



# Multiple Imputation Methods for Latent Profile Analysis in Education and the Behavioral Sciences

## Citation

Waldman, Marcus R. 2020. Multiple Imputation Methods for Latent Profile Analysis in Education and the Behavioral Sciences. Doctoral dissertation, Harvard Graduate School of Education.

## Permanent link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364538>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

**Multiple Imputation Methods for Latent Profile Analysis in Education and the  
Behavioral Sciences**

Marcus R. Waldman

Andrew Ho  
Katherine Masyn  
Dana McCoy

A Thesis Presented to the Faculty  
of the Graduate School of Education of Harvard University  
in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Education

2020

**© 2020**  
**Marcus Richard Waldman**  
**All Rights Reserved**

## Table of Contents

<b>Acknowledgements</b>	vi
<b>Abstract</b>	viii
<b>List of Symbols</b>	x
<b>Chapter 1: Prologue</b> .....	1
Challenge 1: Generating Proper Imputations for LPA.....	5
Challenge 2: Model Selection with Multiply Imputed Data in LPA.....	10
Summary and Contributions to Applied Research.....	11
Figures.....	13
References.....	15
<b>Chapter 2: Evaluating Recursive Partitioning Imputation to Treat Missing Data in Latent Profile Analysis</b> .....	17
Background.....	20
Gaussian Finite Mixture Models.....	21
Inclusive Missing Data Strategy: Limitations with FIML.....	22
Multiple Imputation: Theoretical Advantages and Known Limitations in LPA.....	27
Conditional Mean Imputation by Recursive Partitioning Prediction Models....	29
CART and RF Imputation: A Multiple Group Missing Data Approach.....	31
Simulation Design.....	34
Missing Data Rates Observed in Applied Research.....	35
Number of Indicators and Classes Observed in Applied Research.....	35
Class Separation Values Observed in Applied Research.....	36
Manipulated Conditions.....	36
Overall Design.....	38
Data Generating Mechanism and Missing Data Mechanism.....	38
Results.....	44
Bias and Density Estimation.....	44
95% CI Coverage.....	47
Imputation Efficiency.....	47

Discussion.....	48
Tables.....	53
Figures.....	58
References.....	68
<b>Chapter 3: Investigating the Performance of a Proposed Hybrid Approach to Generate Imputations in a Latent Profile Analysis under Model Uncertainty: An Initial Study.....</b>	<b>73</b>
Background.....	80
Finite Mixture Models.....	80
Bayesianly Proper Imputations.....	82
Normal Imputation as a Case Study for Understanding the FCS and JM.....	85
A Hybrid Imputation Method when the Number of Classes are Known.....	89
Novel Modifications to EM with Sampling.....	89
Hybrid Imputation as an Approximate Posterior Sampler with Data Augmentation.....	91
Multiple Imputation when the Number of Classes is not Known.....	94
Overfitted Mixture Imputation Model.....	95
Reversible Jump MCMC.....	98
Sampling from a Mixture of Competing Models.....	98
Reference Model Selection.....	100
Summary.....	102
Applied Example: Case Study with ECLS-K 1998 Data.....	102
Research Context.....	103
Measures.....	104
Sample and Missing Data Mechanism.....	106
Model Selection and Class Definitions.....	107
Results.....	109
Discussion.....	110
Incorporating Prior Information to Improve Estimation Stability.....	111
Mixture Model Averaging.....	112

Tables.....	114
Figures.....	120
References.....	123
<b>Chapter 4: On Model Selection using Information Criteria with Finite Mixtures in the Presence of Missing Data.....</b>	<b>129</b>
Missing Data Approaches to Estimating Finite Mixture Models.....	134
Notation and Background.....	134
Ignorability and FIML.....	137
Conditional Ignorability and Multiple Imputation.....	139
FIML as the Currently Preferred Missing Data Strategy.....	142
Prevailing Information Criteria Model Selection Practices.....	143
Consequences of Substituting the Observed-Data Likelihood under FIML on the BIC.....	144
Current Practice with Finite Mixture Model Selection under Multiple Imputation.....	147
Sensitivity of Finite Mixture Model Selection Decisions to Missing Data under Current Practices.....	148
Simulation Setup.....	149
Model Selection Results.....	155
Correcting the BIC for Evidence Loss Due to Missing Data under FIML.....	160
Reanalyzing the FIML Simulation Results.....	162
Alternatives to Averaging for Multiple Imputation Model Selection.....	163
Majority Vote.....	164
Stacking.....	164
Consentino & Claeskens' (2010) $D_{LR}$ Procedure.....	165
Reexamining Simulations: Evaluating Alternatives to Averaging.....	166
Discussion.....	166
Figures.....	169
References.....	182
<b>Chapter 5: Epilogue.....</b>	<b>186</b>
Contextualizing the Contributions in this Dissertation.....	186

Guidance to Applied Researchers.....	192
Summary.....	195
References.....	196
<b>Technical Appendix A: Useful Principles, Quantities, and Properties in Missing Data Theory.....</b>	<b>197</b>
<b>Technical Appendix B: Illustrating the Limitations of a Saturated Correlates Approach in Latent Profile Analysis.....</b>	<b>202</b>
<b>Technical Appendix C: Derivation of <math>BIC^{OBS}</math>.....</b>	<b>208</b>

## **Acknowledgments**

This dissertation would not have been possible without the support of so many of my family, friends, and colleagues. To my wife, Caitlin Gemmill: Your love and support has meant everything. Thank you so much for your care and patience, and thank you for editing this dissertation several times over. To my father, Bill Waldman, thank you for your wisdom and guidance, and for instilling in me an ethos of persistence and determination in my early years.

To my in-laws, Karen and Bob Gemmill: Thank you for letting me squat in your house FROG (finished room over garage) as I ran simulations on a series of makeshift computers. Thank you for not getting mad when I caused you to exceed your data limit for your broadband internet, and most of all thank you for encouraging me to keep at it and turn what seemed to be the impossible into something that was not only possible but inevitable. To Margie Stafford, Todd Everson, and Jessica Gemmill: how lucky I have been to have such nice extended relatives that welcomed me in their Atlanta-area homes so that I could meet with a committee member. Your doors have always felt wide open, and I cannot thank you enough for your hospitality.

To my committee, Professors Andrew Ho, Katherine Masyn, and Dana McCoy: Andrew, thank you for always looking forward and for your flexibility. Katherine, coming from another institution, you had no professional incentive to support my intellectual development, but you chose to do so anyways, and in the process, mentor me in this field. Thank you for believing in me and helping get to this point. Dana, you anchored me, even in rough seas. Thank you for always being there while challenging me to become a clearer writer.

I would like to thank Goodwill of Charleston, South Carolina, where I bought many high-quality, affordable supplies to build the homemade computers that I used to conduct the simulation studies contained in this dissertation.

Thank you to all of the first responders, health professionals, and medical researchers as they work tirelessly to keep the public safe during the COVID-19 pandemic.

## Abstract

Researchers in education and the behavioral sciences are increasingly conducting latent profile analysis by focusing investigations on subpopulations of individuals to describe individual differences with greater nuance. At the same time, missing data is practically inevitable, and even modest missingness rates can threaten the validity of inferences. Multiple imputation is a powerful strategy to treat missing data, but it suffers from key limitations in both the imputation and pooling phases when conducting latent profile analysis.

In this dissertation, I conduct three studies to address these gaps. In the first study (Chapter 2), I evaluate whether recursive partitioning imputation algorithms better mitigate nonresponse bias than alternative missing data approaches that are common in practice. I find that recursive partitioning imputation algorithms perform well when sample sizes are large ( $N = 1,200$ ), but not when sample sizes are small ( $N = 300$ ) or when class separation is weak (i.e., entropy  $\approx .74$ ). In response, I propose a hybrid imputation procedure in the second study (Chapter 3); the proposed method embeds a finite mixture model to generate imputations using a joint modeling framework within a larger chained equations procedure. I demonstrate the hybrid imputation procedure using real-world data.

In the final study (Chapter 4), I scrutinize current practices for conducting finite mixture model selection with missing data. I am not aware of any studies evaluating whether model selection decisions are sensitive to real-world missing data problems. I fill this gap in research by studying whether selection decisions are sensitive to missing data if either a full information maximum likelihood (FIML) or a multiple imputation strategy

is employed. Two findings emerge. First, with regards to FIML, the BIC under extracts the true number of classes relative to the complete data condition in the presence of small sample sizes and small classes. Second, with regards to multiple imputation, current practices for pooling information criteria result in model selection decisions that poorly replicate the decisions that would have been made had the data been complete. I propose two remedial procedures for future practice, and, using simulations, I show that these two procedures outperform current practices.

## List of Symbols

### Indices

$i$	Index for individual.
$N$	Total number of individuals.
$j$	Index for profile indicator variable.
$J$	Total number of indicators. Also corresponds to the dimension of the data space.
$p$	Index for the missing data correlate.
$P$	Number of missing data correlates.
$k$	Index for latent class.
$m$	Index for imputed dataset.
$M$	Total number of imputed datasets.
$t$	Iteration within a sampling algorithm.
$T$	Total number of iterations.
$\mathbf{r}_{Y,i}$	Index or indices of observed class indicators for individual $i$ .
$\backslash \mathbf{r}_{Y,i}$	Index or indices of missing class indicators for individual $i$ .

### Variables

$\mathbf{y}_i$	Complete class indicator data for the $i$ th case (i.e., comprised of both observed and missing values).
$Y$	Collection of complete indicator data $\{\mathbf{y}_i: i = 1, \dots, N\}$ .
$\mathbf{y}_i^{\text{obs}}$	Observed indicator data the $i$ th individual.
$Y^{\text{obs}}$	Collection of all observed class indicator data.
$\mathbf{y}_i^{\text{mis}}$	The missing indicator data values for the $i$ th individual that was not observed.
$Y^{\text{mis}}$	Collection of missing class indicator data.
$\mathbf{y}_{i,m}^{\text{imp}}$	The $m$ th imputed data for the $i$ th individual comprised of both the observed and $m$ th set of plausible values.
$Y_m^{\text{imp}}$	Size $N \times J$ imputed dataset matrix.
$\mathbf{x}_i$	Missing data correlate for the $i$ th individual (i.e., missing data correlate). Size $P$ vector.
$X$	Size $N \times (P + 1)$ design matrix for missing data correlates/auxiliary variables. First column is all 1's for an intercept.
$X_{-p}$	Size $N \times (P + 1)$ design matrix without the $p$ th missing data correlate/auxiliary variables.
$\mathbf{x}_i^{\text{obs}}$	Vector with the $i$ th individuals missing data correlate/auxiliary variable values.
$X^{\text{obs}}$	Set of observed missing data correlate/auxiliary variable values.
$X_p^{\text{obs}}$	Size $N$ vector of the values for the $p$ th missing data correlate.
$\mathbf{x}_i^{\text{mis}}$	Vector with the $i$ th individuals missing auxiliary variable values.
$X^{\text{mis}}$	Collection of all missing auxiliary variable values.
$R$	Response matrix.

## Parameters

$K$	Number of mixture components fit to the data.
$\kappa$	Random variable representing latent class.
$\kappa_i$	Class of the $i$ th individual.
$\eta^*$	Underlying latent response influencing missing data pattern.
$\theta$ or $\theta_K$	Vector of all population parameters (of mixture model $\mathcal{M}_K$ ).
$\hat{\theta}$ or $\hat{\theta}_K$	Vector representing estimates of a parameter from a single sample. Unless otherwise stated represents the maximum likelihood estimate, $\hat{\theta}_K^{\text{MLE}}$ (of mixture model $\mathcal{M}_K$ ).
$\hat{\theta}_m$ or $\hat{\theta}_{m,K}$	Maximum likelihood estimate fit to the $m$ th imputed dataset (when fitting mixture model $\mathcal{M}_K$ ).
$\bar{\theta}$ or $\bar{\theta}_K$	Vector representing pooled parameter estimate from Rubin's (1987) rules for model $\mathcal{M}_K$ .
$\mu_k$	Mean vector for the $k$ th Gaussian component.
$\beta_k$	A size $P + 1$ vector of covariates that inform the conditional means such that $\mu_{ik} = \mathbf{x}_i \beta_k$ .
$\Sigma_k$	Covariance matrix for the $k$ th Gaussian component.
$\pi_k$ or $\pi_k(\mathbf{x}_i)$	Marginal class probabilities (i.e., mixture weights) for the $k$ th class.
$\alpha_k$	A size $P + 1$ vector of covariates for the multinomial logistic regression
	$\log \frac{\pi_k(\mathbf{x}_i)}{\pi_K(\mathbf{x}_i)} = \mathbf{x}_i \alpha_k$
$\tau_k(\mathbf{y}_i)$ or $\tau_k(\mathbf{y}_i, \mathbf{x}_i)$	Posterior class probability for the $i$ th individual given $Y_i$ and $X_i$ .
$\psi$	Parameters defining the missing data mechanism.
$\phi$	Parameters defining the univariate regression models for the fully conditional specification.

## Other

$\mathcal{M}$	Set of models fit during enumeration.
$\mathcal{M}_K$	The model fit with $K$ in a sequence of models.
$q_K$	Number of parameters corresponding to the $K$ th model in $\mathcal{M}$
$\mathcal{L}(\theta Y)$	Likelihood function.
$\ell(\theta Y)$	Log-likelihood function. In the presence of missing data, the complete-data loglikelihood, observed-data log likelihood, and missing-data loglikelihood are denoted as $\ell_{\text{com}}(\theta Y)$ , $\ell_{\text{obs}}(\theta Y^{\text{obs}})$ , or $\ell_{\text{mis}}(\theta Y^{\text{mis}})$ , respectively.
IC	Information criterion (e.g., AIC, BIC) .

$\Delta\text{IC}$	Difference in information criteria. Unless otherwise stated this difference is between model $\mathcal{M}_{K+1}$ and $\mathcal{M}_K$ .
LR	Likelihood ratio statistic. Unless otherwise stated this statistic compares the fit between a “reduced model”, $\mathcal{M}_K$ , and a “full” model and $\mathcal{M}_{K+1}$ , i.e., $\text{LR}(\hat{\theta}_K, \hat{\theta}_{K+1} Y) = -2\{\ell(\hat{\theta}_K Y; \mathcal{M}_K) - \ell(\hat{\theta}_{K+1} Y; \mathcal{M}_{K+1})\}.$
$D_{\text{LR}}$	Meng and Rubin’s (1992) statistic D-statistic comparing fit between a “reduced model”, $\mathcal{M}_K$ , and a “full” model and $\mathcal{M}_{K+1}$ .
$I(\hat{\theta} Y)$	Information matrix evaluated at the maximum likelihood estimate given. If data are missing, the information matrix if given the complete data denoted $I_{\text{com}}(\hat{\theta} Y)$ . If given the observed data or the missing data, the corresponding information matrices are $I_{\text{obs}}(\hat{\theta} Y^{\text{obs}})$ or $I_{\text{mis}}(\hat{\theta} Y^{\text{mis}})$ , respectively.
$V(\hat{\theta} Y)$	Asymptotic variance-covariance matrix related to the information matrix by $V(\hat{\theta}) = I^{-1}(\hat{\theta} Y)$ . If data are missing, the complete data, observed data, and missing data asymptotic matrices are denoted $V_{\text{com}}(\hat{\theta} Y)$ , $V_{\text{obs}}(\hat{\theta} Y^{\text{obs}})$ , or $V_{\text{mis}}(\hat{\theta} Y^{\text{mis}})$ , respectively.
$\bar{U}$ or $\bar{U}_K$	Size $q_K \times q_K$ within imputation variance covariance matrix when fitting model $\mathcal{M}_K$ to the imputed datasets. The $K$ subscript is dropped if the number of components fit is implied.
$B$ or $B_K$	Size $q_K \times q_K$ between imputation variance covariance matrix when fitting model $\mathcal{M}_K$ to the imputed datasets. The $K$ subscript is dropped if the number of components fit is implied.
$\hat{T}$ or $\hat{T}_K$	Size $q_K \times q_K$ total imputation variance covariance matrix.
$\mathcal{N}_J(\cdot)$	$J$ -variate normal probability density function.
$\Pr(Y \cdot)$	Probability density of the indicator.
$\Pr(\theta Y, X)$	Posterior density of $\theta$ given data.
$[\cdot \cdot]$	Conditional distribution.
$[\cdot \cdot; \mathcal{M}_K]$	Conditional distribution given a model with $K$ components is fit.
$\mathbf{H}$	Hessian matrix for a vector valued function evaluated at the point $\theta = \hat{\theta}$ . For a general function $f(\theta)$ the elements of the Hessian matrix are given by $\mathbf{H} \circ f(\theta) = \nabla \nabla^T f(\theta).$
	Note the property that the inverse of the observed information matrix is the negative of the inverse Hessian of the log-likelihood function at the maximum likelihood estimate, $I(\theta Y) = -\mathbf{H} \circ \ell(\hat{\theta} Y) = -\mathbf{H} \circ \log \Pr(Y \hat{\theta}).$
$\mathbb{I}$ or $\mathbb{I}_{q_K}$	Identity matrix (with $q_K \times q_K$ elements).

## Chapter 1: Prologue

This dissertation represents the current state of my research agenda focused on effectively treating missing data in person-centered analysis. The origins of this research agenda can be traced back to an impulsive purchase of Professor Craig Enders' (2010) book, *Applied Missing Data Analysis*, while serving as a teaching assistant for Professor Katherine Masyn's *Applied Latent Class Analysis* course at StatsCamp in the summer of 2016. Professor Masyn's course introduced students to the basics of conducting a person-centered analysis where the focus of study is on how individuals in a population are similar to and different from one another. The course has always been quite popular among education and psychological researchers because phenomena in these fields tend to be nuanced and complex, requiring an analytic framework (like person-centered analysis) that describes individual differences (Bergman & Magnusson, 1997; Bergman & Trost, 2006). The course's popularity reflects the general trend of increased adoption of person-centered approaches. In fact, over the past 20 years, publication rates for studies that implement a person-centered analysis are experiencing exponential growth, as indicated by yearly rates doubling every three to four years, on average, for manuscripts in the Web of Science database (Figure 1.1).

Attendees of StatsCamp arrive ready to analyze a real-world dataset of their own by applying the methods they learn in the course. Consistent with the experiences of applied behavioral science researchers everywhere, missing data is prolific in these datasets, and it is important to address the missing data using appropriate statistical techniques because missingness rates as low as 5-10% can threaten the validity of inferences (Little, Jorgensen, Lang, & Moore, 2014). As a teaching assistant, I knew

missing data was an unavoidable topic for which the students would seek my advice, so I was eager to learn more about the topic. “What should I tell my students regarding best practices for treating their missing data in their own person-centered analyses?”, I remember thinking to myself while scarfing down green chile ice cream at a shop in Old Town, Albuquerque. To that point, I had been telling my students that the software we were using defaults to estimating using full information maximum likelihood (FIML), and all that is required with that estimator is that the data be missing at random (MAR) conditional on the observed indicator variables.

But what if the missing at random assumption was not tenable given just the variables that represented the class indicators, as one student pointed out to me? My response was that the student should conduct multiple imputation with the assistance of external variables while using standard software packages such as `mice` (van Buuren & Groothuis-Oudshoorn, 2011) in R (R Core Team, 2020) or the `mi impute` command in Stata (StataCorp, 2019). I soon discovered that I was offering poor advice. As I turned the pages in Professor Enders’ book in the ice cream shop, I came across the sentence that caused my heart to skip. In referring to latent profile analysis (LPA), a type of person-centered analysis, Enders (2010, pp. 268–269) explained, “[M]ultiple imputation can produce biased estimates of the model parameters, even when the data are [missing complete at random] MCAR.” Humbled now, knowing that I had offered the student poor advice, I needed to know for myself why multiple imputation was so inadequate in person-centered analysis, even if its use was prolific in regression analysis, factor analysis, and structural equation modeling. The question also remained: what do I tell this student who is concerned that the MAR assumption requires external variables which

themselves contain missing data? Seeking a satisfactory answer to that question began a research agenda that continues to this day.

Multiple imputation in LPA is problematic for two reasons, but only one has been recognized in the current literature. The first reason, and the reason previously explored in literature, is that the single-class imputation models standard in software fail to produce imputations that reflect the unobserved heterogeneity present in data with latent subpopulations—subpopulations that are assumed to be present in an LPA (Enders, 2010; Enders & Gottschall, 2011; Sterba, 2016). Stated in a more technical way, because single-class imputation models do not reflect the true multiple group structure of the data and result in biased estimates, imputations drawn from these models do not meet statistical requirements to be considered “proper”—neither in a Bayesian sense (Schafer, 1997) nor in a frequentist framework (Rubin, 1987).

Upon leaving Albuquerque, I was determined to identify methods for generating proper imputations in LPA. I had identified a research question that I was going to explore at the next Modern Modeling Methods conference hosted by the University of Connecticut in 2017. Specifically, I was unsatisfied by the fact that the research to date on multiple imputation in LPA had relied on multivariate normal imputation methods to create the imputed datasets, when so many other methods were available that did not make such strong parametric assumptions. In preparing for that conference, I quickly realized that even if I were to identify an imputation method that produced proper imputations, multiple imputation is challenging in a person-centered analysis for a second reason that has not been fully discussed in literature.

The second reason why multiple imputation is problematic concerns model selection. Specifically, researchers conducting LPA typically do not know the number of latent subpopulations supported by the data a priori. To ascertain this value, the researcher enumerates the classes by fitting alternative finite mixture models with different numbers of mixture components specified. Each mixture component represents a latent class which is, in turn, presumed to represent a distinct latent subpopulation in the overall population. To settle on a final model, the researcher is instructed to evaluate the model fit information and balance that information with substantive theory (Masyn, 2013). Simulation studies have shown that relative fit information provided by information criteria, such as the BIC, are useful in deciding on a final model (Nylund, Asparouhov, & Muthén, 2007; Tofighi & Enders, 2008).

The challenge with multiple imputation is that the models are fit separately to each imputed dataset, resulting in multiple information criteria values. Popular software defaults to averaging the information criteria values across the imputed datasets, even though previous literature in the covariate-selection regression context suggests that averaging information criteria is not theoretically justified and is often not an optimal ad-hoc pooling strategy. Indeed, although previous literature had offered different strategies for pooling information criteria, I had not found any studies comparing these alternative strategies to the default averaging technique standard in software. This revelation was concerning because I realized that even if the challenge of creating proper imputations were to be resolved, how an applied researcher conducting person-centered analysis should best conduct model selection would remain an unsettled question.

Thus, the research agenda that came to form this dissertation can be summarized by offering solutions to these two challenges. The first challenge is generating proper imputations that reflect the multiple group structure of the data represented by latent subpopulations. Provided that the first challenge can be satisfactorily resolved, the second challenge is identifying best practices for conducting model selection in an LPA with missing data. The remainder of this prologue provides an intuitive understanding for both these challenges.

### **Challenge 1: Generating Proper Imputations for LPA**

To provide intuition for why multivariate normal imputation—or equivalently, normal regression imputation—does not reflect the multiple group structure in a person-centered analysis, consider that incorporating group information when generating imputations (through, for example, interaction terms) is critical. In fact, multiple studies have shown that if missingness depends on group membership (such as subpopulations), then bias can result if group membership information is not incorporated in the imputation model (Collins, Schafer, & Kam, 2001; Enders & Gottschall, 2011). To avoid bias, researchers should adopt what is termed an inclusive analytic strategy (Graham, 2003; Little et al., 2014; Rubin, 1996; Schafer & Graham, 2002). When there are many groups, this means that the researcher either specifies many more interaction terms with group membership than are hypothesized to exist, or the researcher fits separate imputation models across the groups entirely (Allison, 2002; Enders, 2010).

To clarify concepts, consider the relationship between the amount of strenuous weekly cardiovascular exercise and percent body fat among healthy adults with a substantial portion of individuals missing exercise values. Even though the researcher

may not be interested in how gender may moderate this relationship, it is best practice to either specify interaction terms between gender and daily exercise when imputing missing exercise times, or to specify a multiple group imputation model where exercise values are imputed separately between males and females. Incorporating gender information in the imputation model is important because males and females differ in healthy ranges of percent body fat, so ignoring these gender differences risks biasing estimates if the missingness rates depend on gender. However, even if the missingness does not depend on gender, the inclusion of these interaction terms (or a multiple group imputation model) will still result in unbiased estimates. Thus, methodologists recommend researchers adopt an inclusive strategy by incorporating group information to offset potential sources of bias.

However, unlike in the example above, an individual's group (or subpopulation) is usually assumed to be a latent variable and one that is not directly observable in an LPA. This precludes the adoption of an inclusive analytic strategy that explicitly incorporates group information. The fact that group membership cannot be directly observed has negative consequences when single-class models (such as the normal imputation regression model) are used to impute missing profile indicator values. This is because the imputations generated from a single-class model fail to reproduce subpopulations in the data, resulting in imputations that do not meet the requirements for statisticians to deem them proper (Rubin, 1987; Schafer, 1997).

The problem with single-class imputation models in LPA is easily visualized. Consider a pedagogical example with two well-separated subpopulations of individuals, as displayed in Panel A of Figure 1.2, and many observations missing  $Y_2$  (so that only

$Y_1$  values are observed for these observations). For simplicity, assume the missingness on  $Y_2$  is completely random so that the distribution of the missing observations matches the distribution of the observed observations; thus, a proper imputation procedure should result in the distribution illustrated in Panel A. However, if a normal imputation model is specified, then imputations will be drawn from a single-class distribution resembling Panel B in Figure 1.2. Thus, when the imputed data are combined with the complete cases (see Panel C in Figure 1.2), the resulting distributions of the subpopulations are less well separated and observably different in configuration (i.e., are elliptical and not perfectly circular) than the true distributions in Panel A.

## **Chapter 2**

If the problem with generating proper imputations in LPA with normal regression stems from normal regression models assuming the data are generated from a single class, then it seems logical that methods which do not make this assumption may mitigate any bias due to improper imputations. Clustering algorithms are particularly useful in modeling the unobserved heterogeneity in the data that is implied by latent subpopulations. Fitting finite mixture models is one means for clustering the data and is used ubiquitously in LPA. However, several other clustering methods exist.

Recursive partitioning is one such alternative clustering algorithm, and imputation models that incorporate recursive partitioning are available in mainstream software. In recursive partitioning, the observations are clustered into rectangular partitions in the data, which can be visualized by the rectangular clusters formed in Figure 1.3. The clustering process guarantees that the imputations exhibit some multiple group structure. As can be seen in Figure 1.3, the rectangular clusters formed from recursive partitioning

clustering are different than those implied by a mixture distribution of two Gaussian distributions (separately indicated by the grey and gold dots). Nevertheless, the recursive partitioning process ensures that observations are clustered into increasingly homogenous subgroups with respect to the latent classes. This clustering process clearly avoids the problems of imputing from a single-class model.

I hypothesize that recursive partitioning imputation algorithms will bypass the problems associated with single-class imputation models because they impose a multiple-group structure when generating imputations. In Chapter 2, I evaluate the performance of recursive partitioning imputation when treating missing data in LPA using the fully conditional specification (van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006), where missing values are imputed sequentially by each variable containing missing data. Imputing in this manner is highly flexible in that it allows for external variables of multiple data types (i.e., categorical, ordinal, continuous, etc.) to be specified so that imputations are generated following an inclusive analytic strategy. Consistent with my hypothesis, I find that recursive partitioning imputation algorithms—and classification and regression tree imputation, in particular—bypass the problems associated with single-class imputation models and, under certain conditions, are capable of generating proper imputations.

### **Chapter 3**

Although the recursive partitioning imputation procedure avoids imputing from a single-class model, it is nevertheless uncongenial with the analytic model. Uncongeniality is a concept introduced by Meng (1994) in which the imputation model differs in substantively meaningful ways from the analytic model. For recursive

partitioning imputation, this is visualized in Figure 1.3 by the rectangular clusters which do not match the circular latent classes. Uncongeniality in multiple imputation can result in biased parameter estimates if the imputation model is not sufficiently complex to capture the true data generating model. In contrast, a congenial model is at least as complex as the analytic model and captures the data generating mechanism well. Congenial models are known to mitigate bias even if the additional complexity leads to imputation inefficiency. Because mitigating bias is the primary concern when treating missing data, it is considered best practice to ensure that the imputation model is at least as complex as the analytic model (Murray, 2018; Rubin, 1996). This is done in an inclusive strategy when, for example, more interactions are incorporated into the imputation model than are hypothesized to exist.

In Chapter 3, I perform an initial study to address inherent uncongeniality between a recursive partitioning imputation model and a finite mixture model as the analytic model. To enhance congeniality, I propose and investigate the performance of an imputation procedure where a mixture regression model is used to generate imputations. The mixture regression model allows for AVs to be specified in order to make the MAR assumption more tenable.

I also investigate whether the proposed imputation procedure can perform well under model uncertainty. After all, the number of subpopulations is not known, so the number of components to specify in the mixture model is not known to the imputer. Even outside of the person-centered analysis context, missing data literature is increasingly recognizing that imputations should reflect inherent uncertainty in the true data generating model. In particular, some are now arguing that imputations cannot be proper

if they do not incorporate uncertainty of the true data generating model, in addition to the more traditional sources of uncertainty required for valid inference (Kaplan & Yavuz, 2019). The imputation procedure I propose attempts to ensure that model uncertainty is reflected in the imputations by separately considering competing mixture models specified with different numbers of components at each iteration. I evaluate whether the proposed procedure adequately recaptures key parameters using an empirical dataset. I find that additional work is needed to fine-tune the proposed imputation algorithm before it can be adopted by applied researchers.

Even if these issues can be resolved, the challenge does not end with generating proper imputations for LPA. The imputer is still in a position where they must conduct class enumeration and select a final model. Chapter 4 addresses this second challenge.

### **Challenge 2: Model Selection with Multiply Imputed Data in LPA**

There is currently a gap in literature about how best to pool results from the analysis phase when conducting model selection during class enumeration. Class enumeration refers to a model selection procedure where the researcher determines the number of subpopulations supported by the data using statistical evidence, including nested model tests and information criteria. Best practices for pooling information criteria and conducting hypothesis tests during class enumeration have not been established. Thus, although multiple imputation represents a powerful technique to treat missing data in variable-centered approaches, its application in person-centered approaches will remain limited by unresolved methodological gaps for conducting model selection.

## **Chapter 4**

The third and final study addresses the current gap as it relates to pooling information criteria for model selection in LPA. Through a simulation study, I consider alternative pooling procedures to averaging the information criteria, and I evaluate how these alternatives perform in selecting the correct model.

In addition to completing the simulation study to compare alternative pooling procedures, I also note that the standard BIC formula is only correct if the data do not contain missing values. This is important because applied researchers often prioritize the BIC because the BIC has shown favorable performance in simulations (Nylund et al., 2007). I show that when data are missing, the BIC tends to select a model with too few components (i.e., under extracts the number of classes). The source of this problem has to do with the assumption that  $N$  complete observations are present in the data when calculating the penalty term. In this chapter, I propose an adjustment to the penalty term to be applied when data are missing.

### **Summary and Contributions to Applied Research**

The adoption of multiple imputation to address missing data in LPA is hindered by two challenges. The first challenge is generating proper imputations so that unbiased estimates and valid inferences can be obtained for the parameter estimates that define the latent classes. The second challenge concerns pooling model fit information to conduct model selection given that the number of latent classes supported by the data must be determined empirically. Chapters 2 and 3 investigate methods to address the first challenge. Chapter 4 addresses the challenge with pooling model fit information. In

Chapter 5, I reflect on the lessons learned from each of the studies and provide tangible guidance to applied researchers.

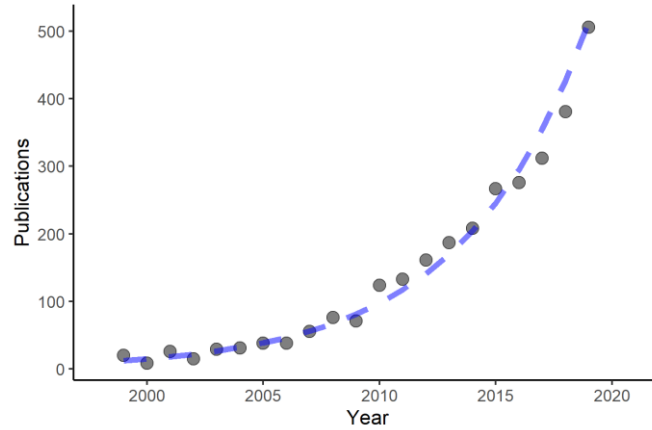
This dissertation provides several contributions to methodological and applied research in education and the behavioral sciences. Specifically, I demonstrate that the current practice of fitting mixture models to profile indicators with missing values using FIML has many limitations. The limitations of current practice threaten the validity of inferences regarding class definitions and subpopulation membership (Chapter 2). Similarly, current practices do not guarantee model selection decisions are robust to missing data (Chapter 4). My fundamental assertion is that multiple imputation can be a viable missing data strategy in LPA, one that is better positioned than FIML to address the limitations that I describe. The simulations contained in this dissertation provide empirical support that, under certain conditions, multiple imputation can better address missing data than FIML.

In summary, my dissertation makes an important contribution by paving the way for a line of inquiry oriented around providing researchers an expanded toolkit to address missing data when conducting person-centered analysis like LPA. Given the limitations of current practices, expanding the methodological toolkit is important to ensure that applied researchers are using statistical techniques that are robust to missing data problems as they seek to advance education and the behavioral sciences.

## Figures

**Figure 1.1**

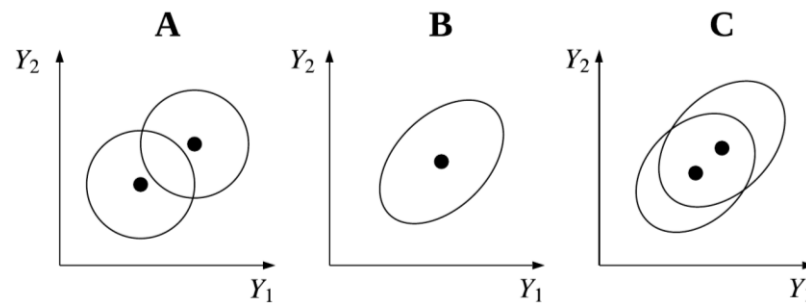
*Yearly Publications*



*Note.* Includes latent profile analysis, latent class analysis, growth mixture modeling, factor mixture modeling, latent transition analysis, or mixture modeling in education and psychology journals listed in the Web of Science database. Best fit line derived from exponential function with a doubling period of 3.75 years.

**Figure 1.2**

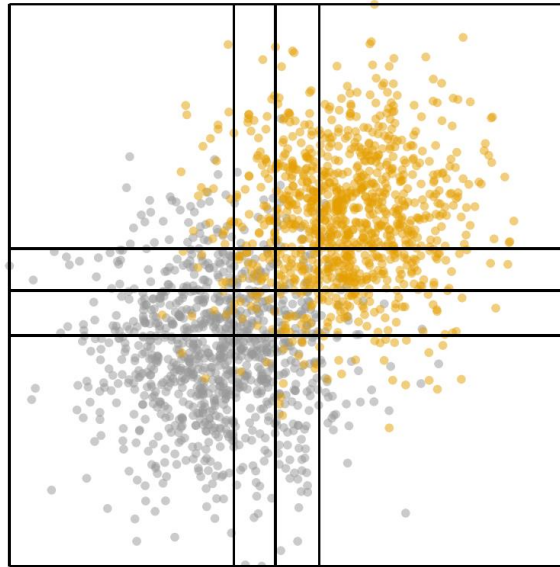
*Problems with Single-Class Imputation Models*



*Notes.* (A) The true distribution of the data (including for the fully observed values and those values missing  $Y_2$  had  $Y_2$  been observed). (B) The distribution of the imputations using a simple linear regression model for the observations missing  $Y_2$  when group membership is not known. (C) The resulting distribution that combines the complete cases and imputed cases (i.e., imputed data).

**Figure 1.3**

*Uncongeniality of Clusters from Recursive  
Partitioning*



*Note.* Rectangular clusters (sixteen in total) formed by recursive partitioning using classification and regression trees overlaid on two well-separated subpopulations generated by a mixture of Gaussians.

## References

- Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology*, 9(2), 291–319. <https://doi.org/DOI: 10.1017/S095457949700206X>
- Bergman, L. R., & Trost, K. (2006). The person-oriented versus the variable-oriented approach: Are they complementary, opposites, or exploring different worlds? *Merrill-Palmer Quarterly*, 52(3), 601–632. <https://doi.org/10.1353/mpq.2006.0023>
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351. <https://doi.org/10.1037/1082-989X.6.4.330>
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.
- Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, 18(1), 35–54.
- Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, 10(1), 80–100.
- Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, 39(2), 151–162. <https://doi.org/10.1093/jpepsy/jst048>
- Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (pp. 551–611). New York, NY: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199934898.013.0025>
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, 14(4), 535–569.
- R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from <https://www.r-project.org/>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Johnson Wiley & Sons. <https://doi.org/10.1002/9780470316696>
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489. <https://doi.org/10.1080/01621459.1996.10476908>
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147.
- StataCorp. (2019). Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.

- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in Latent Variable Mixture Models* (pp. 317–341). Charlotte, NC: Information Age Pub.
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3).

