*plus*, which helps automate the MI process. To aid the imputation process, x_1 and $u_1 - u_5$ were included in the imputation model. Depending on the distribution of x_2 (e.g., normal or Bernoulli) in the condition, x_2 was specified to be either imputed as a normal or a binary variable. In other words, the imputation model was never misspecified. The number of imputations per replication was set to 20, which is in line with previous simulation studies using MI (Enders & Mansolf, 2018; Vera & Enders, 2021). Chain convergence was assessed using the M*plus* default criteria (e.g., the PSRF is close to 1 for each parameter). After the imputation step was complete, the structural model of interest (i.e., "c on x_1 and x_2") was estimated using the ML estimator for each imputed dataset, which had no

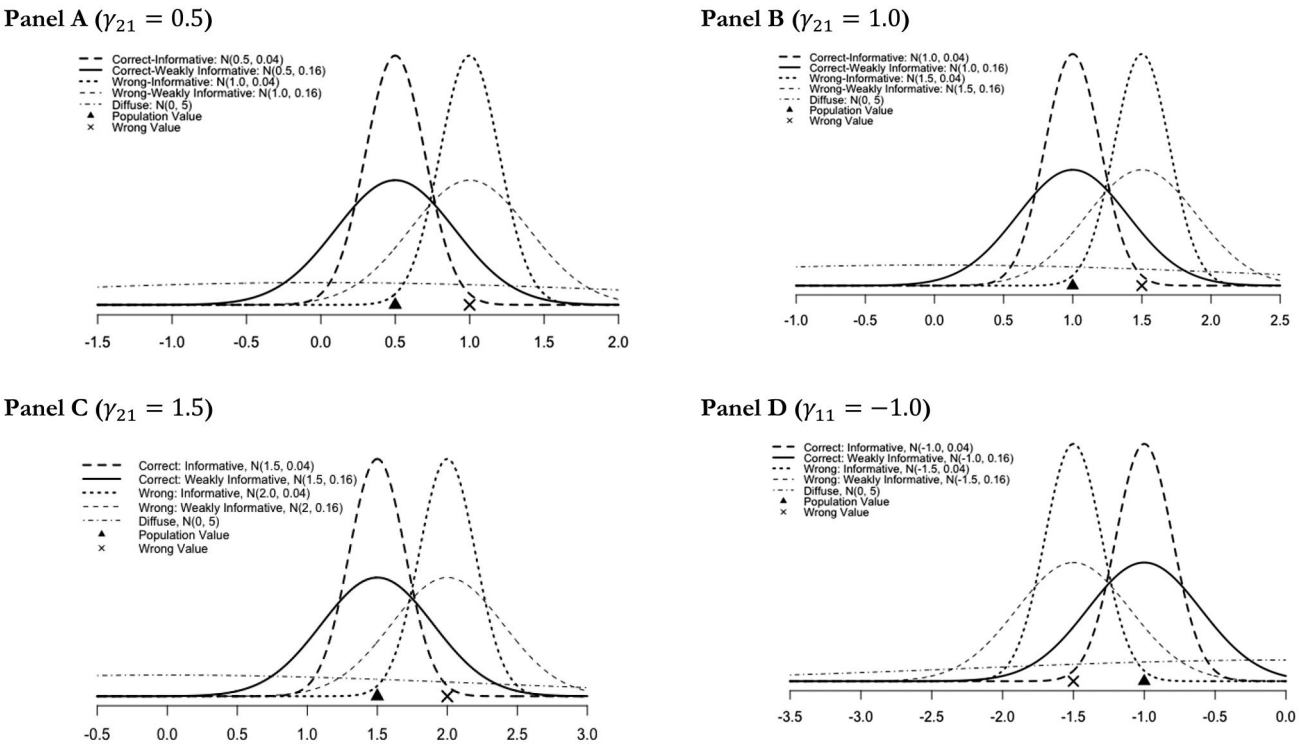

Figure 5. Prior conditions for the regression coefficients in the Bayesian third step.
Panels A, B, and C show the prior specifications for γ_{21}, which have a different population value depending on the condition. Panel D shows the prior specification for γ_{11}, which is held constant across conditions.

missing data. Parameter estimates were then averaged across the imputed datasets.

5. Results

The primary results of interest in this study pertain to the following parameters: γ_{11}, γ_{21}, γ_{01}, which represent the regression coefficients (i.e., "c on x_1" and "c on x_2") and the intercept of the binomial logistic regression, respectively. Bias for each parameter was calculated by subtracting the population value from the mean parameter estimate in the cell. In addition, the mean square error (MSE) was calculated by adding the variance of the estimates across the replications to the squared parameter bias.

5.1. Convergence

To prevent within-chain label switching across replications of the simulation study, a model constraint was included on the latent class indicator u_1 during the first estimation step such that the values adhered to the following order: Class 2 > Class 1. Overall, each cell in the simulation with a continuous covariate converged without issue and a set of stable estimates for the model parameters was obtained, regardless of the estimation strategies in the third step. In conditions using ML estimation (implemented via the EM algorithm) in the third step, M*plus* defaults were utilized, which specifies 20 sets of random starts in the initial stage and 4 final stage optimizations. In each replication, the log-likelihood was replicated indicating convergence. In conditions with Bayesian estimation in the third estimation step, convergence was assessed using PSRF. If PSRF values were less than 1.01, a replication was considered converged. According to this criterion, all replications using Bayesian estimation converged.

5.2. How to Read the Tables

Tables 2–4 display the bias and MSE results for population models with a continuous x_2 covariate and the regression coefficient γ_{21} set to 0.5, 1.0, and 1.5, respectively. Tables 5–7 display the bias and MSE results for population model models with a binary x_2 covariate and the regression coefficient γ_{21} set to 0.5, 1.0, and 1.5, respectively. Each method of addressing missing data in x_2 are listed on the left side of the tables (i.e., Listwise Deletion, FIML, MI, and Bayes). For conditions that utilize Bayesian estimation, five prior specifications were included (i.e., Bayes-Correct Informative, Bayes-Correct Weakly Informative, Bayes-Wrong Informative, Bayes-Wrong Weakly Informative, and Bayes-Diffuse). The six different missing data conditions are presented at the top of the table. Specifically, there is a combination of missing mechanism (MAR vs. MCAR) and missing percentage (15%, 35%, vs. 55%). For each parameter of interest (i.e., γ_{11}, γ_{21}, γ_{01}) the bias and MSE was calculated in each condition. To help illustrate the pattern of results, conditions with ±0.1 bias are bolded in the Tables.