## **Chapter 2: Evaluating Recursive Partitioning Imputation to Treat Missing Data in Latent Profile Analysis**

Marcus R. Waldman & Katherine E. Masyn

Over the last few decades, methodologists have implored researchers in the behavioral sciences to adopt “state of the art” (Schafer & Graham, 2002) strategies to address missing data. Missing at random (MAR; Rubin, 1976) procedures are especially encouraged because they leverage observable information to improve the robustness of conclusions by mitigating nonresponse bias (Hancock, Stapleton, & Mueller, 2019; T. D. Little, Jorgensen, Lang, & Moore, 2014; Wilkinson, 1999). Although valid inference with MAR strategies is technically only guaranteed if the missing data mechanism is ignorable (i.e., the propensity for missingness is explained only by observables), MAR strategies can also protect against nonresponse bias when the mechanism is nonignorable if the missing values are well predicted by observed variables (Enders, 2008, p. 128; van Buuren, 2018, p. 101). The most common MAR strategies employed in psychology and the behavioral sciences include multiple imputation (MI; Rubin, 1987) and full information maximum likelihood (FIML) (Allison, 2002; Enders, 2010; R. J. Little & Rubin, 2002; Molenberghs & Kenward, 2007).<sup>1</sup>

Currently, researchers conducting latent profile analysis (LPA) rely almost exclusively on FIML to treat missing data. To highlight this fact, we identified thirty

---

<sup>1</sup> Weighted adjustment using propensity scores is an alternative MAR strategy to FIML and MI, and it is used heavily in large-scale surveys and in the clinical sciences (Molenberghs & Kenward, 2007). Although some studies in psychology have advocated for a weighted adjustment approach with promising simulation results in attrition contexts (Hayes & McArdle, 2017a; Hayes et al., 2015; McArdle, 2013), weighted adjustment is generally a complete case analysis (R. J. Little & Rubin, 2002) and is limited in its ability to treat multivariate missing data patterns like those observed in LPA (Hayes, 2017). We, therefore, do not consider weighted adjustment approaches.

highly cited LPA studies in the literature to inform the simulation study included in this chapter. We found that FIML was reported as the missing data strategy used to handle missing data in all cases, provided that a missing data strategy was reported at all. None of the authors reported specifying auxiliary variables (AVs) that are external to the LPA to address the missing data problem. This makes clear that researchers conducting LPA are relying heavily on the viability of the assumption that the missing data mechanism is ignorable absent any external information. Such an ignorability assumption is generally considered overly restrictive given the complicated contexts in which behavioral science data are collected and missing data appear.

To mitigate the associated nonresponse bias in observational data, methodologists strongly encourage researchers to adopt an “inclusive analysis strategy” (Collins, Schafer, & Kam, 2001) and incorporate many AVs to treat missing data with MAR procedures (Enders, 2010; T. D. Little et al., 2014; Rubin, 1996; Schafer, 1997; Schafer & Graham, 2002). When analyzing data using covariance structure models (i.e., regression and path analysis, factor models, or structural equation models), an inclusive strategy under FIML can be accomplished using specialized specification rules (Graham, 2003) or two-step procedures (Savalei & Bentler, 2009; Yuan & Bentler, 2000). Both augment the analytic model to factor in information from AVs. However, as we elaborate later, it is not obvious how to appropriately augment the mixture models used in LPA to incorporate AVs, and, even if it is possible, any developed procedure would likely be plagued by several limitations in real-world applications. Indeed, even in covariance structure models, it is common to quickly experience convergence issues when incorporating AVs in a FIML approach (Enders, 2010).

In contrast, multiple imputation is generally considered more flexible and better able to incorporate AVs than FIML because the model used to treat missing data (i.e., the imputation model) can be distinct from the analytic model used to make inferences and answer substantive research questions (Meng, 1994; Rubin, 1996; van Buuren, 2018). Moreover, multiple imputation can accommodate the inevitable complications that arise in the real world, including treating the missingness on AVs, which themselves are comprised of multiple data types (continuous, categorical, etc.). This is especially true if a multivariate imputation by chained equations (MICE) procedure is employed to generate the imputations (Azur, Stuart, Constantine, & Leaf, 2011; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

Some have cautioned against using multiple imputation (Sterba, 2016) because popular models used to generate the imputed datasets assume that the data exhibit a single group structure (Enders, 2010; Enders & Gottschall, 2011). In a joint modeling imputation framework, for example, it is typically assumed that the data are drawn from a (single) multivariate normal distribution (King, Honaker, Joseph, & Scheve, 2001; Schafer, 1997). Equivalently, if using MICE, the linear regression models popular in imputing continuous data assume a single-class structure with imputations drawn from a univariate normal distribution. The single-class assumption is problematic in LPA because such an analysis is person-centered in that the explicit goal is to identify multiple hidden subpopulations of individuals. Thus, creating imputed datasets from single-class imputation models risks obfuscating the very subgroups that the researcher intends to illuminate.

However, imputation models need not assume a single-class structure. For example, recursive partitioning regression models utilize a form of clustering to calculate predicted values for each observation. The clustering process assures that the imputations are drawn from a multiple group structure. Although recursive partitioning regression models are available in software when generating imputations using MICE, the performance of these methods has yet to be evaluated for LPA. This indicates a gap in literature, as these methods may provide a better option than FIML in treating missing data.

The purpose of this study is to fill this gap. Through a simulation study, we show that imputations created from recursive partitioning algorithms can effectively attenuate nonresponse bias and lead to valid inference outside of settings with small class sizes. In this way, we show that recursive partitioning imputation allows for an inclusive missing data strategy and is more effective than FIML in attenuating nonresponse bias outside of small class sizes. Thus, this study opens the door for multiple imputation to be an option for applied researchers treating missing data in person-centered analysis.

### **Background**

In this section, we provide a brief background on the finite mixture models fit in LPA, we highlight the challenges associated with conducting an inclusive strategy with FIML, and we introduce recursive partitioning as a viable imputation strategy for data containing latent subpopulations.

## Gaussian Finite Mixture Models

As typically implemented in LPA, individuals are clustered into homogenous subpopulations (or classes) by fitting Gaussian finite mixture models. It is assumed that the indicators are normally distributed conditional on class membership:

$$[\mathbf{y}_i | \kappa = k] \sim \text{MVN}(\mu_k, \Sigma_k) \quad (1.1)$$

where  $\mathbf{y}_i$  is the vector of latent class indicator data and  $\kappa$  represents the categorical latent variable representing class membership, while  $\mu_k$  and  $\Sigma_k$  represent the within-class mean vector and variance-covariance matrix for class  $k$ , respectively.

Finite mixture models require the researcher to specify the total number of component densities to be fit to the data,  $K$ . Typically, this value is not known a priori, so it must be inferred empirically from the data. The number of components is determined by fitting a sequence of mixture models with increasing numbers of classes until convergence becomes intractable. This process is referred to as enumerating the classes. A model within the sequence is then selected. For our purposes here, we assume that the number of classes is known. Model selection with incomplete or imputed datasets is an important topic that is addressed in Chapter 4.

Assuming that  $K$  is known, then the categorical latent variable,  $\kappa$ , follows a multinomial distribution:

$$\kappa \sim \text{Multinomial}(\pi_1, \dots, \pi_K) \quad (1.2)$$

where  $\pi_k$  represents the likelihood of a randomly selected individual belonging to class  $k$ , and  $k = 1, \dots, K$ . From (1.1) and (1.2), the corresponding likelihood of observing the complete data  $Y$  is given by:

$$\mathcal{L}(Y|\theta) = \prod_{i=1}^N \sum_{k=1}^K \pi_k \mathcal{N}_j(\mathbf{y}_i | \mu_k, \Sigma_k) \quad (1.3)$$

where  $\theta = \{\{\pi_k, \mu_k, \Sigma_k\}: k = 1, \dots, K\}$ .

### **Inclusive Missing Data Strategy: Limitations with FIML**

All missing data strategies can be justified by manipulating and studying the properties of the joint distribution formed by the data and the missing data mechanism. If profile indicator data are missing, then the complete indicator data  $Y$  is composed of both the missing and observed elements so that  $Y = \{Y^{\text{obs}}, Y^{\text{mis}}\}$ . The joint distribution of the data and missing data mechanism can then be written as:

$$\Pr(Y, R; \theta, \psi) = \Pr(Y^{\text{obs}}, Y^{\text{mis}}, R; \theta, \psi) \quad (1.4)$$

where  $Y$  represents the complete data with parameters  $\theta$ , and  $R$  represents the missing data patterns with corresponding parameters  $\psi$  (collectively referred to as the missing data mechanism). All MAR missing data strategies assume that the missing data mechanism is ignorable when estimating the  $\theta$  parameters, resulting in the inferences being drawn from the observed data likelihood. The primary requirement for ignoring  $R$  and  $\psi$  during estimation is that the data must be MAR: that is, the missing profile indicator data ( $Y^{\text{mis}}$ ) is independent of the missing data patterns ( $R$ ) conditional on the observed data ( $Y^{\text{obs}}$ ), as given by

$$Y^{\text{mis}} \perp R \mid Y^{\text{obs}} \quad (1.5)$$

(R. J. Little & Rubin, 2002, pp. 117–119; Rubin, 1976)<sup>2</sup>.

---

<sup>2</sup> An additional requirement is distinctiveness of  $\theta$  and  $\psi$  (R. J. Little & Rubin, 2002, pp. 119–120). Because distinctiveness is not considered as important (Schafer, 1997, p. 11), we do not elaborate.

Violations of the MAR assumption in (1.5) can lead to biased  $\theta$  estimates and invalid inference. In real-world LPA settings where the data are often observational in nature, the MAR assumption is likely untenable. A more tenable assumption is that MAR holds conditional on  $Y^{\text{obs}}$  and other observables. In the latter setting, valid inference can theoretically be obtained by conditioning on additional variables:

$$Y^{\text{mis}} \perp R \mid (Y^{\text{obs}}, X) \quad (1.6)$$

where  $X$  are the additional variables that are incorporated into the model as AVs.

Furthermore, it has been shown that highly predictive AVs can substantially reduce bias, even in settings where data remain missing not at random (Enders, 2010; van Buuren, 2018). For these reasons, methodologists generally implore applied researchers to incorporate many AVs to address missing data as part of an inclusive strategy (Collins et al., 2001; Enders, 2010; Rubin, 1996; Schafer & Graham, 2002).

Outside of mixture modeling, researchers have numerous options to include AVs and adopt an inclusive strategy. If treating missing data with multiple imputation, the researcher can specify AVs as part of the imputation model and remove these external variables when fitting the analytic model to the imputed datasets. Missingness in the AVs is easily treated if using the fully conditional specification (FCS; van Buuren et al., 2006) as part of a MICE imputation procedure.

Likewise, researchers have several options to include AVs if employing FIML estimation to treat missing data for regression analysis, factor analysis, structural equation modeling, and related covariance structure models. An inclusive strategy can be implemented by fitting a “saturated correlates” model (Graham, 2003). A saturated correlates model augments the analytic model with a set of missing data correlates as

AVs. A series of correlations are calculated with the goal of “working the AVs into the analysis without altering the substantive interpretation of the parameters” (Enders, 2010, p. 133). In this way, the MAR assumption is made more tenable by the inclusion of AVs without making parameter interpretations conditional on these variables, as would occur, for example, if the AVs were simply added as predictors in a regression analysis.

The saturated correlates model can be fit to covariance structure models using either a two-stage approach or specifying models according to Graham’s (2003) rules. The two-stage analysis mimics a method-of-moments estimation procedure. In the first stage, one estimates the sufficient statistics of the model (i.e., the variance-covariance matrix) in such a way that the estimates are informed by the AVs. With the variance-covariance matrix estimated, one can then fit the covariance structure model of interest (Savalei & Bentler, 2009; Yuan & Bentler, 2000). Alternatively, a saturated correlates model can be fit by correlating the AVs to any disturbance or uniqueness terms according to Graham’s (2003) rules.<sup>3</sup> By correlating the AVs with these error terms, Graham’s (2003) rules “[transmit] the information [needed for the MAR assumption to be tenable] from the auxiliary variables to the analysis model variables without affecting the interpretation of the parameter estimates” (Enders, 2010, p. 135).

However, saturated correlates models have limitations, even outside of mixture modeling. First, if the AVs themselves contain missing values, then the researcher either must adopt a listwise deletion strategy or specify the AVs as endogenous variables. If the latter approach is taken, then strong parametric assumptions about the distribution of the

---

<sup>3</sup> We refer the reader to Enders (2010, Chapter 5) for illustrative examples of how to implement Graham’s (2003) rules with different types of covariance structure models.

AVs must be made. For example, it is not uncommon to assume that binary AVs are treated as continuous so as to specify the necessary correlations required as part of Graham's (2003) rules. Additionally, convergence issues are common with a saturated correlates approach, and convergence becomes increasingly more difficult to achieve as the number of AVs increases. Between the distributional assumptions implied by endogeneity and the convergence issues, researchers often need to make compromises in deciding which AVs they need to exclude (Enders, 2010). Such compromises are not necessary when conducting multiple imputation according to the FCS because the FCS is highly flexible in treating varied data types and is more robust against nonconvergence (van Buuren, 2018).

Additionally, it remains unclear how to work the AVs into the analysis without causing the definitions of the classes to change and individuals to switch classes in an LPA. In Technical Appendix B, we show that specifying a mixture model according to Graham's (2003) rules causes class-specific mean estimates to change, even when the data are complete. Class means are often used to define the classes when naming the classes. Therefore, specifying a mixture model according to Graham's (2003) rules risks substantively altering the very definitions of the classes. In addition, there is evidence that individuals switched classes when the mixture was specified according to Graham's (2003) rules. This is shown by nonequivalent marginal probability estimates across the classes between a mixture model fit per the usual specification as opposed to a mixture model specified according to Graham's (2003) rules.

Why do Graham's (2003) rules alter key parameter interpretations when the latent variable is categorical (i.e., mixture models) and not when the latent variable is

continuous (e.g., factor analysis)? In brief, implementing Graham's (2003) rules with categorical latent variables results in the AVs effectively becoming additional class indicators; the corresponding fitted mixture model results in parameter estimates that best explain differences in relations between the AVs and the class indicators. However, AVs are external variables, so content validity considerations demand that they should not influence the definitions of the classes. What is clear is that by altering the parameters that define the classes, Graham's (2003) rules do not accomplish their fundamental goal of preserving interpretations in mixture models.

If Graham's (2003) rules fail to preserve key interpretations, it seems reasonable to question whether there is a two-stage procedure that could be developed to accomplish this goal, as is available with covariance structure models. Unfortunately, we do not consider a two-step approach possible. Mixture models require person-level data and cannot be fit to a covariance matrix, unlike covariance structure models. This is because mixture models estimate higher order moments in the data than simply the means and covariance matrices modeled in covariance structure models (Masyn, 2013; McLachlan & Peel, 2004). Consequently, there is no obvious way to translate the first step from the two-step approach by estimating the sufficient statistics.

In summary, it remains unclear how to conduct an inclusive missing data strategy when fitting a mixture model using FIML without changing the definitions of the classes. Even if a clever set of rules akin to Graham's (2003) rules could be identified for mixtures that accomplish this goal, there are still practical limitations that make an inclusive strategy with FIML unpalatable. These include increased convergence issues or invoking strong parametric assumptions to treat missing values in the AVs. Multiple

imputation is more flexible in treating a variety of data types and is less prone to practical problems like convergence issues, making the method more amenable to an inclusive strategy.

### **Multiple Imputation: Theoretical Advantages and Known Limitations in LPA**

As an alternative to FIML, MI is better equipped to incorporate AVs (Enders, 2010; van Buuren, 2018). Unlike FIML, MI's multiple phase approach separates and distinguishes the problem of treating missing data (imputation phase) from how substantive questions would be answered had the data been complete (the analysis model). Specifically, the model used to treat the missing data through generating imputations can be "uncongenial" (Meng, 1994) with the finite mixture model that would ultimately be fit to make inferences in LPA. This explicit separation of the missing data problem from the analysis problem protects the AVs from unduly influencing class membership above and beyond their role in treating the missing data. Additionally, saturated correlates models exhibit convergence difficulties when fit with a large number of AVs, a problem that is not observed with MI (Enders, 2010). Finally, unlike a saturated correlates approach, which places strict parametric assumptions on the AVs if these variables themselves contain missingness, missingness in the AVs can be addressed using nonparametric MI algorithms that are also more robust to misspecification issues (van Buuren, 2018). Thus, from a theoretical perspective, MI is often a more attractive alternative for dealing with missing data than FIML. Nevertheless, nuanced issues remain that threaten the validity of inferences drawn from mixture models when MI is used to treat missing data that contain latent subpopulations.

To understand these nuanced issues, it is important to consider some of the theoretical underpinnings that make MI a justifiable missing data strategy. MI, like FIML, requires that the missing data mechanism be ignorable. Unlike FIML, however, MI takes a Bayesian perspective in treating the missing data and relies on the Bayesian Central Limit theorem for valid frequentist inference. In particular, Little and Rubin (2002, p. 210) show that the posterior density for parameters  $\theta$  given the observed data is given by

$$\Pr(\theta|Y^{\text{obs}}) \propto \int_{Y^{\text{mis}}} \Pr(\theta|Y^{\text{mis}}, Y^{\text{obs}}) \Pr(Y^{\text{mis}}|Y^{\text{obs}}, \theta) dY^{\text{mis}} \quad (1.7)$$

where  $Y^{\text{obs}}$  and  $Y^{\text{mis}}$  represent observed and missing data, respectively. AVs can be incorporated in (1.7) to ensure that the ignorability requirement is met or to mitigate the effect of violations to the MAR assumption if the missing data mechanism is nonignorable:

$$\Pr(\theta|Y^{\text{obs}}) \propto \int_{Y^{\text{mis}}} \Pr(\theta|Y^{\text{mis}}, Y^{\text{obs}}) \Pr(Y^{\text{mis}}|Y^{\text{obs}}, X, \theta) dY^{\text{mis}} \quad (1.8)$$

where  $X$  is the set of AVs. The integral in (1.8) implies that valid inferences about  $\theta$  can be accomplished through a marginalizing process over all plausible values for  $Y^{\text{mis}}$  under the assumption that true posterior predictive distribution for the missing data (i.e.,  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$ ) can be sampled independently (R. J. Little & Rubin, 2002; Rubin, 1987). Sampling from the true posterior predictive distribution  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$  is guaranteed to lead to “Bayesianly proper” (Schafer, 1997, pp. 105–106) imputed datasets and valid Bayesian inference. In large samples and when the regularity conditions are met, the Bayesian Central Limit theorem ensures that Bayesian inference is consistent with frequentist inference. However, imputations sampled from a distribution other than

the true posterior predictive distribution are not theoretically guaranteed to lead to valid Bayesian inference, but, in many cases, they can perform well in real-world situations (van Buuren, 2018). Unfortunately, LPA is not one of those situations.

Multiple studies have shown that if missingness depends on group membership (such as subpopulations), then bias can result if group membership information is not incorporated in the imputation model (Collins et al., 2001; Enders & Gottschall, 2011). This is because the imputed datasets may not reflect key features of the data's structure if group membership information is not incorporated into the imputation models. As part of an inclusive strategy, this means that the researcher either specifies many interaction terms with group membership information, or the researcher fits separate imputation models across the groups entirely (Allison, 2002; Enders, 2010).

However, in a person-centered analysis, an individual's group (i.e., subpopulation) is assumed to be a latent quantity and not known a priori. This precludes the adoption of inclusive analytic strategies that explicitly incorporate group information. The fact that group membership cannot be directly observed has negative consequences when single-class imputation models are used to construct plausible values. This is because the imputations generated from a single-class model fail to reproduce the multiple group structure of the data, resulting in imputations that are not Bayesianly proper. This was discussed in the pedagogical example in Chapter 1.

### **Conditional Mean Imputation by Recursive Partitioning Prediction Models**

Recursive partitioning refers to a stratification algorithm by which partitions are formed by successively splitting up the data in a hierarchical fashion, ultimately leading to rectangular clusters of observations (Breiman, 1984). Specifically, an automated

procedure identifies a splitting variable and then selects the optimal cut point on that variable in a manner that results in the maximum reduction on a so-called impurity metric. This variable selection and splitting process repeats itself until a termination condition is met: either the change on the impurity measure no longer sufficiently decreases, or the number of observations in a resulting partition reaches some minimum value (Strobl, Malley, & Tutz, 2009). The final partition (or node) for any given observation can be determined by a set of decision rules that follow a tree-like structure, often referred to as a decision tree. An example of a decision tree and the corresponding rectangular clusters is illustrated in Figure 2.1.

In regression contexts, Breiman's (1984) classification and regression tree algorithm (CART) is a highly popular application of recursive partitioning for prediction. Specifically, upon termination of the recursive partitioning process, a summary statistic (e.g., mean, median, etc.) using only outcome values from observations in the terminal node are utilized (see Strobl, Malley, & Tutz, 2009 for more details). Because the same variable is chosen as the splitting variable multiple times, CARTs can accommodate nonlinear and even discontinuous functions. Moreover, CARTs automatically search for the presence of high-order interaction terms (Burgette & Reiter, 2010; Doove, van Buuren, & Dusseldorp, 2014) in the recursive partitioning process.

The machine learning community has since extended Breiman's (1984) CART algorithm with post hoc procedures to improve predictions coming from recursive partitioning regression algorithms. The random forest (RF; Breiman, 2001) algorithm represents one important extension. RFs are known to produce more accurate predictions

that better generalize to out-of-sample data than predictions from CART with pruning alone (Hastie, Tibshirani, & Friedman, 2009).

The machine learning community was also first to employ CART and RF predictions for missing values as part of a conditional mean imputation procedure (Bárcena & Tusell, 2000; Conversano & Cappelli, 2002; Conversano & Siciliano, 2009; Creel & Krotki, 2006; D'Ambrosio, Aria, & Siciliano, 2012; Ishwaran, Kogalur, Blackstone, & Lauer, 2008; Stekhoven & Bühlmann, 2011). Although the predicted value may best approximate the true value on average, it is well known that conditional mean imputation does not lead to valid statistical inference because it does not account for sampling variability or other important sources of variability (van Buuren, 2018). In particular, conditional mean imputation leads to speciously narrow confidence intervals, inflated Type-I error rates, and, critically in multivariate application, positively biased point estimates among association parameters that define the relations among the variables (R. J. Little & Rubin, 2002, Chapter 4). One may expect that biased association parameters would translate to bias across parameters of substantive interest, such as the class-specific means of the profile indicators.

### **CART and RF Imputation: A Multiple Group Missing Data Approach**

To resolve the problems associated with conditional mean imputation, several researchers have extended recursive partitioning imputation algorithms to reflect the important sources of variability around predicted values (Burgette & Reiter, 2010; Doove et al., 2014; Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014; Sovilj et al., 2016). In order to capture important sources of variability, values are randomly drawn from a donor pool that includes all values that would have otherwise been used to inform

the aggregate summary statistic to construct a predicted value. Randomly sampling from the donor pool implies that the plausible values are constructed by sampling from an empirical distribution. Valid inference technically requires that this empirical distribution sufficiently approximates the true posterior predictive distribution of the missing values,  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$  in (1.8), so that the imputations are Bayesianly proper. However, previous empirical studies from imputation methods that utilize donor pools to sample from empirical distributions have shown strong performance so long as the empirical distribution approximates the true distribution well (van Buuren, 2018)

Imputation algorithms that incorporate random partitioning through CART or RF models work in much the same way as when the goal is prediction. Specifically, CART and RF regression models can be embedded in a MICE procedure in multivariate missing data settings. MICE proceeds by sweeping across the data and sequentially fitting univariate CART or RF regression models to construct a donor pool for the missing values. At each sequential step, imputations are then drawn from the donor pool and substituted for the missing values. Sweeping across the dataset in this way repeats multiple times so that a Markov chain Monte Carlo (MCMC) is formed (van Buuren et al., 2006). Multiple imputed datasets are then constructed by sampling from the corresponding MCMC chain. We refer to the corresponding CART and RF MICE procedures as MICE-CART or MICE-RF, respectively.

In the R (R Core Team, 2020) computing environment, researchers can implement MICE-CART using the `mice` (van Buuren & Groothuis-Oudshoorn, 2010) package. MICE-RF can be implemented using either the `mice` or `CALIBERrfimpute` (Shah, Bartlett, Hemingway, Nicholas, & Hingorani, 2014) packages. For convenience,

only the `mice` package is employed in this study. It is important to note that neither implementation prunes to increase parsimony in the decision trees used to construct the donor pools. The advantages and disadvantages of not pruning are detailed later in the Discussion section.

MICE-CART and MICE-RF are promising alternatives to imputing with single-class models. Doove et al. (2014) and Shah et al. (2014) showed that generating imputed data from MICE-CART and MICE-RF models substantially reduces bias when high-order interactions in multiple regression contexts are present but not modeled, even compared to alternative nonparametric methods, such as predictive mean matching. These findings are consistent with other literature demonstrating the utility of recursive partitioning algorithms to detect interactions (Jacobucci, Grimm, & McArdle, 2017; Morgan & Sonquist, 1963; Strobl et al., 2009).

The favorable performance of MICE-CART and MICE-RF with unmodeled interaction terms has direct implications for LPA. In the hypothetical event that class membership was known, one could specify appropriate interactions with dummy variables to generate proper imputations. Thus, the challenge of imputing class indicators without knowing subpopulation membership is highly related to the challenge of imputing data when interactions are present but are left unmodeled, a scenario where CARTs and RFs excel.

Another reason why MICE-CART and MICE-RF are expected to perform well is because they incorporate clustering when constructing imputations (van Buuren, 2018). Specifically, the stratification process that results in rectangular clusters results in increasingly homogenous subgroups ultimately used to form the donor pool (see Figure

2.1). The clustering process in the first step guarantees that the imputations have some multiple group structure. Although the structure implied by imputed data generated from MICE-CART and MICE-RF may be qualitatively different than the true structure of data with subpopulations in LPA, it nevertheless avoids the problems of a single-class model.

Finally, an important application of mixture modeling separate from latent profile analysis is semiparametric density estimation where a multivariate density is approximated using finite mixture models. There, the goal is to simply approximate the joint distribution for the true data in a flexible and nonparametric manner. MICE-CART and MICE-RF are both nonparametric methods, implying that the imputation models are more aligned with the goal of semiparametric density estimation than traditional, parametric imputation models.

### **Simulation Design**

To study the performance of MICE-CART and MICE-RF in addressing missing data when the missingness depends on external variables, we conducted a simulation study. To inform the simulation conditions and to set parameters to typical values observed in applied literature, we used the Web of Science database to identify thirty frequently cited articles employing LPA that were published between 2008-2018 in developmental and educational psychology journals. From each study, we recorded the sample size of the analytic dataset, the number of classes selected, the number of indicators used to construct the profiles, the proportion of observations in the smallest profile, and the observed entropy value (Table 2.1). We also noted typical rates of missingness and how missing data were treated for each study.

### **Missing Data Rates Observed in Applied Research**

The thirty LPA studies provided limited information regarding missing data rates, when such information was provided at all. Indeed, only about one-third (11 of the 30 studies) reported missingness information. Of those reporting indicator missingness information, the reporting range of missing value rates across individual indicators was the most common reporting method (seven of 11 studies), followed by reporting the proportion of observations with complete data (three of 11 studies). A single study reported covariance coverage rates, which is only appropriate in covariance structure models.

The reporting observations with complete cases would have been ideal to inform the simulation design. However, predicting this value from either the range of missing data rates or the covariance coverage rates proved difficult. For example, one typical study included five indicators and reported missingness rates ranging from 10.7% to 25%. The percentage of observations with complete data in this study could range anywhere between 0-75%. In the simulation study, we fixed the missing data rate such that 50% of observations are missing at least one indicator value. With the chosen missing data mechanism (discussed below) this corresponds to marginal missingness rates for each variable of approximately 25%.

### **Number of Indicators and Classes Observed in Applied Research**

The modal number of profiles in the selected LPA articles was found to be  $J = 4$  indicators (Panel A in Figure 2.2). Correspondingly, four classes were the most frequently reported number of classes in the 30 selected LPA articles (Panel B in Figure 2.2). However, we chose to simulate data from a three-class model ( $K = 3$ ) to make the

simulation study time-feasible and to minimize convergence issues encountered with fitting mixture models with more components.

### **Class Separation Values Observed in Applied Research**

The 25th and 75th quantiles of the reported entropy values were .74 and .88, respectively (Panel C in Figure 2.2). To achieve entropy values that reflect these typical values, the Mahalanobis distances (MD) between the class means were manipulated so that the entropy values from the simulation roughly matched the 25th and 75th quantile values from the LPA studies (Figure 2.3). In conducting this manipulation exercise, the unequal mixing condition was chosen because that condition was highly represented in the LPA studies. We found that the corresponding MD values associated with the 25<sup>th</sup> and 75<sup>th</sup> entropy quantiles were MD = 2.84 and MD = 3.70, respectively.

### **Manipulated Conditions**

#### ***Primary Manipulated Factors***

We manipulated three primary factors in the simulations. First, we manipulated the sample size because we believe sample size influences the quality of the donor pools. This is because the donor pool represents an empirical distribution that approximates a posterior distribution. As sample size increases, the empirical distribution is expected to converge to the true posterior distribution for  $Y^{obs}$ , leading to improved imputations. Thus, we manipulated the sample size between  $N = 300$  (small sample) and  $N = 1,200$  (large sample), according to the interquartile range that is typically observed in applied research as indicated by the sample of our 30 studies.

Along the same lines, we also manipulated mixing proportions because we anticipated that the quality of the donor pool would be hindered by small class sizes. We

expected that the donor pool would be a particularly poor reflection of the true posterior predictive distribution for  $Y^{\text{mis}}$  with unequal mixing proportions and small class sizes. To test this hypothesis, we manipulated the mixing proportions between the condition with equal mixing and a condition with unequal mixing, such that the smallest class represented about 10% of the sample. This value is slightly less than would be predicted by the best fit line in Panel D of Figure 2.3, but it is certainly consistent with proportions observed in literature.

Finally, we manipulated class separation between weakly separated ( $MD = 2.86$ ; entropy  $\approx .74$ ) and highly separated ( $MD = 3.70$ ; entropy  $\approx .88$ ) values that were observed from the manipulation exercise described above. We anticipated that recursive partitioning algorithms would outperform imputation procedures that assume a single class and do not embed any clustering in this setting. However, hard clustering (like that formed by the rectangular stratum in recursive partitioning) is less appropriate when classes are weakly separated. As a result, performance was expected to decline in weakly separated settings because recursive partitioning algorithms do not account for uncertainty in classification.

### ***Secondary Manipulated Factors***

In addition to sample size, mixing proportions, and class separation, we also manipulated additional quantities to enhance the generalizability of the conclusions. This includes: (1) whether there are mean differences by classes as would occur when the auxiliary variable is a distal outcome, and (2) whether profile membership moderates the correlation between the auxiliary variable and profile indicators. Specifically, in the mean differences condition, the mean value between the reference class and two nonreference

classes will be separated by  $\delta = 0.5$  SD, while the nonreference classes will be separated by  $\delta = 1$  SD. Finally, in the moderation condition, the correlation between the AV and the profile indicators is set to  $r = .40$  for the first class,  $r = 0$  for the second class, and  $r = -.40$  for the last class.

### **Overall Design**

In total, 32 simulation conditions were manipulated. The fixed and manipulated conditions are summarized in Table 2.2. We conducted 500 replications across each condition with  $M = 100$  imputed datasets.

### **Data Generating Mechanism and Missing Data Mechanism**

We generated complete data using a three-class Gaussian mixture model as a template (Panel A of Figure 2.4). Specifically, data were generated from this template by first randomly drawing profile memberships for each observation from a multinomial distribution and then subsequently drawing values for the profile indicators and AVs from a multivariate normal distribution with a profile-specific population mean vector and a unitary variance-covariance matrix. Drawing values for the AVs jointly with the profile indicators separately across profiles allowed for mean differences across profiles, and it also allowed for different covariance specifications with the indicators across the profiles.

We identified indicator data that were set to missing in a manner which ensured that the propensity for missingness was informed by the auxiliary variable. Observations could be missing up to three indicator values, resulting in 15 possible missing data patterns total (including the pattern corresponding to no missingness). Missing data patterns were assigned using a latent response variable formulation:

$$\eta^* = AV + \epsilon$$

where  $\epsilon \sim N(0, 0.1)$ . We manipulated cut points for the  $\eta^*$  latent variable so that the marginal missingness rates across each indicator value averaged approximately 25% each, while half of the observations contained complete data.

### Analytic Model

We fit the standard latent profile model diagrammed in Panel B in Figure 2.4 to either the observed dataset of profile indicators or to the imputed datasets using Mplus version 8.0 (Muthén & Muthén, 2017). Subsequently, results were exported to R using the `MplusAutomation` package (Hallquist & Wiley, 2018). We evaluated the performance of FIML in estimating parameters from the observed data likelihood function. We also evaluated the performance of a multivariate normal imputation procedure using the `Amelia` package (Honaker, King, & Blackwell, 2011) in R. The multivariate normal imputation is denoted EMS-MVN because the expectation-maximization with sampling (EMS; King et al., 2001) imputation algorithm is employed by `Amelia`. We assessed the performance of several MICE imputation procedures, including (a) predictive mean matching (denoted MICE-PMM), and (b) MICE-CART, and (c) MICE-RF using the `mice` R package.

### Evaluative Criteria

All pooling was conducted using Rubin's (1987) rules, which assume that the posterior distribution in (1.8) is approximated by a normal distribution, i.e.

$$[\theta | Y^{\text{obs}}, X] \sim \mathcal{N}_{q_K}(\bar{\theta}, \hat{T}), \quad (1.9)$$

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_m \quad (1.10)$$

$$\hat{T} = \bar{U} + \left(1 + \frac{1}{M}\right) B \quad (1.11)$$

where  $q_K$  is the number of parameters in  $\theta$  and depends on the number of mixture components  $K$  fit to the data. The posterior distribution is centered at the mean of the maximum likelihood estimates  $\hat{\theta}_m$  obtained from fitting the mixture model to each  $m = 1, \dots, M$  imputed datasets. Finally,  $\hat{T}$  is the observed-data posterior variance; it is a composite of the within imputation variance covariance matrix ( $\bar{U}$ ) and the between imputation variance covariance matrix ( $B$ ), each given by

$$\bar{U} = \sum_{m=1}^M I^{-1}(\hat{\theta}_m | Y_m^{\text{imp}}) \quad (1.12)$$

$$B = \frac{1}{M-1} (\hat{\theta}_m - \bar{\theta})(\hat{\theta}_m - \bar{\theta})^T \quad (1.13)$$

where  $I^{-1}(\hat{\theta}_m | Y_m^{\text{imp}})$  is the observed information matrix obtained by fitting the model to the  $m$ th imputed dataset  $Y_m^{\text{imp}}$ .

### ***Recovery of $\mu_k$ and $\pi_k$ Parameters (i.e., Bias)***

We analyzed the absolute and relative bias of the class-specific mean estimates as well as the marginal class probability estimates. We defined relative bias as the ratio:

$$\text{Relative Bias} = \frac{\text{Bias}(\hat{\theta}^{\text{IMP}})}{\text{Bias}(\hat{\theta}^{\text{FIML}})} \quad (1.14)$$

where “IMP” refers to one of the missing data procedures tested in this simulation study (e.g., EMS-MVN, MICE-PMM, MICE-CART, or MICE-RF). Relative bias values near zero indicate complete attenuation of the bias, whereas relative bias values greater than 1 or less than -1 imply that the bias is greater in magnitude for the alternative procedure

relative to FIML. Finally, negative values are indicative that the direction of the bias is opposite from that observed in FIML.

***KL-Divergence (i.e., Density Estimation)***

Whereas the proceeding section focused on recovering select statistics of typical interest to applied researchers (e.g.,  $\mu_k$  and  $\pi_k$ ) to answer substantive research questions, this section studies the performance of each MAR strategy in recovering the full joint distribution of the true data generating mechanism. We used the Kullback-Leibler (KL) divergence to quantify the degree to which the distribution implied by fitting a mixture model to the observed (FIML) or imputed (EMS-MVN, MICE-CART, etc.) data matched the joint distribution in the population. In this way, we were interested in the KL-divergence to evaluate how well the implied multivariate distribution estimated from one of the missing data strategies approximated the true data generating mechanism. Indeed, one important application of finite mixture models is semiparametric applications where a joint density is approximated (e.g., density estimation). The KL-divergence is a metric for evaluating how well the density estimation is proceeding.

The KL-divergence between the model-estimated distribution and the population distribution is given by:

$$KL(\theta||\hat{\theta}) = \int_Y \Pr(Y|\theta) \log \frac{\Pr(Y|\theta)}{\Pr(Y|\hat{\theta})} dY \quad (1.15)$$

where  $\Pr(Y|\theta)$  is the density for a  $K = 3$  class Gaussian mixture model and  $\hat{\theta}$  is the estimate of the sufficient statistics from the sample after employing one of the MAR strategies. We used Monte Carlo integration with 50,000 sampled nodes to approximate the integral in (1.15).

The KL-divergence is always greater than zero and takes on the value zero if the estimated joint distribution exactly equals the true data generating distribution. However, the value of KL-divergence is not on an interpretable scale. Therefore, we mapped the KL-divergence on an interpretable scale and calculated the percent reduction in the KL-divergence relative to the KL-divergence observed from the FIML parameter estimates. This percent reduction is given by:

$$\text{Percent Reduction KL} = 100 \times \left( 1 - \frac{\text{KL}(\theta || \hat{\theta}^{\text{IMP}})}{\text{KL}(\theta || \hat{\theta}^{\text{FIML}})} \right) \% \quad (1.16)$$

where  $\hat{\theta}^{\text{FIML}}$  is the FIML estimate and  $\hat{\theta}^{\text{IMP}}$  is the estimate obtained from an imputation procedure. A near 100% reduction in the KL-divergence corresponds to estimates that approximate the true data generating mechanism extremely well and reflects a holistic attenuation of bias across all sufficient statistics estimated by fitting a mixture model to a sample. In contrast, a negative percent reduction corresponds to a set of estimates that, relative to FIML, fails to well-approximate the full joint distribution.

### ***95% CI Coverage Rates***

To assess coverage, we took the mean of the individual 95% confidence interval coverage rates across replications. We did not analyze the mean of the coverage rates of the profile variance estimates, as these quantities are generally not of substantive interest to applied researchers. Following Muthén & Muthén (2002), we considered coverage rates falling between .91 to .98 as acceptable because they neither unduly risked invalid inference through narrow intervals nor were they too conservative to be considered needlessly inefficient.