5.3. Continuous Covariate with Missing Data

5.3.1. Bias

Tables 2–4 provide the bias for the regression coefficients, γ_{11} and γ_{21}, and the intercept, γ_{01}. Across all three levels of covariate strength (i.e., 0.5, 1.0, 1.5), several important

Table 2. The bias and MSE in the regression coefficients for *c* on x_1 (i.e., γ_{11}), *c* on x_2 (i.e., γ_{21}), and the intercept (i.e., γ_{01}), under different missing data conditions for $\gamma_{21} = 0.5$ and a continuous x_2.

Missing mechanism			MAR						MCAR					
Missing percentage			15%		35%		55%		15%		35%		55%	
Approach	Parameter	Pop. Value	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Listwise deletion	γ_{11}	−1	−.015	.029	−.036	.046	−.044	.060	−.021	.031	−.032	.041	−.037	.066
	γ_{21}	0.5	.008	.023	.013	.032	.023	.044	.010	.022	.020	.029	.019	.046
	γ_{01}	0	**−.137**	.048	**−.371**	.179	**−.610**	.425	.001	.026	.005	.036	−.015	.043
FIML	γ_{11}	−1	−.010	.025	−.022	.028	−.010	.026	−.020	.024	−.022	.027	−.020	.030
	γ_{21}	0.5	.008	.023	.009	.030	.010	.041	.009	.022	.017	.028	.013	.044
	γ_{01}	0	.011	.026	−.004	.029	−.002	.024	−.001	.024	.007	.025	−.009	.027
MI	γ_{11}	−1	−.007	.024	−.016	.027	.000	.025	−.017	.023	−.015	.025	−.011	.028
	γ_{21}	0.5	−.006	.021	−.012	.026	−.037	.034	−.004	.021	−.013	.022	−.034	.035
	γ_{01}	0	.011	.025	−.005	.029	−.004	.024	−.001	.023	.009	.025	−.004	.026
Bayes-correct informative	γ_{11}	−1	−.010	.009	−.017	.010	−.010	.009	−.017	.009	−.017	.010	−.015	.010
	γ_{21}	0.5	.007	.010	.005	.010	.006	.010	.008	.009	.012	.009	.007	.010
	γ_{01}	0	.012	.026	−.002	.029	.002	.024	−.001	.024	.008	.025	−.005	.026
Bayes-correct weakly informative	γ_{11}	−1	−.024	.019	−.036	.021	−.026	.020	−.034	.019	−.036	.021	−.035	.023
	γ_{21}	0.5	.015	.019	.015	.022	.019	.026	.017	.018	.025	.021	.024	.029
	γ_{01}	0	.012	.026	−.003	.030	.000	.025	−.001	.024	.008	.026	−.005	.027
Bayes-wrong informative	γ_{11}	−1	**−.260**	.079	**−.276**	.089	**−.282**	.091	**−.269**	.084	**−.277**	.089	**−.288**	.096
	γ_{21}	0.5	**.229**	.065	**.270**	.086	**.333**	.123	**.232**	.066	**.278**	.089	**.329**	.121
	γ_{01}	0	.011	.034	−.010	.041	−.012	.035	.000	.032	.008	.034	−.005	.037
Bayes-wrong weakly informative	γ_{11}	−1	**−.113**	.035	**−.132**	.042	**−.132**	.041	**−.125**	.037	**−.132**	.041	**−.142**	.047
	γ_{21}	0.5	.092	.030	**.116**	.040	**.160**	.058	.095	.029	**.125**	.040	**.161**	.061
	γ_{01}	0	.011	.029	−.006	.034	−.005	.029	−.001	.027	.009	.029	−.006	.031
Bayes-diffuse	γ_{11}	−1	−.031	.027	−.047	.032	−.039	.030	−.042	.027	−.047	.030	−.051	.036
	γ_{21}	0.5	.019	.025	.023	.033	.033	.048	.022	.024	.035	.031	.041	.052
	γ_{01}	0	.012	.027	−.004	.031	−.002	.026	−.001	.025	.009	.026	−.006	.028

Note. Conditions with ±0.1 bias are bolded.

Table 3. The bias and MSE in the regression coefficients for c on x_1 (i.e., γ_{11}), c on x_2 (i.e., γ_{21}), and the intercept (i.e., γ_{01}), under different missing data conditions for $\gamma_{21} = 1$ and a continuous x_2.

Missing mechanism			MAR						MCAR					
Missing percentage			15%		35%		55%		15%		35%		55%	
Approach	Parameter	Pop. Value	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Listwise deletion	γ_{11}	−1	−.032	.037	−.024	.049	−.056	.078	−.023	.032	−.036	.046	−.047	.075
	γ_{21}	1	.028	.037	.039	.050	.068	.086	.023	.032	.041	.053	.055	.082
	γ_{01}	0	**−.146**	.054	**−.368**	.176	**−.623**	.448	.006	.030	.004	.038	−.018	.048
FIML	γ_{11}	−1	−.028	.032	−.018	.032	−.014	.035	−.023	.026	−.028	.033	−.034	.043
	γ_{21}	1	.027	.037	.033	.049	.040	.072	.022	.032	.037	.053	.049	.081
	γ_{01}	0	.002	.028	.004	.029	.000	.031	.003	.027	.001	.031	−.014	.031
MI	γ_{11}	−1	−.016	.029	.011	.027	.028	.028	−.010	.024	.001	.028	.012	.031
	γ_{21}	1	−.009	.032	−.048	.037	−.082	.052	−.013	.028	−.043	.038	−.080	.050
	γ_{01}	0	.001	.027	.003	.028	−.005	.029	.003	.026	.007	.029	−.007	.028
Bayes-correct informative	γ_{11}	−1	−.018	.010	−.011	.009	−.007	.009	−.016	.008	−.017	.010	−.016	.010
	γ_{21}	1	.015	.011	.017	.010	.016	.010	.015	.009	.018	.011	.017	.010
	γ_{01}	0	.003	.028	.008	.029	.006	.030	.003	.027	.005	.030	−.008	.030
Bayes-correct weakly informative	γ_{11}	−1	−.040	.023	−.032	.022	−.028	.022	−.036	.019	−.040	.023	−.042	.026
	γ_{21}	1	.038	.026	.044	.030	.051	.037	.036	.023	.048	.033	.054	.037
	γ_{01}	0	.002	.029	.007	.030	.003	.031	.004	.028	.005	.032	−.008	.031
Bayes-wrong informative	γ_{11}	−1	**−.304**	.105	**−.311**	.108	**−.324**	.115	**−.304**	.103	**−.316**	.111	**−.333**	.123
	γ_{21}	1	**.319**	.115	**.363**	.143	**.416**	.184	**.319**	.113	**.362**	.144	**.411**	.180
	γ_{01}	0	−.002	.040	−.001	.042	−.013	.046	.006	.039	.004	.044	−.010	.045
Bayes-wrong weakly informative	γ_{11}	−1	**−.157**	.052	**−.161**	.053	**−.178**	.060	**−.153**	.046	**−.169**	.056	**−.193**	.069
	γ_{21}	1	**.165**	.059	**.203**	.077	**.267**	.117	**.163**	.054	**.207**	.082	**.264**	.112
	γ_{01}	0	.001	.033	.003	.036	−.007	.039	.005	.032	.005	.037	−.009	.039
Bayes-siffuse	γ_{11}	−1	−.056	.037	−.051	.039	−.060	.046	−.050	.030	−.061	.039	−.080	.057
	γ_{21}	1	.055	.043	.070	.059	**.105**	.102	.051	.037	.077	.065	**.112**	.109
	γ_{01}	0	.002	.030	.006	.031	−.001	.034	.004	.028	.005	.033	−.009	.034

Note. Conditions with ±0.1 bias are bolded.

Table 4. The bias and MSE in the regression coefficients for c on x_1 (i.e., γ_{11}), c on x_2 (i.e., γ_{21}), and the intercept (i.e., γ_{01}), under different missing data conditions for $\gamma_{21} = 1.5$ and a continuous x_2.

Missing mechanism			MAR						MCAR					
Missing percentage			15%		35%		55%		15%		35%		55%	
Approach	Parameter	Pop. Value	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Listwise deletion	γ_{11}	−1	−.016	.037	−.033	.061	−.044	.092	−.020	.035	−.047	.055	−.058	.083
	γ_{21}	1.5	.027	.052	.052	.087	.069	.125	.037	.056	.072	.102	.083	.151
	γ_{01}	0	**−.137**	.051	**−.399**	.204	**−.610**	.438	.011	.036	.003	.047	−.025	.057
FIML	γ_{11}	−1	−.017	.033	−.025	.047	−.022	.052	−.020	.029	−.035	.041	−.040	.055
	γ_{21}	1.5	.026	.051	.041	.087	.038	.119	.035	.056	.064	.099	.069	.143
	γ_{01}	0	.018	.029	−.011	.035	.021	.038	.007	.033	.003	.039	−.020	.040
MI	γ_{11}	−1	.007	.030	.033	.034	.061	.038	.007	.026	.028	.031	.052	.036
	γ_{21}	1.5	−.042	.043	**−.113**	.061	**−.193**	.095	−.035	.045	−.098	.063	**−.184**	.093
	γ_{01}	0	.015	.028	−.014	.032	.008	.031	.007	.031	.011	.035	−.009	.033
Bayes-correct informative	γ_{11}	−1	−.012	.009	−.012	.010	−.012	.010	−.012	.008	−.014	.010	−.016	.010
	γ_{21}	1.5	.017	.009	.018	.009	.028	.009	.020	.009	.024	.011	.020	.009
	γ_{01}	0	.019	.029	−.005	.034	−.005	.034	.007	.033	.009	.039	−.012	.038
Bayes-correct weakly informative	γ_{11}	−1	−.032	.022	−.035	.026	−.035	.027	−.034	.020	−.041	.025	−.044	.028
	γ_{21}	1.5	.049	.030	.058	.036	.061	.042	.056	.032	.072	.045	.071	.047
	γ_{01}	0	.018	.030	−.007	.036	.023	.039	.008	.034	.009	.040	−.012	.040
Bayes-wrong informative	γ_{11}	−1	**−.320**	.113	**−.334**	.123	**−.352**	.134	**−.322**	.114	**−.337**	.125	**−.356**	.138
	γ_{21}	1.5	**.395**	.165	**.430**	.194	**.468**	.226	**.399**	.170	**.435**	.200	**.468**	.228
	γ_{01}	0	.018	.042	−.020	.051	.010	.056	.010	.048	.009	.057	−.016	.058
Bayes-wrong weakly informative	γ_{11}	−1	**−.174**	.056	**−.196**	.070	**−.221**	.081	**−.178**	.056	**−.203**	.071	**−.232**	.087
	γ_{21}	1.5	**.239**	.092	**.292**	.128	**.356**	.175	**.250**	.099	**.305**	.145	**.365**	.187
	γ_{01}	0	.018	.036	−.016	.044	.013	.050	.009	.041	.008	.049	−.015	.052
Bayes-diffuse	γ_{11}	−1	−.050	.039	−.070	.060	−.090	.073	−.054	.035	−.080	.054	**−.107**	.080
	γ_{21}	1.5	.077	.064	**.116**	.119	**.166**	.191	.089	.070	**.143**	.139	**.193**	.226
	γ_{01}	0	.018	.031	−.011	.038	.019	.044	.008	.035	.008	.043	−.014	.046

Note. Conditions with ±0.1 bias are bolded.

patterns of results emerged. Listwise deletion produced unbiased regression coefficient estimates, regardless of the conditions. However, the intercept, γ_{01}, was consistently biased when the missing data mechanism was MAR. Using FIML to address the missing x_2 data resulted in unbiased regression coefficients and intercepts, regardless of the strengths of the regression coefficient and missing data conditions. Similarly, MI produced unbiased parameter estimates for γ_{11} and γ_{01} estimates, across all conditions. When the γ_{21} regression coefficient strength was weak or moderate (i.e., $\gamma_{21} = 0.5$ and $\gamma_{21} = 1.0$), MI also produced unbiased γ_{21} estimates. However, there were several conditions with a strong regression coefficient (i.e., $\gamma_{21} = 1.5$) that had biased γ_{21} estimates. As evidenced by Table 4, the γ_{21} regression

coefficient was biased when using MI in conditions with 55% missing data. In addition, γ_{21} was biased when using MI when 35% missing data when the missing data mechanism was MAR.

When using Bayesian estimation in the third step, prior specification impacted results for the regression coefficients, γ_{11} and γ_{21}. Informative and weakly informative priors centered on the correct regression coefficient population value produced unbiased parameter estimates, regardless of the missing data conditions, and regression coefficient strengths. In contrast, informative and weakly informative priors centered on the wrong value biased the regression coefficient estimates. Diffuse priors always resulted in unbiased regression coefficients when the missing data percentage was 15%. However, diffuse priors biased γ_{21} parameter estimates when the percentage of missing data increased (i.e., 35% and 55%) and the strength of the regression coefficient increased (1.0 and 1.5), regardless of the missing data mechanisms. The only exception to this trend was when the missing percentage was 35% and $\gamma_{21} = 1.0$. Notably, γ_{11} was also biased when the missing percentage was 55% and $\gamma_{21} = 1.5$.