### *Imputation Efficiency*

Finally, we analyzed the increase in variance attributable to the missingness. A typical statistic for this quantity is the average relative increase in variance (ARIV), which can be roughly thought of as the mean increase in the standard errors squared due to between-imputation uncertainty. The ARIV is calculated as

$$\text{ARIV} = \frac{\text{tr}(\bar{U}^{-1}\hat{T})}{M} \quad (1.17)$$

which mathematically translates to the mean of the proportional increase in the eigenvalues observed in the within- vs. total- imputation asymptotic covariance matrices. It is important to note that the ARIV results in an unbiased estimate of the (average) increase in variance attributable to missingness only if  $\bar{U}$  is an unbiased estimate of the complete-data asymptotic variance-covariance matrix. Bayesianly proper imputations are sufficient to guarantee that this condition holds (Schafer, 1997), but in simulation studies, the complete-data asymptotic variance-covariance matrix can be obtained directly. Thus, we calculated the average relative increase in variance from this matrix directly using the formula:

$$\text{ARIV} = \frac{\text{tr}(I(\hat{\theta}|Y)\hat{T})}{M} \quad (1.18)$$

where  $I(\hat{\theta}|Y)$  is the observed information matrix using the complete data, and  $\hat{T}$  is the covariance matrix of the posterior distribution  $[\theta|Y^{\text{obs}}, X]$  approximated by Rubin's (1987) rules during the pooling phase.

We note that imputation efficiency is not to be confused with sampling variability. Imputation efficiency relates to the variability due to sampling different

imputations (e.g., between-imputation variability). The sampling variability refers to both the within- and between- imputation variability.

## Results

We now present the results according to our evaluative criteria: bias, density estimation, 95% CI coverage, and imputation efficiency.

### Bias and Density Estimation

#### *Large Sample ( $N = 1,200$ ) Recovery of $\mu_k$ and $\pi_k$*

Figures 2.5 and 2.6 display the change in absolute bias in the class indicators across latent classes when implementing FIML (start of arrow) as opposed to one of the alternative MAR strategies (end of arrow) for the large sample size condition. In recovering class means, MICE-CART and MICE-RF generally outperformed FIML, EMS-MVN, and MICE-PMM in large sample, equal mixing conditions. This was true in both weakly and strongly separated classes, as exhibited by the end of the arrow terminating near the line indicating nominal bias in Figure 2.6.

However, when classes were unequally mixed, imputing using recursive partitioning imputation algorithms did not always exhibit minimal bias. Specifically, although both MICE-CART and MICE-RF resulted in decreased bias when classes were strongly separated, these algorithms were associated with an increase in bias for the small class (i.e., class  $k = 3$ ) in the weakly separated condition (see Figure 2.5).

The relative bias estimates (Tables 2.3 and 2.4) coincided with the conclusions drawn from analyzing change in absolute bias (Figures 2.5 and 2.6). In terms of relative bias for the profile indicator means and class proportions, MICE-CART generally outperformed MICE-RF in large samples with equal mixing and strong separation.

However, performance between these two methods was roughly equivalent when classes were weakly separated. Additionally, both recursive partitioning methods outperformed EMS-MVN and MICE-PMM methods in the large sample, equal missing condition, regardless of the class separation condition. Similarly, these methods most often outperformed FIML as indicated by values between -1 and 1; exceptions represented the special circumstance when the magnitude of the bias exhibited by FIML was extremely small.

In summary, in large samples, recursive partitioning algorithms outperformed FIML and alternative imputation algorithms in recovering class means and proportions. The situation where there was a small class and where the classes were weakly separated represented a notable exception to this rule. Although MICE-CART mitigated nonresponse bias in large samples if class separation was strong (MD = 3.70; entropy values averaging near .88), when class separation was weak (MD = 2.84; entropy values averaging near .74), MICE-CART only mitigated nonresponse bias across all mean and marginal class probability parameters when classes were equally mixed.

#### ***Small Sample (N = 300) Recovery of $\mu_k$ and $\pi_k$***

Figures 2.7 and 2.8 display the change in absolute bias when implementing FIML compared to the alternative missing data procedures in the small sample condition. Similar to that found in large samples, we found that the recursive partitioning algorithms (and especially MICE-CART) most frequently outperformed FIML. Moreover, the degree to which recursive partitioning algorithms attenuated bias appeared more sensitive to class separation in the small sample condition than in the large sample condition when classes were equally mixed. Relative to other imputation methods such as EMS-MVN

and MICE-PMM, recursive partitioning methods demonstrated equally poor or even worse performance in recovering class means and proportions for the small class in unequal mixing. This was especially true under weak separation conditions.

***Large Sample ( $N = 1,200$ ) Percent Reduction in the KL-Divergence***

We found that MICE-CART consistently resulted in parameter estimates that minimized the KL-divergence between the parameter estimates and the population distribution. This is illustrated in Figure 2.9, which plots the median and interquartile range of the percent reduction in the KL-divergence, marginalized across all manipulated conditions except sample size. When analyzed separately by mixing and separation conditions (Table 2.5), the KL-divergence relative to FIML was quite consistent across the manipulated conditions, ranging between 71.3% to 73.9%. Nevertheless, the difference in performance between MICE-CART and MICE-RF was substantively small compared to the difference in performance between recursive partition imputation algorithms and MAR procedures that did not implement recursive partitioning. Specifically, while the percent reduction in the KL-divergence ranged between 66.38-72.76% for MICE-RF, this range was observed to be -8.95-57.50% and 10.21-59.80% for EMS-MVN and MICE-PMM, respectively. In summary, recursive partitioning imputation algorithms outperformed alternative MAR procedures in recovering the joint distribution of the complete data when sample sizes were large.

***Small Sample ( $N = 300$ ) Percent Reduction in the KL-Divergence***

The strong performance of recursive partitioning imputation algorithms to replicate the true data generating distribution observed in large samples ( $N = 1,200$ ) did not transfer to small samples ( $N = 300$ ) (see Figure 2.10). As reflected in Table 2.5,

recursive partitioning algorithms performed as poorly and often worse than alternative imputation algorithms in reducing the KL-divergence (MICE-CART: Range = 36.02-50.91%; MICE-RF: Range = 48.11-56.30%; EMS-MVN = 36.02-79.90%; MICE-PMM = 33.87-67.39%). Moreover, whereas MICE-CART uniformly outperformed MICE-RF in large samples, the reverse was true in small samples. We discuss causes and implications of these results in the Discussion section.

### **95% CI Coverage**

Figure 2.10 illustrates the average 95% CI coverage rates across profile indicators by sample size, marginalized across all other simulation conditions. In general, MICE-CART and MICE-RF outperformed FIML in estimating the 95% CIs that demonstrate acceptable coverage.

Table 2.5 reports the 95% CI coverage rates averaged across (a) class means and (b) marginal class probabilities by sample size, mixing, and separation condition. In large samples, we found that the MICE-CART procedure resulted in acceptable coverage rates across all mixing and separation conditions. In contrast, MICE-RF resulted in too narrow of confidence intervals consistently across class means (Range = .86-.92) and frequently across class proportions (Range = .87-.96). All other missing data procedures, and especially FIML (Range = .54-.88), resulted in unacceptably poor coverage across the conditions and in large samples.

### **Imputation Efficiency**

In large samples, we found that the ARIV differed modestly across the imputation procedures and ranged between 0.60-0.99. Thus, compared to the standard errors of the complete data, the standard errors estimated after pooling were 36-98% wider. MICE-

CART is the least efficient of the imputation strategies, resulting in the largest increase in variance (ARIV: Range = 0.63-0.99), whereas MICE-RF resulted in increases in variance that were competitive with the most efficient procedure, MICE-PMM.

In small samples, we found that imputation results in extremely large increases in variance across procedures when classes were weakly separated and a small class was present, with ARIV ranging between 5.68-6.71. Increases in variance were relatively more modest in the unequal mixing and strongly separated condition (Range = 1.30-1.61) and in the equal mixing and weakly separated conditions (Range = 1.09-1.75). In the equal mixing and strongly separated condition, ARIV values mirrored those observed in the large sample conditions (Range = 0.71-0.88).

As with what was observed in large samples, MICE-CART resulted in the least efficient imputations of the imputation procedures (Range = 0.71-6.71), while MICE-RF was relatively more efficient than MICE-CART, in general (Range = 0.78-6.42), but less efficient than EMS-MVN (Range = 0.88-6.44). Finally, MICE-PMM was frequently found to result in the smallest increase in variance, even in small samples (Range = 0.74-5.68).

## **Discussion**

This simulation study suggests that when the data are MAR conditional on AVs, recursive partitioning imputation models (and MICE-CART, in particular) can outperform FIML and single-class imputation alternatives in terms of mitigating nonresponse bias and recovering the joint distribution of the data generating mechanism. In particular, the simulation study suggests that MICE-CART imputation uniformly performs well in large samples ( $N = 1,200$ ) when class separation is strong (entropies

average near .88). When class separation is weak (entropy  $\approx .74$ ), the strong performance in large samples is sensitive to whether or not the classes are equally mixed; if a small class is present (one that represents 10% of the population) even when sample sizes are large, then MICE-CART only uniformly mitigates nonresponse bias when the classes are strongly separated.

In small samples ( $N = 300$ ), recursive partitioning imputation methods only demonstrated superior performance when classes were strongly separated and equally mixed. Because it is rather unlikely that applied researchers experience data conditions where classes are both equally mixed and strongly separated in practice, we conclude that neither MICE-CART nor MICE-RF demonstrated adequate performance in small sample settings in treating the missing data.

Why do the recursive partitioning methods perform poorly in small samples or when there exists a small class and the classes are not strongly separated? This finding can be explained by the degradation of the donor pool used to sample the imputations in these settings. A high-quality donor pool is one whose empirical distribution well approximates the posterior predictive distribution of the missing values,  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$ , and is effective at matching missing values with observations from the same latent class. As implemented in the `mice` R package, MICE-CART and MICE-RF create decision trees with stopping criteria determined by a minimum node size of five observations. One would expect that for a fixed node size, a decision tree would include more cut points in large samples than in small samples. Thus, in large samples, observations are more likely to be stratified into clusters so that the donor pools are more homogenous with respect to the joint distribution of the profile indicators and auxiliary

variables and, therefore, are clustered together into the same latent class. The improved homogeneity implies that observations with missing values are more finely matched to observations that are similar with respect to the joint distribution of the profile indicators and AVs. In contrast, decision trees constructed with small observations with a fixed terminating criterion lead to fewer cut points, decreased stratification, and clusters that are less homogenous. The matching process is, therefore, coarser, resulting in an empirical distribution that does not approximate the true posterior predictive distribution of the missing values,  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$ , well.

As demonstrated by the simulations, high-quality donor pools for observations in a small class (one that represents 10% of the population) are difficult to construct, even in large samples, if the classes are weakly separated so that entropy values average near .74. This is not surprising, given that the number of observations belonging to the small class is not large. Compounded by greater overlap between the classes induced by weak separation, it is expected that the quality of donor pools for small classes will degrade.

Future studies should investigate how to improve the quality of the donor pools, especially in the settings where MICE-CART and MICE-RF fail to uniformly perform well (i.e., small sample settings or large sample settings where classes are weakly separated [entropy  $\approx .78$ ]). Implementing common pruning techniques to improve predictions and to increase the size of the donor pool may seem attractive, but pruning is likely counterproductive in attenuating nonresponse bias. Again, the goal of the donor pool is to construct an empirical distribution that well approximates the posterior predictive distribution of the missing values with respect to all moments of the  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$  distribution. Pruning may improve prediction by better estimating the

first moment of  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$ . However, mixture models model higher order moments. Combined with the fact that pruning leads to coarser matches due to fewer cut points, decreased stratification, and less homogenous clusters, one would expect that any gain in approximating the first moment of  $[Y^{\text{mis}}|Y^{\text{obs}}, X, \theta]$  is offset by declines across higher order moments. In fact, previous studies have shown that the superior performance of pruning in prediction contexts does not necessarily translate to superior performance in treating missing data (Hayes & McArdle, 2017b; Hayes, Usami, Jacobucci, & McArdle, 2015).

Improving the quality of the donor pool in LPA likely requires researchers to rely on parametric assumptions to sample from a known distribution rather than an empirical distribution (as is done in forming donor pools). There are several avenues by which this can proceed. First, covariance structure models can be imbedded within a recursive partitioning framework so that, at each terminal node, a covariance and mean structure are estimated. This “SEM Trees” (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013) approach is currently an area of active research with promising results.

Alternatively, regression mixture models can be modeled directly to generate imputations. This is also an area of active research. A Bayesian nonparametric imputation model utilizing Dirichlet processes to model class membership in order to fit an “infinite” mixture model has recently received attention in the machine learning community (Sovilj et al., 2016). Alternatively, EM with sampling imputation algorithms with finite mixtures have also been developed and utilized for imputation in the behavioral sciences (Vidotto, Vermunt, & Kaptein, 2015). A critical challenge with a finite mixture approach, however, is that the number of subpopulations is not known a priori, and the number of classes

supported by the data may be sensitive to nonresponse bias. Regardless of the mixture modeling approach, it is necessary to make these algorithms flexible enough to incorporate many AVs which themselves contain missing data to reflect the multiple data types that appear in the real world. In the next chapter, we perform an initial study to investigate whether Bayesian model averaging can effectively address model uncertainty.

Finally, we acknowledge that this study only focuses on the challenges associated with the imputation phase and leaves the pooling phase unaddressed. Before multiple imputation can become mainstream or be recommended over FIML, appropriate strategies for pooling model fit information must be identified. We evaluate appropriate pooling procedures for information criteria in Chapter 4.

In summary, we have shown that multiple imputation can outperform FIML in the more realistic situation where the MAR assumption is only tenable with the inclusion of AVs. As expected, FIML results in biased parameter estimates in real-world data conditions. Thus, applied researchers should test the MAR assumption before conducting an LPA. This can be done by evaluating if variables available to the researcher in the dataset explain missing data patterns above and beyond that explained by observed profile indicator values. If predictive variables are found, researchers employing FIML to treat missing data with sample sizes above  $N = 1,200$  should conduct a sensitivity analysis using MICE-CART imputation. Such a sensitivity analysis can provide valuable evidence regarding whether conclusions about class definitions and inferences regarding class proportion are sensitive to violations of the MAR assumption.

## Tables

**Table 2.1**

*Descriptive Statistics of Metadata in Reviewed Articles*

	<b># of Studies</b>	<b>M</b>	<b>SD</b>	<b>Min</b>	<b>0.25 Quantile</b>	<b>Median</b>	<b>0.75 Quantile</b>	<b>Max</b>
Sample size, N	30	932.5	886.8	137	292	640	1167	4417
# of indicators, J	30	5.2	2.44	2	4	5	6	12
# of components, K	30	3.9	1.37	2	3	4	4	8
Entropy	27	0.82	0.09	0.63	0.74	0.85	0.88	0.98
Min. class prop., $\pi_{min}$	29	0.11	0.1	0.02	0.06	0.09	0.13	0.47

**Table 2.2***Data Generating and Manipulated Simulation Conditions (500**Replications/Condition)*

	Value	# of Conditions	Notes
<b><u>Fixed Simulation Conditions</u></b>			
# of indicators, J	4	1	Represents the modal number in the literature review.
# of classes, K	3	1	Although $K=4$ represented the modal value in the literature review, $K = 3$ was the second most common and chosen because it was the second most common value and greatly reduced computational burden.
Imputed datasets, M	100	1	
Obs. with complete data	50%	1	Corresponds to a missing data rate of approximately 25% across each indicator. This value is the upper limit of those found in the literature review.
Correlation, r	0.40	1	
<b><u>Primary Manipulated Conditions</u></b>			
Sample size	Small sample vs. Large sample	2	Small sample size ( $N = 300$ ) and large sample size ( $N = 1,200$ ) correspond to the 25 <sup>th</sup> and 75 <sup>th</sup> percentile of sample sizes found in the literature review.
Mixing proportions	Not Equal vs. Equal	2	Not equal corresponds to $\pi = [0.45, 0.45, 0.10]$ . $\pi_3 = 0.1$ was identified as typical given a $K = 3$ class model in the literature review.
Class separation	Weak vs. Strong	2	Weak separation corresponds to entropy value of approximately .74 in not equal mixing proportions condition. Strong correlation corresponds to an entropy value of approximately .88. These values correspond to the 25 <sup>th</sup> and 75 <sup>th</sup> percentile of entropies found in the literature review.
<b><u>Secondary Manipulated Conditions</u></b>			
Mean differences in AVs by class	No Mean Differences vs. Mean Differences	2	In the mean difference setting, the mean value between the reference class and the two nonreference classes is 0.5 SD. The nonreference classes is separated by 1 SD.
Profile membership moderates relationship between AVs and indicators	No Moderation vs. Moderation	2	In moderation condition the correlations between the AVs and indicators are $r = .4$ , $r = 0$ , and $r = -.4$ , respectively across the $K = 3$ classes.
<b>Total Simulation Conditions</b>		<b>32</b>	

**Table 2.3**

*Large Sample ( $N = 1,200$ ) Relative Bias*

	Unequal Mixing						Equal Mixing					
	Weakly Separated			Strongly Separated			Weakly Separated			Strongly Separated		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
<b>Means (<math>\mu_k</math>), <math>Y_1</math></b>												
Complete	0.11	-0.02	-0.01	0.03	-0.05	-0.05	-0.05	-0.06	0.01	0.22	0.00	0.01
MVN	0.15	0.60	0.65	-1.52	0.17	0.40	-2.81	0.32	0.90	-8.23	0.02	0.78
PMM	0.50	0.41	0.12	-0.47	0.39	-0.09	-2.16	0.34	0.71	-5.93	0.44	0.60
CART	0.57	-0.04	0.03	-0.35	-0.08	0.17	-0.92	-0.10	0.28	-2.50	-0.03	0.29
RF	-0.03	-0.38	0.04	-1.70	-0.73	0.05	-2.86	-0.40	0.35	-7.17	-0.63	0.29
<b>Means (<math>\mu_k</math>), <math>Y_2</math></b>												
Complete	0.13	0.10	0.01	-0.16	-0.03	0.00	-0.09	-0.02	0.02	0.10	-0.01	0.03
MVN	-1.70	-0.37	1.26	-9.76	-0.89	1.13	-4.17	-1.30	0.76	11.93	-1.65	0.68
PMM	0.26	0.22	1.21	-3.30	0.12	0.92	-3.93	-1.06	0.60	13.03	-1.14	0.35
CART	0.69	0.38	0.49	-0.60	0.12	0.39	-0.07	-0.11	0.33	-1.46	0.00	0.28
RF	0.36	0.31	0.81	-1.59	0.04	0.81	-0.55	-0.32	0.62	-2.92	-0.27	0.57
<b>Means (<math>\mu_k</math>), <math>Y_3</math></b>												
Complete	0.06	-0.05	0.04	0.04	0.01	-0.04	0.00	-0.01	0.05	0.16	0.00	-0.04
MVN	1.87	0.40	0.71	1.55	0.47	0.56	1.29	0.05	0.81	0.93	0.28	0.71
PMM	1.81	0.38	0.34	1.58	0.41	0.31	1.49	0.10	0.76	1.25	0.30	0.68
CART	1.16	0.12	0.14	0.44	0.20	0.05	0.65	0.26	0.25	0.20	0.31	0.08
RF	1.55	0.37	0.33	0.86	0.40	0.19	0.95	0.57	0.43	0.43	0.58	0.25
<b>Means (<math>\mu_k</math>), <math>Y_4</math></b>												
Complete	0.07	0.00	-0.36	0.05	-0.04	-0.03	0.07	0.02	0.01	0.46	-0.04	0.01
MVN	-0.10	0.70	-1.20	-1.56	0.78	-0.65	-10.54	0.83	0.46	51.29	1.05	0.44
PMM	1.49	0.28	-0.30	0.93	0.27	0.40	-4.91	0.54	0.92	21.06	0.72	1.02
CART	1.12	0.20	-1.50	0.19	0.20	0.06	-1.39	0.29	0.30	8.42	0.30	0.41
RF	1.48	0.46	-1.56	0.58	0.43	-0.23	-2.34	0.61	0.41	13.10	0.60	0.46
<b>Marginal Class Probabilities (<math>\pi_k</math>)</b>												
Complete	0.11	-0.12	2.73	0.02	-0.01	-0.11	0.05	0.04	0.04	0.12	0.00	0.01
MVN	-0.57	-0.58	-0.46	-1.15	-0.43	1.62	-31.72	-0.06	2.10	32.07	-0.11	2.28
PMM	-0.23	-0.57	3.67	-0.90	-0.51	0.57	-27.05	-0.03	1.81	28.29	-0.23	1.89
CART	0.36	-0.39	9.01	-0.27	-0.11	0.32	-8.12	0.07	0.63	10.32	-0.02	0.75
RF	0.38	-0.32	8.53	-0.14	-0.11	-0.03	-11.07	0.18	0.95	12.80	0.03	0.98

**Table 2.4**

*Small Sample ( $N = 300$ ) Relative Bias*

	Unequal Mixing						Equal Mixing					
	<u>Weakly Separated</u>			<u>Strongly Separated</u>			<u>Weakly Separated</u>			<u>Strongly Separated</u>		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
<b>Means (<math>\mu_k</math>), <math>Y_1</math></b>												
Complete	0.52	0.25	-0.27	0.31	-0.01	-0.06	0.91	0.36	-0.08	-1.79	0.19	-0.01
MVN	0.98	0.67	-0.79	-0.15	0.29	0.07	-0.20	0.57	0.79	-23.74	0.24	0.72
PMM	1.16	0.22	-1.62	0.31	0.30	-0.37	0.39	0.28	0.61	-17.60	0.44	0.57
CART	1.69	-0.41	-3.07	0.86	-0.13	-0.30	1.52	-0.41	0.26	-7.03	0.01	0.39
RF	0.85	-0.68	-2.36	-0.90	-1.16	-0.50	0.18	-0.64	0.31	-28.95	-0.58	0.29
<b>Means (<math>\mu_k</math>), <math>Y_2</math></b>												
Complete	0.24	0.42	0.07	0.17	0.12	0.00	-1.70	0.14	-0.09	0.13	0.03	0.00
MVN	-0.82	-0.01	0.89	-3.77	-0.40	1.05	-12.67	-0.52	0.72	-7.37	-1.49	0.66
PMM	0.26	0.20	0.74	-0.84	0.28	0.83	-12.13	-0.69	0.60	-7.95	-1.17	0.40
CART	1.87	-0.37	0.30	1.58	0.17	0.59	1.86	-0.47	0.42	-0.45	-0.24	0.41
RF	0.73	-0.01	0.70	-0.19	0.34	1.00	-2.64	-0.47	0.77	-2.23	-0.58	0.76
<b>Means (<math>\mu_k</math>), <math>Y_3</math></b>												
Complete	0.23	-0.89	-0.15	-0.03	0.06	-0.06	0.40	-0.02	0.02	-0.06	-0.11	-0.03
MVN	1.51	-0.63	0.17	1.76	0.39	0.33	1.11	0.00	0.80	0.96	0.22	0.71
PMM	1.55	-0.49	-0.24	1.62	0.40	0.08	1.16	0.10	0.69	1.18	0.36	0.65
CART	2.24	-1.23	-0.60	1.61	0.29	-0.06	1.40	0.08	0.37	0.80	0.51	0.24
RF	1.99	-0.57	-0.36	1.56	0.49	-0.07	1.19	0.62	0.51	0.47	0.85	0.35
<b>Means (<math>\mu_k</math>), <math>Y_4</math></b>												
Complete	0.42	-0.28	1.02	-0.01	-0.02	-2.96	6.01	-0.25	-0.25	0.38	-0.13	-0.10
MVN	1.01	0.59	2.34	-1.74	0.75	-12.69	-47.33	0.80	0.21	11.71	1.13	0.29
PMM	1.73	0.27	2.22	1.07	0.36	-6.51	-25.40	0.56	0.63	5.79	0.81	0.95
CART	2.70	0.18	3.03	1.96	0.36	-9.88	-1.30	0.44	-0.09	2.62	0.54	0.42
RF	2.27	0.52	2.84	1.66	0.53	-13.33	-14.41	0.79	0.11	4.52	0.86	0.43
	0.42	-0.28	1.02	-0.01	-0.02	-2.96	6.01	-0.25	-0.25	0.38	-0.13	-0.10
<b>Marginal Class Probabilities (<math>\pi_k</math>)</b>												
Complete	0.45	1.43	0.69	0.19	-0.21	1.13	2.68	-0.23	-0.89	0.05	-0.13	-0.10
MVN	0.59	2.78	1.13	-0.31	-1.19	1.82	-12.78	-0.44	2.38	11.39	-0.31	2.19
PMM	0.97	2.41	1.32	-0.04	-1.07	2.42	-7.32	0.00	1.68	9.81	-0.29	1.87
CART	1.88	1.99	1.91	0.88	-0.43	4.01	5.03	0.66	-0.34	3.74	0.07	0.85
RF	1.53	2.34	1.72	0.80	-0.86	4.76	1.97	0.69	0.40	4.79	0.12	1.12

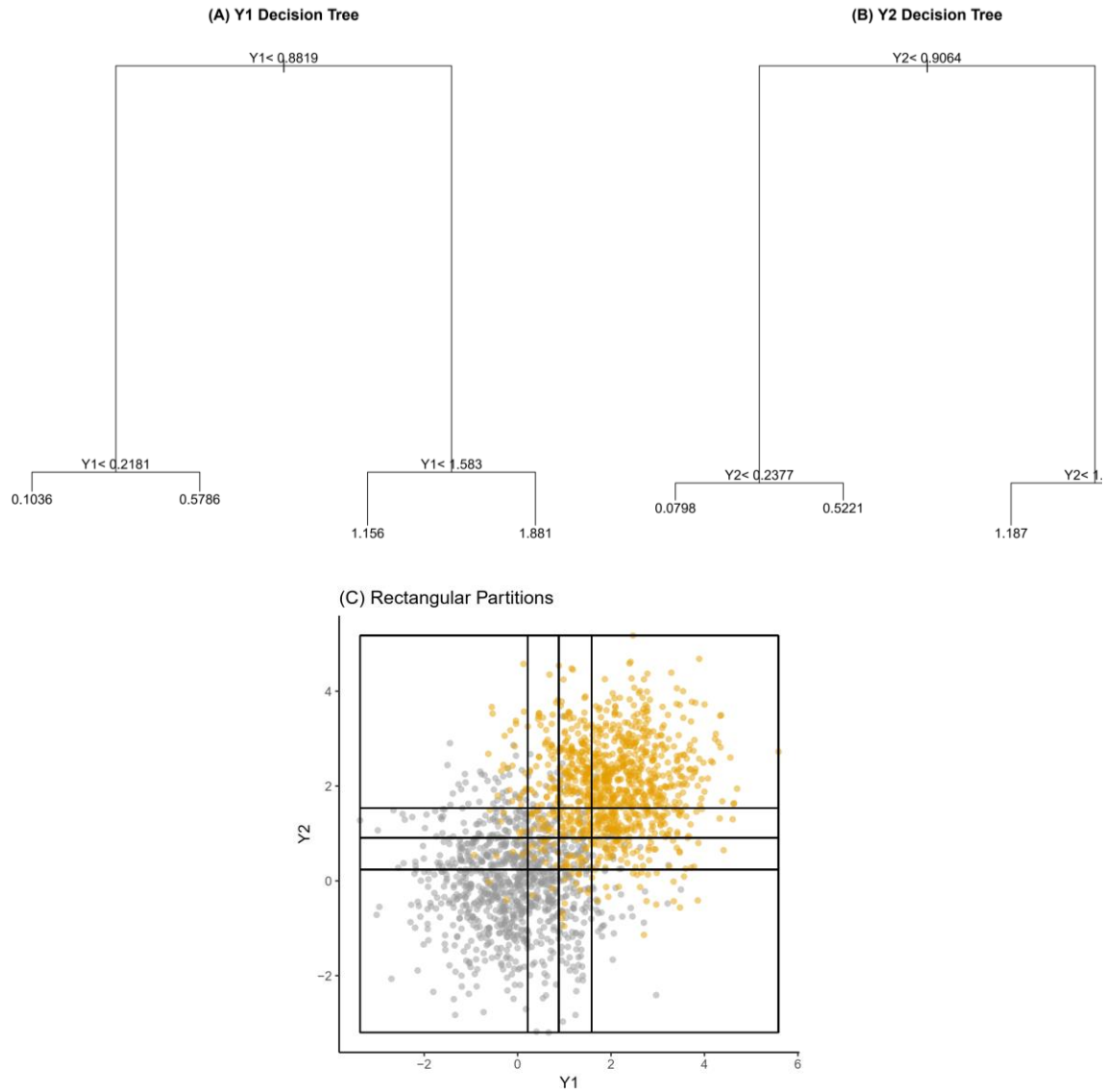
**Table 2.5***Recovery of the Joint Distribution, Imputation Efficiency, and 95% CI**Coverage*

	Small Sample (N = 300)				Large Sample (N = 1,200)			
	% Reduc. KL Div.	ARIV	Avg. $\mu_k$ Coverage	Avg. $\pi_k$ Coverage	% Reduc. KL Div.	ARIV	Avg. $\mu_k$ Coverage	Avg. $\pi_k$ Coverage
<b>Unequal Mixing &amp; Weakly Separated</b>								
Complete	100	-	.87	.78	100	-	.94	.94
FIML	0	-	.76	.70	0	-	.68	.85
MVN	79.90	6.44	.96	.98	57.50	0.65	.88	.90
PMM	67.39	5.68	.95	.96	59.79	0.64	.85	.92
CART	36.02	6.71	.95	.93	73.34	0.99	.94	.96
RF	51.89	6.42	.96	.94	72.76	0.69	.92	.96
<b>Unequal Mixing &amp; Strongly Separated</b>								
Complete	100	-	.93	.96	100	-	.95	.99
FIML	0	-	.82	.92	0	-	.61	.88
MVN	57.0	1.4	.95	.99	14.3	0.7	.83	.74
PMM	57.42	1.30	.93	.99	37.18	0.64	.80	.85
CART	50.91	1.60	.94	.98	73.63	0.73	.93	.96
RF	56.30	1.54	.95	.98	68.28	0.78	.90	.95
<b>Equal Mixing &amp; Weakly Separated</b>								
Complete	100	-	.92	.93	100	-	.95	.99
FIML	0	-	.82	.87	0	-	.64	.82
MVN	74.67	1.09	.96	.98	42.04	0.60	.82	.81
PMM	60.68	1.20	.95	.98	40.74	0.54	.81	.85
CART	44.73	1.75	.95	.98	71.27	0.63	.93	.95
RF	57.47	1.47	.95	.98	68.89	0.52	.89	.91
<b>Equal Mixing &amp; Strongly Separated</b>								
Complete	100	-	.94	1	100	-	.94	1
FIML	0	-	.82	.95	0	-	.54	.78
MVN	36.02	0.88	.93	.93	-8.95	0.77	.74	.70
PMM	33.87	0.74	.91	.93	10.21	0.60	.72	.75
CART	46.09	0.71	.93	.97	73.93	0.63	.91	.92
RF	48.11	0.78	.93	.97	66.38	0.59	.86	.87

## Figures

**Figure 2.1**

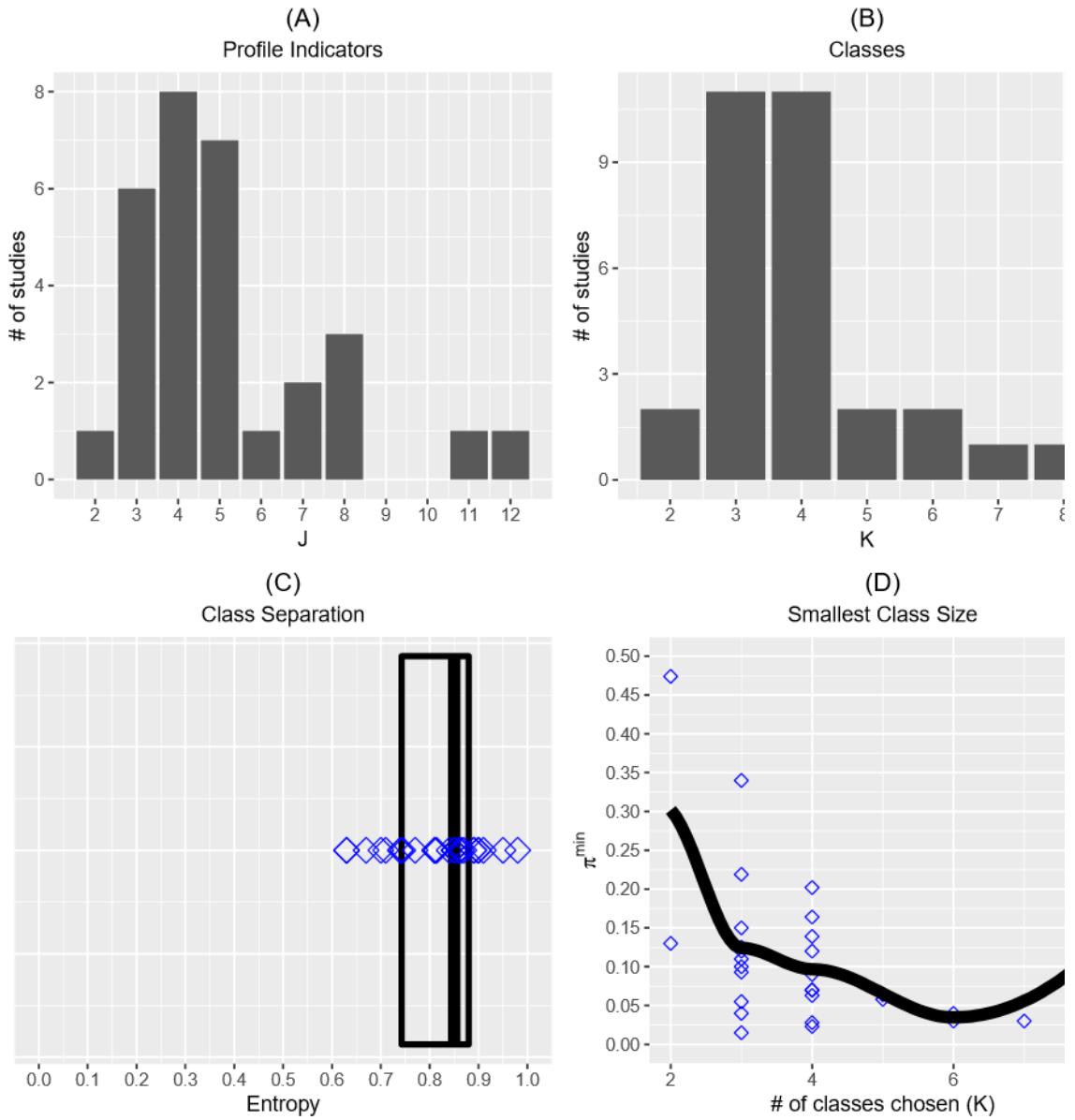
*Recursive Partitioning Decision Trees and Rectangular Partitions*



*Notes.* Illustration showing decision tree (Panels A & B) and corresponding rectangular partitions (Panel C). Colored points represent observations drawn from a  $K=2$  class mixture model.

**Figure 2.2**

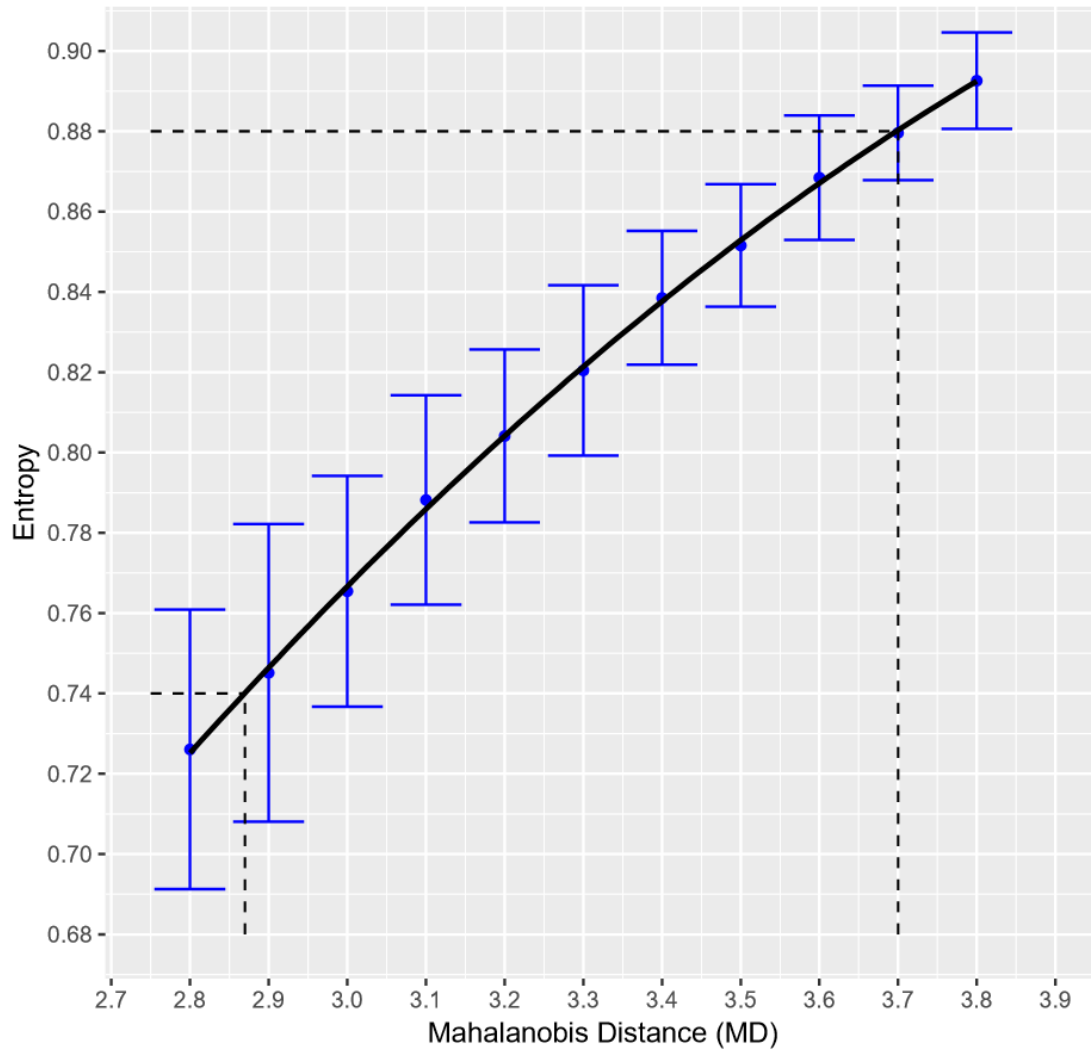
*Metadata Collected on Applied Studies which Informed the Simulations*



*Notes.* Univariate information and bivariate relationships among the metadata collected from 30 frequently cited studies that employed LPA.