5.3.2. Mean Square Error

Tables 2–4 display the MSE for the regression coefficients, γ_{11} and γ_{21}, and the intercept, γ_{01}. The pattern of MSE results were similar across all three levels of covariate strength (i.e., 0.5, 1.0, 1.5); however, the MSE values tended to be higher when $\gamma_{21} = 1.5$. Regardless of the conditions, similar MSE values were obtained for the regression coefficients, γ_{11} and γ_{21}, when using listwise deletion, FIML, and MI. Across these three methods for addressing missing x_2 data, the MSE tended to increase as the percentage of missing data increase from 15% to 55%. Despite the similarities in the regression coefficient results, listwise deletion had inflated MSE values for the intercept, γ_{01}, when the missing data mechanism was MAR. In contrast, FIML and MI had lower MSE values for the intercept.

When using Bayesian estimation, the MSE for the regression coefficients were largely dependent on the prior specifications. Across all conditions with a continuous x_2, addressing missing data with a Bayesian third step using informative priors correctly centered on the population value resulted in the lowest MSE in the regression coefficients, γ_{11} and γ_{21}. In contrast, the highest MSE values for the regression coefficients were seen in conditions with informative priors centered on the wrong population value, regardless of missing data conditions, and the regression coefficient strengths. Weakly informative priors correctly centered on the population value were comparable to FIML and MI, whereas diffuse priors and weakly informative priors centered on the wrong value tended to have higher MSE than FIML and MI. For the regression coefficients, diffuse priors produced comparable MSE values to FIML and MI when $\gamma_{21} = 0.5$, but the MSE values were higher than FIML and MI when $\gamma_{21} = 1.0$ and $\gamma_{21} = 1.5$. Overall, the Bayesian estimator has the potential to produce the lowest and highest MSE for γ_{11} and γ_{21}, depending on the prior specifications.

5.4. Categorical Covariate with Missing Data

5.4.1. Bias

Tables 5–7 provide the bias for the regression coefficients, γ_{11} and γ_{21}, and the intercept, γ_{01}. The patterns of bias in the binary x_2 conditions were similar to the patterns seen in the continuous x_2 conditions. Regardless of the simulation

Table 5. The bias and MSE in the regression coefficients for c on x_1 (i.e., γ_{11}), c on x_2 (i.e., γ_{21}), and the intercept (i.e., γ_{01}), under different missing data conditions for $\gamma_{21} = 0.5$ and a binary x_2.

Missing mechanism			MAR						MCAR					
Missing percentage			15%		35%		55%		15%		35%		55%	
Approach	Parameter	Pop. Value	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Listwise deletion	γ_{11}	−1	−.022	.030	−.013	.039	−.029	.068	−.017	.030	−.027	.039	−.033	.063
	γ_{21}	0.5	−.001	.068	−.009	.088	.009	.145	.008	.081	−.002	.092	.009	.148
	γ_{01}	0	**−.156**	.073	**−.368**	.195	**−.600**	.451	.003	.049	.001	.055	−.013	.078
FIML	γ_{11}	−1	−.019	.024	−.012	.026	−.010	.028	−.016	.025	−.024	.025	−.015	.024
	γ_{21}	0.5	.000	.069	−.005	.087	.006	.136	.008	.081	−.004	.091	.001	.144
	γ_{01}	0	−.006	.045	−.002	.047	−.001	.062	−.001	.046	.004	.048	−.007	.058
MI	γ_{11}	−1	−.018	.024	−.012	.025	−.009	.027	−.015	.025	−.022	.024	−.013	.024
	γ_{21}	0.5	−.009	.064	−.032	.081	−.037	.126	−.005	.076	−.035	.080	−.046	.123
	γ_{01}	0	−.002	.043	.010	.046	.017	.059	.006	.045	.020	.046	.018	.051
Bayes-correct informative	γ_{11}	−1	−.015	.009	−.010	.010	−.007	.010	−.012	.009	−.017	.009	−.009	.009
	γ_{21}	0.5	.000	.009	−.002	.008	.000	.007	.003	.010	−.001	.008	.001	.007
	γ_{01}	0	−.006	.030	−.001	.026	.005	.028	.001	.028	.004	.028	−.005	.025
Bayes-correct weakly informative	γ_{11}	−1	−.030	.019	−.024	.020	−.020	.021	−.027	.019	−.034	.019	−.024	.018
	γ_{21}	0.5	.003	.033	−.002	.035	.005	.040	.010	.039	.002	.036	.006	.042
	γ_{01}	0	−.005	.037	.000	.034	.005	.037	.001	.036	.005	.035	−.006	.034
Bayes-wrong informative	γ_{11}	−1	**−.245**	.071	**−.240**	.070	**−.240**	.071	**−.241**	.070	**−.248**	.073	**−.243**	.070
	γ_{21}	0.5	**.352**	.133	**.381**	.153	**.423**	.186	**.354**	.136	**.381**	.154	**.420**	.184
	γ_{01}	0	**−.156**	.062	**−.168**	.061	**−.182**	.069	**−.145**	.055	**−.159**	.060	**−.187**	.067
Bayes-wrong weakly informative	γ_{11}	−1	**−.114**	.035	**−.110**	.035	**−.111**	.037	**−.111**	.034	**−.121**	.036	**−.114**	.035
	γ_{21}	0.5	**.176**	.066	**.205**	.079	**.266**	.113	**.181**	.074	**.207**	.082	**.262**	.114
	γ_{01}	0	−.082	.046	−.093	.045	**−.116**	.054	−.074	.044	−.087	.045	**−.121**	.051
Bayes-diffuse	γ_{11}	−1	−.039	.027	−.033	.028	−.034	.030	−.036	.028	−.045	.027	−.039	.027
	γ_{21}	0.5	.002	.069	−.005	.087	.006	.140	.011	.082	.001	.091	.009	.146
	γ_{01}	0	−.003	.046	.003	.047	.005	.063	.002	.047	.007	.048	−.005	.058

Note. Conditions with ±0.1 bias are bolded.

Table 6. The bias and MSE in the regression coefficients for c on x_1 (i.e., γ_{11}), c on x_2 (i.e., γ_{21}), and the intercept (i.e., γ_{01}), under different missing data conditions for $\gamma_{21} = 1$ and a binary x_2.

Missing mechanism			MAR						MCAR					
Missing percentage			15%		35%		55%		15%		35%		55%	
Approach	Parameter	Pop. Value	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Listwise deletion	γ_{11}	−1	−.004	.030	−.003	.037	−.026	.067	−.020	.032	−.019	.040	−.031	.069
	γ_{21}	1	−.015	.085	−.011	.124	.001	.160	.008	.092	−.011	.102	.028	.173
	γ_{01}	0	**−.133**	.060	**−.357**	.191	**−.611**	.470	.004	.051	.002	.055	−.011	.081
FIML	γ_{11}	−1	−.001	.025	−.016	.024	−.013	.031	−.019	.028	−.015	.025	−.011	.027
	γ_{21}	1	−.006	.087	.006	.128	.006	.158	.007	.092	−.015	.101	.014	.167
	γ_{01}	0	.013	.041	.007	.052	−.004	.062	.002	.047	.004	.048	−.005	.061
MI	γ_{11}	−1	.002	.025	−.011	.023	−.004	.029	−.015	.027	−.008	.024	.000	.025
	γ_{21}	1	−.032	.081	−.049	.110	−.076	.141	−.019	.085	−.074	.091	−.080	.139
	γ_{01}	0	.023	.040	.031	.050	.028	.060	.014	.046	.032	.046	.040	.055
Bayes-correct informative	γ_{11}	−1	−.003	.009	−.012	.008	−.008	.011	−.013	.009	−.013	.009	−.007	.009
	γ_{21}	1	−.002	.009	.000	.009	.001	.007	.002	.010	−.004	.008	.004	.007
	γ_{01}	0	.013	.028	.013	.030	.003	.030	.004	.031	.002	.030	.002	.027
Bayes-correct weakly informative	γ_{11}	−1	−.014	.018	−.027	.017	−.023	.023	−.028	.020	−.027	.019	−.020	.019
	γ_{21}	1	.002	.038	.008	.044	.011	.041	.011	.040	−.001	.036	.018	.044
	γ_{01}	0	.015	.033	.013	.037	.003	.038	.005	.037	.005	.036	.001	.036
Bayes-wrong informative	γ_{11}	−1	**−.252**	.075	**−.265**	.080	**−.264**	.083	**−.262**	.081	**−.265**	.081	**−.262**	.080
	γ_{21}	1	**.377**	.152	**.409**	.176	**.445**	.205	**.381**	.155	**.404**	.171	**.446**	.206
	γ_{01}	0	**−.112**	.047	**−.130**	.055	**−.161**	.066	**−.121**	.053	**−.138**	.058	**−.153**	.058
Bayes-wrong weakly informative	γ_{11}	−1	**−.113**	.035	**−.131**	.038	**−.133**	.045	**−.128**	.040	**−.130**	.039	**−.130**	.040
	γ_{21}	1	**.199**	.082	**.242**	.108	**.300**	.135	**.209**	.087	**.232**	.094	**.304**	.139
	γ_{01}	0	−.057	.039	−.077	.046	**−.114**	.056	−.066	.045	−.083	.046	**−.110**	.051
Bayes-diffuse	γ_{11}	−1	−.021	.027	−.039	.027	−.039	.035	−.039	.031	−.038	.028	−.036	.031
	γ_{21}	1	.000	.087	.011	.129	.018	.162	.014	.093	−.005	.101	.032	.169
	γ_{01}	0	.019	.042	.015	.052	.004	.062	.008	.048	.011	.048	.000	.061

Note. Conditions with ±0.1 bias are bolded.

Table 7. The bias and MSE in the regression coefficients for c on x_1 (i.e., γ_{11}), c on x_2 (i.e., γ_{21}), and the intercept (i.e., γ_{01}), under different missing data conditions for $\gamma_{21} = 1.5$ and a binary x_2.

Missing mechanism			MAR						MCAR					
Missing percentage			15%		35%		55%		15%		35%		55%	
Approach	Parameter	Pop. Value	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
Listwise deletion	γ_{11}	−1	−.002	.033	−.003	.047	.004	.065	−.014	.035	−.015	.043	−.023	.072
	γ_{21}	1.5	.006	.097	−.036	.122	−.047	.171	.003	.113	.006	.135	.046	.215
	γ_{01}	0	**−.138**	.069	**−.348**	.178	**−.572**	.425	.010	.053	.006	.057	−.007	.085
FIML	γ_{11}	−1	−.002	.028	−.004	.033	.000	.033	−.011	.031	−.009	.030	−.002	.032
	γ_{21}	1.5	.026	.101	.004	.131	.001	.189	.000	.112	−.001	.130	.022	.202
	γ_{01}	0	.007	.048	.009	.048	.001	.071	.007	.049	.004	.051	−.009	.063
MI	γ_{11}	−1	.004	.027	.010	.031	.019	.030	−.005	.030	.006	.027	.020	.028
	γ_{21}	1.5	−.023	.092	−.097	.109	**−.137**	.163	−.039	.103	−.098	.115	**−.124**	.164
	γ_{01}	0	.024	.047	.046	.046	.050	.070	.023	.048	.044	.050	.053	.058
Bayes-correct informative	γ_{11}	−1	−.003	.009	−.004	.010	−.003	.010	−.009	.009	−.009	.009	−.001	.009
	γ_{21}	1.5	.009	.008	.001	.007	.002	.006	.000	.009	.000	.007	.007	.007
	γ_{01}	0	.017	.033	.014	.031	.008	.037	.008	.033	.008	.034	.003	.031
Bayes-correct weakly informative	γ_{11}	−1	−.015	.019	−.016	.022	−.014	.022	−.023	.021	−.022	.020	−.013	.020
	γ_{21}	1.5	.028	.039	.012	.039	.015	.040	.011	.043	.012	.040	.029	.043
	γ_{01}	0	.014	.039	.015	.036	.008	.045	.010	.040	.009	.039	.000	.038
Bayes-wrong informative	γ_{11}	−1	**−.272**	.085	**−.276**	.089	**−.280**	.090	**−.277**	.088	**−.281**	.090	**−.278**	.088
	γ_{21}	1.5	**.420**	.185	**.436**	.198	**.467**	.224	**.410**	.177	**.436**	.197	**.471**	.228
	γ_{01}	0	−.083	.049	**−.102**	.050	**−.127**	.064	−.091	.050	**−.105**	.054	**−.123**	.055
Bayes-wrong weakly informative	γ_{11}	−1	**−.132**	.041	**−.139**	.047	**−.146**	.048	**−.139**	.045	**−.145**	.045	**−.144**	.046
	γ_{21}	1.5	**.263**	.111	**.284**	.124	**.341**	.160	**.244**	.107	**.282**	.125	**.351**	.170
	γ_{01}	0	−.052	.045	−.068	.044	**−.100**	.060	−.055	.049	−.072	.048	**−.101**	.053
Bayes-diffuse	γ_{11}	−1	−.024	.030	−.028	.037	−.029	.037	−.033	.034	−.034	.033	−.031	.036
	γ_{21}	1.5	.037	.102	.017	.136	.025	.194	.013	.114	.015	.134	.053	.209
	γ_{01}	0	.016	.048	.020	.048	.012	.071	.016	.050	.014	.051	.000	.063

Note. Conditions with ±0.1 bias are bolded.

conditions, Listwise deletion produced unbiased parameter estimates for the regression coefficients. However, listwise deletion severely biased the intercept parameter when the missing data mechanism was MAR. As the percentage of missing data increased, the intercept became increasingly biased when using listwise deletion. Both FIML and MI produced unbiased γ_{11} and γ_{01} estimates, across simulation condition. Despite these similarities, FIML produced unbiased γ_{21} estimates, whereas MI had some conditions with biased γ_{21} parameters. As seen in Table 7, MI introduced bias to the γ_{21} parameter when the missing data percentage was 55% and $\gamma_{21} = 1.5$, regardless of the missing data mechanisms.