**Figure 2.3**

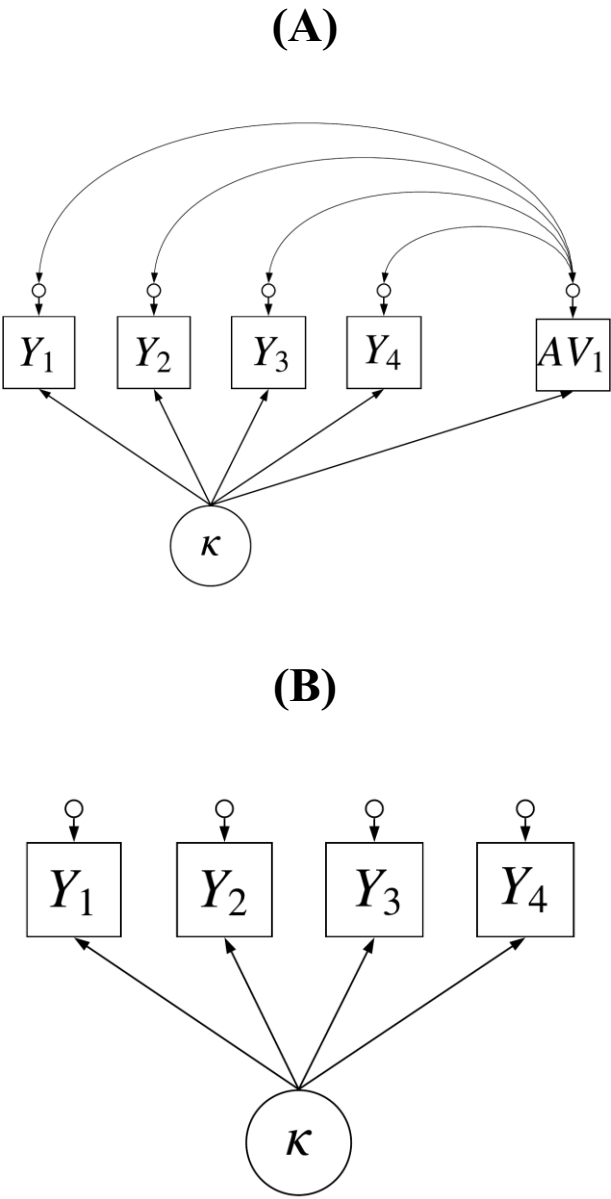
*Simulated Entropy Values*



*Notes.* Observed entropy values at various class separation values. Each separation value was replicated 500 times. Dashed lines represent the predicted MD required for the entropy value to correspond the 25<sup>th</sup> and 75<sup>th</sup> percentile recorded in the literature review, on average.

**Figure 2.4**

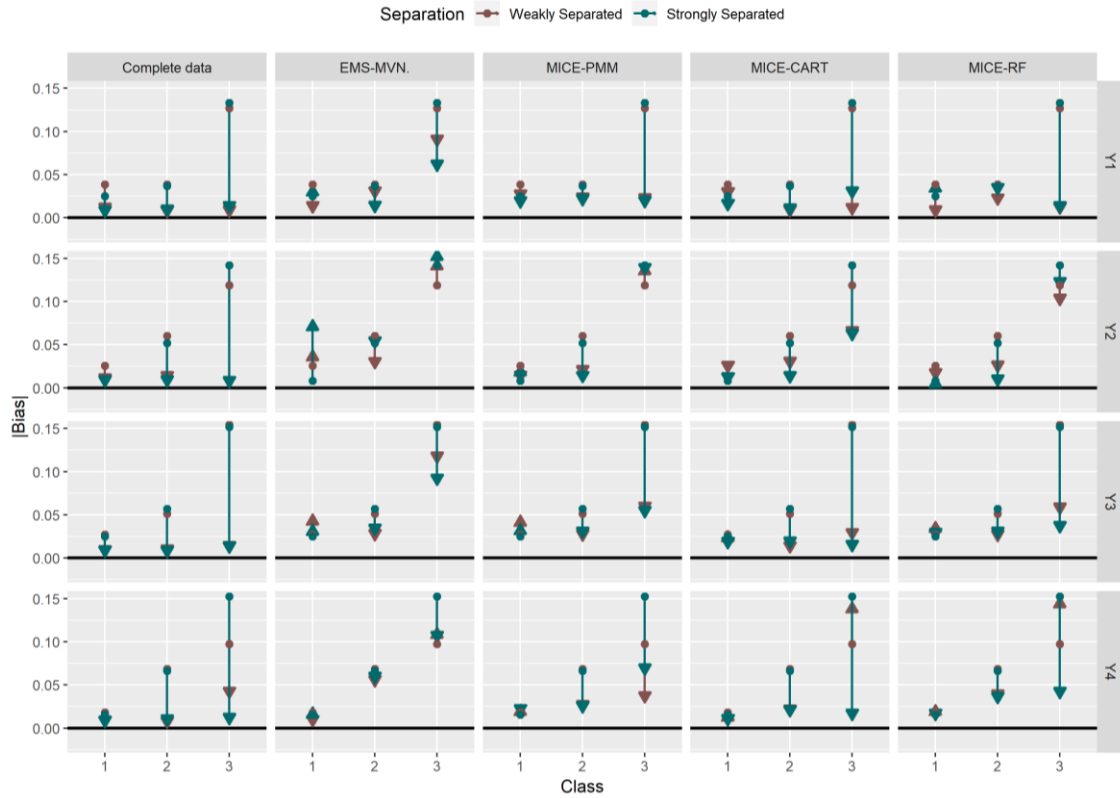
*Template and Analytic LPA Models*



*Notes.* Latent profile models reflecting (A) the data generating mechanism for simulating complete data, and (B) the analytic models fit to either the observed or the imputed data.

**Figure 2.5**

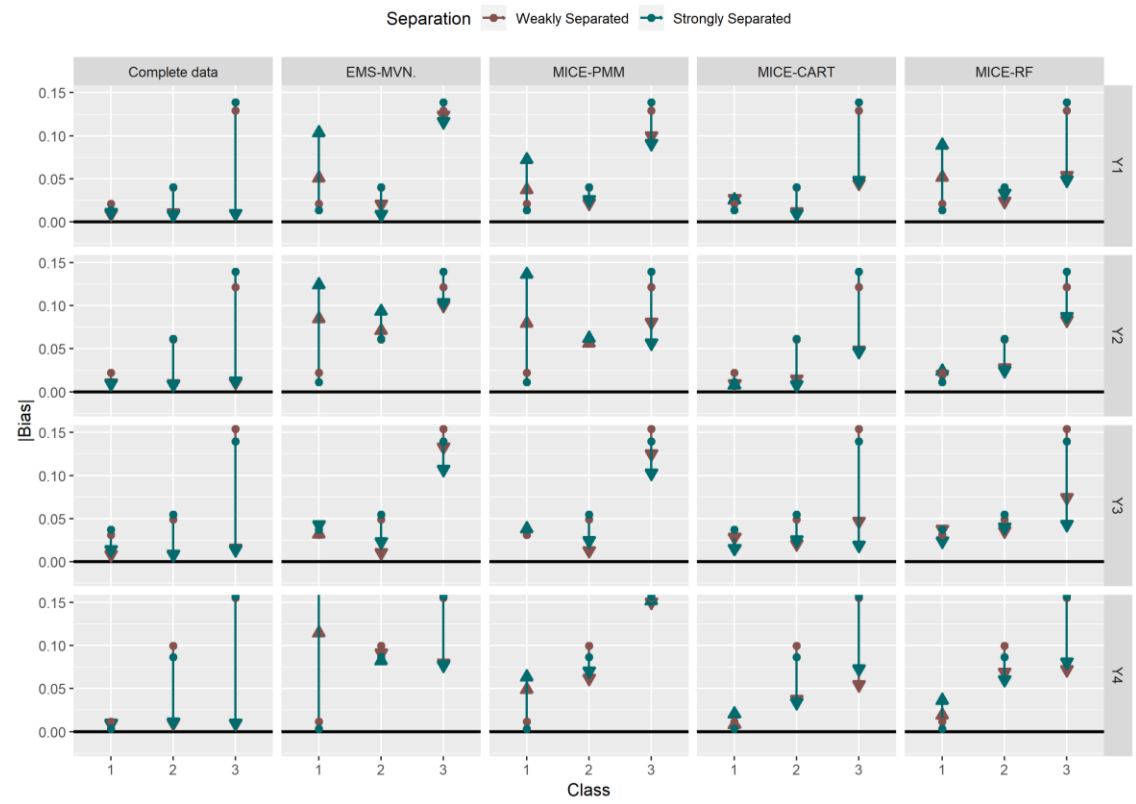
*Absolute Bias Reduction in Large Sample ( $N = 1,200$ ) and Unequal Mixing Conditions*



*Notes.* Change in absolute bias relative to FIML in the large sample and unequal mixing conditions, displayed separately by weakly separated (crimson) and strongly separated (teal) conditions.

**Figure 2.6**

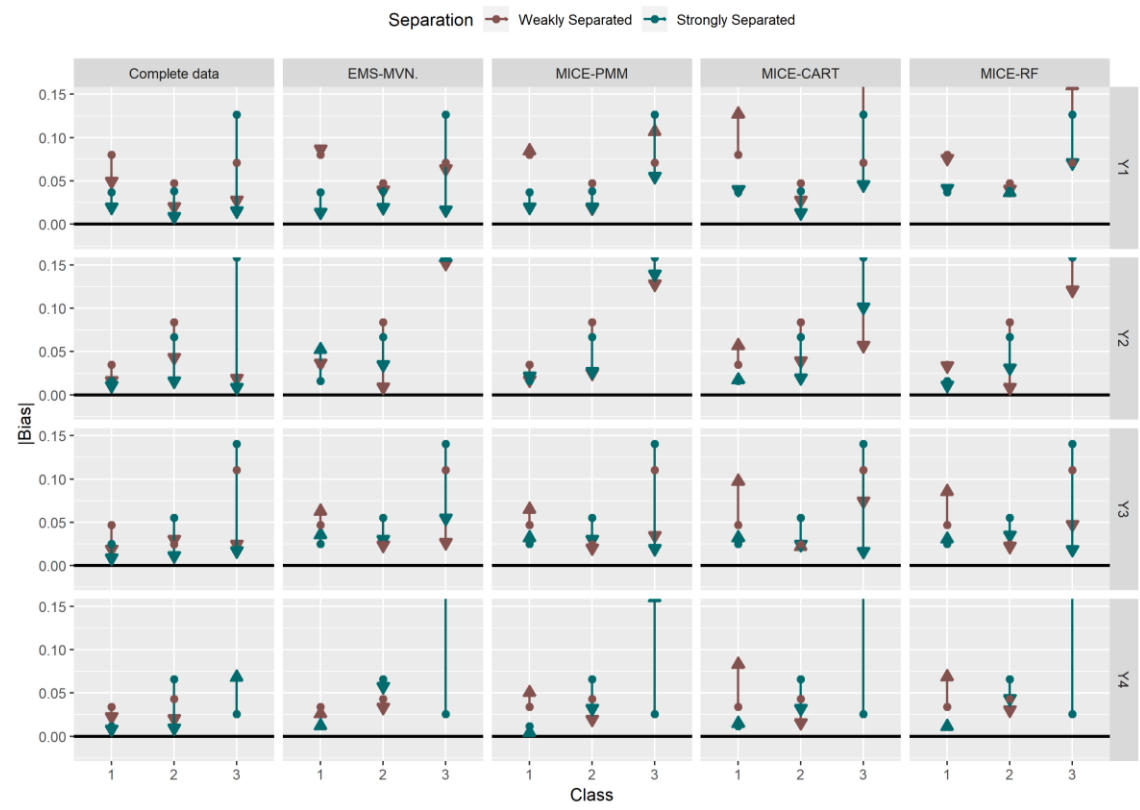
*Absolute Bias Reduction in Large Sample ( $N = 1,200$ ) and Equal Mixing Conditions*



*Notes.* Change in absolute bias relative to FIML in the large sample and equal mixing conditions, displayed separately by weakly separated (crimson) and strongly separated (teal) conditions.

**Figure 2.7**

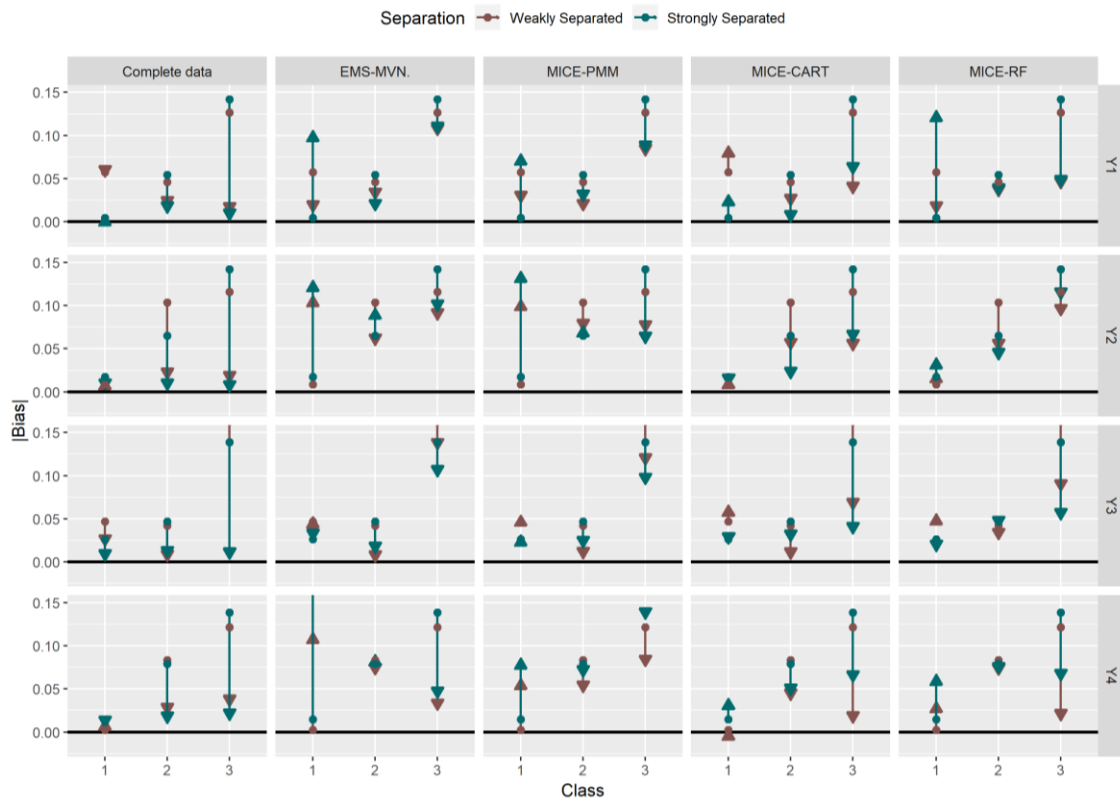
*Absolute Bias Reduction in Small Sample ( $N = 300$ ) and Unequal Mixing Conditions*



*Notes.* Change in absolute bias relative to FIML in the small sample and unequal mixing conditions, displayed separately by weakly separated (crimson) and strongly separated (teal) conditions.

**Figure 2.8**

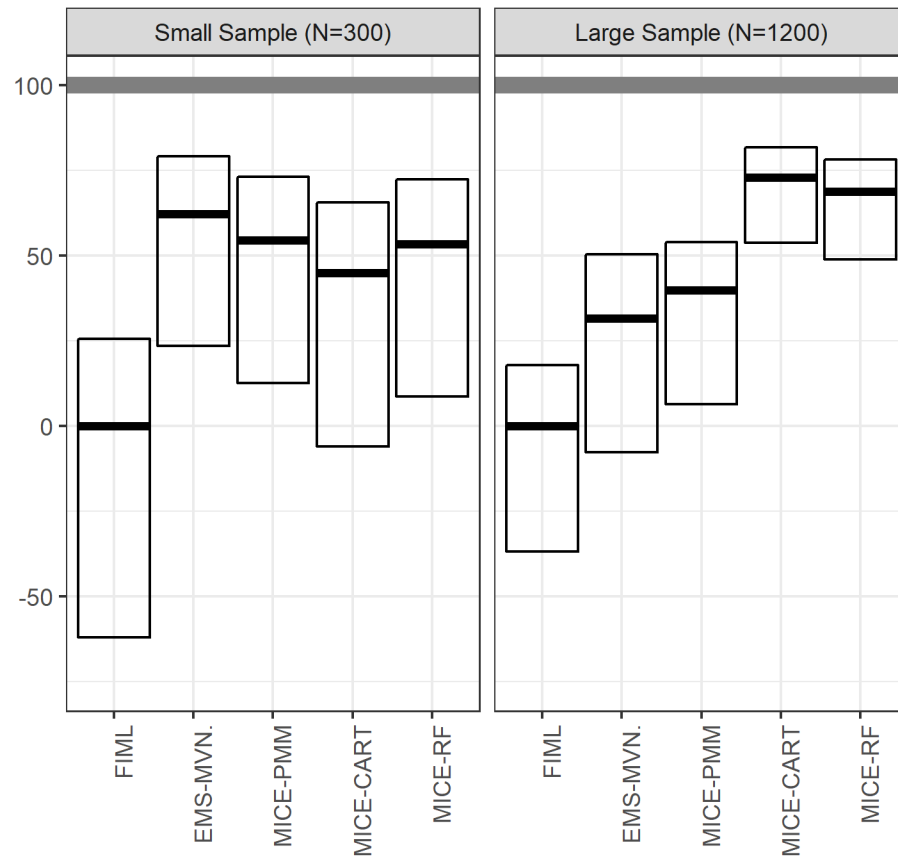
*Absolute Bias Reduction in Small Sample ( $N = 300$ ) and Equal Mixing Conditions*



*Notes.* Change in absolute bias relative to FIML in the small sample and equal mixing conditions, displayed separately by weakly separated (crimson) and strongly separated (teal) conditions.

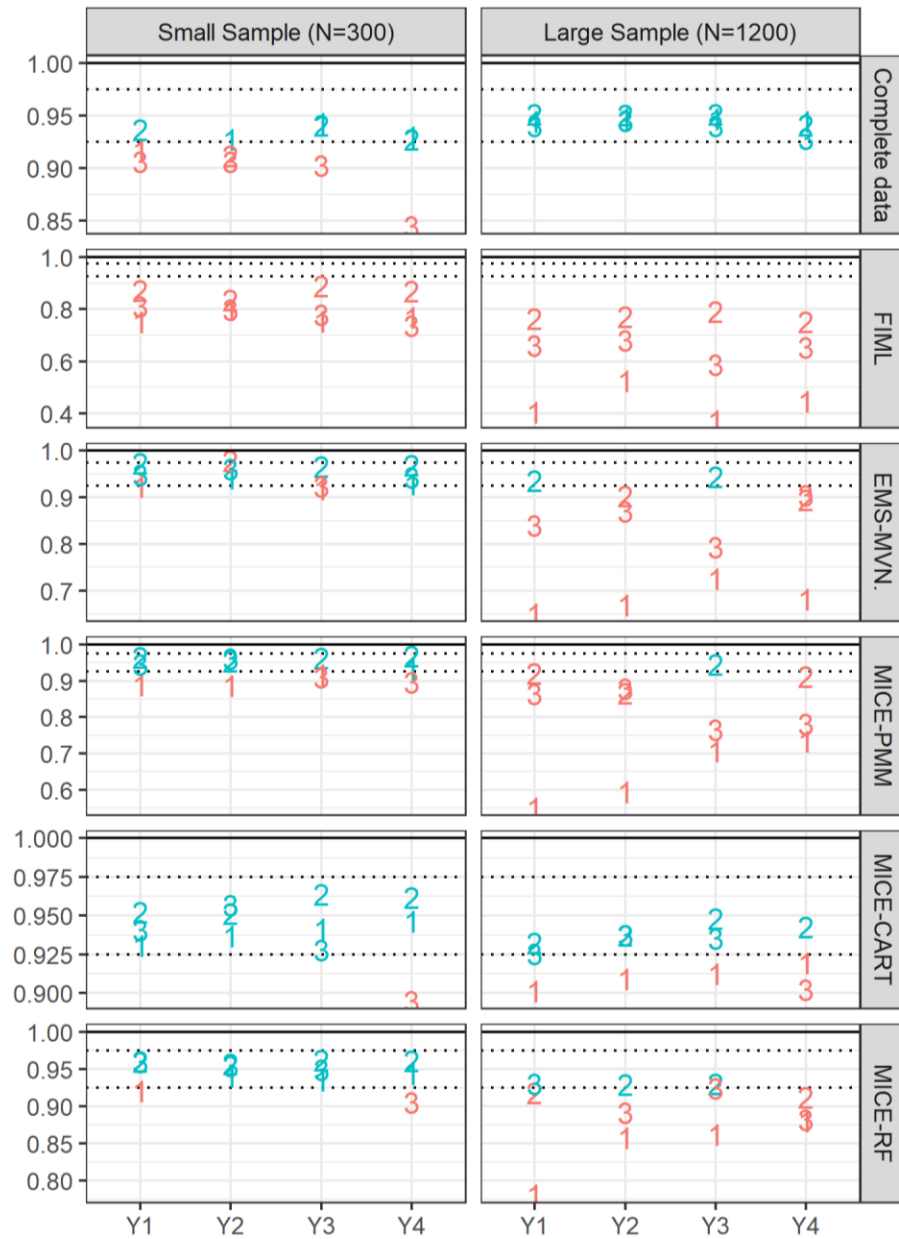
**Figure 2.9**

*Percent Reduction in KL-Divergence*



**Figure 2.10**

*95% CI Coverage*



## References

- Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: SAGE Publications.
- Azur, M. J., Stuart, E. A., Constantine, F., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1), 40–49. <https://doi.org/10.1002/mpr.329>
- Bárcena, M. J., & Tusell, F. (2000). Tree-based algorithms for missing data imputation. *Proceedings in Computational Statistics*, 193–198.
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, 18(1), 71–86. <https://doi.org/10.1037/a0030001>
- Breiman, L. (1984). *Classification and regression trees*. Belmont, CA, USA: Wadsworth International Group.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9), 1070–1076.
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351. <https://doi.org/10.1037/1082-989X.6.4.330>
- Conversano, C., & Cappelli, C. (2002). Missing data incremental imputation through tree based methods. In *Compstat* (pp. 455–460). Springer.
- Conversano, C., & Siciliano, R. (2009). Incremental tree-based missing data imputation with lexicographic ordering. *Journal of Classification*, 26(3), 361–379. <https://doi.org/10.1007/s00357-009-9038-8>
- Creel, D. V., & Krotki, K. (2006). Creating imputation classes using classification tree methodology. *American Statistical Association (Hg.): Proceedings of the Section on Survey Research Methods*, 2884–2887.
- D'Ambrosio, A., Aria, M., & Siciliano, R. (2012). Accurate tree-based missing data imputation and data fusion within the statistical learning paradigm. *Journal of Classification*, 1–32.
- Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, 92–104.
- Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling*, 15(3), 434–448.

- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.
- Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, 18(1), 35–54.
- Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, 10(1), 80–100.
- Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621–638.
- Hancock, G. R., Stapleton, L. M., & Mueller, R. O. (2019). *The reviewer's guide to quantitative methods in the social sciences* (2nd ed.). New York, NY: Routledge.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.
- Hayes, T. (2017). *Using classification and regression trees (CART) and random forests to address missing data*. [Doctoral dissertation, University of Southern California]. ProQuest Dissertations and Theses. Retrieved from <http://search.proquest.com.ezp-prod1.hul.harvard.edu/docview/2071284125?accountid=11311>
- Hayes, T., & McArdle, J. J. (2017a). Evaluating the performance of CART-based missing data methods under a missing not at random mechanism. *Multivariate Behavioral Research*, 52(1), 113–114. <https://doi.org/10.1080/00273171.2016.1264287>
- Hayes, T., & McArdle, J. J. (2017b). Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables. *Computational Statistics & Data Analysis*, 115, 35–52. <https://doi.org/10.1016/j.csda.2017.05.006>
- Hayes, T., Usami, S., Jacobucci, R., & McArdle, J. J. (2015). Using classification and regression trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychology and Aging*, 30(4), 911–929. <https://doi.org/10.1037/pag0000046>
- Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1–47.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 841–860.
- Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2017). A comparison of methods for uncovering sample heterogeneity: structural equation model trees and finite mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(2), 270–282. <https://doi.org/10.1080/10705511.2016.1250637>
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political*

- Science Review*, 95(1), 49–69.
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, N.J.: John Wiley & Sons.
- Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, 39(2), 151–162.  
<https://doi.org/10.1093/jpepsy/jst048>
- Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The oxford handbook of quantitative methods* (pp. 551–611). New York, NY: Oxford University Press.  
<https://doi.org/10.1093/oxfordhb/9780199934898.013.0025>
- McArdle, J. J. (2013). Dealing with longitudinal attrition using logistic regression and decision tree analyses. *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, 282–311.
- McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York, NY: John Wiley & Sons.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558. <https://doi.org/10.1214/ss/1177010269>
- Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies* (Vol. 61). New York, NY: John Wiley & Sons.
- Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302), 415–434.  
<https://doi.org/10.1080/01621459.1963.10500855>
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Teacher's Corner. *Structural Equation Modeling*, 9(4), 599–620. [https://doi.org/10.1207/S15328007SEM0904\\_8](https://doi.org/10.1207/S15328007SEM0904_8)
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, California: Muthén & Muthén. Retrieved from [https://www.statmodel.com/html\\_ug.shtml](https://www.statmodel.com/html_ug.shtml)
- R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from <https://www.r-project.org/>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Johnson Wiley & Sons. <https://doi.org/10.1002/9780470316696>
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489.  
<https://doi.org/10.1080/01621459.1996.10476908>

- Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 477–497.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147.
- Shah, A. D., Bartlett, J., Hemingway, H., Nicholas, O., & Hingorani, H. (2014). CALIBERrfimpute: Imputation in MICE using Random Forest.
- Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology*, 179(6), 764–774. <https://doi.org/10.1093/aje/kwt312>
- Sovilj, D., Eirola, E., Miche, Y., Björk, K.-M., Nian, R., Akusok, A., & Lendasse, A. (2016). Extreme learning machine for missing data using multiple imputations. *Neurocomputing*, 174(PA), 220–231. <https://doi.org/10.1016/j.neucom.2015.03.108>
- Stekhoven, D. J., & Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118.
- Sterba, S. K. (2016). Cautions on the use of multiple imputation when selecting between latent categorical versus continuous models for psychological constructs. *Journal of Clinical Child & Adolescent Psychology*, 45(2), 167–175.
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC Press.
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.
- Vidotto, D., Vermunt, J. K., & Kaptein, M. C. (2015). Multiple imputation of missing categorical data using latent class models: State of the art. *Psychological Test and Assessment Modeling*, 57(4), 542–576.
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594.

Yuan, K., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30(1), 165–200.

### **Chapter 3: Investigating the Performance of a Proposed Hybrid Approach to Generate Imputations in a Latent Profile Analysis under Model Uncertainty: An Initial Study**

Marcus R. Waldman & Katherine E. Masyn

A central goal in person-centered analysis is to identify subpopulations of individuals for which membership cannot be directly observed. These subpopulations are of substantive interest because they may explain important individual differences that may not otherwise be identified using traditional variable-centered analyses that assume a single, overall relation exists among a set of variables (e.g., ANOVA, factor analysis, and structural equation modeling). Common examples of person-centered analysis include latent class analysis, latent profile analysis, latent transition analysis, and growth mixture modeling. The focus of this study is on latent profile analysis (LPA), whereby researchers cluster individuals into homogenous subgroups using a carefully chosen set of continuous indicator variables. The clustering process is completed by fitting finite mixture models (FMMs), and the resulting clusters are referred to as latent classes. Each latent class is presumed to represent a distinct subpopulation, and the classes are then named based on distinguishable features that define that cluster. For example, classes are often distinguished based on class-specific means for the indicator variables.

The naming process is only meaningful if the mean estimates are unbiased so that valid inferences are made during the naming process. Biased estimates may distort between-class differences and complicate the ensuing inference that naming of the classes represents. It is well established that missing data can threaten the validity of inferences if not adequately treated using appropriate missing data strategies. For

example, listwise deletion may result in nonresponse bias if the missingness is not completely at random. In contrast, nonresponse bias can be mitigated if the researcher employs an estimator that assumes the data are missing at random (MAR; Rubin, 1976). Full information maximum likelihood (FIML) and multiple imputation (MI; Rubin, 1978, 1987) are two MAR estimators used ubiquitously in traditional variable-centered analyses. The data are said to be MAR if the missing data patterns are independent of the missing values conditional on the observed values. In practice, the MAR assumption is made more tenable through the inclusion of auxiliary variables (AVs) that would not otherwise appear in the main analysis but are specified to inform the missing values (Meng, 1994; Rubin, 1996; Schafer, 1997; Schafer & Graham, 2002). An inclusive missing data strategy is one in which the researcher liberally incorporates AVs while employing a MAR estimator (Collins, Schafer, & Kam, 2001; Enders, 2010) such as FIML or MI.

Despite the effectiveness of FIML and MI at mitigating nonresponse bias in variable-centered approaches, mitigating bias remains a challenge for both estimators in person-centered analysis such as LPA. In Chapter 2, we detailed limitations of FIML in a person-centered analysis. In brief, FIML is not conducive to an inclusive missing data strategy because the corresponding AVs effectively act as additional indicators of the latent classes. In a person-centered analysis, it is undesirable for external variables (e.g., AVs) to directly define the latent classes; the only variables that should define the latent classes should be the carefully chosen set of indicator variables (Asparouhov & Muthén, 2014; Vermunt, 2010).

In contrast, MI explicitly separates the treatment of the missing data from the analysis of the data by using a stagewise approach which ensures that the AVs do not unduly influence the class definitions. In the first stage of MI, the researcher creates multiple copies of completed datasets by substituting the missing data with imputed values. In the second stage, the imputed datasets are analyzed as if they were complete (Enders, 2010; Rubin, 1987; van Buuren, 2018). An inclusive strategy in MI can be accomplished by incorporating AVs in the first stage but not including these external variables when fitting the mixture models in the second stage. Thus, the explicit separation between treating the missing data problem and fitting an analytic model allows the researcher to include AVs to treat the missing data without having these variables enter the analysis stage where the class definitions are determined.

Despite the conceptual benefits of separating the missing data problem from the intended analysis, methodologists have previously cautioned against employing MI in LPA (Enders, 2010; Enders & Gottschall, 2011; Sterba, 2016). This is because default methods in software assume that the population is comprised of a single, overall class. These methods, therefore, do not generate imputations that can reflect multiple latent subpopulations. We showed in Chapter 2 that imputing from single-class models results in biased class-specific mean estimates of the indicators. Bias in class-specific means threatens inference implied by the class naming process because it may distort between-class differences.

Using alternative imputation models that do not assume the data are composed of a single class partly addresses the bias in class-specific means. In Chapter 2, we found that when sample sizes were large ( $N = 1,200$ ) and classes were well separated so that

entropies averaged near .88, recursive partitioning imputation models reduced nonresponse bias and generated proper imputations. Recursive partitioning imputation does not assume a single-class model because the method partitions the data into rectangular clusters during model fit (Doove, van Buuren, & Dusseldorp, 2014), with each rectangular cluster effectively representing a distinct subpopulation from which imputations are drawn. Moreover, recursive partitioning imputation is readily available in the `mice` (van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2010) package in R (R Core Team, 2020).

However, in Chapter 2, we found that recursive partitioning imputation failed to perform well when sample sizes were small ( $N = 300$ ). Additionally, in large samples ( $N = 1,200$ ) with a small-sized class that represented only 10% of the population, recursive partitioning imputation failed to mitigate nonresponse bias if class separation was weak so that entropy values averaged approximately .74. Thus, a remaining challenge is identifying an imputation procedure that mitigates nonresponse bias in these small-sample or small-class, weakly separated data conditions.

The perspective of this study is that of an applied researcher conducting a person-centered analysis seeking to implement an inclusive missing data strategy in the presence of small class sizes and in sample sizes that are not large ( $N = 300$ -600). These data conditions are very common in LPA studies in education and psychology. We investigate whether enhancing congeniality between the imputation model and the analytic model by generating imputations from a FMM may improve the performance of multiple imputations in small-sample and small-class size settings. According to Meng's definition (1994), uncongeniality means that the "analysis procedure does not correspond

to the imputation model” (p. 539). Bias can result from uncongeniality if the imputation model fails to adequately reflect the complexity of the true data generating mechanism. For example, single-class imputation models fail to capture important sources of heterogeneity when the overall population is comprised of multiple subpopulations. Similarly, recursive partitioning imputation models fail to adequately model the true data generating mechanism in small samples because the resulting clusters are shaped differently than the elliptical clusters implied by a mixture model. Thus, we prioritize greater congeniality between the imputation model and the analytic model in our proposed imputation method.

Specifically, we propose a “hybrid imputation” (van Buuren, 2018) procedure. As we will discuss in the following sections, our hybrid imputation procedure is a chained equations imputation approach where the indicator variables are imputed simultaneously by fitting a compatible FMM, while the fully conditional specification (FCS; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006) is reserved for treating the incomplete AVs. Joint modeling (JM) imputation refers to multiple variables being imputed simultaneously (Schafer, 1997), rather than sequentially. By incorporating a JM block where the indicator variables are imputed from fitting a finite mixture model, we enhance the congeniality between the imputation model and the analytic model, resulting in decreased bias and improved efficiency. At the same time, we preserve the flexibility of the FCS to treat the complex missing data problems that arise when employing an inclusive analysis strategy with many AVs. In this sense, our hybrid imputation procedure combines the strengths of JM with the flexibility of the FCS.

Model uncertainty is an unavoidable complication in a person-centered analysis, and it is increasingly being recognized as a source of uncertainty that should be reflected in the imputations (Kaplan & Yavuz, 2019). Applied researchers rarely know the number of classes supported by the data a priori; the value must be inferred from the data by comparing the fit of alternative models with different numbers of latent classes specified. Therefore, we augment the hybrid imputation procedure to incorporate a model selection step at each iteration. This is done by fitting a sequence of models to the data specified with increasingly many classes. A reference model is then selected with a probability value based on model fit information. By sampling different models from a set of alternatives, the imputations are drawn from a mixture of competing models in order to reflect model uncertainty. We clarify how our procedure is similar to and different from Kaplan and Yavuz's (2019) Bayesian model averaging imputation approach.

The investigated method builds on previous literature. While JM imputation procedures that incorporate mixture models have been proposed in literature (Razzak & Heumann, 2019; Si & Reiter, 2013; Sovilj et al., 2016; van der Palm, van der Ark, & Vermunt, 2016; Vermunt, van Ginkel, van der Ark, Andries, & Sijtsma, 2008; Vidotto, Vermunt, & van Deun, 2018), these methods have mainly been evaluated in applications where categorical variables need to be imputed in large-scale surveys with thousands of observations, such as in item response theory contexts. In these setting, sample sizes are much larger than what is found in a person-centered analysis. Moreover, the number of classes that can be fit to the data can be as high as 70 (Vidotto, Vermunt, & Kaptein, 2015). This contrasts with the two to 10 classes typically supported by the data for research studies in education and psychology, where sample sizes are frequently in the

hundreds. Thus, none of the proposed imputation procedures have been specifically tailored to the person-centered analysis context where sample sizes are generally much smaller, and the number of classes supported by the data is also much smaller. Therefore, identifying imputation methods that produce proper imputations in these common data conditions is important in order for multiple imputation to be useful for applied researchers.

Moreover, we are unaware of any procedure that employs a hybrid imputation strategy with an FMM JM block. Thus, unlike the proposed JM procedures, our procedure is designed to exhibit greater congeniality to the analytic models employed in person-centered analysis, while also being flexible enough to accommodate complex missing data problems in the AVs themselves. Finally, this study aligns with the growing sentiment that “Bayesianly proper” (Schafer, 1997) imputations must account for model uncertainty. In line with Kaplan and Yavuz (2019), we attempt to account for model uncertainty by sampling from a mixture of competing models when generating imputations.

The organization of this chapter is as follows. We provide a brief background discussion on FMMs, proper imputations, and important distinctions and commonalities between the FCS and JM imputation procedures. Next, we turn to the challenge of generating proper imputations. To scaffold concepts, we first assume that the number of classes is known a priori in order to clarify concepts regarding hybrid imputation. Specifically, we discuss how imputations for the class indicators can be sampled using the expectation-maximization (EM) with sampling algorithm (King, Honaker, Joseph, &

Scheve, 2001), and how the EM with sampling algorithm can be embedded in a chained equation procedure that employs the FCS for the auxiliary variables.

Having clarified how hybrid imputation can be implemented when the number of classes is known, we next consider how to appropriately address model uncertainty induced by this number not being known. We discuss limitations with popular strategies, such as setting the number of classes to a large value. We then propose a sampling strategy that treats the number of classes as a random variable to be estimated. We apply the proposed imputation method to a real-world example using the ECLS-K dataset. We end with a discussion for future directions.

## **Background**

### **Finite Mixture Models**

Finite mixture modeling exploits the law of total probability to model the joint density of individual  $i$ 's indicator data vector,  $\mathbf{y}_i$ , as a weighted average across  $K$  component densities,

$$\Pr(\mathbf{y}_i) = \sum_{k=1}^K \Pr(\kappa = k) \Pr(\mathbf{y}_i | \kappa = k) \quad (3.1)$$

where  $\kappa$  is a categorical latent variable representing a latent class, and the family of the conditional distribution  $[\mathbf{y}_i | \kappa = k]$  is assumed to be known. The specific title for the person-centered analysis that is being performed may differ depending on the family of this conditional distribution. For a latent class analysis, the indicators are dichotomous, and the conditional density is assumed to be a Bernoulli distribution. Latent profile analysis (LPA), on the other hand, is generally appropriate when the  $J$  indicators are

continuous variables. The assumption is that the data are a mixture of  $K$  multivariate normal component densities,

$$[\mathbf{y}_i | \kappa = k] \sim \mathcal{N}_J(\boldsymbol{\mu}_k, \Sigma_k) \quad \forall k = 1, \dots, K \quad (3.2)$$

where  $\boldsymbol{\mu}_k$  is the  $k$ th component mean vector and  $\Sigma_k$  is the  $k$ th component variance covariance matrix. In LPA, classes are usually defined and given names based on examining how  $\boldsymbol{\mu}_k$  differ across the classes.