The bias patterns in the binary x_2 conditions were highly influenced by the prior when using the Bayesian third step.

Informative and weakly informative priors centered on the correct population value produced unbiased estimates for the regression coefficients and intercept, regardless of the simulation conditions. Diffuse priors also produced unbiased parameter estimates. In contrast, informative and weakly informative priors centered on the wrong value biased the regression coefficients, γ_{11} and γ_{21}. Notably, using priors centered on the wrong value for the regression coefficients also biased the intercept, γ_{01}, when the missing percentage was 55%.

5.4.2. Mean Square Error

Tables 5–7 provide the MSE for the regression coefficients, γ_{11} and γ_{21}, and the intercept, γ_{01}. Listwise deletion, MI, and FIML tended to have similar MSE values for the regression coefficients in conditions with a binary x_2, regardless of the covariate strengths (i.e., 0.5, 1.0, 1.5) and missing data conditions. In contrast, Listwise Deletion had some the highest MSE values for the intercept. When using listwise deletion, the combination of a higher percentage of missing x_2 data and the MAR missing data mechanism resulted in the highest MSE values for the intercept, γ_{01}.

When using Bayesian estimation, the MSE values for the regression coefficients was highly dependent on the prior specifications. The lowest MSE values for the regression coefficients were seen in conditions using informative priors that were correctly centered on the population values. The MSE was much lower in these conditions than FIML and MI. However, informative priors centered on the wrong population values tended to inflate the MSE for the regression coefficients. Weakly-informative priors correctly centered on the population value also produced lower MSE values for the regression coefficients when compared to FIML and MI. In contrast, weakly informative priors centered on the wrong population value tended to have similar MSE values for the regression coefficients when compared to FIML and MI. When using diffuse priors for the regression coefficients and intercept, Bayesian estimation produced MSE values that were comparable to FIML and MI. Thus, the advantages of Bayesian estimator were lost when using more diffuse priors.

5.5. Additional Points

Regardless of whether the x_2 variable was continuous or binary, Bayesian estimation with correct informative prior specifications produced unbiased estimates and the lowest MSE values for the regression coefficients, γ_{11} and γ_{21}. When the x_2 variable was continuous, the MSE values for the intercept, γ_{01}, tended to be equivalent for FIML, MI, and Bayesian estimation with informative priors correctly centered on the population value. In contrast, when the x_2 variable was binary, Bayesian estimation with informative priors correctly centered on the population value had the lowest MSE values for the intercept. One possible (and likely) explanation for the Bayesian third step not always having the lowest MSE values for the intercept is that a diffuse prior was used on the intercept for Bayesian analysis models. If a more informative prior correctly centered on the intercept population value was used, it is likely Bayesian estimation would produce the lowest MSE values for the intercept, regardless of conditions. When using the ML three-step approach, the user would have information about the intercept from class enumeration in the first estimation step that could be used to help specify a prior in the third estimation step.

6. Discussion

The primary aim of this simulation study was to explore the performance of available methods for addressing incomplete covariate data when using the ML three-step approach. To accomplish this aim, we generated data with different missing data mechanisms (MCAR and MAR), percentages of missing data (15%, 35%, and 55%), covariate distributions (standard normal and Bernoulli), and strengths of the covariate effect (weak, moderate, and strong). Next, we analyzed the datasets with four different methods for addressing the incomplete covariate data: listwise deletion, FIML, MI, and Bayesian estimation. When using Bayesian estimation, a variety of prior specifications were considered. No prior simulation study has compared these methods for addressing incomplete covariate data when using the ML three-step approach.

Statistical software such as M*plus* defaults to listwise deletion when using the ML three-step approach. Previous methodological research suggests listwise deletion is a poor method for addressing missing data because it reduces sample size and statistical power (Little, 1992; Little & Zhang, 2011). In addition, listwise deletion can bias parameter estimates when the missing data mechanism was not MCAR (Little, 1992; Little & Zhang, 2011). The results from the current study illustrate this point; listwise deletion introduced bias into the intercept, γ_{01}, when the missing data mechanism was MAR. This is especially problematic because bias in the intercept suggests the latent class measurement model is not remaining intact during the third estimation step. For this reason, listwise deletion is the worst option for addressing incomplete covariate data when using the ML three-step approach.

In contrast to listwise deletion, FIML consistently produced unbiased parameter estimates, regardless of the missing data mechanism and the percentage of missing data. Despite misspecifying the categorical x_2 as continuous, FIML still produced unbiased parameter estimates for the regression coefficients. For this reason, we would recommend using FIML in modeling situation with a single covariate with missing data. If the user includes several covariates with missing data in the model, FIML may not be the best option because computation time increases with each additional covariate, see Asparouhov and Muthén (2021) for more details. Another important factor to consider when using FIML is the variability in the estimates. Although parameter estimates obtained from FIML were unbiased, they exhibited greater variability compared to

estimates obtained through Bayesian estimation with informative and weakly-informative priors. This trend was especially evident in conditions with a categorical x_2 variable.

The MI results were somewhat surprising. MI produced unbiased parameter estimates in all conditions with a covariate strength of $\gamma_{21} = 0.5$ and $\gamma_{21} = 1.0$. However, MI underestimated the γ_{21} regression coefficient in conditions with high percentage of missing data when the covariate strength was $\gamma_{21} = 1.5$. In conditions with a continuous x_2, MI underestimated γ_{21} when 55% of the x_2 data was missing and when 35% of the x_2 data was MAR. In conditions with a binary x_2, MI underestimated γ_{21} when 55% of the x_2 data was missing. One possible explanation for why MI produced biased regression coefficients in conditions with higher percentages of missing data is the number of imputations used. Specifically, we set the number imputations across conditions to 20, which is considered standard in simulation research (Enders & Mansolf, 2018; Vera & Enders, 2021). However, some past methodological research suggests the optimal number of imputations should reflect the percentage of missing data (e.g., 55 imputations for 55% missingness; Von Hippel, 2009; White, Royston, & Wood, 2011). Future methodological research should consider whether a greater number of imputations would improve the performance of MI. Another factor to consider is the variability in the parameter estimates. When the x_2 variable was categorical, parameters obtained through MI had much greater variability than those obtained from Bayesian estimation with informative and weakly-informative priors. Based on these findings, MI may not be the best choice when there is a high percentage of incomplete covariate data. Alternative methods (e.g., FIML and Bayesian estimation) provide unbiased γ_{21} estimates under similar conditions.

Bayesian estimation has the potential to be the best or worst method for addressing incomplete covariate data in the third step, depending on the prior specification. When using informative priors correctly centered on the population value of the regression coefficient, the parameter estimates were consistently unbiased and the variability in the parameter estimates was very low, regardless of the strengths of the covariate, distributions of the covariate, and the missing data conditions. Similarly, weakly-informative priors correctly centered on the population value produced unbiased parameter estimates, across conditions. However, when using informative or weakly informative priors that are centered on the wrong value, we see some of the highest levels of bias in the regression coefficients. For this reason, we would recommend the use of informative and weakly-informative priors in the third estimation step if *a priori* knowledge about the relationship between the covariate and latent class variable is available. When using Bayesian estimation in applied settings, it is important to always perform a prior sensitivity analysis. For an example of how to implement a prior sensitivity analysis, refer to the work by Depaoli et al. (2020).

Considering wrong priors can have a dramatic impact on model results, users may be tempted to use diffuse priors for the regression coefficients in the third estimation step. When using diffuse priors for all parameters, Bayesian estimation tended to produce unbiased estimates in most conditions. The primary exception to this trend is when the incomplete covariate data is continuous, the strength of the covariate is either 1 or 1.5, and the missing percentage is high 55%. The reason these conditions tended to be trickier for diffuse priors is that the diffuse prior is centered on zero, whereas the population value was higher (i.e., 1 or 1.5). Although diffuse priors may be appealing to applied researchers who do not have much knowledge about the relationship between the covariate and the latent class variable, most of the advantages of Bayesian estimation disappear when using diffuse priors. One would be as well off using FIML or MI in these situations. Regardless of the prior specifications on the regression coefficients, all Bayesian estimation conditions had a diffuse prior on the latent class intercept. It is likely the Bayesian estimator would have lower variability in the intercept if more informative priors were specified. In applied settings, the user can incorporate prior information about the class proportions by specifying more informative priors on the latent class intercept and regression coefficients.

The current study was not without limitations. The simulation study only explored the performance of the methods for addressing incomplete covariate data in conditions with moderate class separation and equal class proportions. These factors can have a dramatic impact on the performance of the ML three-step approach and may also influence the performance of the methods for addressing missing data. Bayesian estimation may be especially helpful in situations with poor class separation or a minority latent class (Depaoli, 2012; 2013; Kim, 2014; Lu et al., 2011; Nylund, Asparouhov, et al., 2007; Tueller & Lubke, 2010). Future research should examine the possible benefits of the Bayesian third step in these modeling situations. The current study was also limited to the conditional LCA model. These results may not be applicable for addressing incomplete covariate data in latent growth mixture models and mixture confirmatory factor analysis models. MI worked well for many conditions in the simulation study, but previous research suggests MI does not typically work well in mixture models (Enders & Gottschall, 2011; Sterba, 2014; 2017).

Overall, results from the current study suggest a variety of methods can be used to address the incomplete covariate data when using the ML three-step approach. Based on our findings, we would recommend the use of Bayesian estimation when it is possible to use informative or weakly-informative priors on the regression coefficients. In situations where informative priors are not possible, FIML could work well with one or a small number of incomplete covariates; and MI could be use when the missing data percentage is not large or with an increased number of imputed datasets. Listwise deletion should be avoided whenever possible.

References

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. http://statmodel.com/download/Bayes3.pdf

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling*, *21*, 329–341. https://doi.org/10.1080/10705511.2014.915181

Asparouhov, T., & Muthén, B. (2021). *Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary second model*. Mplus Webnote No.21. Muthén & Muthén. https://www.statmodel.com/examples/webnotes/webnote21.pdf

Bakk, Z., Oberski, D., & Vermunt, J. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, *22*, 520–540. https://doi.org/10.1093/pan/mpu003

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*, 272–311. https://doi.org/10.1177/0081175012470644

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*, *23*, 20–31. https://doi.org/10.1080/10705511.2014.955104

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375–1386. https://doi.org/10.1080/01621459.1997.10473658

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3–27. https://doi.org/10.1093/pan/mph001

Clogg, C. C., & Goodman, L. A. (1985). Simultaneous latent structure analysis in several groups. In N. B. Tuma (Ed.), *Sociological methodology* (pp. 81–110). Jossey-Bass. https://doi.org/10.2307/270847

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and Health Sciences*. Wiley. https://doi.org/10.1002/9780470567333

Collins, L. M., Schafer, J. L., & Kam, C. K. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, *6*, 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent class models. *Journal of the American Statistical Association*, *83*, 173–178. https://doi.org/10.2307/2288938

Depaoli, S. (2012). Measurement and structural model class separation in mixture CFA: ML/EM versus MCMC. *Structural Equation Modeling*, *19*, 178–203. https://doi.org/10.1080/10705511.2012.659614

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*, 186–219. https://doi.org/10.1037/a0031609

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling*, *21*, 239–252. https://doi.org/10.1080/10705511.2014.882686

Depaoli, S., Winter, S. D., & Visser, M. (2020). The importance of prior sensitivity analysis in Bayesian Statistics: Demonstrations using an interactive Shiny App. *Frontiers in Psychology*, *11*, 608045. https://doi.org/10.3389/fpsyg.2020.608045

Di Mari, R., & Bakk, Z. (2018). Mostly harmless direct effects: A comparison of different latent Markov modeling approaches. *Structural Equation Modeling*, *25*, 467–483. https://doi.org/10.1080/10705511.2017.1387860

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Enders, C. K., & Gottschall, A. C. (2011). The impact of missing data on the ethical quality of a research study. In Panter and Serba (Eds.), *Handbook of Ethics in quantitative methodology* (pp. 357–381). Taylor and Francis Group.