Applied researchers typically do not know the value of  $K$  or the structure of the component variance-covariance matrix,  $\Sigma_k$ . Two common specifications for  $\Sigma_k$  include the class-varying conditional independence model, where  $\Sigma_k$  is a diagonal matrix, as well as the class-varying unrestricted model, where all elements in  $\Sigma_k$  are freely estimated. We refer the reader to Masyn (2013) for additional specifications, such as specifications that assume the estimates do not vary across the classes.

To simplify notation, we assume that the researcher specifies the least restrictive configuration, in that all elements in  $\Sigma_k$  are freely estimated and may differ across the classes. The sequence of models fit to the data can be written as  $\boldsymbol{\mathcal{M}} = \{\mathcal{M}_K: K = 1, 2, \dots\}$  where  $K$  is the number of components. The likelihood of observing the collection of complete indicator values,  $\mathbf{y}_i$ , for all  $N$  individuals, denoted  $Y$ , given a model with  $K$  components is

$$\mathcal{L}(\theta_K | Y, \mathcal{M}_K) = \prod_{i=1}^N \sum_{k=1}^K \pi_k \mathcal{N}_J(\mathbf{y}_i | \boldsymbol{\mu}_k, \Sigma_k) \quad (3.3)$$

where  $\theta_K = \{(\boldsymbol{\mu}_k, \Sigma_k, \pi_k): k = 1, \dots, K\}$  and  $\pi_k$  is referred to as the marginal class probabilities or mixture weights. The equation in (3.3) is the complete-data likelihood because all elements in the data matrix  $Y$  are complete. Software relies heavily on the

EM algorithm (Dempster, Laird, & Rubin, 1977) to calculate the maximum likelihood estimates,  $\hat{\theta}_K$ . We refer the interested reader to the vast theoretical literature regarding the EM algorithm (R. J. Little & Rubin, 2002; McLachlan, 2008; Schafer, 1997), including its extensions and applications to mixture modeling (Frühwirth-Schnatter, Celeux, & Robert, 2019, Chapter 2; McLachlan & Peel, 2004).

### **Bayesianly Proper Imputations**

When data are missing, the complete data is a superset of the observed and missing data  $Y = \{Y^{\text{obs}}, Y^{\text{mis}}\}$ . Briefly stated, an imputation procedure is said to generate proper imputations for  $Y^{\text{mis}}$  if the procedure results in valid inferences when given the observed data and a set of AVs (denoted  $X$ ) necessary for an inclusive missing data strategy.

Two different perspectives have been offered in previous literature depending on whether one takes a Bayesian (Schafer, 1997) or frequentist approach (Rubin, 1987) to conducting inference with multiply imputed data. However, only a Bayesian perspective allows for uncertainty in the underlying imputation model (Kaplan & Yavuz, 2019). For this reason, we take a Bayesian perspective in defining proper imputations.

From a Bayesian perspective, uncertainty in the parameter estimates caused by missing data is fully captured by the observed data posterior distribution, given by

$$\Pr(\theta_K | Y^{\text{obs}}, X; \mathcal{M}_K) \propto \int_{Y^{\text{mis}}} \Pr(\theta_K | Y^{\text{obs}}, Y^{\text{mis}}; \mathcal{M}_K) \Pr(Y^{\text{mis}} | Y^{\text{obs}}, X; \mathcal{M}_K) dY^{\text{mis}} \quad (3.4)$$

where  $[\theta_K | Y^{\text{obs}}, Y^{\text{mis}}; \mathcal{M}_K]$  is referred to as the complete data posterior distribution and  $[Y^{\text{mis}} | Y^{\text{obs}}, X; \mathcal{M}_K]$  is the posterior predictive distribution for  $Y^{\text{mis}}$  (R. J. Little & Rubin,

2002, p. 210; Rubin, 1987; van Buuren, 2018, pp. 41–44). The posterior predictive distribution for  $Y^{\text{mis}}$  is further given by

$$\Pr(Y^{\text{mis}}|Y^{\text{obs}}, X; \mathcal{M}_K) = \int_{\theta_K} \Pr(Y^{\text{mis}}|Y^{\text{obs}}, X, \theta_K; \mathcal{M}_K) \Pr(\theta_K|Y^{\text{obs}}) d\theta_K. \quad (3.5)$$

As defined by Schafer (1997, p. 105), if imputations are independent samples from the posterior predictive distribution for  $Y^{\text{mis}}$ , then the imputations are said to be Bayesianly proper.

Uncongeniality between the imputation and analysis models can result in invalid inference, particularly if the imputation model is less complex than the analysis model. In other words, valid inference is only guaranteed if the imputation model is at least as complex as the analysis model and is well enough aligned with the reality of the underlying data generating mechanism for the missing values (Collins et al., 2001; Meng, 1994). In this way, the validity of inference may be highly dependent on the imputation model being fit because the imputation model must be sufficiently aligned with the reality of the true data generating mechanism. Such a situation is difficult in settings such as LPA where the number of components is usually not known.

Kaplan & Yavuz (2019) argue that inference can become more robust to violations of congeniality by incorporating Bayesian model averaging imputation. They argue that when a researcher is in a situation where the true data generating mechanism is unknown, then the imputation model must account for the resulting induced uncertainty to generate imputations that are Bayesianly proper. Through Bayesian model averaging, model dependence is effectively marginalized out of the posterior predictive distribution by considering the posterior distribution as a mixture of competing models. Although not explicitly defined by the authors, it stands to reason that imputations for  $Y^{\text{mis}}$  are said to

be Bayesianly proper under Kaplan & Yavuz’s (2019) definition if  $Y^{\text{mis}}$  is an independent sample from the marginal distribution  $[Y^{\text{mis}}|Y^{\text{obs}}, X]$ , which is a mixture across competing models, i.e.

$$\Pr(Y^{\text{mis}}|Y^{\text{obs}}, X) = \sum_{\mathcal{M}_K} \Pr(Y^{\text{mis}}|Y^{\text{obs}}, X; \mathcal{M}_K) \Pr(\mathcal{M}_K|Y^{\text{obs}}, X) \quad (3.6)$$

where  $\Pr(\mathcal{M}_K|Y^{\text{obs}}, X)$  is referred to as the posterior model probability for model  $\mathcal{M}_K$ .

The authors show that Bayesian model averaging can minimize the discrepancy between the true posterior distribution of  $Y^{\text{mis}}$  and the distribution from which imputations are actually drawn when uncongeniality is unavoidable because the true data generating mechanism is not known.

Rather than averaging across competing models, our proposed hybrid imputation method addresses model uncertainty by sampling directly from the mixture model implied in (3.6). At each iteration in our imputation procedure, we first sample a reference model using the posterior model probabilities for each competing model. Next, we sample imputation from the corresponding predictive distribution conditional on the reference model being selected.

We sample from the mixture model in (3.6) directly because we are not in a position to employ Bayesian model averaging. The challenge with including a Bayesian model averaging strategy directly in FMMs is that the parameters must have the same meaning across the different models. In the linear regression context considered by Kaplan and Yavuz (2019), for example, this occurs by creating a full model that contains all possible predictors and interaction terms. Each competing model is then simply a nested (or constrained) model specified by fixing some of the regression estimates to

zero. The result of Bayesian model averaging is a large set of estimates for all of the parameters that define the full model.

For FMMs, however, there is no direct or obvious correspondence to map classes obtained from a model with  $K$  components to a model with  $K + 1$  or more components. The classes are not exchangeable across different models. That said, combining classes in a principled manner has previously been explored in literature (Baudry, Raftery, Celeux, Lo, & Gottardo, 2010; Hennig, 2010), with some finding that such an endeavor improves classification quality (Wei & McNicholas, 2015). In the discussion, we consider possible future directions to map the parameters across competing models so that a model averaging approach more aligned with Kaplan and Yavuz’s (2019) recommendations can be employed.

### **Normal Imputation as a Case Study for Understanding the FCS and JM**

To provide intuition on how FCS and JM modeling differ in their imputation approach, we consider the special case where the data originate from a single-class, multivariate normal model. Valid inference requires that the imputations reflect variability from three sources. These sources are: (1) individual differences in predicted values, (2) variability due to sampling error, and (3) variability due to random noise (van Buuren, 2018). The FCS takes a “variable-by-variable” approach to construct an imputed dataset that reflects the three sources of variability. In particular, imputations for each variable are drawn separately by fitting univariate regression models with each variable as an outcome and the remaining variables (and AVs) as predictors. JM, in contrast, generates imputations for all variables simultaneously.

### ***Fully Conditional Specification***

There are several procedures available to ensure that imputations from the FCS are proper and reflect the three sources of variability needed for valid inference. To reflect sampling variability, a common procedure is to bootstrap the data before fitting the univariate regression model. Once the univariate model is fit, predicted values are then calculated by using the corresponding coefficient estimates. Random noise is then added to the predicted values in a principled manner so that the imputations appropriately reflect the third and final source of variability. For normal imputation, random noise can be appropriately added to the predicted values by sampling from a normal distribution using the residual variance estimate obtained from a linear regression model fit to the bootstrapped data. This three-stage procedure of bootstrapping, predicting missing values, and then adding random noise repeats itself for all variables in the dataset. One sweep across the dataset results in a single sample from the posterior predictive distribution of  $Y^{\text{mis}}$ . Additional samples are obtained by conducting multiple iterations of the FCS procedure to form a Markov chain Monte Carlo (MCMC) (van Buuren, 2018; van Buuren et al., 2006).

In many cases, the FCS is akin to Gibbs sampling (Casella & George, 1992; Gelfand & Smith, 1990) in Bayesian estimation whereby the fully joint distribution is difficult to sample, so samples are drawn from a constituent set of easier-to-sample conditional distributions. As in Gibbs sampling, the FCS results in an MCMC chain and, upon convergence, users obtain imputed datasets by sampling from the MCMC chain. We refer the reader to the existing literature for a technical discussion on necessary

conditions for the FCS to asymptotically approximate Gibbs sampling (Liu, Gelman, Hill, Su, & Kropko, 2013; van Buuren, 2018, pp. 119–124).

One advantage of the FCS is the tremendous flexibility the researcher has in specifying the univariate regression models to create imputations. The FCS is unrestricted in the types of regression models that can be fit to each variable and is amenable to a large number of parametric and nonparametric regression methods. Thus, researchers have great flexibility to treat missing data across the varied data types that appear in real-world settings.

### ***Joint Modeling by EM with Sampling***

The primary difference between JM and FCS is that in JM, the imputations are created simultaneously across all of the variables. We focus on the EM with sampling imputation algorithm because this algorithm shares many similarities with the three-step FCS procedure described above, but it can easily incorporate a finite mixture imputation model. The first step in the EM with sampling algorithm is to bootstrap the data to ensure that the imputed datasets reflect sampling variability. Next, a multivariate normal regression model is fit to observed data. Specifically, parameters are estimated by maximizing the observed data likelihood using the EM algorithm. Finally, imputed values are sampled in two stages. First, the *sweep* operator<sup>1</sup> is applied to obtain the predictive density of the missing data conditional on the observed data; in a normal regression, the conditional distribution will either be a univariate normal (if only one value is missing) or multivariate normal (if an observation is missing more than one value) distribution. Next, imputations are simulated from the corresponding conditional

---

<sup>1</sup> For an accessible introduction to the *sweep* operator, we refer the reader to Little & Rubin (2002).

distribution. Researchers can conduct EM with sampling assuming a multivariate normal imputation model, as is done in the `Amelia` R package (Honaker, King, & Blackwell, 2011). Alternatively, LatentGold (Vermunt & Magdison, 2016) generalizes the EM with sampling algorithm to include multiple latent classes by estimating a finite mixture of Bernoulli or multinomial distributions, instead of the single-class multivariate normal model assumed by `Amelia`. The EM with sampling algorithm used by LatentGold has shown to be useful for imputing categorical item response data (Vermunt et al., 2008; Vidotto et al., 2015). We summarize the steps for the EM with sampling algorithm procedure in Table 3.1.

The fact that the EM with sampling algorithm allows for the incorporation of a mixture model to generate imputations may be advantageous for person-centered analysis, as it allows for greater congeniality between the imputation model and the analysis model. The disadvantage, however, is that JM may not be ideal for treating incomplete AVs; greater flexibility is needed given the mix of data types (e.g., continuous, categorical, etc.) that comprise the set of AVs. Hybrid imputation (i.e., imputing a block of variables using joint modeling as part of a larger chained equations procedure) may offer a useful compromise between JM and FCS for person-centered analysis. It allows greater congeniality between the imputation model and analysis model for better performance when small classes are present. At the same time, it allows for flexibility in treating the missingness of the AVs which are not of substantive interest. We now turn to how our proposed hybrid imputation procedure would be implemented if the number of classes is known. Although this is generally not realistic, we present the material in this scaffolded format to clarify concepts.

### **A Hybrid Imputation Method when the Number of Classes are Known**

The hybrid imputation procedure that we propose is a chained equation procedure that includes a JM block to impute the class indicators and an FCS block to impute any AVs. Algorithm 1 in Table 3.2 contains the exact implementation steps. In this section, we detail how our hybrid procedure differs from previously proposed imputation algorithms that use FMMs by proposing two novel modifications. Next we give a Bayesian justification for Algorithm 1 as a proper imputation procedure that approximately samples from a posterior distribution using data augmentation (Tanner & Wong, 1987).

#### **Novel Modifications to EM with Sampling**

We make two novel modifications to the EM with sampling procedure described in Table 3.1 that have not previously been employed in literature. Specifically, we employ Bayesian bootstrapping in Step 2, instead of relying on sampling with replacement, in order to mitigate convergence issues when class sizes are small. Second, we fit a mixture regression model in Step 3 that allows for the incorporation of AVs that concomitantly predict class membership and missing indicator values.

#### ***The Bayesian Bootstrap***

We use the Bayesian bootstrap (Rubin, 1981) as an alternative to the traditional bootstrap conducted by sampling observations with replacement. The Bayesian bootstrap is a simple procedure to reweight observations in order to emulate resampling. Despite its name, the Bayesian bootstrap can be conducted in a manner that is completely nonparametric and noninformative. The Bayesian bootstrap is best understood by comparing it with the traditional bootstrap. In fact, sampling with replacement is simply a

means to weight observations. Indeed, by sampling with replacement, the user is effectively weighting observations by a value that is in the natural numbers (e.g., if the observation was never sampled, then its weight is zero; if it was sampled once, then the weight is one; if sampling with replacement resulted in an observation being selected twice, then the weight is two; etc.). The sum of the weights is then equal to the total number of observations. The Bayesian bootstrap is a procedure to sample an observation's weight, rather than sampling the observations themselves. Weights are drawn from a Dirichlet distribution with  $N$  total categories. A uniform prior can be specified so that each observation has an equal chance of being assigned a given weight. Thus, the uniform prior is completely noninformative and parallels sampling with replacement. The uniform prior is implemented by specifying unit concentration parameters for the Dirichlet distribution.

Both the Bayesian bootstrap and the traditional bootstrap sample from the empirical CDF. Thus, both lead to very similar inferences in samples. In fact, sampling with replacement can be seen as simply a special case of the Bayesian bootstrap in that the weights are restricted to the natural numbers. By allowing the weights to take on the full set of positive real numbers, the Bayesian bootstrap effectively smooths the empirical CDF, which is advantageous when sample sizes are small, so the histogram of the empirical CDF can appear quite discrete (Chernick, 2011, Chapter 6). We choose the Bayesian bootstrap because we find in practice that it leads to improved convergence rates when small class sizes are present compared to sampling with replacement.

### *Mixture Regression Models*

Our goal in including the JM step is to enhance the congeniality between the imputation model and the analysis model, so that valid inferences can be made even in small class sizes. To do that, our goal is to sample from a posterior predictive density that is itself a mixture model. We propose a mixture regression model (B. O. Muthén & Asparouhov, 2009; Wedel, 2002), also called a mixture of experts (Jacobs, Jordan, Nowlan, & Hinton, 1991) in machine learning literature, as the multivariate model fit in Step 3 of Table 3.1. A mixture regression model allows for the specification of AVs so that an inclusive strategy can be employed. In particular, mixture regression models allow a set of background covariates to predict class probabilities, as well as the class indicators, so that the probability of observing the data is given by

$$\Pr(\mathbf{y}_i) = \sum_{k=1}^K \pi_k(\mathbf{x}_i) \mathcal{N}_J(\mathbf{y}_i | \mathbf{x}_i \boldsymbol{\beta}_k, \Sigma_k) \quad (3.7)$$

where  $\pi_k(\mathbf{x}_i)$  is modeled using a multinomial logistic regression given by

$$\pi_k(\mathbf{x}_i) = \begin{cases} \frac{\exp(\mathbf{x}_i \boldsymbol{\alpha}_k)}{1 + \sum_{k=1}^{K-1} \exp(\mathbf{x}_i \boldsymbol{\alpha}_k)} & \text{if } k < K \\ \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\mathbf{x}_i \boldsymbol{\alpha}_k)} & \text{if } k = K \end{cases} \quad (3.8)$$

and  $\mathbf{x}_i$  is a size  $P$  vector of the AVs in  $X$ . Mixture regression models can be fit under maximum likelihood using the EM algorithm (Gormley & Frühwirth-Schnatter, 2019) in software such as LatentGold and Mplus (L. K. Muthén & Muthén, 2017).

### **Hybrid Imputation as an Approximate Posterior Sampler with Data Augmentation**

Having discussed our modification to the EMs procedure, we now detail our hybrid imputation procedure given in Algorithm 1 (see Table 3.2) and justify it as an approximate Bayesian estimation algorithm to sample from the posterior distribution

$$[\{\phi_p: p = 1, \dots, P\}, \{\alpha_k, \beta_k, \Sigma_k: k = 1, \dots, K\} | Y^{\text{obs}}, X^{\text{obs}}; \mathcal{M}_K] \quad (3.9)$$

where  $\phi_p$  are the univariate regression parameters defining the FCS when imputing  $X^{\text{mis}}$ . Our procedure uses the Bayesian bootstrap to draw independent samples from a sampling distribution that approximates a posterior distribution. In missing data literature, approximately sampling a posterior by bootstrapping is common (Efron, 1994; R. J. Little & Rubin, 2002; Rubin, 1987; Rubin & Schenker, 1986). Because the posterior distribution approximates a posterior with completely noninformative priors, this type of posterior sampling is truly nonparametric. Indeed, nonparametric bootstrapping sampling algorithms are increasingly gaining attention in the machine learning community because they are more efficient at sampling the multimodal posterior distributions that occur in mixtures due to label switching (Fong, Lyddon, & Holmes, 2019). Traditional sampling algorithms, such as the Gibbs sampler or Metropolis Hastings algorithms, are inefficient because they often fail to switch between modes of the posterior distribution (Celeux, Kamary, Malsiner-Walli, Marin, & Robert, 2018, pp. 75–81).

Our hybrid imputation procedure is designed to sample the posterior distribution in (3.9) using three separate data augmentation steps to treat the missing information in class membership, missing class indicator data, and missing data in the AVs. The data augmentation steps break the full joint distribution down into conditional distributions from which samples can easily be obtained. Data augmentation is a common Bayesian estimation procedure to sample from an observed-data posterior distribution in a two-step, iterative fashion. In the most basic application, the two steps involve constructing a single imputed dataset (i.e., “I-step”) and a posterior step (i.e., “P-step”). The I-step involves sampling from the posterior predictive distribution given a current set of

parameter values. The parameters are then updated in the P-step by sampling new parameter values conditional on the imputed data points in the I-step.

In the first data augmentation step, we augment the data matrix to include a column for class membership,  $\kappa$ , for each individual. This value is unknown, so it must be imputed using data augmentation. Augmenting the data matrix in this way and treating  $\kappa$  as a value that is “missing” is the foundation of all Bayesian estimation algorithms; with FMMs this is because it allows sampling to occur using conditional distributions that are easier to sample (Gelman et al., 2013; McLachlan & Peel, 2004).

To implement the first data augmentation step, a set of parameter values at iteration  $t$  can be obtained (e.g.,  $\{\alpha_k^{(t)}, \beta_k^{(t)}, \Sigma_k^{(t)}: k = 1, \dots, K\}$ ) by fitting a mixture regression model using the EM with sampling procedure given a current imputed dataset for the AVs. Upon convergence of the fitted model, posterior class membership can then be sampled by first calculating a set of posterior class probabilities given by

$$\Pr(\kappa_i = k) = \frac{\pi_k(\mathbf{x}_i) \mathcal{N}_j(y_i | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^K \pi_k(\mathbf{x}_i) \mathcal{N}_j(y_i | \mathbf{x}_i \boldsymbol{\beta}_k, \Sigma_k)}$$

and then sampling class membership using a multinomial distribution using the posterior class probabilities. Having imputed  $\kappa_i$  values for each individual in the augmented data matrix, the first data augmentation step is completed.

The second data augmentation step involves sampling the missing profile indicator data  $Y^{\text{mis}}$ , given the individual’s  $\kappa_i$  value. Because in LPA the component densities are multivariate normal, the posterior predictive density conditional on the current class membership value is also normally distributed. Therefore, sampling can be accomplished with the assistance of the *sweep* operator. Upon sampling from the

posterior predictive density, the result is a single imputed dataset for the class indicators at the current iteration in the MCMC chain,  $Y^{(t)}$ .

The third and final data augmentation step treats the missing data in the AVs. Our proposed method attempts to accomplish this by using the FCS to sequentially update the imputations for missing AVs. In particular, the  $p$ th AV is imputed by specifying the univariate regression model to include the imputed class membership at  $\kappa^{(t)}$  from the first data augmentation step, the imputed indicator data  $Y^{(t)}$  from the second data augmentation step, and the remaining imputed AVs (denoted  $X_{-p}^{(t)}$ ).

In summary, by conducting three data augmentation steps and by iteratively sampling  $\{\alpha_k^{(t)}, \beta_k^{(t)}, \Sigma_k^{(t)}: k = 1, \dots, K\}$  and  $\{\phi_p^{(t)}: p = 1, \dots, P\}$ , our procedure can be viewed as a Bayesian sampling procedure to approximately sample the posterior distribution in (3.9) with missing class membership values, missing class indicator values, and incomplete AVs. An MCMC chain is formed by iteratively repeating the data augmentation steps. The independent samples of the posterior predictive distribution for  $Y^{\text{mis}}$  needed for the imputed datasets to be proper can be obtained by conducting a sufficient number of iterations for the chain to converge and then directly sampling from the MCMC chain. Model uncertainty is a remaining challenge that we have so far not addressed. We turn to this topic next.

### **Multiple Imputation when the Number of Classes is not Known**

Rarely are the number of classes,  $K$ , known when an applied researcher conducts a person-centered analysis. As a result, some argue that there is inherent model uncertainty that should be reflected in the imputations in order for the imputations to be Bayesianly proper (Kaplan & Yavuz, 2019). We present three alternative options for

dealing with model uncertainty. The first option is to intentionally specify an imputation model that is an overfitted mixture model, meaning that the number of components specified is sufficiently large that it is, ostensibly, greater than the number of components that can be supported by the data. Reversible jump MCMC (Green, 1995) is an alternative option to switch between models when sampling from the MCMC chain. We discuss the limitations of both these options and justify our choice for employing a third option—sampling from a mixture of competing models using Akaike weights.

### **Overfitted Mixture Imputation Model**

Uncongeniality threatens inference when the imputation model is less complex than the true data generating mechanism. In mixture models, an imputation model fit with too few components will lead to uncongeniality and threatens the validity of inferences. However, overfitting the number of components is not necessarily problematic for inference. This is because fitting a more complex mixture model than the true, simpler mixture model still results in the multivariate density being consistently estimated.

To understand why overfitting leads to consistent density estimation, consider the situation where a finite mixture model is fitted with  $K = 3$  total components (i.e.,  $\mathcal{M}_3$ ) with data generated from a  $K = 2$  component model (i.e.,  $\mathcal{M}_2$ ). The parameters in  $\mathcal{M}_2$  can be recaptured by the parameters in  $\mathcal{M}_3$  by either setting one of the three mixture weights,  $\pi_k$  in (3.3), to zero or by setting any two of the component densities to be the same (Frühwirth-Schnatter, 2006; McLachlan & Peel, 2004, p. 28). In terms of estimation, an overfitted mixture model leads to identifiability issues in the sense that the maximum likelihood estimate is no longer unique. For example, setting  $\pi_3$  to zero allows the  $\boldsymbol{\mu}_3$  and  $\Sigma_k$  to take on any arbitrary value. Reflecting the problems with identifiability,

the Fisher information increasingly becomes singular, regardless of whatever optimum is reached as the sample size approaches infinity (Drton & Plummer, 2017), and the likelihood function takes the shape of a ridge. Despite issues with identifiability for the maximum likelihood estimates, any estimates that are obtained that result in the likelihood reaching its maximum value will result in the joint distribution of the data being consistently modeled as arising from a two-component model. In summary, although identifiability of the parameters itself is an issue with overfitting a mixture model, overfitting is not necessarily problematic when the goal of fitting the finite mixture model is limited to approximating a multivariate density.

The fact that overfitting still leads to consistent density estimates has direct implications for generating proper imputations when  $K$  is unknown. All that is necessary for proper imputations in the mixture context is to fit a mixture model that has more components than the true data generating mechanism because the resulting predictive density implied from the imputation model will consistently estimate the true posterior predictive density for  $Y^{\text{mis}}$ . For instance, in the context where finite mixture models are employed in a JM framework to impute categorical data in large-scale surveys, it is recommended to generate imputations from an imputation model with up to 70 classes (Vidotto et al., 2015). Others have recommended that the number of classes be determined through an enumeration step conducted before imputing by using the AIC to decide on the number of classes. This is because the AIC tends to over extract the number of classes (Vermunt et al., 2008; Vidotto et al., 2015).

Simply setting the number of classes to a large value has several drawbacks. First, even if it were possible to know a priori that a large value of  $K$  is beyond what the data

supports, imputation efficiency is not optimized. Imputation efficiency may be less of a concern in large-scale survey or assessment data, but in a person-centered analysis with small samples, power is at a premium. Moreover, practical problems, such as instability during estimation due to unbounded likelihoods, convergence issues, and the EM algorithm getting stuck in suboptimal stationary points, are likely to occur if the number of classes is set too large.

In addition, if the number of classes is determined by the AIC before any imputation is completed, then the missingness of the AVs cannot be treated in an optimal manner when deciding on the number of classes. This is important because many of the AVs may be distal outcomes in a subsequent analysis, so missingness in the AVs and the number of classes should mutually inform one another. Finally, deciding on the number of classes before drawing imputations prohibits imputations being drawn from the mixture of competing models as in (3.6). Thus, uncertainty in the selection of the final model is never reflected in the resulting imputation, arguably compromising the degree to which the imputations are Bayesianly proper (Kaplan & Yavuz, 2019).

In summary, imputing using an overfitted finite mixture model is considered less detrimental to inference than using underfitted mixture model. This is because an overfitted mixture model still consistently approximates the density of the predictive distribution. Nevertheless, overfitting should be done judiciously because it does not maximize imputation efficiency—an important consideration in the small-sample and small-class size settings present in the educational, behavioral, and social sciences.

## Reversible Jump MCMC

As an alternative to fitting an overfitted mixture imputation model, the reversible jump MCMC sampler (Green, 1995) is an alternative procedure for sampling posterior distributions between alternative mixture models with different numbers of components. Paralleling the Metropolis Hastings algorithm, at each step in the sampler, a  $K$  value is proposed, and the corresponding model is either accepted or rejected with a probability value calculated given the current state of the parameter estimates. The reversible jump MCMC sampler is appealing in theory because it treats  $K$  as an unknown parameter in the model for which a posterior distribution needs to be sampled (McLachlan & Peel, 2004). By sampling a  $K$  value at each iteration, the reversible jump MCMC sampler incorporates model uncertainty in a principled and direct manner.

Despite the theoretical advantages of the reversible jump MCMC sampler, the procedure fails to perform well in practice. Calibrating proposals for  $K$  is an extremely arduous task. It is not uncommon, for example, for acceptance rates of the proposal to be as small as 1% (Celeux, Fruewirth-Schnatter, & Robert, 2018). Such small acceptance rates would unduly extend the computational time required to reach convergence of the MCMC chain. Thus, the practical challenges of calibrating the reversible jump MCMC sampler are too great to overcome.

## Sampling from a Mixture of Competing Models

In our attempt to account for model uncertainty, we investigate whether sampling from a mixture of competing models as described in (3.6) results in valid inference even in small class settings. In the Background, we offered a Bayesian justification for sampling from a mixture of competing models at each iteration of the MCMC chain. Two

unresolved issues remain: (1) identifying the set of competing models at each iteration of the MCMC chain (referred to as “class enumeration” in person-centered analysis literature), and (2) calculating the posterior model probabilities  $\Pr(\mathcal{M}_K|Y^{obs}, X)$  in (3.6) so that sampling weights for each competing model can be calculated (referred to as the reference model selection).

### ***Class Enumeration***

Model selection with mixture models in a person-centered analysis proceeds first by enumerating the classes. This involves fitting a sequence of models with increasingly many components until convergence issues become intractable (Masyn, 2013). We have found in practice that this threshold usually occurs well before a  $K = 10$  component model is fit to the data when all parameters in the  $\Sigma_K$  are freely estimated across the classes. Thus, we expect relatively few alternative models will need to be considered at each iteration of the MCMC chain.

We propose enumerating the classes fitting a sequence of competing mixture regression models in (3.7) at each joint modeling iteration of the hybrid imputation procedure. First, a single-class mixture regression model with  $K = 1$  components is fit to the indicator data given the current state of the imputed auxiliary variables,  $(Y^{obs}, X^{(t)})$ . Next, a  $K = 2$  mixture regression model is fit to the data. This process continues until convergence becomes intractable.

We specify all elements in  $\Sigma_K$  as freely estimated when enumerating the classes, even if the researcher is only ultimately fitting an analysis model with exclusively diagonal elements in the class-specific variance-covariance matrices. This is because conditioning on auxiliary variables that predict two indicators induces a correlation in the

residuals between those two indicators, even when those two indicators are conditionally independent given class membership. Moreover, fitting a mixture regression model with a simpler  $\Sigma_k$  risks unnecessary uncongeniality if  $\Sigma_k$  requires that the covariance estimates be freely estimated to reflect the true data generating mechanism. On the other hand, congeniality is guaranteed if the  $\Sigma_k$  configuration is more complex than reality. In summary, the enumeration step we propose consists of fitting a sequence of mixture regression models with freely estimated residual variance-covariance structures. Models are fit until convergence issues become intractable.

### Reference Model Selection

Once the classes have been enumerated, a reference must be selected so that imputations can be drawn for  $Y^{\text{mis}}$ , and data augmentation can commence for a particular iteration of the MCMC chain. We propose selecting a reference model from a mixture of the competing models enumerated in the previous step. A reference model is selected from the sequence of models enumerated where each model is weighted by the posterior model probability,  $\Pr(\mathcal{M}_K | Y^{\text{obs}}, X^{(t)})$ .

Literature is rife with options for calculating the posterior model probability. Fully Bayesian estimation allows this value to be calculated directly by using the posterior distribution and calculating the corresponding Bayes factors (Berger, 1985; Kass & Raftery, 1995). Although such an approach may theoretically be ideal, it is computationally demanding. To calculate the Bayes factors, the posterior distribution would need to be sampled until convergence for each model in the enumeration sequence. This would require thousands of samples be drawn at each iteration of the hybrid imputation procedure. Clearly, this is not feasible to implement.

Alternatives include approximations to the Bayes factor using either Laplace's method (Kass & Raftery, 1995; Tierney & Kadane, 1986) or the BIC (Schwarz, 1978). Laplace's method relies on asymptotic theory in that the log-posterior distribution is assumed to be well approximated by a quadratic approximation centered at the maximum likelihood estimate with a curvature given by the observed information matrix. The BIC is a further approximation for the posterior model probability. The BIC approximation does not require that the observed information matrix be calculated, which makes it the easiest to implement in practice. The posterior class probability given the BIC for a model fit with  $K$  components is given as

$$\Pr(\mathcal{M}_K | Y^{\text{obs}}, X) \approx \frac{\exp\left(-\frac{1}{2} \text{BIC}(\mathcal{M}_K)\right)}{\sum_{\mathcal{M}_{K^*}} \exp\left(-\frac{1}{2} \text{BIC}(\mathcal{M}_{K^*})\right)} \quad (3.10)$$

where  $\text{BIC}(\mathcal{M}_K)$  are the BIC values obtained by fitting a model with  $K$  components to observed indicators  $Y^{\text{obs}}$  conditional on the auxiliary variables at a given iteration,  $X^{(t)}$ .

Despite the simplicity of the BIC approximation, it is known to be a poor approximation for the posterior model probability. The label switching issue implies that the posterior is not unimodal even in large samples, as is assumed by the BIC approximation. Additionally, when an overfitted mixture model is fit to the data, the quadratic approximation is no longer tenable because the observed information matrix becomes singular (Drton & Plummer, 2017; Gelman, Hwang, & Vehtari, 2014; Yamazaki & Watanabe, 2003). In practice, these complications imply that model selection based on the BIC only performs well when the true model is one considered in the set of models fit to the data. Moreover, as we will show in Chapter 4, we find that the

BIC tends to under extract the number of classes when the model is fit to the observed data, as is done in our proposed hybrid imputation procedure. Under extraction is especially problematic because it would lead to imputations being drawn from an uncongenial model. In contrast to the BIC, the AIC (Akaike, 1974) results in consistent density estimation and is not prone to under extracting the number of classes (Frühwirth-Schnatter et al., 2019, p. 123; McLachlan & Peel, 2004, p. 201; Nylund, Asparouhov, & Muthén, 2007; Tofghi & Enders, 2008). For this reason alone, we approximate the posterior class probabilities using Akaike weights given by

$$\Pr(\mathcal{M}_K | Y^{\text{obs}}, X) \approx \frac{\exp\left(-\frac{1}{2} \text{AIC}(\mathcal{M}_K)\right)}{\sum_{\mathcal{M}_{K^*}} \exp\left(-\frac{1}{2} \text{AIC}(\mathcal{M}_{K^*})\right)}. \quad (3.11)$$

## Summary

In summary, in this section we modified the hybrid imputation procedure with the intention to adequately reflect model uncertainty in the imputations. Specifically, our proposed method incorporates an enumeration step to define a set of competing models. A reference model is then selected among the competing models using Akaike weights. This hybrid imputation procedure is detailed in Algorithm 2 (Table 3.4).

## Applied Example: Case Study with ECLS-K 1998 Data

We now apply our hybrid imputation procedure to a real-world dataset in order to conduct an initial investigation of its performance. In a pedagogical introduction to LPA for applied researchers, Berlin, Williams, & Parra (2014) demonstrated how to identify different profiles characterizing nutritional, physical activity, and sedentary behaviors among Black, non-Hispanic adolescents in the eighth grade. We build on and modify this example to investigate the performance of our proposed hybrid imputation procedure.

As with Berlin et al. (2014), we utilized data from the Early Childhood Longitudinal Study, Kindergarten (ECLS-K) 1998 cohort to serve as an illustrative example. The ECLS-K was a nationally representative, longitudinal study following more than 21,000 children enrolled in either full-time or part-time Kindergarten in the 1998-1999 school year. The ECLS-K study followed the children to the end of eighth grade, with observations on a wide range of achievement, behavioral, psychological, and school-environment outcomes being collected up to two times per school year.

### **Research Context**

Berlin et al. (2014) point out that over the past several decades, many studies have examined how nutrition and physical activity relate to body mass index (BMI) and differ across populations of adolescents, with several studies focusing on differences across sex, race, and ethnicity. As stated by Berlin et al. (2014), previous studies have found that Black, non-Hispanic youth are at an increased risk of being considered overweight or being diagnosed as obese (Davison & Birch, 2001; Ogden, Carroll, Kit, & Flegal, 2012). Medical and public health researchers have been interested in examining the environmental factors that contribute to these observations, including some hypothesizing that school-based determinants influence these disparities. School-based determinants in previous research include the availability of unhealthy foods and drinks, participation in physical education, and access to extracurricular intramurals or sports. Each of these determinants have previously been found to be associated with BMI in adolescents (Dennison, Erb, & Jenkins, 2002; Feng, Reed, Esperat, & Uchida, 2011; Fox, Dodd, Wilson, & Gleason, 2009; Hollar et al., 2010; Janssen & Leblanc, 2010).

Beyond school-based determinants such as PE, researchers are also investigating how an individual's psychological functioning within the school environment contributes to between-group differences in obesity rates. The hypothesis considered by Berlin et al. (2014) is that the population of eighth graders represented in the ECLS-K is comprised of subpopulations defined by differences in physical activity, sedentary behaviors, and healthy dietary intake; the authors tested this hypothesis using LPA.

## **Measures**

### ***Indicator Variables***

Indicator variables included responses to items on a Likert scale that measure weekly physical activity, sedentary behaviors, and dietary intake. Three items measured physical activities, including participation in school sports (three categories), participation in non-school sports (four categories), days exercised in the past seven days (0-7 days; eight categories), and average days in PE per week (0-5 days; six categories). Six items measured sedentary behaviors by asking the number of hours per day (0-24 hours; 25 categories) spent watching TV, playing videogames, or using the internet. Nine items measured dietary intake by asking the number of days (0-7 days; eight categories) specified foods (e.g., carrots, potatoes, fruit, fast food, etc.) or drinks (e.g., a glass of milk, a glass of juice, drank soda, etc.) were consumed.