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods*, *23*, 76–93. https://doi.org/10.1037/met0000102

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100. https://doi.org/10.1207/S15328007SEM1001_4

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Taylor & Francis.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231. https://doi.org/10.1093/biomet/61.2.215

Haberman, S. J. (1979). *Analysis of qualitative data, Volume 2: New developments*. Academic Press.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. SAGE. https://doi.org/10.2307/1165391

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. SAGE. https://doi.org/10.4135/9781412984850

Hallquist, M. N., & Wiley, J. F. (2018). *MplusAutomation*: An R package for facilitating large-scale latent variable analyses in M*plus*. *Structural Equation Modeling: a Multidisciplinary Journal*, *25*, 621–638. https://doi.org/10.1080/10705511.2017.1402334

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling: a Multidisciplinary Journal*, *17*, 193–215. https://doi.org/10.1080/10705511003659342

Huh, J., Riggs, N. R., Spruijt-Metz, D., Chou, C., Huang, Z., & Pentz, M. (2011). Identifying patterns of eating and physical activity in children: A latent class analysis of obesity risk. *Obesity*, *19*, 652–658. https://doi.org/10.1038/oby.2010.228

Janssen, J. H. M., van Laar, S., de Rooij, M. J., Kuha, J., & Bakk, S. (2019). The detection and modeling of direct effects in latent class analysis. *Structural Equation Modeling*, *26*, 280–290. https://doi.org/10.1080/10705511.2018.1541745

Kamakura, W. A., Wedel, M., & Agrawal, J. (1994). Concomitant variable latent class models for conjoint analysis. *International Journal of Research in Marketing*, *11*, 451–464. https://doi.org/10.1016/0167-8116(94)00004-2

Kim, S. (2014). Determining the number of latent classes in single- and multiphase growth mixture models. *Structural Equation Modeling*, *21*, 263–279. https://doi.org/10.1080/10705511.2014.882690

Kolb, R. R., & Dayton, C. M. (1996). Correcting for nonresponse in latent class analysis. *Multivariate Behavioral Research*, *31*, 7–32. https://doi.org/10.1207/s15327906mbr3101_2

Li, L., & Hser, Y. I. (2011). On inclusion of covariates for class enumeration of growth mixture models. *Multivariate Behavioral Research*, *46*, 266–302. https://doi.org/10.1080/00273171.2011.556549

Little, R. J. (1992). Regression with Missing X's: A review. *Journal of the American Statistical Association*, *87*, 1227–1237. https://doi.org/10.1080/01621459.1992.10476282

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis of missing data* (2nd ed.). Wiley.

Little, R. J., & Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society Series C*, *60*, 591–605. https://doi.org/10.1111/j.1467-9876.2011.00763.x

Lloyd, J., Doll, H., Hawton, K., Dutton, W. H., Geddes, J. R., Goodwin, G. M., & Rogers, R. D. (2010). How psychological symptoms relate to different motivations for gambling: An online study of internet gamblers. *Biological Psychiatry*, *68*, 733–740. https://doi.org/10.1016/j.biopsych.2010.03.038

Lu, Z. L., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent class dependent missing data. *Multivariate Behavioral Research*, *46*, 567–597. https://doi.org/10.1080/00273171.2011.589261

Lubke, G., & Muthén, B. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling*, *14*, 26–47. https://doi.org/10.1080/10705510709336735

Masyn, K. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 2, pp. 551–611). Oxford University Press.

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling*, *24*, 180–197. https://doi.org/10.1080/10705511.2016.1254049

Merkle, E. C. (2011). A comparison of imputation methods for Bayesian factor analysis models. *Journal of Educational and Behavioral Statistics*, *36*, 257–276. https://doi.org/10.3102/1076998610375833

Moss, H. B., Chen, C. M., & Yi, H. (2007). Subtypes of alcohol dependence in a nationally representative sample. *Drug and Alcohol Dependence*, *91*, 149–158. https://doi.org/10.1016/j.drugalcdep.2007.05.016

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81–117. https://doi.org/10.2333/bhmk.29.81

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's Guide* (8th ed.). Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*, 535–569. https://doi.org/10.1080/10705510701575396

Nylund, K., Bellmore, A., Nishina, A., & Graham, S. (2007). Subtypes, severity, and structural stability of peer victimization: What does latent class analysis say? *Child Development*, *78*, 1706–1722. PMID: 17988316 https://doi.org/10.1111/j.1467-8624.2007.01097.x

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling*, *26*, 967–985. https://doi.org/10.1080/10705511.2019.1590146

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results from a simulation study exploring the impact of Misspecified effects on class enumeration. *Structural Equation Modeling*, *23*, 782–797. https://doi.org/10.1080/10705511.2016.1221313

Petras, H., & Masyn, K. (2010). General growth mixture analysis with ante- cedents and consequences of change. In A. R. Piquero & D. Weisburd (Eds.), *Handbooks of quantitative criminology* (pp. 69–100). Springer.

Quirk, M., Nylund-Gibson, K., & Furlong, M. (2013). Exploring patterns of Latino/a children's school readiness at kindergarten entry and their relations with grade 2 achievement. *Early Childhood Research Quarterly*, *28*, 437–449. https://doi.org/10.1016/j.ecresq.2012.11.002

R Core Team. (2019). *R: A language and environment for statistical computing*. R Core Team.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592. https://doi.org/10.2307/2335739

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Skrondal, A., & Rabe-Hesketh, S. (2014). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society Series C*, *63*, 211–237. https://doi.org/10.1111/rssc.12023

Sterba, S. K. (2014). Handling missing covariates in conditional mixture Models under missing at random assumptions. *Multivariate Behavioral Research*, *49*, 614–632. https://doi.org/10.1080/00273171.2014.950719

Sterba, S. K. (2017). Pattern mixture models for quantifying missing data uncertainty in longitudinal invariance testing. *Structural Equation Modeling*, *24*, 283–300. https://doi.org/10.1080/10705511.2016.1250635

Tueller, S., & Lubke, G. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling*, *17*, 165–192. https://doi.org/10.1080/10705511003659318

Vera, J. D., & Enders, C. K. (2021). Is item imputation always better? An investigation of wave-missing data in growth models. *Structural Equation Modeling*, *28*, 506–517. https://doi.org/10.1080/10705511.2020.1850289

Vermunt, J. K. (1997). *Log-linear models for event histories*. SAGE.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*, 450–469. https://doi.org/10.1093/pan/mpq025

Vermunt, J. K., & Magidson, J. (2016). *Technical Guide for Latent GOLD 5.1: Basic, advanced, and syntax*. Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2021). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling*, *28*, 356–364. https://doi.org/10.1080/10705511.2020.1818084

Von Hippel, P. T. (2009). How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, *39*, 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*, 377–399. https://doi.org/10.1002/sim.4067

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219–240, 269–281). Lawrence Erlbaum Associates Publishers.

Yamaguchi, K. (2000). Multinomial logit latent class regression models: An analysis of predictors of gender role attitudes among Japanese women. *American Journal of Sociology*, *105*, 1702–1740. https://doi.org/10.1086/210470