We made the following modification to Berlin's (2014) pedagogical example: instead of relying on the raw responses to the items, we instead parceled items to construct three profile indicators that measure overall physical activity, sedentary behaviors, and dietary intake. We made this decision because LPA assumes the indicators are continuous and are not ordinal (as is the format of the raw responses). We constructed

the ACTIVITY parcel by standardizing all items before aggregating. The resulting ACTIVITY parcel exhibited an average inter-item correlation of  $r = .19$  (range = .13-.26;  $\alpha = .48$ ). Although the reliability for the ACTIVITY score would not be sufficient in a substantive research study, for the purposes of illustrating a statistical technique with a pedagogical example, we believe the reliability is adequate. We constructed the SEDENTARY and DIETARY parcels using the items in their raw scales because all items were on the same scale (i.e., hours per day or days per week). Further, all of the items used to construct the SEDENTARY and DIETARY parcels exhibited positive inter-item correlations (SEDENTARY:  $M = .36$ , range = .35-.37 ; DIETARY:  $M = .20$ , range = .19-.22), and we concluded that both exhibited acceptable levels of internal consistency reliability for the purposes of an empirical example (SEDENTARY:  $\alpha = .77$ , DIETARY:  $\alpha = .59$ ). Histograms of parcel scores for the analytic sample are shown in Figure 3.1. Of note are bumps in density function on the positive end of the tails of the distribution for SEDENTARY and DIETARY, as well as the bump in the density function on the negative end of the tails for the ACTIVITY distribution. These bumps may be suggestive of a subpopulation higher than average on SEDENTARY and DIETARY parcel scores and lower than average on ACTIVITY parcel scores.

### *Auxiliary Variables*

We included several AVs which either demonstrated that they were predictive of profile membership in the previous LPA studies conducted by Berlin et al. (2017, 2014), or external variables that were observed to have an association with one or more of the profile parcel indicators. AVs include FEMALE, indicating whether the student identified as female; BMI, providing the body mass index of the student; TVROOM, indicating

whether the student had a television in their bedroom (1=yes; 0=no); and SES, providing a continuous measure of the student's family's socioeconomic status. School-level AVs include school urbanicity (URBAN) and indicators for whether high-sugar drinks (DRINKS), sweets and candy (SWEETS), and other unhealthy snacks (SNACKS) were available at school.

We also included other psychosocial measures as AVs, including continuous measures of externalizing (SDQEXT) and internalizing (SDQINT) behaviors. Finally, we included measures of socioemotional constructs, including the student's self-concept (CONCPT), perceived locus of control (LOCUS), and feeling of fitting in at school (FITIN) as additional AVs.

### **Sample and Missing Data Mechanism**

Following Berlin (2014), only data for individuals who identify as Black, non-Hispanic in eighth grade were analyzed in the illustrative example. However, because the intent of this study is to compare missing data strategies, we made the following inclusion criterion: observations required complete item-level responses on the profile indicators. We made no exclusions based on missingness on the AVs because the AVs should not directly inform class membership, even if they are predictive of the latent classes. Applying these inclusion and exclusion criteria results in a sample size of  $N = 608$  girls and boys. In total, 20.7% of the observations were complete across all AVs. Missingness in the individual AVs ranged from 0-66.3% ( $M = 6.4\%$ ). For the empirical example, missingness was induced on the ACTIVITY, DIET, and SEDENTARY parcels using the `ampute` function as part of the `mice` R package. The missing data mechanism was specified so that missing indicator values were dependent on the student's BMI. An

overall missingness rate of the indicator variables was set at 50%, with each variable missing approximately one-quarter of its observations.

### **Model Selection and Class Definitions**

Best practices for the class enumeration process and the decision on the number of subpopulations supported by the data continues to be an active area of research. Some simulation studies purport that the BIC (Nylund et al., 2007) and parametric bootstrap likelihood ratio tests (McLachlan, 1987; McLachlan & Peel, 2004) perform best in identifying the number of classes; others have reported that the aBIC is superior (Tofighi & Enders, 2008). Given the ambiguities from simulation results, methodologists recommend that researchers settle on the number of classes by taking a holistic approach, informed by balancing statistical considerations, such as information criteria and hypothesis tests, with substantive considerations, such as the formation of very small and uninteresting classes if fitting a model with too many mixture components.

Although applied researchers have generally embraced the concept of balancing multiple statistical criteria with substantive considerations in published manuscripts employing LPA, there has been little consideration with regards to the within-class covariance structure. Masyn (2013) suggests that four alternative structures be examined, including (1) class-invariant, diagonal, (2) class-varying diagonal, (3) class-invariant, unrestricted, and (4) class-varying, unrestricted covariance matrices. Thus, the model selection process involves identifying the number of classes supported by the data by considering different within-class covariance structures. To provide structure to such a process, Masyn (2013, p. 591) recommends a two-stage procedure for model selection. In the first stage, classes are enumerated separately across the four different covariance

structures to arrive at four candidate models. The second stage involves deciding among the four candidate models.

Information criteria, likelihood ratio tests, and other relevant information for model selection across the four covariance structures is displayed in Table 3.6. Also provided is a brief explanation about why each model was selected. We settled on  $K = 3$  classes within each structure, with the exception of the class-varying, diagonal structure, for which we settled on  $K = 4$  classes.

Next, to settle on a final model, we first employed chi-squared difference testing to test which of the three covariance structures was best supported among the three candidates, for which we settled on a  $K = 3$  classes (see Table 3.7). Although chi-square difference testing can be accomplished for testing nested models fit with the same number of classes, such a test cannot be performed for models with a different number of classes because of violations of the regularity conditions (McLachlan & Peel, 2004). The data supported a class-varying, unrestricted within-class covariance structure  $K = 3$  solution over a class-invariant, diagonal structure ( $p < .001$ ) or a class-invariant, unrestricted structure ( $p < .001$ ). Thus, the final selection decision rested on choosing among the  $K = 4$  solution with a class-varying, diagonal structure or the  $K = 3$  decision with a class-varying, unrestricted structure.

We selected the  $K = 3$  solution with a class-varying, unrestricted structure as the final model. Although the model specified with a  $K = 4$  class-varying diagonal structure displayed better model fit, the  $K = 3$  solution resulted in classes that were substantively more meaningful. Profiles with class-specific means plotted across the ACTIVITY,

SEDENT, and DIET parcels in the complete data case are displayed in Figure 3.2. We name our classes as follows:

1. “Balanced” class (63.5%). Characterized by about-average levels of physical activity but exhibited both less sedentary behaviors and less dietary intake than the active-recovery class.
2. “Active-recovery” class (31.8%). Higher-than-average physical activity levels and sedentary behaviors.
3. “At risk” class (4.6%). Characterized by very high sedentary behaviors, high dietary intake, and substantially lower-than-average activity levels.

## **Results**

The profiles of estimated class-specific means when using a FIML estimation strategy, CART multiple imputation, and the proposed hybrid imputation procedure are displayed in Figure 3.2. As expected, we found a substantial discrepancy in the “at risk” class between estimated means when the missing data are treated with FIML compared to the complete data solution. In particular, the mean estimate for dietary consumption and activity is substantially more positive for the FIML estimate as compared to the mean estimates when fitting the complete data in the “at risk” class. As expected, due to the small sample size, the performance of MI by CART imputation was mixed. CART imputation better recovered mean values in dietary consumption in the “at risk” class; however, it failed to adequately recover the mean values for activity and sedentary behaviors.

Surprisingly, the performance of the MI by hybrid imputation was also mixed. As with CART imputation, the proposed hybrid imputation strategy better recovered mean

estimates for the “at risk” class for dietary intake. However, hybrid imputation failed to recapture corresponding mean values for activity levels and sedentary behaviors. Additionally, Table 3.8 displays the discrepancies between the complete data standardized means and the estimate obtained from each model. While hybrid imputation was superior for recovering means across the indicators for the “Balanced” class, surprisingly, FIML generally better recovered means of the indicators for the two other classes. We discuss the implication of these mixed findings next.

## **Discussion**

We have investigated a proposed hybrid imputation strategy that incorporates a mixture regression model to impute class indicators that selects the number of classes iteratively so that model uncertainty is reflected in the imputations. We compared the performance of our proposed strategy to FIML and CART imputation using an empirical example from the ECLS-K dataset. We found surprisingly mixed results, with hybrid imputation better recapturing class-specific means on some indicators but FIML better recapturing the means for the indicators in most of the cases.

To understand these results, we first ruled out that the poor performance may be the result of data being generated from an uncongenial model. In particular, we would have expected poor performance if imputations were being generated from a mixture model with fewer than the true number of components. This is because such a model is less complex than the true data generating mechanism, and congeniality requires that the imputation model be at least as complex as the data generating mechanism. Figure 3.3 illustrates the number of classes selected at each iteration of the MCMC chain. Our model selection procedure resulted in a model that is at least as complex in the vast

majority of iterations (99.3%). This result is to be expected because the AIC is known to over extract the number of classes. In summary, the mixed results were not explained by selecting a model with too few components.

We believe there are several modifications we can make in future work to improve the performance of the hybrid imputation strategy. This includes incorporating prior information to circumvent unbounded likelihoods and mixture model averaging to utilize all information from overfit mixture models. We discuss each of these two modifications in more detail.

### **Incorporating Prior Information to Improve Estimation Stability**

We often encountered convergence issues, in that the likelihood function behaved as if it were unbounded and unstable when fitting mixture models at each iteration of the MCMC chain. An unbounded likelihood is a known issue when the within-class variance covariance matrix,  $\Sigma_k$ , is not constrained (McLachlan & Peel, 2004, p. 4). In future work, we will specify weakly informative priors on the class-specific variance-covariance matrix in order to remedy this issue and allow for more stable estimation to proceed.

In addition, priors can be specified on  $\alpha$  parameters that define the weights of the mixture,  $\pi(\mathbf{x}_i)$ . In fact, the posterior distribution of parameters from an overfitted mixture has much more stable behavior than the likelihood function when the priors on the marginal probability parameters are sufficiently shrunk towards zero (Celeux, Fruewirth-Schnatter, et al., 2018, p. 142; Rousseau & Mengersen, 2011). Specifying priors may lead to more stable behavior in sampling the mixture regression model parameters obtained for the joint modeling step.

## Mixture Model Averaging

In addition to incorporating prior information, we plan to modify our proposed hybrid imputation procedure in a manner that incorporates mixture model averaging (Wei & McNicholas, 2015). Mixture model averaging involves selecting a reference model and then incorporating all available information from models fit with more components to the reference model to improve classification quality. In this way, all relevant information could inform class membership during each iteration of the MCMC chain.

Mixture model averaging proceeds by mapping (or merging) the clusters from the more complex mixture model to the clusters from the simpler reference model. This is done by considering different merging combinations and selecting the combination that minimizes the Rand index. Posterior class probabilities following mixture model averaging are then computed through a weighted average of the class probabilities from each overfitted model, weighted by the posterior model probabilities. Wei & McNicholas (2015) find that mixture model averaging improves the classification quality of the resulting clusters. The improvement in classification quality likely would result in improved imputation efficiency, which is important to counteract the inefficiency caused by relying on the AIC, as it results in an overfit (but congenial) mixture model. Mixture model averaging is a Bayesian model averaging procedure and, therefore, is a more direct extension of the Kaplan & Yavuz (2019) imputation approach to account for model uncertainty than the hybrid procedure we have proposed.

In conclusion, identifying methods that produce proper imputations in small-sample settings (e.g.,  $N = 300$ ) or in large samples (e.g.,  $N = 1,200$ ) when a small class is present (e.g., a class that represents 10% of the population) and class separation is weak

(e.g., entropy is approximately .74) is an important step for multiple imputation to become a mainstream strategy for treating missing data in person-centered analyses like LPA. We investigated whether a proposed a hybrid imputation procedure that incorporates mixture modeling would perform well in these settings. We were surprised to find that the proposed procedure poorly recaptured the class-specific means in an illustrative example using the ECLS-K. In response, we outlined several modifications we plan to make for future work

## Tables

**Table 3.1**

*Summary of EM with Sampling*

EM with Sampling
1. Identify a multivariate distribution to model the complete data, $Y X$ .
2. Bootstrap the data.
3. Fit the multivariate model in Step 1 using the EM algorithm to the bootstrapped data.
4. Sample from the conditional distribution $Y^{\text{mis}} Y^{\text{obs}}, X$ using the parameter estimates obtained from Step 3.
5. Repeat Steps 2-4 many times to form an MCMC chain.

**Table 3.2**

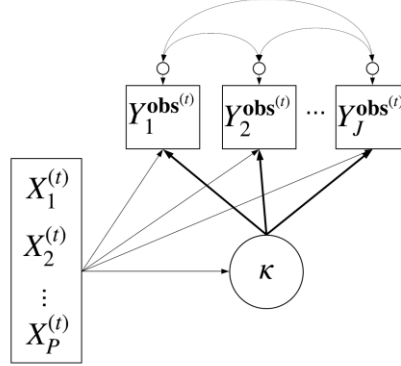
*Algorithm 1: Proposed Hybrid MICE Procedure when  $\mathcal{M}_K$  is Known*

- 
1. Specify an imputation model for the missing data correlates,  $[X_p^{\text{mis}} | X_p^{\text{obs}}, X_{-p}, Y, \kappa, \phi_p, R]$  for each  $p = 1, \dots, P$
  2. **Initialize  $\kappa$ .** Fit a  $k$  class mixture model to  $Y^{\text{obs}}$ . Save the corresponding parameter estimates  $\hat{\theta}^{(0)}$ . Use  $\hat{\theta}^{(0)}$  to obtain the estimated posterior class probability vector for each individual,  $\tau_i^{(0)}$ . For each individual, sample  $\kappa_i^{(0)} \sim \text{Multinomial}(\tau_i^{(0)})$  to obtain  $\kappa^{(0)}$ .
  3. **Initialize  $Y^{\text{mis}}$ .** Using  $\hat{\theta}^{(0)}$  and  $\kappa^{(0)}$ , sample from the conditional predictive distribution  $Y^{\text{mis}(0)} \sim [Y^{\text{mis}} | Y^{\text{obs}}, \hat{\theta}^{(0)}, \kappa^{(0)}]$  using EM with sampling or alternative multivariate imputation procedure. Note that in the initialization step, the predictive distribution need not be conditional on the missing data correlates.
  4. **Initiate  $X$ .** Stratified by classes as specified by  $\kappa^{(0)}$ , fill in starting imputations  $X_p^{(0)}$  by random draws from the observed values  $X_p^{\text{obs}}$ .
  5. Repeat for iterations  $t = 0, \dots, T$ 
    - a. **JM step.** Impute the class indicators using the EM with sampling procedure outlined in Subroutine 1.
    - b. **FCS step.** For each  $p = 1, \dots, P$ 
      - i. Sample weights from  $\mathbf{w}_p^{(t+1)} \sim \text{Dirichlet}(N, \mathbf{1})$ .
      - ii. Fit the imputation model identified in Step 1 with sampling weights  $\mathbf{w}_p^{(t+1)}$  to obtain  $\hat{\phi}_p^{(t+1)}$
      - iii. From the fitted model, impute  $X_p^{\text{mis}}$  by sampling from the predictive distribution implied in Step 1.
-

**Table 3.3**

*Subroutine 1: Single Iteration of EM with Sampling Procedure for Sampling  $Y^{\text{mis}}$*

1. **Sample weights** from  $\mathbf{w}_0^{(t+1)} \sim \text{Dirichlet}(N, \mathbf{1})$ .
2. **Sample sufficient statistics** by fitting the mixture regression model below with  $\mathbf{w}_0^{(t+1)}$  specified as sampling weights and obtaining the corresponding MLE,  $\hat{\theta}^{(t+1)}$ .



3. Obtain posterior class probabilities  $\hat{t}_{ik}^{(t+1)}$  for each observation using  $\hat{\theta}^{(t+1)}$  and  $X^{(t)}$  via

$$\hat{t}_{ik}^{(t+1)} = \frac{\hat{\pi}_{ik}^{(t+1)} \mathcal{N}_j(\hat{\mu}_{ik}^{(t+1)}, \hat{\Sigma}_k^{(t+1)})}{\sum_{k=1}^K \hat{\pi}_{ik}^{(t+1)} \mathcal{N}_j(\hat{\mu}_{ik}^{(t+1)}, \hat{\Sigma}_k^{(t+1)})}$$

where

$$\hat{\pi}_{ik}^{(t+1)} = \begin{cases} \frac{\exp(\mathbf{x}_i^{(t+1)} \hat{\mathbf{a}}_k^{(t+1)})}{1 + \sum_{k=1}^{K-1} \exp(\mathbf{x}_i^{(t+1)} \hat{\mathbf{a}}_k^{(t+1)})} & \text{if } k < K \\ \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\mathbf{x}_i^{(t+1)} \hat{\mathbf{a}}_k^{(t+1)})} & \text{if } k = K \end{cases}$$

and

$$\hat{\mu}_{ik}^{(t+1)} = \mathbf{x}_i^{(t+1)} \hat{\boldsymbol{\beta}}_k^{(t+1)}$$

4. **Sample class membership.** For each individual, sample  $\kappa_i^{(t+1)} \sim \text{Multinomial}(\hat{t}_i^{(t+1)})$
5. **Sample  $Y^{\text{mis}}$ .** Using  $\hat{\theta}^{(t+1)}$  and  $\kappa_i^{(t+1)}$ , sample plausible values for  $Y^{\text{mis}}$  by sampling from the posterior predictive distribution  $[Y^{\text{mis}} | Y^{\text{obs}}, \hat{\theta}^{(t+1)}, \boldsymbol{\kappa}^{(t+1)}, X^{(t)}, R]$  where

$$\begin{aligned} [Y_i^{\text{mis}} | Y_i^{\text{obs}}, \hat{\theta}^{(t+1)}, \kappa_i^{(t+1)}, X^{(t)}, R] &\sim \mathcal{N}_j(\hat{\mu}_{ik}^{\text{mis}}, \hat{\Sigma}_{ik}^{\text{mis}}), \\ \hat{\mu}_{ik}^{\text{mis}} &= \hat{\mu}_{i,k}^{(t+1)}[\setminus \mathbf{r}_{Y,i}] \\ &\quad + \hat{\Sigma}_k^{(t+1)}[\setminus \mathbf{r}_{Y,i}, \mathbf{r}_{Y,i}] (\hat{\Sigma}_k^{(t+1)})^{-1} [\mathbf{r}_{Y,i}, \mathbf{r}_{Y,i}] (\mathbf{y}_i^{\text{obs}} \\ &\quad - \hat{\mu}_{i,k}^{(t+1)}[\mathbf{r}_{Y,i}]) \\ \hat{\Sigma}_{ik}^{\text{mis}} &= \hat{\Sigma}_k^{(t+1)}[\mathbf{r}_{Y,i}, \mathbf{r}_{Y,i}] - \hat{\Sigma}_k^{(t+1)}[\setminus \mathbf{r}_{Y,i}, \mathbf{r}_{Y,i}] (\hat{\Sigma}_k^{(t+1)})^{-1} \hat{\Sigma}_k^{(t+1)}[\mathbf{r}_{Y,i}, \setminus \mathbf{r}_{Y,i}] \end{aligned}$$

and  $\mathbf{r}_{Y,i}$  is a scalar or vector indexing the observed responses so that  $\setminus \mathbf{r}_{Y,i}$  refers to the index of missing responses for individual  $i$ .

**Table 3.4**

*Algorithm 2: Proposed BMA-Hybrid MICE Procedure when  $\mathcal{M}_K$  is Unknown*

---

1.	Conduct Steps (1)-(5) in Algorithm 1
2.	Repeat for iterations $t = 0, \dots, T$ <ol style="list-style-type: none"> <li>a. <b>JM with BMA Step.</b> Impute the class indicators using the EM with sampling procedure outlined in Subroutine 2.</li> <li>b. <b>FCS step.</b> For each <math>p = 1, \dots, P</math> <ol style="list-style-type: none"> <li>i. Sample weights from <math>\mathbf{w}_p^{(t+1)} \sim \text{Dirichlet}(N, \mathbf{1})</math>.</li> <li>ii. Fit the imputation model identified in Step 1 with sampling weights <math>\mathbf{w}_p^{(t+1)}</math> to obtain <math>\hat{\phi}_p^{(t+1)}</math></li> <li>iii. From the fitted model, impute <math>X_p^{\text{mis}}</math> by sampling from the predictive distribution implied in Step 1.</li> </ol> </li> </ol>

---

**Table 3.5**

*Subroutine 2: Proposed BMA-Hybrid MICE Procedure when  $\mathcal{M}_K$  is Unknown*

---

1.	<b>Sample weights</b> from $\mathbf{w}_0^{(t+1)} \sim \text{Dirichlet}(N, \mathbf{1})$ .
2.	<b>Enumeration.</b> Define Occam's window by enumerating the classes and fitting regression mixture models weighted by $\mathbf{w}_0^{(t+1)}$ until convergence issues are reached. This will lead to the set $\{\theta_K^{(t+1)}: 1 \leq K \leq G\}$
3.	<b>Sample Reference Model, <math>H</math>.</b> Among the models in Occam's window, sample the reference model, $\mathcal{M}_H$ , with probability <div style="text-align: center;"> <math display="block">\Pr(H = K) = \frac{\text{AIC}^{(t+1)}(\mathcal{M}_K)}{\sum_{K^*=1}^G \text{AIC}^{(t+1)}(\mathcal{M}_{K^*})}, \quad K = 1, 2, \dots, G</math> </div>
4.	<b>Sample <math>Y^{\text{mis}}</math>.</b> Using the estimates from the reference model, complete Steps (3)-(5) in Subroutine 1.

---

**Table 3.6**

*Class Enumeration: ECLS-K Illustrative Example*

Classes (K)	<i>npar</i>	<i>LL</i>	AIC	CAIC	BIC	adj-BIC	Adj. LMR- LRT <i>p</i> - value	Entropy	Min class proportion ( $\pi_k^{min}$ )
<u>Class-invariant, diagonal</u>									
1	6	-2779.32	5570.65	5603.11	5597.11	5578.06			
2	10	-2673.30	5366.60	5420.70	5410.70	5378.95	0.551	0.97	0.037
3	14	-2622.97	5273.93	5349.68	5335.68	5291.23	0.015	0.94	0.017
4	18	-2584.73	5205.46	5302.85	5284.85	5227.70	0.234	0.93	0.008
5	22	-2561.06	5166.11	5285.13	5263.13	5193.29	0.032	0.90	0.007
6	26	-2539.27	5130.54	5271.20	5245.20	5162.66	0.022	0.88	0.005
7	30	-2527.78	5115.57	5277.87	5247.87	5152.63	0.301	0.89	0.005
8	34	-2521.80	5111.59	5295.54	5261.54	5153.60	0.691	0.76	0.005
<u>Class-varying, diagonal</u>									
1	6	-2779.32	5570.65	5603.11	5597.11	5578.06			
2	13	-2588.33	5202.67	5273.00	5260.00	5218.73	<0.001	0.61	0.219
3	20	-2552.54	5145.08	5253.29	5233.29	5169.79	0.251	0.59	0.070
4	27	-2530.60	5115.20	5261.27	5234.27	5148.55	0.035	0.60	0.044
5	34	-2511.23	5090.46	5274.40	5240.40	5132.46	<0.001	0.67	0.022
<u>Class-invariant, unrestricted</u>									
1	9	-2769.51	5557.02	5605.72	5596.72	5568.14			
2	13	-2663.00	5352.00	5422.34	5409.34	5368.06	0.501	0.98	0.031
3	17	-2611.84	5257.68	5349.65	5332.65	5278.68	0.006	0.94	0.016
4	21	-2578.24	5198.48	5312.09	5291.09	5224.42	0.155	0.91	0.016
5	25	-2557.79	5165.58	5300.84	5275.84	5196.47	0.043	0.91	0.005
6	29	-2537.03	5132.07	5288.96	5259.96	5167.89	0.032	0.88	0.005
<u>Class-varying, unrestricted</u>									
1	9	-2769.51	5557.02	5605.72	5596.72	5568.14			
2	19	-2578.23	5194.45	5297.25	5278.25	5217.92	<0.001	0.68	0.232
3	29	-2532.32	5122.63	5279.53	5250.53	5158.46	0.150	0.76	0.046
4	39	-2509.31	5096.63	5307.62	5268.62	5144.81	0.270	0.67	0.046

*Notes.* Highlight indicates candidate models across the class-specific variance-covariance specifications.

**Table 3.7***Chi-square Difference Tests: ECLS-K Illustrative Example*

	H0 Model	$\chi^2$	df	p
<u>2 Profiles</u>				
1	Class-invariant, diagonal	43.38	9	<0.001
2	Class-varying, diagonal	27.62	6	<0.001
3	Class-invariant, unrestricted	34.86	6	<0.001
<u>3 Profiles</u>				
4	Class-invariant, diagonal	426.62	15	<0.001
5	Class-varying, diagonal	228.69	9	<0.001
6	Class-invariant, unrestricted	376.31	12	<0.001
<u>4 Profiles</u>				
7	Class-invariant, diagonal	282.62	21	<0.001
8	Class-varying, diagonal	39.69	12	<0.001
9	Class-invariant, unrestricted	294.11	18	<0.001

Notes. With the exception of (1) and (3) which employed a Wald test due to negative test statistic values, Satorra-Bentler scaled chi-square statistic was employed for all hypothesis testing. Model selected through enumeration highlighted.

**Table 3.8***Discrepancies Against Complete Data Estimates:**ECLS-K Illustrative Example*

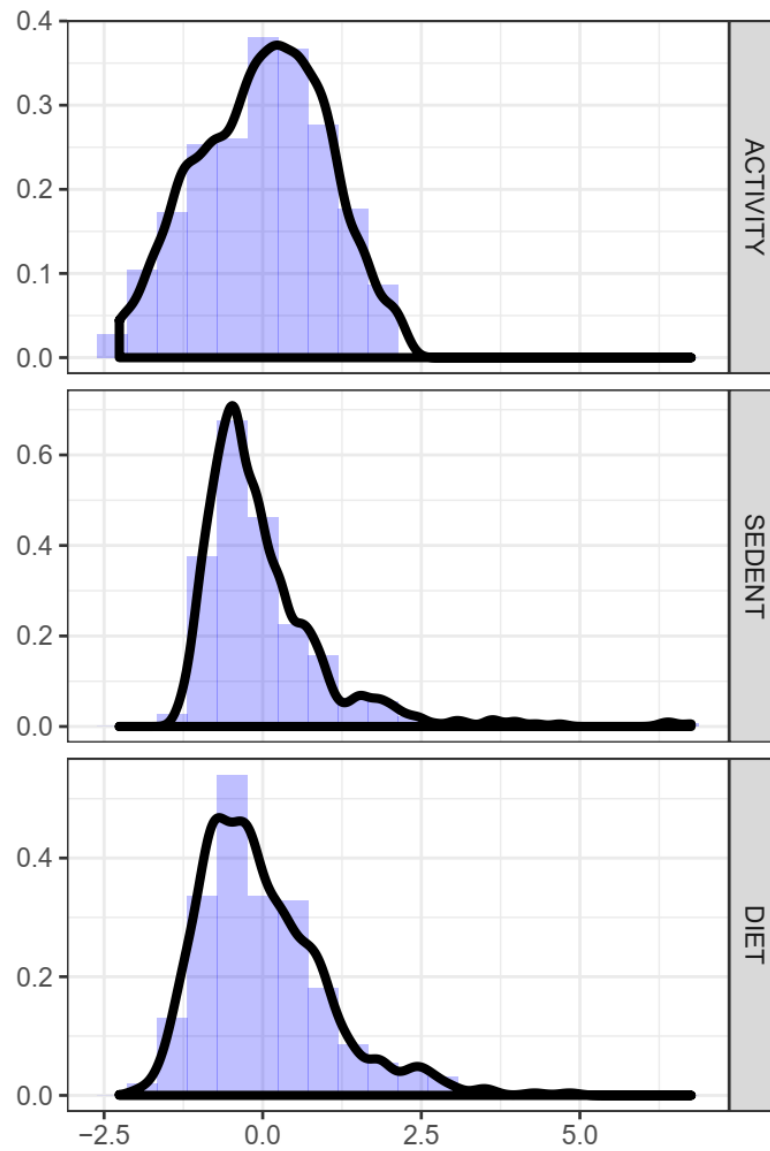
	ACTIVITY	SEDENT	DIET
<u>Class 1: "Balanced"</u>			
FIML	-0.032	-0.052	-0.081
CART	-0.021	-0.049	-0.093
Hybrid	<b>-0.005</b>	<b>0.016</b>	<b>-0.011</b>
<u>Class 2: "Active-recovery"</u>			
FIML	<b>-0.043</b>	<b>-0.008</b>	<b>-0.133</b>
CART	-0.163	-0.079	-0.224
Hybrid	-0.069	-0.136	0.187
<u>Class 3: "At risk"</u>			
FIML	<b>0.506</b>	<b>-0.098</b>	1.467
CART	1.098	-0.75	0.66
Hybrid	0.738	-1.232	<b>-0.41</b>

Notes. Standardized units. Best performing model is bolded.

## Figures

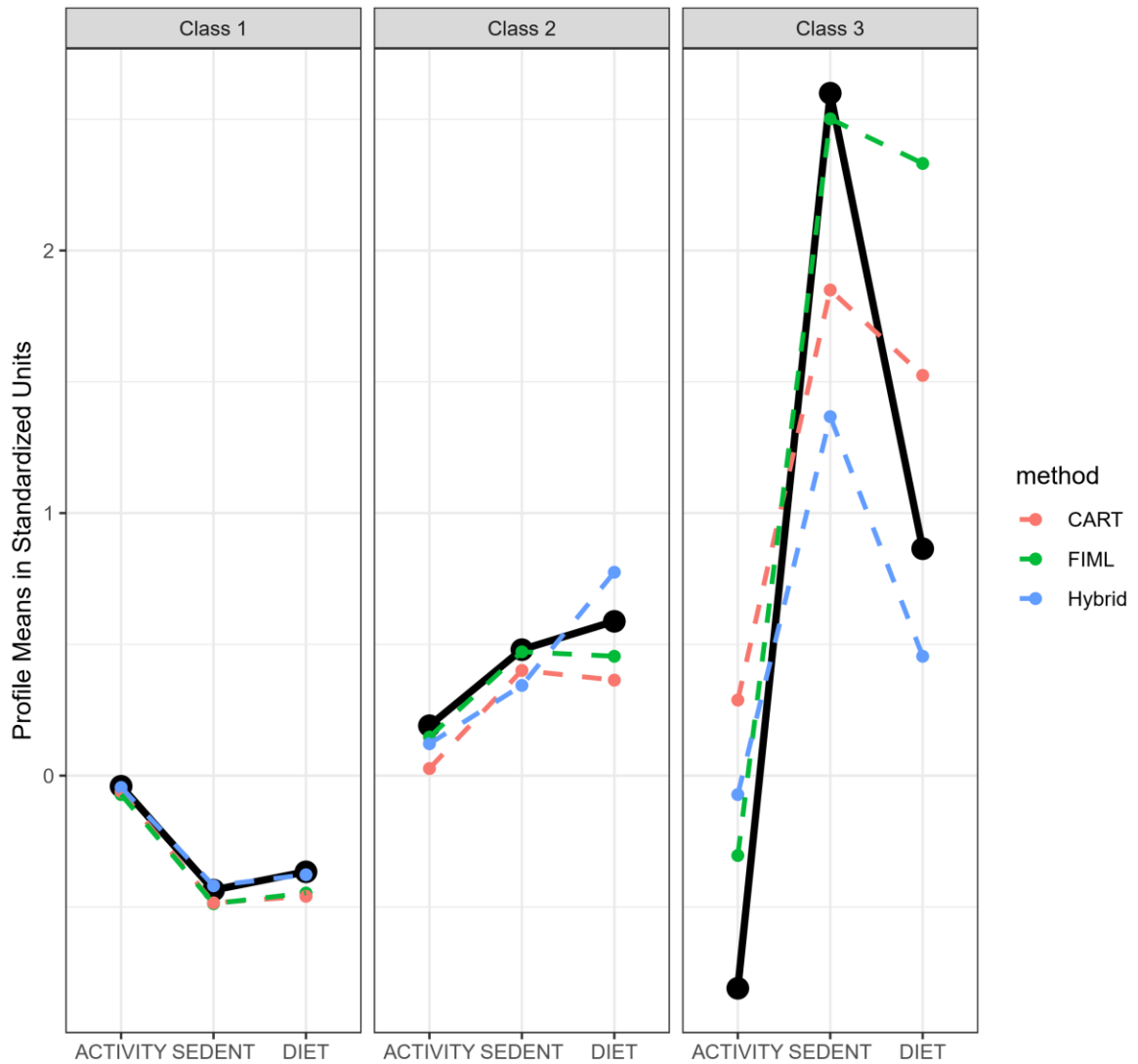
**Figure 3.1**

*Standardized Parcel Scores: ECLS-K Illustrative Example*



**Figure 3.2**

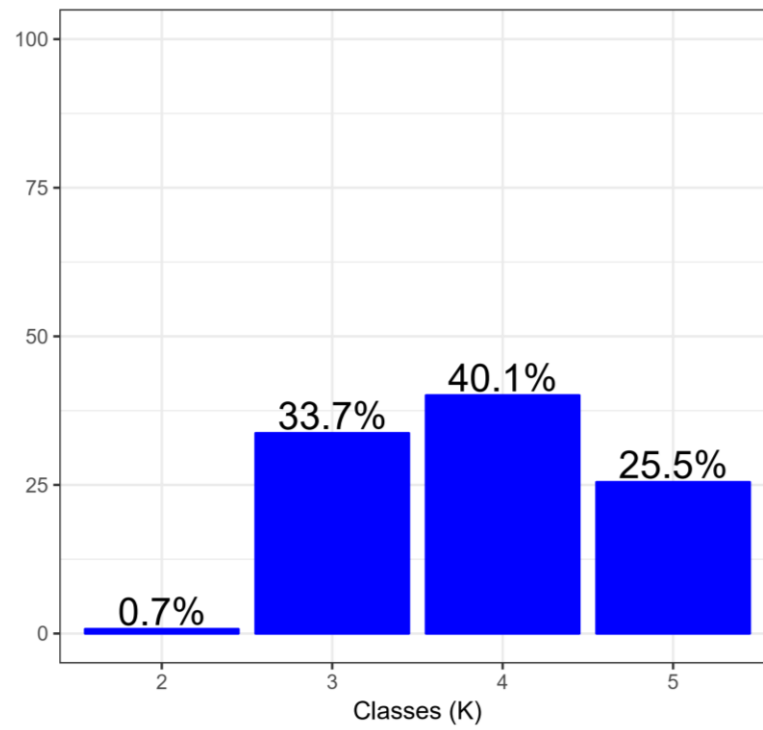
*Profile Plot by Imputation Method: ECLS-K Illustrative Example*



*Notes.* Profiles constructed by plotting class-specific means across the parcels. Class 1 corresponds to the “Balanced” profile, Class 2 corresponds to the “Active-recovery” profile, and Class 3 corresponds to the “At-risk” profile. Complete data results shown by solid black line. Dashed lines indicate solutions given missing data according to an imputation method.

**Figure 3.3**

*AIC Imputation Model Selection*



*Notes.* MCMC chain contained 500 iterations.

## References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Asparouhov, T., & Muthén, B. O. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3), 329–341.
- Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2), 332–353. <https://doi.org/10.1198/jcgs.2010.08111>
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. (J. O. Berger, Ed.) (2nd ed.). New York: Springer-Verlag.
- Berlin, K. S., Kamody, R. C., Thurston, I. B., Banks, G. G., Rybak, T. M., & Ferry, R. J. (2017). Physical activity, sedentary behaviors, and nutritional risk profiles and relations to body mass index, obesity, and overweight in eighth grade. *Behavioral Medicine*, 43(1), 31–39. <https://doi.org/10.1080/08964289.2015.1039956>
- Berlin, K. S., Williams, N. A., & Parra, G. R. (2014). An introduction to latent variable mixture modeling (part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology*, 39(2), 174–187. Retrieved from <http://dx.doi.org/10.1093/jpepsy/jst084>
- Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3), 167–174. <https://doi.org/10.2307/2685208>
- Celeux, G., Fruewirth-Schnatter, S., & Robert, C. P. (2018). Model selection for mixture models - perspectives and strategies. In S. Frühwirth-Schnatter, G. Celeux, & C. P. Robert (Eds.), *Handbook of mixture analysis*. New York, NY: CRC Press.
- Celeux, G., Kamary, K., Malsiner-Walli, G., Marin, J.-M., & Robert, C. P. (2018). Computational solutions for bayesian inference in mixture models. In *Handbook of mixture analysis*. CRC Press.
- Chernick, M. R. (2011). *An introduction to bootstrap methods with applications to R*. (R. A. LaBudde, Ed.). Hoboken, New Jersey: Wiley.
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351. <https://doi.org/10.1037/1082-989X.6.4.330>
- Davison, K. K., & Birch, L. L. (2001). Childhood overweight: a contextual model and recommendations for future research. *Obesity Reviews*, 2(3), 159–171. <https://doi.org/10.1046/j.1467-789x.2001.00036.x>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.
- Dennison, B. A., Erb, T. A., & Jenkins, P. L. (2002). Television viewing and television in

- bedroom associated with overweight risk among low-income preschool children. *Pediatrics*, 109(6), 1028–1035. <https://doi.org/10.1542/peds.109.6.1028>
- Doove, L. L., van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72, 92–104.
- Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2), 323–380.
- Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89(426), 463–475. <https://doi.org/10.1080/01621459.1994.10476768>
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.
- Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, 18(1), 35–54.
- Feng, D., Reed, D. B., Esperat, M. C., & Uchida, M. (2011). Effects of TV in the bedroom on young hispanic children. *American Journal of Health Promotion*, 25(5), 310–318. <https://doi.org/10.4278/ajhp.080930-QUAN-228>
- Fong, E., Lyddon, S., & Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. *ArXiv.Org*.
- Fox, M. K., Dodd, A. H., Wilson, A., & Gleason, P. M. (2009). Association between school food environment and practices and body mass index of US public school children. *Journal of the American Dietetic Association*, 109(2, Supplement), S108–S117. <https://doi.org/https://doi.org/10.1016/j.jada.2008.10.065>
- Frühwirth-Schnatter, S. (2006). *Finite mixture and markov switching models* (1st ed. 20). New York, NY: Springer New York : Imprint: Springer.
- Frühwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2019). *Handbook of mixture analysis*. Boca Raton: CRC Press, Taylor & Francis Group.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409. <https://doi.org/10.2307/2289776>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016. <https://doi.org/10.1007/s11222-013-9416-2>
- Gormley, I. C., & Frühwirth-Schnatter, S. (2019). Mixture of experts models. In S. Frühwirth-Schnatter, G. Celeux, & C. P. Robert (Eds.), *Handbook of mixture analysis* (pp. 271–307). CRC Press. <https://doi.org/10.1201/9780429055911-12>
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and

- Bayesian model determination. *Biometrika*, 82(4), 711–732.  
<https://doi.org/10.2307/2337340>
- Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification*, 4(1), 3–34. <https://doi.org/10.1007/s11634-010-0058-3>
- Hollar, D., Messiah, S. E., Lopez-Mitnik, G., Hollar, T. L., Almon, M., & Agatston, A. S. (2010). Healthier options for public school children improves weight and blood pressure in 6- to 13-year-olds. *Journal of the American Dietetic Association*, 110(2), 261–267. <https://doi.org/https://doi.org/10.1016/j.jada.2009.10.029>
- Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1–47.
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87.  
<https://doi.org/10.1162/neco.1991.3.1.79>
- Janssen, I., & Leblanc, A. G. (2010). Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *International Journal Of Behavioral Nutrition And Physical Activity*, 7(1). <https://doi.org/10.1186/1479-5868-7-40>
- Kaplan, D., & Yavuz, S. (2019). An approach to addressing multiple imputation model uncertainty using Bayesian model averaging. *Multivariate Behavioral Research*, 1–15. <https://doi.org/10.1080/00273171.2019.1657790>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), 49–69.
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, N.J.: John Wiley & Sons.
- Liu, J., Gelman, A., Hill, J., Su, Y.-S., & Kropko, J. (2013). On the stationary distribution of iterative imputations. *Biometrika*, 101(1), 155–173.  
<https://doi.org/10.1093/biomet/ast044>
- Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The oxford handbook of quantitative methods* (pp. 551–611). New York, NY: Oxford University Press.  
<https://doi.org/10.1093/oxfordhb/9780199934898.013.0025>
- McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36(3), 318.  
<https://doi.org/10.2307/2347790>
- McLachlan, G. (2008). *The EM algorithm and extensions*. (T. (Thriyambakam) Krishnan, Ed.) (2nd ed.). Hoboken, N.J.: Wiley-Interscience.

- McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York, NY: John Wiley & Sons.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558. <https://doi.org/10.1214/ss/1177010269>
- Muthén, B. O., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3), 639–657. <https://doi.org/10.1111/j.1467-985X.2009.00589.x>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, California: Muthén & Muthén. Retrieved from [https://www.statmodel.com/html\\_ug.shtml](https://www.statmodel.com/html_ug.shtml)
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14(4), 535–569.
- Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2012). Prevalence of obesity and trends in body mass index among US children and adolescents, 1999–2010. *JAMA*, 307(5), 483–490. <https://doi.org/10.1001/jama.2012.40>
- R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from <https://www.r-project.org/>
- Razzak, H., & Heumann, C. (2019). Hybrid multiple imputation in a large scale complex survey. *Statistics in Transition*, 20(4). <https://doi.org/10.21307/stattrans-2019-033>
- Rousseau, J., & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73(5), 689–710. Retrieved from <http://www.jstor.org/stable/41262270>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1978). Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association* (Vol. 1, pp. 20–34). American Statistical Association.
- Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1), 130–134. Retrieved from <http://www.jstor.org/stable/2240875>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Johnson Wiley & Sons. <https://doi.org/10.1002/9780470316696>
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489. <https://doi.org/10.1080/01621459.1996.10476908>
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366.

- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Si, Y., & Reiter, J. P. (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5), 499–521. <https://doi.org/10.3102/1076998613480394>
- Sovilj, D., Eirola, E., Miche, Y., Björk, K.-M., Nian, R., Akusok, A., & Lendasse, A. (2016). Extreme learning machine for missing data using multiple imputations. *Neurocomputing*, 174(PA), 220–231. <https://doi.org/10.1016/j.neucom.2015.03.108>
- Sterba, S. K. (2016). Cautions on the use of multiple imputation when selecting between latent categorical versus continuous models for psychological constructs. *Journal of Clinical Child & Adolescent Psychology*, 45(2), 167–175.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.
- Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393), 82–86. <https://doi.org/10.2307/2287970>
- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Charlotte, NC: Information Age Pub.
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC Press.
- van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76(12), 1049–1064.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.
- van der Palm, D. W., van der Ark, L. A., & Vermunt, J. K. (2016). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification*, 33(1), 52–72. <https://doi.org/http://dx.doi.org/10.1007/s00357-016-9195-5>
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4), 450–469. <https://doi.org/10.1093/pan/mpq025>
- Vermunt, J. K., & Magdison, J. (2016). Upgrade Manual for Latent GOLD 5.1. Retrieved from <https://www.statisticalinnovations.com/wp-content/uploads/UpgradeManual5.1.pdf>

- Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., Andries, L., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38(1), 369–397.
- Vidotto, D., Vermunt, J. K., & Kaptein, M. C. (2015). Multiple imputation of missing categorical data using latent class models: State of the art. *Psychological Test and Assessment Modeling*, (4), 542–576.
- Vidotto, D., Vermunt, J. K., & van Deun, K. (2018). Bayesian multilevel latent class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics*, 43(5), 511–539. <https://doi.org/10.3102/1076998618769871>
- Wedel, M. (2002). Concomitant variables in finite mixture models. *Statistica Neerlandica*, 56(3), 362–375. <https://doi.org/10.1111/1467-9574.t01-1-00072>
- Wei, Y., & McNicholas, P. D. (2015). Mixture model averaging for clustering. *Advances in Data Analysis and Classification*, 9(2), 197–217. <https://doi.org/10.1007/s11634-014-0182-6>
- Yamazaki, K., & Watanabe, S. (2003). Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16(7), 1029–1038.

## **Chapter 4: On Model Selection using Information Criteria with Finite Mixtures in the Presence of Missing Data**

Marcus R. Waldman & Katherine E. Masyn

Education and behavioral science researchers conduct person-centered analysis to identify homogenous subpopulations within a larger heterogenous population. Through the process of identifying these subpopulations, researchers are investigating sources of unobserved heterogeneity that cannot be obtained by traditional variable-centered analytic strategies, such as linear regression models, factor analysis, or structural equation modeling. Some examples of person-centered analysis performed by behavioral researchers include the identification of developmental trajectories for socio-emotional difficulties (McCoy, Jones, Roy, & Raver, 2017), the discovery of emergent patterns of ADHD psychopathologies in school-aged children (Ostrander, Herman, Sikorski, Mascendaro, & Lambert, 2008), and the investigation of distinct typologies of students who choose to drop out of high school (Bowers & Sprott, 2012). The rapid adoption of person-centered analysis in the last decade demonstrates that advancing behavioral research requires statistical methods that do not assume all individuals follow a single relationship, process, or trajectory. Instead, person-centered analysis represents an advancement because the focus of study is on exploring how individuals are systematically similar or different from one another.

Researchers rely heavily on finite mixture models to conduct person-centered analysis. Finite mixture models partition the overall population into a specified number of latent classes; each latent class is assumed to represent a distinct subpopulation of individuals within the larger population. The classes are “latent” because class

membership cannot be ascertained directly, and the mixture models are “finite” because the researcher must specify the number of classes to be modeled. Because the number of classes is rarely known a priori, it must be inferred from the sample through a model selection process. Model selection involves fitting a sequence of finite mixture models with increasingly more classes and then proceeding to select a final model from the sequence by analyzing the model fit information.

The model selection process remains controversial and is a longstanding challenge, even after more than 30 years of sustained research. Over this time period, there has been a myriad of proposed strategies, including traditional frequentist information criteria (IC), newer criteria specifically designed for mixture models (e.g., Drton & Plummer, 2017), *k*-fold cross-validation procedures (Grimm, Mazza, & Davoudzadeh, 2017; He & Fan, 2019), specialized nested model tests (Lo, Mendell, & Rubin, 2001; McLachlan, 1987), measures of classification quality (Celeux & Soromenho, 1996), and alternative ad-hoc approaches that are purported to work well in practice absent a theoretical basis (Celeux, Fruewirth-Schnatter, & Robert, 2018). No single procedure has proven best in simulations, and the current recommendation is that applied researchers synthesize all available statistical information and augment that information with substantive theory to justify and settle on a final model (Masyn, 2013; Ram & Grimm, 2009).

In vetting all the available information for model selection, applied researchers tend to use the Akaike information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978), and the adjusted BIC (aBIC; Sclove, 1987) to justify a final model. Multiple ICs are required because no single IC is best at

accomplishing two complementary but distinct goals: identifying the correct number of subpopulations present in the data (i.e., consistency) and maximizing out-of-sample fit of the multivariate distribution that is implied by the fitted mixture model (i.e., prediction). In particular, the BIC is consistent in that it will lead to correct model selection decisions if the true model is one that is under consideration when sample sizes are sufficiently large. However, the AIC leads to selection decisions that minimize fit to out-of-sample data if the true model is not in the sequence of models actually fit to the data. Therefore, applied researchers must balance the competing goals of consistency and prediction in real-world settings by synthesizing the evidence from each IC when settling on a final model (Masyn, 2013; Ram & Grimm, 2009).

However, the ICs all make simplifying assumptions, rely on asymptotic principles, and assume the data are complete (i.e., non-missing). Consequently, methodologists have conducted simulation studies that are designed to evaluate how consistency and prediction play out in real-world conditions experienced by researchers conducting a person-centered analysis (e.g., Drton & Plummer, 2017; Nylund, Asparouhov, & Muthén, 2007; Tofghi & Enders, 2008). As evidenced by Nylund et al.'s (2007) paper demonstrating the consistency properties of the BIC being cited over 5,000 times since its publication, these simulation studies have played a prominent role in shaping model selection decisions and have complemented our understanding of how consistency and prediction manifest in real-world data.

Even so, simulation studies conducted to date have assumed that data are complete, although missing data is almost inevitable in the behavioral sciences. Presumably, model selection decisions made when data are missing adequately replicate

the decisions that would have been made if the data were fully observed, provided that the researcher employed appropriate missing data strategies. We challenge this assumption and assert that it remains an open question whether model selection decisions will replicate, even if appropriate missing data strategies are employed. Indeed, if model selection decisions fail to replicate in the presence of missing data, then we must understand the causes behind such a phenomenon in order to update our best practices.

There are many reasons why model selection decisions may fail to replicate when the data are missing. Nonresponse bias is the most obvious. If the missing data mechanism is not ignorable because the data are missing not at random (MNAR; Rubin, 1976), then the parameters that define the subpopulations may themselves be biased. As a result of nonresponse bias, the latent classes may exhibit decreased separation, with the limiting case occurring when the classes collapse into a single class. Intuitively, bias that results in decreased class separation, especially collapse, would lead to systematically under extracting the true number of classes, thereby potentially obfuscating important sources of heterogeneity.

A less obvious reason why model selection decisions may not replicate involves sample size. In fact, in small samples with missing data treated with FIML, it is well established that hypothesis tests exhibit inflated Type I error rates. As pointed out by McNeish & Harring (2017), this is because a reference distribution with  $N$  complete observations is assumed. When data are missing, the reference distribution fails to capture the added variability induced by uncertainty associated with the missing data. The problems associated with inference when data are missing and sample sizes are small are not limited to hypothesis tests, however. For example, the penalty term for the BIC

multiplies the number of parameters by  $\log N$ . This penalty term assumes that  $N$  observations are complete. Consequently, in sample sizes experienced by real-world researchers ( $N = 300$ - $1,200$ ), the BIC may over-penalize models when data are missing. We show that the relative magnitude of this over-penalization depends on the fraction of missing information—a prominent quantity in missing data theory. Our simulations suggest that the BIC tends to under extract the true number of classes, especially when sample sizes are small or there exists a small-sized class.

Finally, model selection decisions may not replicate if a multiple imputation strategy is employed and ICs are averaged across the imputed datasets, as is done in leading person-centered software such as Mplus (Muthén & Muthén, 2017) or LatentGOLD (Vermunt & Magdison, 2016). This is because averaging assumes that the average of the model deviances obtained from the imputed datasets is an unbiased estimate for the complete data model deviance. We show that averaging leads to a biased estimate of the complete data deviance. Several alternative model selection procedures to averaging have been proposed (Chaurasia & Harel, 2012; Consentino & Claeskens, 2010; van Buuren, 2012; Wood, White, & Royston, 2008); however, no strategy has been shown to be universally superior (van Buuren, 2018), and none have been evaluated in the context of finite mixture models. To fill this gap in research, we evaluate several alternative model selection procedures when fitting finite mixture models. We show that a stacking strategy, whereby a final model is selected by appending all imputed datasets into a single flat file, results in model selection decisions that better capture the decision that would have been made had the data been complete.

In summary, our simulations indicate that model selection decisions using ICs are not robust to missing data, regardless of whether FIML or multiple imputation is employed. This finding highlights that missing data is an area that has been too long ignored in person-centered analysis. It further identifies an important area for further methodological inquiry, given that researchers in education and psychology are increasingly using mixture models to identify sources of unobserved heterogeneity in individual outcomes, trajectories, and processes.

To be clear, our central argument is not that FIML or multiple imputation are fundamentally limited as missing data strategies in person-centered analysis. Instead, our central argument is that model selection is currently conducted using practices with strong assumptions that are easily violated and highly sensitive to missing data. This is true regardless of whether a FIML or multiple imputation strategy is employed. By gaining a deeper understanding of how the underlying causes of these assumption violations manifest in practice, we seek to bring attention to this issue so that best practices can be updated and remedial measures, like the two we propose, can be taken.

## **Missing Data Approaches to Estimating Finite Mixture Models**

### **Notation and Background**

Finite mixture modeling exploits the law of total probability to model the joint density of individual  $i$ 's data vector,  $\mathbf{y}_i$ , as a weighted average across  $K$  component densities

$$\Pr(\mathbf{y}_i) = \sum_{k=1}^K \Pr(\kappa = k) \Pr(\mathbf{y}_i | \kappa = k) \quad (4.1)$$

where  $\kappa$  is a categorical latent variable representing a latent class, and the family of the conditional distribution  $\Pr(\mathbf{y}_i|\kappa = k)$  is assumed to be known. In this study, we consider a finite mixture of Gaussians where the conditional distribution is assumed to be a multivariate normal distribution

$$[\mathbf{y}_i|\kappa = k] \sim \mathcal{N}_J(\boldsymbol{\mu}_k, \Sigma_k) \quad \forall k = 1, \dots, K \quad (4.2)$$

where the data vector  $\mathbf{y}_i$  is of size  $J$ ,  $\boldsymbol{\mu}_k$  is the  $k$ th component mean vector, and  $\Sigma_k$  is the  $k$ th component variance covariance matrix. A finite mixture of Gaussians is frequently fitted when applied researchers conduct a latent profile analysis (LPA) to cluster individuals and identify homogenous subpopulations within a larger population. Here, the  $J$  “indicators” of the clusters contained in  $\mathbf{y}_i$  and the indicators are generally treated as a continuous random variable.

In LPA, applied researchers typically do not know the value of  $K$  or the structure of the component variance-covariance matrix,  $\Sigma_k$ . It is often observed in practice that if  $\Sigma_k$  includes freely estimated covariance terms, then convergence issues result because the likelihood can be unbounded. Unboundedness is not present and convergence issues are less prevalent in conditional independence models where the component variance-covariance matrix is diagonal and of the form

$$\Sigma_k = \text{diag}(\boldsymbol{\sigma}_k^2) \quad \forall k \in \{1, 2, \dots, K\} \quad (4.3)$$

where  $\boldsymbol{\sigma}_k^2 = [\sigma_{1k}^2, \sigma_{2k}^2, \dots, \sigma_{Jk}^2]$ . For notational convenience, we assume that indicators are conditionally independent, except in the case of single-class models where the indicators can freely covary. We note, however, that conditional independence is a strong assumption, and alternative specifications of the component variance-covariance matrices should be tested when conducting a person-centered analysis in practice (Masyn, 2013).

Under conditional independence, the sequence of models fit to the data can be written as  $\mathcal{M} = \{\mathcal{M}_K: K = 1, 2, \dots\}$  where  $K$  is the number of components. The likelihood of observing the collection of complete  $\mathbf{y}_i$  for all  $N$  individuals, denoted  $Y$ , given a model with  $K$  components is

$$\mathcal{L}(\theta_K|Y, \mathcal{M}_K) = \prod_{i=1}^N \sum_{k=1}^K \pi_k \prod_{j=1}^J \phi(y_{ij}|\mu_{jk}, \sigma_{jk}^2) \quad (4.4)$$

where  $\theta_K = \{\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2, \pi_k: k = 1, \dots, K\}$ ,  $\pi_k$  is a parameter for the marginal probability  $\Pr(\kappa = k)$ , and  $\phi$  is the univariate normal density. The equation in (4.4) is the complete-data likelihood because all elements in the data matrix  $Y$  are complete. The corresponding complete-data loglikelihood is

$$\ell(\theta_K|Y, \mathcal{M}_K) = \log \mathcal{L}(\theta_K|Y, \mathcal{M}_K) = \sum_{i=1}^N \log \sum_{k=1}^K \pi_k \prod_{j=1}^J \phi(y_{ij}|\mu_{jk}, \sigma_{jk}^2). \quad (4.5)$$

Software relies heavily on the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) to calculate the maximum likelihood estimates,  $\hat{\theta}_K$ , although gradient-based optimization methods can also be used to assist in obtaining faster convergence. We refer the interested reader to the vast theoretical literature regarding the EM algorithm (R. J. Little & Rubin, 2002; McLachlan & Krishnan, 2008; Schafer, 1997), including its extensions and applications to mixture modeling (Frühwirth-Schnatter, Celeux, & Robert, 2019, Chapter 2; McLachlan & Peel, 2004). Briefly, the EM algorithm is a two-step procedure. In the E-step, an individual's probability of belonging to each of the  $\kappa = 1, \dots, K$  class is calculated. These values are referred to as posterior class probabilities. In the M-step, the parameters in  $\theta_K$  are updated by what the  $\theta_K$  values would have been if  $\kappa$  had been observed and weighted by the posterior class probabilities.

Thus, even when the indicator data are complete, the EM algorithm treats the estimation problem as a missing data problem such that each individual's  $\kappa$  value is missing. The E-step provisionally treats the missing data problem using the current  $\theta_K$  estimates in the cycle, while the M-step updates the parameters as if the data were complete. The E-step and M-step repeat until the parameter estimates converge.

### **Ignorability and FIML**

In the presence of missing indicator data,  $Y$  can be partitioned into observed and missing subsets (i.e.  $Y = \{Y^{\text{obs}}, Y^{\text{mis}}\}$ ). The observed data loglikelihood for model  $\mathcal{M}_K$  is calculated by including only the observed elements in the likelihood calculation, i.e.

$$\ell_{\text{obs}}(\theta_K | Y^{\text{obs}}, \mathcal{M}_K) = \sum_{i=1}^N \log \sum_{k=1}^K \pi_k \prod_{\{j: y_{ij} \in Y^{\text{obs}}\}} \phi(y_{ij} | \mu_{jk}, \sigma_{jk}^2). \quad (4.6)$$

As is the case when the indicator data were fully observed, estimates for  $\theta_K$  in the presence of missing data can be obtained through either the EM algorithm or through direct likelihood approaches using gradient-based optimization algorithms. In particular, the E-step is augmented such that the missing data problem in  $Y$  is also provisionally treated, in addition to provisionally treating the missing  $\kappa$  values. Alternatively, a direct likelihood approach applies root finding optimization techniques to maximize the observed data likelihood. Whether maximized by the EM algorithm or by direct likelihood, all available elements in  $Y^{\text{obs}}$  inform the calculation of the corresponding observed-data likelihood function and, therefore, influence the final parameter estimates. This procedure, FIML, therefore gets its name because all available information is utilized, unlike in complete case analysis.

To produce parameter estimates free of nonresponse bias, FIML requires that the so-called ignorability assumption be tenable. Ignorability means that the missing data mechanism can be ignored when maximizing the observed-data loglikelihood function (Rubin, 1976). To understand why, consider that theoretical justifications for all treatments of missing data, including FIML, begin by studying the joint distribution of the complete data and the missing data mechanism

$$\Pr(Y^{\text{obs}}, Y^{\text{mis}}, R | \theta_K, \psi, \mathcal{M}_K) \quad (4.7)$$

where the complete data  $Y$  is segmented into observed  $Y^{\text{obs}}$  and missing  $Y^{\text{mis}}$  components,  $R$  is the pattern of missingness, and  $\psi$  are a set of nuisance parameters describing the missing data mechanism.

As can be seen in (4.6), the missing data mechanism is ignored in the observed-data likelihood because no terms depend on  $R$ . Ignoring the missing data mechanism requires that the data be MAR and that the  $\theta_K$  parameters be “distinct” from the  $\psi$  parameters, with the latter assumption being tenable under general conditions in practice (Schafer, 1997, p. 11). Formally, the data are MAR provided that the missing data mechanism is independent of the missing values, conditional on the observed values, i.e.

$$\Pr(R | Y^{\text{obs}}, Y^{\text{mis}}, \psi, \mathcal{M}_K) = \Pr(R | Y^{\text{obs}}; \psi, \mathcal{M}_K). \quad (4.8)$$

The MAR assumption is not fully testable because the missing data mechanism may depend on unobservables. To make the ignorability assumption more tenable, methodologists recommend that applied researchers employ a “fully inclusive” (Collins, Schafer, & Kam, 2001) approach by incorporating many auxiliary variables (AVs) as missing data correlates to attenuate any resulting nonresponse bias (Enders, 2010; T. D. Little, Jorgensen, Lang, & Moore, 2014; Schafer & Graham, 2002). Simulations have

shown that incorporating AVs that explain at least 16% of the variance in the observed indicator data can significantly attenuate nonresponse bias (Enders, 2010). Although several methods exist to incorporate AVs with the covariance structure models used for variable-centered analyses, no such procedures have been developed for mixture models. In fact, a current restriction of FIML for person-centered analysis is the inability to accommodate missing data correlates in a manner that does not sacrifice the interpretability of the classes themselves (see Chapter 2). As a result, parameter estimates exhibit bias in real-world settings where missingness most often depends on variables other than the observed values of the indicators themselves.

### **Conditional Ignorability and Multiple Imputation**

In practice, it is often the case that the ignorability assumption is not tenable without the inclusion of AVs. Specifically, conditional ignorability relaxes the assumption that missingness is independent of the missing indicator data. A less restrictive assumption is that the missing data mechanism is independent of the missing values conditional on the observed data and a set of AVs,  $X$ ,

$$\Pr(R|Y^{\text{obs}}, Y^{\text{mis}}, X, \psi, \mathcal{M}_K) = \Pr(R|Y^{\text{obs}}, X, \psi, \mathcal{M}_K). \quad (4.9)$$

Multiple imputation is generally more flexible in accommodating AVs so that an inclusive missing data strategy is employed and the less restrictive conditional ignorability assumption is made (Enders, 2010). This is especially true in mixture settings because the multiple imputation procedure completely separates the treatment of the missing data from the person-centered analysis itself. In Chapter 2, we explained how this separation protects the definition of the classes from being influenced by the AVs beyond what is required to treat the missing data.

Multiple imputation is conducted in three separate phases, which consist of (1) an imputation phase, (2) an analysis phase, and (3) a pooling phase. In the imputation phase, the researcher generates  $m = 1, \dots, M$  completed (i.e., imputed) datasets by substituting the missing data with plausible values drawn from probability distribution implied by the imputation model. We denote the  $m$ -th imputed dataset as  $Y_m^{\text{imp}}$ . In the analysis phase, the researcher fits the analysis model as if the data were complete to each of the  $M$  imputed datasets. Finally, the estimates from the  $M$  analyses must be pooled for inference.

Multiple imputation is justified under a Bayesian framework. Accordingly, researchers seeking to make frequentist inferences using multiple imputation must rely on the asymptotic properties guaranteed by the Bayesian central limit theorem. Briefly, the Bayesian central limit theorem states that Bayesian and frequentist inference will agree under many conditions if the sample size is large enough (Freedman, 1963, 1965; Kaplan, 2014, pp. 26–30; Le Cam, 1986, pp. 618–621). From a Bayesian perspective, uncertainty in the parameter estimates caused by missing data is fully captured by the observed data posterior distribution given by

$$\Pr(\theta_K | Y^{\text{obs}}, X; \mathcal{M}_K) \propto \int_{Y^{\text{mis}}} \Pr(\theta_K | Y^{\text{obs}}, Y^{\text{mis}}; \mathcal{M}_K) \Pr(Y^{\text{mis}} | Y^{\text{obs}}, X) dY^{\text{mis}} \quad (4.10)$$

where  $[\theta_K | Y^{\text{obs}}, Y^{\text{mis}}, \mathcal{M}_K]$  is referred to as the complete data posterior distribution and  $[Y^{\text{mis}} | Y^{\text{obs}}, X]$  is the posterior predictive distribution for  $Y^{\text{mis}}$  (R. J. Little & Rubin, 2002, p. 210; Rubin, 1987; van Buuren, 2018, pp. 41–44). The integral in (4.10) implies that valid inferences for  $\theta_K$  can be accomplished through a marginalizing process over all plausible values for  $Y^{\text{mis}}$ , provided that independent samples from the posterior predictive distribution can be obtained (R. J. Little & Rubin, 2002; Rubin, 1987). As defined by

Schafer (1997, p. 105), if imputations are “Bayesianly proper” in that they are independent samples from the posterior predictive distribution for  $Y^{\text{mis}}$ , then valid Bayesian inference is assured.

The posterior distribution for  $\theta_K$  in (4.10) must be approximated by numerical integration techniques. Combined, the imputation and analysis phase of multiple imputation is a Monte Carlo numerical integration strategy to approximate the integral. A common simplifying assumption is that the observed data posterior follows a normal distribution, which is generally guaranteed when the sample size is sufficiently large.

Rubin (1987) provided a set of rules for pooling the parameter estimates obtained after analyzing the  $m = 1, \dots, M$  datasets. Rubin (1987) showed that the posterior distribution is centered around the mean of the maximum likelihood estimates (MLEs) taken across the  $M$  imputed datasets (denoted  $\bar{\theta}_K$ ) and with a variance that accounts for both within- and between- imputation variability, i.e.

$$[\theta | Y^{\text{obs}}; \mathcal{M}_K] \sim N(\bar{\theta}_K, \hat{T}_K) \quad (4.11)$$

where

$$\bar{\theta}_K = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_{m,K} \quad (4.12)$$

and  $\hat{\theta}_{m,K}$  is the maximum likelihood estimate when fitting  $\mathcal{M}_K$  to the  $m$ -th imputed dataset. Furthermore,  $\hat{T}_K$  is the total variance reflecting the within-imputation and between-imputation components. Technical specifics for pooling parameter estimates and conducting hypothesis tests can be found in outside texts (Enders, 2010; R. J. Little & Rubin, 2002; Schafer, 1997; van Buuren, 2018). We emphasize that the  $\hat{\theta}_{m,K}$  estimates are averaged only to obtain an estimate of the mean of the posterior distribution in (4.10).

However, this does not imply that valid inference results if functions of the  $\hat{\theta}_{m,K}$  parameters (e.g., loglikelihood functions) are averaged. We return to this point when discussing pooling information criteria by averaging across the values obtained from the imputed datasets.

### **FIML as the Currently Preferred Missing Data Strategy**

FIML is by far the most frequently adopted missing data strategy in person-centered analysis. This may be in part because it is implemented by default in software. However, challenges with generating proper imputations have limited the adoption of multiple imputation when conducting person-centered analysis. Specifically, popular imputation software traditionally assumes that the data were generated from a single-class model (Enders & Gottschall, 2011). Improper imputations from single-class models have been shown in simulations to cause biased parameter estimates (see Chapter 2) and under extraction of the classes (Sterba, 2016). Under extraction is problematic in applied settings because it obfuscates the potentially important sources of heterogeneity and individual differences which motivate a person-centered analysis in the first place.

Recently developed nonparametric imputation models show great promise in generating proper imputations. In Chapter 2, we showed that recursive partitioning imputation algorithms do not make a single-class assumption and are available in the `mice` R package (van Buuren & Groothuis-Oudshoorn, 2010). These imputation models greatly resolve the parameter bias when in large sample sizes ( $N = 1,200$ ) and classes are well separated (i.e., an entropy value near .88). Furthermore, they outperform FIML when the missingness depends on AVs. In Chapter 3, we highlighted methodological work that is currently underway in order to construct truly Bayesianly proper imputations

for person-centered analysis in the smaller sample sizes that are also commonly observed in practice (e.g.  $N \approx 300-600$ ) or when classes are less well separated (e.g., an entropy near .74) in large samples ( $N = 1,200$ ). Therefore, although FIML has been the dominant missing data approach adopted by applied researchers conducting person-centered analysis, multiple imputation shows great potential for researchers conducting a person-centered analysis to incorporate an inclusive strategy.

### **Prevailing Information Criteria Model Selection Practices**

Having provided a brief background on FIML and multiple imputation, we now discuss current practices in finite mixture model selection. Finite mixture model selection in a person-centered analysis is strongly informed by the AIC, the BIC, and the aBIC.

Each of the three ICs take on the form

$$\text{IC}(\mathcal{M}_K) = -2\ell(\hat{\theta}_K|Y, \mathcal{M}_K) + \text{penalty}(q_K, N) \quad (4.13)$$

where  $-2\ell(\hat{\theta}_K|Y)$  is referred to as the model deviance and is the loglikelihood value at the maximum likelihood estimate. The penalty depends on the number of parameters,  $q_K$ , in  $\mathcal{M}_K$  and, in some cases, the sample size,  $N$ . The penalty terms for the various ICs considered in this study are  $2q_K$  corresponding to the AIC,  $q_K \log N$  corresponding to the BIC, and  $q_K \log \frac{N+2}{24}$  corresponding to the aBIC. Thus, we write the expressions for the AIC, BIC, and aBIC for a model fit with  $K$  as

$$\text{AIC}(\mathcal{M}_K) = -2\ell(\hat{\theta}_K|Y, \mathcal{M}_K) + 2q_K \quad (4.14)$$

$$\text{BIC}(\mathcal{M}_K) = -2\ell(\hat{\theta}_K|Y, \mathcal{M}_K) + q_K \log N \quad (4.15)$$

$$\text{aBIC}(\mathcal{M}_K) = -2\ell(\hat{\theta}_K|Y, \mathcal{M}_K) + q_K \log \frac{N+2}{24} \quad (4.16)$$

The expressions for the information criteria listed above are all formulated using the complete-data likelihood in order to ascertain the model deviance. When data are missing, however, the model deviance from the complete-data likelihood cannot be calculated directly. Under FIML, it is common practice to substitute the observed data likelihood for the complete-data likelihood. Under multiple imputation, it is common practice to substitute the average of the model deviances across the imputed datasets. We scrutinize these practices.

### **Consequences of Substituting the Observed-Data Likelihood under FIML on the BIC**

When missing data are treated with FIML, the IC values are usually calculated using the observed-data likelihood value at the maximum likelihood estimate,

$$\text{IC}^{\text{FIML}}(\mathcal{M}_K) = -2\ell(\hat{\theta}_K | Y^{\text{obs}}, \mathcal{M}_K) + \text{penalty}(q_K, N). \quad (4.17)$$

We designate this procedure as producing a FIML-IC value,  $\text{IC}^{\text{FIML}}$ . Note that, in general, this value does not equal the value that would have been obtained had the data been complete because the observed-data and complete-data likelihoods are not equal. Despite substituting the observed-data likelihood, the penalty term remains the same and takes on the same value, regardless of whether the data are complete or missing. Because the observed-data model deviance is usually less than the complete-data model deviance in a finite mixture of Gaussians, the effect is that the model deviance is penalized relatively more harshly when there are missing data.

We argue that failing to adjust the penalty values when relying on the observed-data likelihood is not appropriate when the penalty term depends on the sample size,  $N$ . Such penalty terms wrongly assume that  $N$  complete observations are present in the

sample. This can be seen in the BIC expression typically calculated using FIML, given by

$$\text{BIC}^{\text{FIML}}(\mathcal{M}_K) = -2\ell(\hat{\theta}_K | Y^{\text{obs}}, \mathcal{M}_K) + q_K \log N. \quad (4.18)$$

In Technical Appendix C we derive a correction that accounts for evidence loss due to missing data. Briefly, the derivation proceeds by noting that the variances obtained from the Fisher information matrix for the observed-data likelihood are larger than the corresponding variances obtained from the Fisher information matrix for the complete-data likelihood. Applying Laplace’s approximation to estimate the model evidence with the inflated variance term leads to our BIC expression. By assuming that the variance of the Fisher information matrix is inflated, we correct for the fact that not all  $N$  observations are complete. Our proposed BIC expression simplifies to a correction that can be applied to FIML-BIC. The correction depends on the fraction of missing information,  $\text{FMI}(\hat{\theta}_K)$ , a quantity that is prominent in missing data theory because it governs the rate of convergence for the EM algorithm (Dempster et al., 1977), as well as the number of imputed datasets required to reach near-optimal power in frequentist inference. In particular, our proposed corrected BIC expression is as follows:

$$\text{BIC}^{\text{obs}}(\mathcal{M}_K) = \text{BIC}^{\text{FIML}} + \log|\mathbb{I}_{q_K} - \text{FMI}(\hat{\theta}_K)|. \quad (4.19)$$

We refer to the correction term,  $\log|\mathbb{I}_{q_K} - \text{FMI}(\hat{\theta}_K)|$ , as the evidence loss due to missing data because this value is always negative and represents the degree to which missing data diminishes the model evidence approximated by the BIC. We explain four properties by studying our corrected BIC value in (4.19) because these properties lead to hypotheses about why model selection decisions for the BIC are sensitive to sample size.

1. (Property 1) If the fraction of missing information is trivial, then model selection decisions will not be sensitive to whether  $\text{BIC}^{\text{FIML}}$  or  $\text{BIC}^{\text{obs}}$  is used. This is true regardless of sample size.
2. (Property 2) If the fraction of missing information is nontrivial, the sensitivity of model selection decisions to using  $\text{BIC}^{\text{FIML}}$  instead of  $\text{BIC}^{\text{obs}}$  will decrease as sample size increases.
3. (Property 3) If the fraction of missing information is nontrivial and sample size is small, model selection decisions will be sensitive to the choice of  $\text{BIC}^{\text{obs}}$  versus  $\text{BIC}^{\text{FIML}}$ .
4. (Property 4) In the presence of missing data, model selection using the  $\text{BIC}^{\text{FIML}}$  will tend to favor more parsimonious models than if  $\text{BIC}^{\text{obs}}$  were used for model selection.

We note that Property 1 holds because trivial missingness leads to  $\log|\mathbb{I} - \text{FMI}(\hat{\theta}_K)| \approx 0$ . In the limiting case when there is no missingness, then  $\text{BIC}^{\text{obs}} = \text{BIC}^{\text{FIML}}$ , so model selection decisions will be equivalent. Property 2 holds because the fraction of missing information does not scale with sample size, so

$$|-2\ell(\hat{\theta}_K|Y^{\text{obs}}; \mathcal{M}_K) + q_K \log N| \gg |\log|\mathbb{I} - \text{FMI}(\hat{\theta}_K)||.$$

Thus, as sample size increases,  $\text{BIC}^{\text{FIML}}$  converges to  $\text{BIC}^{\text{obs}}$ , implying that model selection decisions are less sensitive to missing data if using the  $\text{BIC}^{\text{FIML}}$  when sample sizes are large. In contrast, Property 3 holds because the fraction of missing information is not trivial relative to the magnitude to  $\text{BIC}^{\text{FIML}}$  when sample sizes are small in the presence of missing data.

Finally, we expect that  $\text{BIC}^{\text{FIML}}$  will tend to result in selection decisions where a more parsimonious finite mixture model is selected than  $\text{BIC}^{\text{obs}}$  when missing data are present (Property 4). We note that the magnitude of the evidence loss scales with the number of parameters because it is the product of the diagonal elements of the  $\mathbb{I} - \text{FMI}(\hat{\theta}_K)$  matrix, and each of these diagonal elements are less than 1. In fact, as the number of parameters increases,  $\log|\mathbb{I} - \text{FMI}(\hat{\theta}_K)|$  takes on an increasingly negative value. Because this negative correction value is added to the traditional BIC penalty term, the effect is that the missing data increasingly offset the traditional penalty term as model complexity increases. Consequently, compared to the final model selected using  $\text{BIC}^{\text{obs}}$ ,  $\text{BIC}^{\text{FIML}}$  will tend to result in a selection decision for a final model specified with fewer classes.

In summary, we delineate these properties in order to construct a hypothesis for how model selection decisions may differ if employing FIML in contrast to if the data had been complete. In particular, we expect that the FIML-BIC calculated from (4.18) will lead to an under extraction of the number of classes in small samples. When the evidence loss due to missing data is small in magnitude relative to the likelihood and penalty terms, however, we expect that model selections will be less sensitive to missingness.

### **Current Practice with Finite Mixture Model Selection under Multiple Imputation**

Unlike FIML, model selection procedures with multiple imputation make use of the complete-data likelihood to calculate model deviances given the  $m = 1, \dots, M$  imputations. The challenge is that the researcher must reconcile  $M$  distinct IC values. We denote an IC obtained by the  $m$ -th imputed dataset for model  $\mathcal{M}_K$  as

$$\text{IC}_m^{\text{MI}}(\mathcal{M}_K) = -2 \ell(\hat{\theta}_K | Y_m^{\text{imp}}; \mathcal{M}_K) + \text{penalty}(q_K, N). \quad (4.20)$$

Currently, popular software used in person-centered analysis employs an averaging strategy to pool the model deviances. Specifically, the average across the model deviances at the maximum likelihood estimate for the  $m$ th imputed dataset is taken. Thus, under averaging, the pooled criterion is

$$\text{IC}_{\text{avg}}^{\text{MI}}(\mathcal{M}_K) = -2 \frac{\sum_{m=1}^M \ell_{\text{imp}}(\hat{\theta}_{m,K} | Y_m^{\text{imp}}, \mathcal{M}_K)}{M} + \text{penalty}(q_K, N). \quad (4.21)$$

As pointed out in previous literature, there is no theoretical justification for averaging (Consentino & Claeskens, 2010; Meng & Rubin, 1992). The implicit assumption underlying current practice is that  $\text{IC}_{\text{avg}}^{\text{MI}}(\mathcal{M}_K)$  is an unbiased estimate of the corresponding complete data criteria value,  $\text{IC}(\mathcal{M}_K)$ . However, this can only be true if the averages of the model deviances in (4.21) are unbiased estimates of the complete data values. We evaluated this assumption in our simulations, which we turn to now.

### **Sensitivity of Finite Mixture Model Selection Decisions to Missing Data under Current Practices**

To highlight the shortcomings of current finite mixture model selection practices, we conducted a simulation study. We used the Web of Science database to identify 30 frequently cited articles employing LPA that were published between 2008-2018 in developmental and educational psychology journals in order to inform the simulation conditions and set parameters to typical values observed in applied literature. We focused exclusively on LPA that fit finite mixtures of Gaussian models to make inferences. From each study, we recorded the sample size of the analytic dataset, the number of classes selected, the number of indicators used to construct the profiles, the proportion of

observations in the smallest profile, and the observed entropy value. We also noted typical rates of missingness across each indicator variable.

## **Simulation Setup**

### ***Manipulated Conditions: Sample Size, Mixing Proportions, and Class Separation***

We manipulated three primary factors in the simulations according to values that appear in literature. This includes the sample size, the mixing proportions, and the separation between the classes. To demonstrate that model selection using the BIC is sensitive to sample size, we varied the sample size between  $N = 300$  (small sample) and  $N = 1,200$  (large sample). These values corresponded to the interquartile range that we observed in applied research from our sample of 30 studies.

Sample sizes can be relatively large, but we also hypothesized that the presence of a small class could influence the performance of IC-based selection decisions. To test this hypothesis, we manipulated the mixing proportions between a condition with equal mixing and a condition with unequal mixing, such that the smallest class represented about 10% of the sample. This value was well within the interquartile range of the smallest class proportions observed in literature.

Finally, model selection decisions are highly sensitive to the separation of the classes. When classes are extremely well-separated, for example, all ICs consistently identify the true model if given a large enough sample size. Disagreement among the ICs occurs when the classes exhibit the decreased separation that occurs in practice. The 25th and 75th quantiles for the reported entropy values were .74 and .88, respectively. To achieve entropy values that reflect these typical values, the Mahalanobis distances (MD) between the class means were manipulated so that the entropy values from the simulation

roughly matched the 25th and 75th quantile values from the LPA studies.<sup>1</sup> As a result, we manipulated class separation between weakly separated ( $MD = 2.86$ ; entropy = .74) and highly separated ( $MD = 3.70$ ; entropy = .88) values that were observed from the literature review.

***Fixed Conditions:  $J$ ,  $K$ , Missingness Rate, &  $M$***

The number of indicator variables,  $J$ , the number of classes,  $K$ , the missingness rate, and the total number of imputations,  $M$ , were held constant throughout the simulations. We set the total number of indicator variables to  $J = 4$ , the modal value that appears in the 30 selected LPA studies. We chose to simulate data from a three-class model ( $K = 3$ ) to make the simulation study and class enumeration time-feasible; this differs from the four-class model that appears most frequently, but it is well within the typical range.

In establishing a fixed missingness rate, we attempted to discern common rates reported by applied researchers conducting LPA. However, the 30 studies provided limited information regarding missing data rates, when such information was provided at all. Only about one-third (11 of the 30 studies) reported missingness information. Of those reporting indicator missingness information, the reporting range of missing value rates across individual indicators was the most common reporting method (seven of 11 studies), followed by reporting the proportion of observations completed (three of 11 studies). A single study reported covariance coverage rates.

We note that the reporting of complete cases would be ideal to inform the simulation. However, predicting this value from either the range of missing data rates or

---

<sup>1</sup> For a given Mahalanobis distance, the entropy value will fluctuate depending on whether equal or unequal mixing is present. In conducting this manipulation exercise, the unequal mixing condition was chosen to calibrate the Mahalanobis distance because that condition was highly represented in the LPA studies.

the covariance coverage rates proved difficult. For example, one typical study included five indicators and reported missingness rates ranging from 10.7% to 25%. The percentage of observations with complete data in this study ranged anywhere from 0-75%. In the simulation study, we fixed the missing data rate such that 50% of observations were missing at least one indicator value. With the chosen missing data mechanism (discussed below), this corresponded to missingness rates for each variable of approximately 25%.

Finally, following Sterba (2016), we set the total number of imputed datasets to  $M = 100$  when studying model selection procedures under multiple imputation. Although this value is larger than what is often necessary in practice, we decided to be conservative to minimize the risk that model selections were the result of a too-small  $M$  value. That is, if problems appear with  $M = 100$  datasets, then they most certainly appear in the smaller values (e.g.,  $M = 10$  or  $20$ ) chosen by applied researchers. Nevertheless, we do point out that computational speed and storage are no longer barriers like they were in the past, when precedents on the number of imputations were set. Thus, we encourage researchers to impute more datasets than what has been traditionally seen as necessary.

### ***Data Generating Model and Missing Data Mechanism***

We generated complete data using  $K = 3$  classes as a template with  $J = 4$  indicator variables and one AV. The data generating model is presented in Figure 4.2. Data were generated from this template by first randomly drawing class memberships for each observation from a multinomial distribution and subsequently drawing values for the class indicators and the AV from a finite mixture model with a class-specific population mean vector and a unitary variance-covariance matrix. For convenience, the

AV was sampled jointly with the profile indicators such that there were no mean differences in the AV between classes. Within classes, the correlation between the AV and each indicator variable was set to  $r = .40$ ; Enders (2010) reports that such a correlation is beneficial to enhance statistical power.

We identified indicator data that were set to missing in a manner that ensures that the propensity for missingness is informed by the AV. Observations could be missing up to three indicator values, resulting in 15 possible missing data patterns total (including the pattern corresponding to no missingness). Missing data patterns were assigned using a latent variable formulation:

$$\eta_i^* = AV_i + \epsilon_i$$

where  $\epsilon_i \sim \mathcal{N}(0,0.1)$ . We manipulated cut points for the  $\eta^*$  latent variable so that the marginal missingness rates across each indicator value averaged approximately 25% each, while half of the observations contained complete data.

Critically, this choice of missing data mechanism did not lead to the collapsing of the classes or decreased class separation under any condition. Figure 4.1 contrasts the class separation value displayed with the complete data versus the separation implied by the FIML estimates. The difference is explained by nonresponse bias because the AV is excluded from the analysis, making the missing data mechanism nonignorable, as would be consistent in practice with external variables. One can see that nonresponse bias led to increased class separation with FIML. Thus, we would have expected any under extraction results to be exacerbated had the missing data mechanism resulted in decreased class separation.

### ***Imputation Procedure***

Imputed datasets were constructed by following an EM with sampling (EMS) procedure. EMS is a joint, multivariate imputation procedure originally proposed by King et al. (2001). It results in proper imputations provided the model is correct and the missing data mechanism is ignorable. EMS has subsequently been adopted in mixture settings (Vermunt, van Ginkel, van der Ark, Andries, & Sijtsma, 2008; Vidotto, Vermunt, & Kaptein, 2015), a natural extension of its use, given that parameter estimation for finite mixture models relies heavily on the EM algorithm.

EMS is a three-step procedure to simulate plausible values and construct imputed datasets. The first step accounts for between-imputation variability by bootstrapping the data. The EM algorithm is then used to fit the imputation model to the bootstrapped data. The final step accounts for within-imputation variability. Here, imputations are simulated using the parameter estimates from the second step augmented by the information presented by the observed data. A single imputed dataset results at the end of the third step. The user continues the three-step cycle until the desired number of imputed datasets is constructed.

Traditionally, the bootstrapping step is conducted by sampling observations with replacement. We found this frequently resulted in convergence issues in small samples. Through experimentation, we also found that substituting the Bayesian bootstrap (Rubin, 1981) for sampling with replacement greatly decreased convergence issues. We therefore modified step 1 of the EMS procedure by implementing the Bayesian bootstrap.

The Bayesian bootstrap is a simple procedure to weight observations, and it can be conducted in a manner so that the prior is noninformative. It is best understood by

comparing it with the traditional bootstrap, in that sampling with replacement is simply a reweighting procedure, as well. Indeed, by sampling with replacement, the user is effectively weighting observations by a value that is in the natural numbers (e.g., if the observation was never sampled, then its weight is zero; if it was sampled once, then the weight is one; if sampling with replacement resulted in an observation being selected twice, then the weight is two; etc.). The sum of the weights then equals the total number of observations. The Bayesian bootstrap is a procedure to sample an observation's weight, rather than the observations themselves. Weights are drawn from a Dirichlet distribution with  $N$  total categories. A uniform prior can be specified so that each observation has an equal chance of being assigned a given weight. Thus, the uniform prior is completely noninformative and parallels sampling with replacement. The uniform prior is implemented by specifying unit concentration parameters for the Dirichlet distribution.

Both the Bayesian bootstrap and the traditional bootstrap sample from the empirical CDF. Thus, both lead to similar inferences. In fact, sampling with replacement can be seen as simply a special case of the Bayesian bootstrap in that the weights are restricted to the natural numbers. By allowing the weights to take on the positive real numbers, the Bayesian bootstrap effectively smooths the empirical CDF, which is advantageous when sample sizes are small, and the histogram of the empirical CDF can appear quite discrete (Chernick, 2011, Chapter 6). We credit the improved convergence to the smoothing of the empirical CDF in that all observations generally inform the parameter estimates, even if they are down-weighted.

### ***Analytic Model***

We fit a conditionally independent, finite mixture of Gaussians model diagrammed in Figure 4.3 to model the  $J = 4$  profile indicators using only the observed data (to study selection under FIML) or to the imputed datasets (to study selection when averaging information criteria under multiple imputation). The mixture models were fit using Mplus version 8.0 (Muthén & Muthén, 2017). Subsequently, results were exported to R (R Core Team, 2020) using the `MplusAutomation` package (Hallquist & Wiley, 2018).

### **Model Selection Results**

Figures 4.4-4.6 plot the proportion of the replications that resulted in the  $\mathcal{M}_1$  to  $\mathcal{M}_4$  that were selected according to minimizing the AIC, BIC, and aBIC, respectively. The model selection proportions were disaggregated according to the complete data results, the observed data results (in which FIML estimation was applied to the indicator data), and the imputed data results. We discuss the results in detail below, but summarize three central findings: (1) our results with the complete data replicated what has previously been documented in literature; (2) consistent with our hypothesis, BIC model selection with FIML resulted in substantial under extraction for several of the simulated conditions; and (3) with imputed data, model selection conducted with an averaging pooling strategy poorly replicated the complete data selections for the AIC or aBIC. Taken together, these findings suggest that finite mixture model selection is highly sensitive to missing data.

### ***Complete Data Model Selection***

We found that the complete data results were consistent with previous research. Consistent with Nylund et al. (2007), we found that the BIC resulted in optimal model selection decisions across the simulated conditions (see Figure 4.5). For most conditions, the BIC almost uniformly selected the true three-class model,  $\mathcal{M}_3$ , across the replications. However, using the BIC resulted in under extraction decisions when the sample size was small ( $N = 300$ ) and class separation was weak (entropy values averaged near .74). This under extraction was not present in large samples ( $N = 1,200$ ), regardless of class separation. Thus, our findings are consistent with the asymptotic consistency properties of the BIC: with large sample sizes ( $N = 1,200$  in our study), the BIC selected the correct model when the true model was contained within the set of models being evaluated.

As expected, model selection based on the AIC resulted in over extraction of the classes across all simulated conditions (Figure 4.4); approximately 90% of replications resulted in the four-class model,  $\mathcal{M}_4$ , being selected. The tendency of the AIC to over extract the number of classes has been well documented in literature (Frühwirth-Schnatter et al., 2019, p. 123; McLachlan & Peel, 2004, p. 201) and in previous simulation studies (Nylund et al., 2007; Tofighi & Enders, 2008).

Finally, model selection using the aBIC resulted in over extraction decisions in small samples with approximately three-quarters of the replications resulting in the selection of  $\mathcal{M}_4$  (Figure 4.6). However, in large samples, using the aBIC resulted in a correct model decision nearly three-quarters of the time. In summary, model selection decisions were most accurate using the BIC, followed by the aBIC. Model selection decisions using the AIC were the least accurate.

### ***Model Selection using FIML in the Presence of Missing Data***

Consistent with our hypothesis, the BIC led to a substantial under extraction for the number of classes when sample sizes were small and missing data were treated with FIML (Figure 4.5). The degree of the under extraction was also substantial and noteworthy. For example, under extraction was present in at least nine out of 10 replications when sample sizes were small ( $N = 300$ ) and classes were weakly separated (entropy values averaged near .74). Though less severe, under extraction was also present when separation was strong and mixing was unequal so that a small class was present; in that case, nearly 25% of the replications resulted in under extraction. Accordingly, we found that the performance of the BIC was highly sensitive to missing data.

### ***Model Selection using an Averaging Pooling Strategy with Multiply Imputed Data***

Except for when using the BIC, we found that averaging the ICs obtained by fitting a mixture model to multiply imputed datasets poorly replicated the complete-data selection decisions. Specifically, we found that averaging uniformly resulted in  $\mathcal{M}_4$  being selected across all of the simulated conditions when the AIC was used for model selection. In other words, averaging exacerbated the over extraction already problematic with the AIC when the data were complete.

When the aBIC was used to select a final model, averaging also poorly replicated complete-data decisions. As with the AIC, averaging exacerbated over extraction problems when the data were complete and sample sizes were small, as demonstrated by the aBIC uniformly resulting in  $\mathcal{M}_4$  being selected. Interestingly, however, averaging led to the correct model being selected at a far greater rate compared to the complete data in the large sample condition. It would be a mistake, however, to conclude that averaging

performed well in large samples with the aBIC because the goal of any missing data problem should be to replicate inference had the data been complete.

Moreover, Figure 4.8 shows that the implicit assumption for averaging (i.e., that the average of the model deviances across the imputed datasets is an unbiased estimate) was violated because the averaged model deviance was found to be negatively biased. In fact, bias in model deviance was found to be most pronounced for  $\mathcal{M}_1$ , but the magnitude and direction of bias remained almost constant thereafter. At first glance, such a finding may suggest that the bias should not have resulted in different selection decisions provided that the final model contained more than one class. However, such a conclusion would conflate aggregated patterns with replication-specific trajectories in the model deviances. In other words, we risked committing the ecological fallacy if we assumed trends of the bias represented trends of replication-specific deviance trajectories from the model deviances had the data been complete.

To avoid committing an ecological fallacy, we conducted a trajectory analysis by studying bias in rates of change of the model deviances. The first forward difference is shown in Figure 4.9. If averaging the model deviances matched the rate of change, we would have expected similar model selection results. Correspondingly, if the second-central difference obtained by averaging matched the complete data at  $K = 3$ , we would have expected that the correct model,  $\mathcal{M}_3$ , would have been selected as frequently when averaging as compared to the complete data.

We found substantial differences in the first- and second-order rates of change. As can be seen in Figure 4.9, the rate of change was strongly negatively biased when comparing model deviances between  $\mathcal{M}_1$  and  $\mathcal{M}_2$ . This resulted in a decreased preference

for selecting  $\mathcal{M}_1$  and was evidenced by the BIC; with small samples and weakly separated classes, model selections using the BIC were correct in approximately 40% of the replications with complete data, while the correct selection rate for averaging was found to be less than 30%. We also found substantial negative bias in the second-order rates of change for  $\mathcal{M}_3$  and  $\mathcal{M}_4$ , implying that the overall trajectory pattern was less concave at  $K = 2$  and  $K = 3$ . Thus, a minimum value was less likely to be reached in the set of  $K$  values observed. We observed this with the AIC where  $\mathcal{M}_4$  was consistently selected over  $\mathcal{M}_3$  at a much greater rate than what was observed had the data been complete.

In summary, model selection decisions made by averaging information criteria across the imputed datasets did not match those made with complete data. This was because there were systematic differences in rates of change between the averaged model deviance trajectories and the trajectories that would have been observed had the data been complete.

### ***Discussion of Current Practice Results***

Clearly, model selection decisions are not robust when data are missing. This is true regardless of whether a FIML approach or multiple imputation approach is adopted. Therefore, the best practices shaped by simulations assuming data are complete do not translate to real-world settings where data are most often missing values and researchers adopt the two most popular approaches to treating missing data.

Given the sensitivity of model selections to missing data, it is then natural to ask, “What are the underlying causes?” Accordingly, we can provide guidance to address this issue in practice. We have demonstrated that the problem is rooted in suboptimal

statistical procedures being used to calculate information criteria (e.g., the BIC calculation in the case with FIML) or in choosing a pooling strategy (e.g., averaging ICs across imputed datasets). In the next section, we discuss a practical remedial strategy to correct the BIC so that it accounts for missing information. The following section evaluates alternative model selection strategies under multiple imputation.

### **Correcting the BIC for Evidence Loss Due to Missing Data under FIML**

We propose a practical, easy-to-calculate correction to  $\text{BIC}^{\text{FIML}}$  so that model selection decisions in the presence of missing data are more consistent with the decisions that would have been made had the data been complete. The problem with correcting the BIC for evidence loss due to missing data using (4.19) directly is that, unlike with covariance structure models (e.g., Savalei & Rhemtulla, 2012), the  $\text{FMI}(\hat{\theta})$  is a nontrivial matrix to obtain empirically in finite mixture models. This is because finite mixtures estimate higher order moments in the data and require an expanded set of sufficient statistics relative to those needed for covariance structure models. As a result, structural equation modeling software cannot simply be provided with a mean vector, covariance matrix, sample size, and have the software estimate the fraction of missing information, as is done in Savalei & Rhemtulla's (2012) procedure to estimate the fraction of missing information in covariance structure models.

However, McNeish & Harring (2017) demonstrated that the proportion of elements in the data matrix that are missing can be used to generate a suitable approximation of the fraction of information matrix,  $\text{FMI}(\hat{\theta}_K)$ . In their study, such an approximation resolved issues with inflated Type I error rates for multiparameter hypothesis tests in small samples in latent growth curve models. McNeish & Harring's

(2017) approximation is grounded in Enders' (2010, p. 204) claim that this proportion represents an upward bound in the fraction of missing information for a given parameter, as well as Wagner's (2010) finding that this value approximates the fraction of missing information in real-world survey data well.

Following McNeish & Harring (2017), we propose that  $\text{FMI}(\hat{\theta}_K)$  be approximated as

$$\text{FMI}(\hat{\theta}_K) = (1 - c_K)\mathbb{I}_{q_K} \quad (4.22)$$

where  $c_K$  is a scalar given by

$$c_K = \frac{\text{\# of elements observed in data matrix}}{N \times (J + K - 1)}. \quad (4.23)$$

Note that  $c_K$  represents the proportion of non-missing values in the dataset augmented to include posterior class probabilities for a model with  $K$  classes. The last class is discarded (hence, the  $K - 1$  term) because the probabilities must sum to one.

We argue that it is incorrect to disregard the posterior probabilities in approximating the fraction of missing information. The slow convergence of the EM algorithm, even when indicator data are complete, is evidence in favor of our argument. Specifically, it is known that  $\text{FMI}(\hat{\theta}_K)$  governs the convergence rate of the EM algorithm (Dempster et al., 1977). Therefore, if it were true that we could disregard the posterior probabilities, then complete indicator data should lead to the EM algorithm converging to a stationary point in one step if provided a reasonable starting position near the maximum likelihood estimate so that the quadratic approximation is tenable (Schafer, 1997). We know in practice that the EM algorithm takes many iterations to converge, even when the indicator data are complete and the starting position is near the maximum likelihood estimate. Indeed, the slow convergence of the EM algorithm near the MLE for finite

mixture models is a well-documented problem that has resulted in many proposed accelerators (McLachlan & Peel, 2004, p. 52). The many iterations required for convergence when near the maximum likelihood estimate are an indication that the missing information is nontrivial, even when indicator data are complete. Consequently, the missing posterior class probabilities must contribute significantly to the overall amount of information missing.

Substituting the approximation in (4.23) into (4.19) and simplifying results in our proposed BIC value corrected for information loss:

$$\text{BIC}^{\text{obs}} \approx \text{BIC}^{\text{FIML}} + q_K \log c_K = -2\ell_{\text{obs}}(\hat{\theta}_K | Y^{\text{obs}}, \mathcal{M}_K) + q_K \log c_K N. \quad (4.24)$$

The  $c_K N$  term on the right hand side suggests that, under McNeish & Harring's (2017) approximation, the correction for missing information can simply be viewed as appropriately down-weighting the sample size because not all  $N$  observations are complete. We now reanalyze the simulations to evaluate whether the proposed  $\text{BIC}^{\text{obs}}$  approximation better replicates model selection decisions had the data been complete.

### **Reanalyzing the FIML Simulation Results**

The model selection results using the proposed  $\text{BIC}^{\text{obs}}$  approximation in (4.24) are shown in Figure 4.7. Overall, we found that our proposed approximation better replicated the complete data selection decisions than if the BIC was not adjusted for evidence loss. For example, the correct model was selected in approximately 15% of replications with complete data when mixing was unequal, sample sizes were small, and the classes were weakly separated. This value was matched when  $\text{BIC}^{\text{obs}}$  was used for model selection. In contrast, less than 3% of the replications resulted in  $\mathcal{M}_3$  being selected when  $\text{BIC}^{\text{FIML}}$  was used for model selection.

Additionally, we found that the proposed  $\text{BIC}^{\text{obs}}$  approximation was less susceptible to the under extraction problem as compared to  $\text{BIC}^{\text{FIML}}$ . For example,  $\text{BIC}^{\text{FIML}}$  resulted in  $\mathcal{M}_1$  being selected nearly 60% of the time when classes were weakly separated, sample sizes were small, and mixing proportions were equal. The corresponding value was 30% with the  $\text{BIC}^{\text{obs}}$ , a value that nearly matched the 37% found with complete data. Finally, under the large sample size conditions when the BIC and  $\text{BIC}^{\text{obs}}$  uniformly resulted in correct model selection,  $\text{BIC}^{\text{obs}}$  exhibited similar correct selection rates. Taken together, these findings suggest that correcting the BIC for missing information resulted in model selection decisions that better replicated decisions that would have been made had the data been complete.

### **Alternatives to Averaging for Multiple Imputation Model Selection**

An unresolved issue in methodological literature is how best to conduct model selection using information criteria with multiply imputed data. Several alternatives to averaging have been proposed and tested using simulations in the variable selection context with linear regression models. Even in this simpler context, the unresolved nature of model selection with multiple imputation is highlighted by the lack of clearly defined guidelines for applied researchers (see, e.g., van Buuren, 2018, p. 154).

The alternatives to averaging can be classified as ad-hoc or theoretically-based alternatives, depending on whether the approach is based on probability theory (van Buuren, 2018). Ad-hoc procedures for model selection include selecting based on a majority vote (Brand, 1999; van Buuren, 2012) or pooling the  $M$   $\text{IC}_m^{\text{MI}}(\mathcal{M}_K)$  estimates by data stacking (Wood et al., 2008). Alternatively, Consentino and Claeskens (2010) proposed a theoretically based alternative using Meng and Rubin's (1992)  $D_{LR}$  statistic,

which approximates a likelihood ratio statistic. We discuss each of these alternatives in more detail below.

### Majority Vote

The majority procedure was first proposed by Brand (1999) and maintains current support because it has been found to work well in practice and because it provides insights about how sensitive model selection decisions are to the missing data (van Buuren, 2012, 2018). Selecting a model by majority vote is a two-step procedure that is distinct, as it involves no pooling at all. In the first step, the researcher forms a set with  $M$  elements, where each element is the model that minimized the information criteria for a given imputed dataset. In the second step, the researcher selects the model that appears most frequently in the set from the first step. This majority vote process is summarized by the mathematical expression,

$$\mathcal{M}_{\text{maj}}^* = \mathbf{mode} \left\{ \arg \min_{\mathcal{M}_K} \text{IC}_m^{\text{MI}}(\mathcal{M}_K) : m = 1, \dots, M \right\} \quad (4.25)$$

where  $\text{IC}_m^{\text{MI}}(\mathcal{M}_K)$  is given in (4.13). In the case of a multimodal distribution, the most parsimonious model is selected to conform to Occam's razor.

### Stacking

As an ad-hoc strategy, stacking was first proposed by Wood et al. (2008). They found that the strategy performed well in small samples where power might be an issue for variable selection. Stacking involves creating one large flat file out of each of the  $m = 1, \dots, M$  datasets, denoted  $Y^{\text{stack}} = \{Y_m^{\text{imp}} : m = 1, \dots, M\}$ ,

$$\text{IC}_{\text{stack}}^{\text{MI}}(\mathcal{M}_K) = -\frac{2}{M} l(\hat{\theta}^{\text{stack}} | Y^{\text{stack}}; \mathcal{M}_K) + \text{penalty}(q_K, N) \quad (4.26)$$

where  $\hat{\theta}^{\text{stack}}$  is the maximum likelihood estimate fit to the stacked flat file. The selected model is the model that minimizes  $\text{IC}^{\text{stack}}(\mathcal{M}_K)$ ,

$$\mathcal{M}_{\text{stack}}^* = \arg \min_{\mathcal{M}_K} \{\text{IC}^{\text{stack}}(\mathcal{M}_K): K = 1, 2, \dots\}. \quad (4.27)$$

### **Consentino and Claeskens' (2010) $D_{\text{LR}}$ Procedure**

Consentino and Claeskens (2010) proposed a theoretically-based pooling procedure to conduct model selection. We note that model selection using information when models are nested can be accomplished by studying differences. In particular, the more parsimonious, restricted model is rejected if the difference between the full-model and the restricted-model is negative because this indicates that the IC for the full model is less than the IC for the more restricted model. When enumerating the classes under complete data, for example, it can be shown that the difference in the IC values between the restricted model  $\mathcal{M}_K$  and the full model  $\mathcal{M}_{K+1}$  is given by

$$\Delta\text{IC}(\mathcal{M}_K) = -\text{LR}_{K,K+1} + \text{penalty}(q_{K+1}, N) - \text{penalty}(q_K, N) \quad (4.28)$$

where  $\text{LR}_{K,K+1}$  refers to the likelihood ratio statistic which, in the case of the complete-data likelihood, is defined as

$$\text{LR}_{K,K+1} = -2\{\ell(\hat{\theta}_K|Y; \mathcal{M}_K) - \ell(\hat{\theta}_{K+1}|Y; \mathcal{M}_{K+1})\}. \quad (4.29)$$

Consentino and Claeskens (2010) point out that the likelihood ratio statistic given complete data can be approximated with multiply imputed data using Meng and Rubin's (1992)  $D_{\text{LR}}$  statistic. Thus, the approximation for  $\Delta\text{IC}$  that Consentino and Claeskens (2010) propose is

$$\Delta\text{IC} \approx -D_{\text{LR}} + \text{penalty}(q_{K+1}, N) - \text{penalty}(q_K, N). \quad (4.30)$$

Enders (2010, pp. 240–242) provides an accessible overview for how to calculate the  $D_{LR}$  statistic.<sup>2</sup> This procedure demonstrated modest improvements in performance to ad-hoc alternatives such as averaging in simulations. However, it has never been studied when deciding on the number of classes in finite mixture models.

### **Reexamining Simulations: Evaluating Alternatives to Averaging**

Selection decisions by the AIC, BIC, and aBIC using majority vote, stacking, and Consentino and Claeskens' (2010)  $D_{LR}$  procedure as alternatives to averaging are illustrated in Figures 4.11-4.13, respectively. We found that model selection based on the  $D_{LR}$  statistic poorly replicated the model selection decisions had the data been complete. However, selections based on majority vote or stacking better replicated the complete data model selection decisions than averaging. For example, if the aBIC was used for model selection, both stacking and majority vote selected the correct model about as often as when the data were complete, whereas averaging very rarely resulted in the true model being selected. Similar results were also found when the AIC was used for model selection, where averaging almost always resulted in  $\mathcal{M}_4$  being selected. All three strategies replicated the complete data model selection decisions when the BIC was used (Figure 4.12). In conclusion, stacking and majority vote were less susceptible to over extraction, a problem that was most pervasive when an averaging strategy was applied.

### **Discussion**

We have demonstrated that current finite mixture model selection practices using information criteria lead to suboptimal decisions when data are missing and a FIML or multiple imputation strategy is employed. In the case of FIML, the traditional formula

---

<sup>2</sup> Enders (2010) refers to the  $D_{LR}$  statistic as  $D_3$ .

used to calculate the BIC fails to account for evidence loss due to missing data, resulting in an under extraction of classes in many real-world conditions. In the case of multiple imputation, pooling information criteria by averaging leads to biased estimates of the complete data model deviance value, resulting in a tendency for a more complex model to be selected compared to the model that would have been selected in the hypothetical situation that the data were complete.

We have evaluated some easy-to-implement remedial strategies to address the shortcomings in current practice. Specifically, we propose that the sample size in the BIC be corrected to account for the fact that not all  $N$  observations are complete when the observed data loglikelihood is used to calculate the model deviance, as is the case with FIML. Additionally, we have shown through simulations that both the stacking and majority vote pooling strategies lead to model selections that are frequently more consistent with complete-data decisions and are less prone to extraction than averaging in multiply imputed data. Although stacking and majority vote pooling strategies both outperformed averaging in replicating selection decisions had the data been complete, we recommend that researchers adopt a stacking strategy for practical reasons. This is because stacking only requires that the researcher fit the model to one dataset and, therefore, makes model comparison more manageable so that researchers are better able to synthesize the evidence provided by the AIC, BIC, and aBIC as a whole. Moreover, we have found that convergence issues are less prevalent with stacked datasets, and time-to-converge when many imputations are used (e.g.,  $M = 100$ ) is only marginally increased compared to fitting the model to a single imputed dataset. Thus, stacking represents a practical remedial strategy that performs better than averaging.

Our proposed remedial strategies represent an important step in addressing issues of model selection in the presence of missing data—issues that have not been given due consideration to this point in previous methodological research. Nevertheless, our proposals represent only a first step in this area of inquiry, and there are many avenues for improvement. For example, our correction relies on an approximation for the fraction of missing information, but it remains unknown how well our approximation reflects the true net evidence loss.

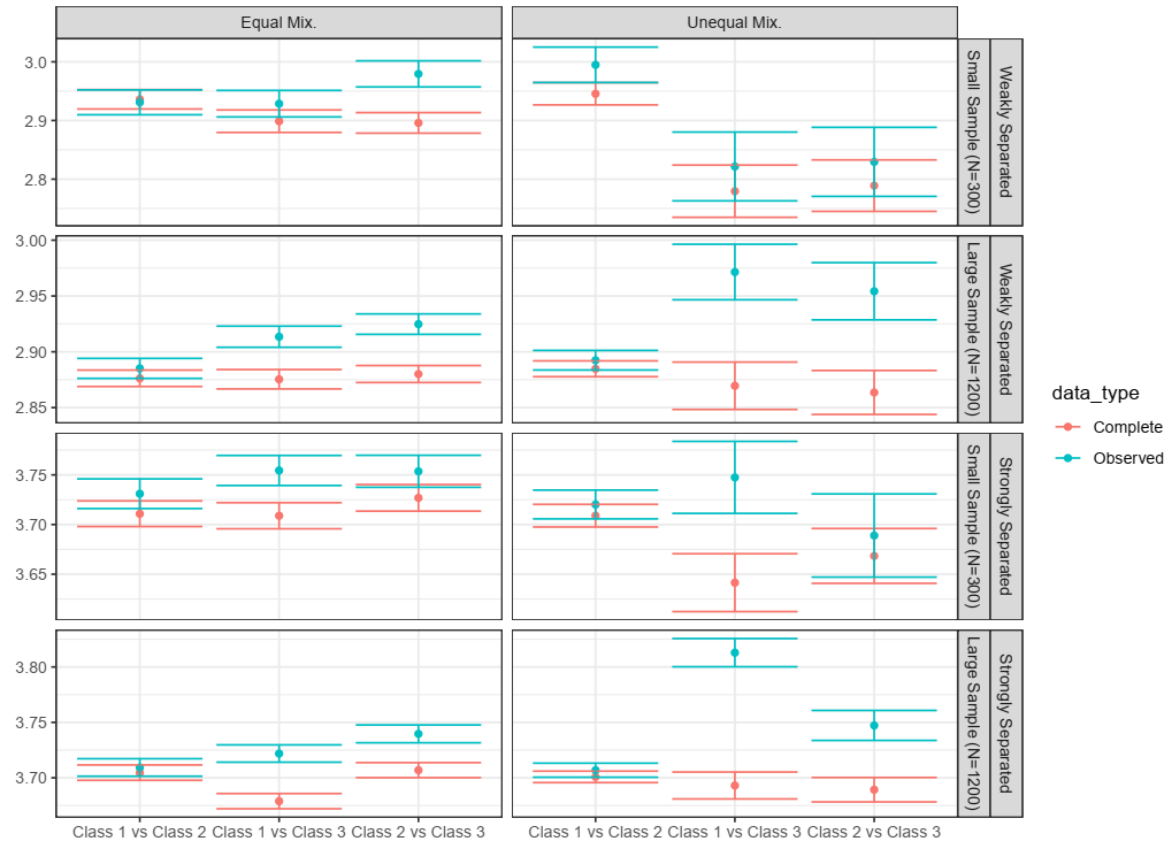
Future research should explore practical methods for extracting better estimates of the amount of information lost due to missing data. Following the approach by Rubin and Schenker (1986), this could be done by simulating plausible values given the maximum likelihood estimates. However, it would be most time efficient if, for example, such information could be extracted by studying the rate of convergence of the EM algorithm near the maximum likelihood estimate.

Additionally, future research should identify theoretically justified statistical procedures for unbiasedly estimating the complete data model deviance if given multiply imputed data. Although stacking resulted in model selections consistent with complete-data decisions in our simulations, it is limited because it is an ad-hoc strategy which fails to be founded in missing data theory. Identifying a theoretically justified pooling strategy is especially important because stacking has not been found to perform well outside of the finite mixture model selection context, such as with covariate selection in linear regression models. Future research should evaluate whether a stacking approach generalizes to other mixture contexts, such as latent class models, latent growth curve models, factor mixture models, and mixed indicator models.

## Figures

**Figure 4.1**

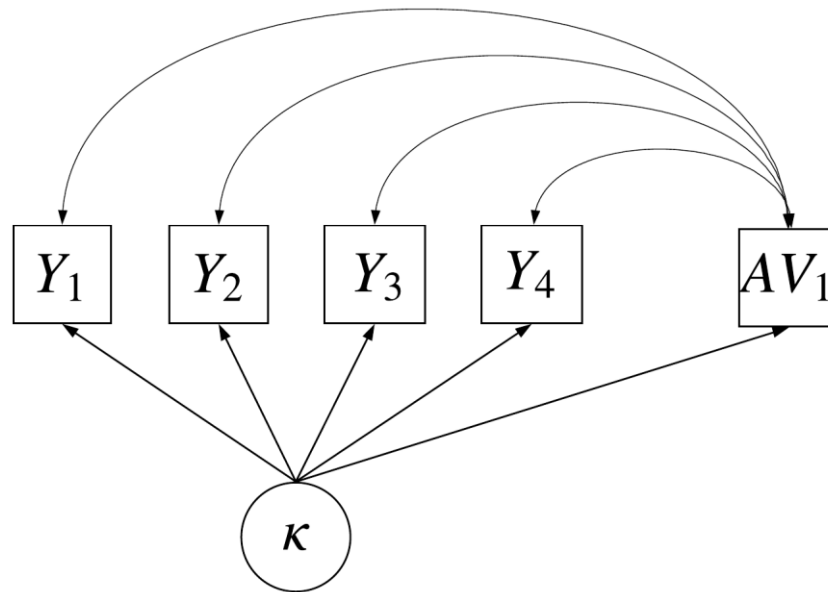
*Mahalanobis Distance Across Simulated Conditions*



*Notes.* Class separation implied by fitted models to complete indicator-only data or observed indicator-only data. A FIML strategy is employed with the observed data. The increased class separation with the observed data treated with FIML are the result of nonresponse bias.

**Figure 4.2**

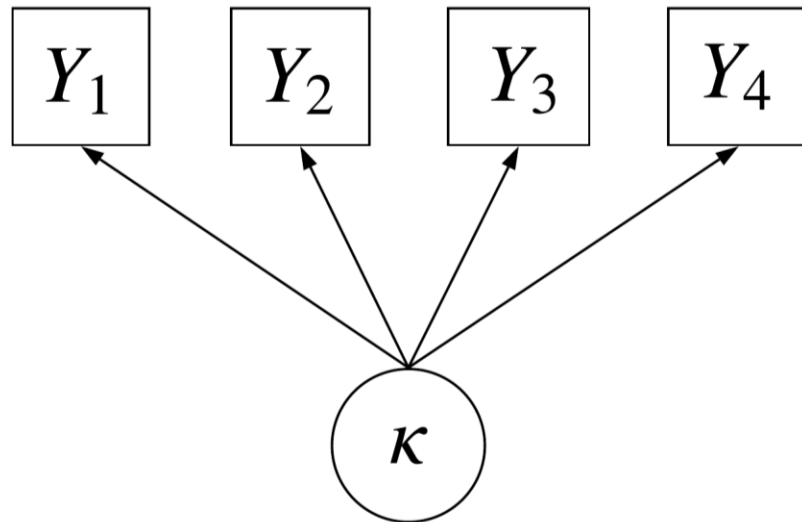
*Template Model*



*Notes.* Data generating (i.e., template) finite mixture of Gaussians model to construct complete and imputed data for the simulation studies. AV denotes auxiliary variable.

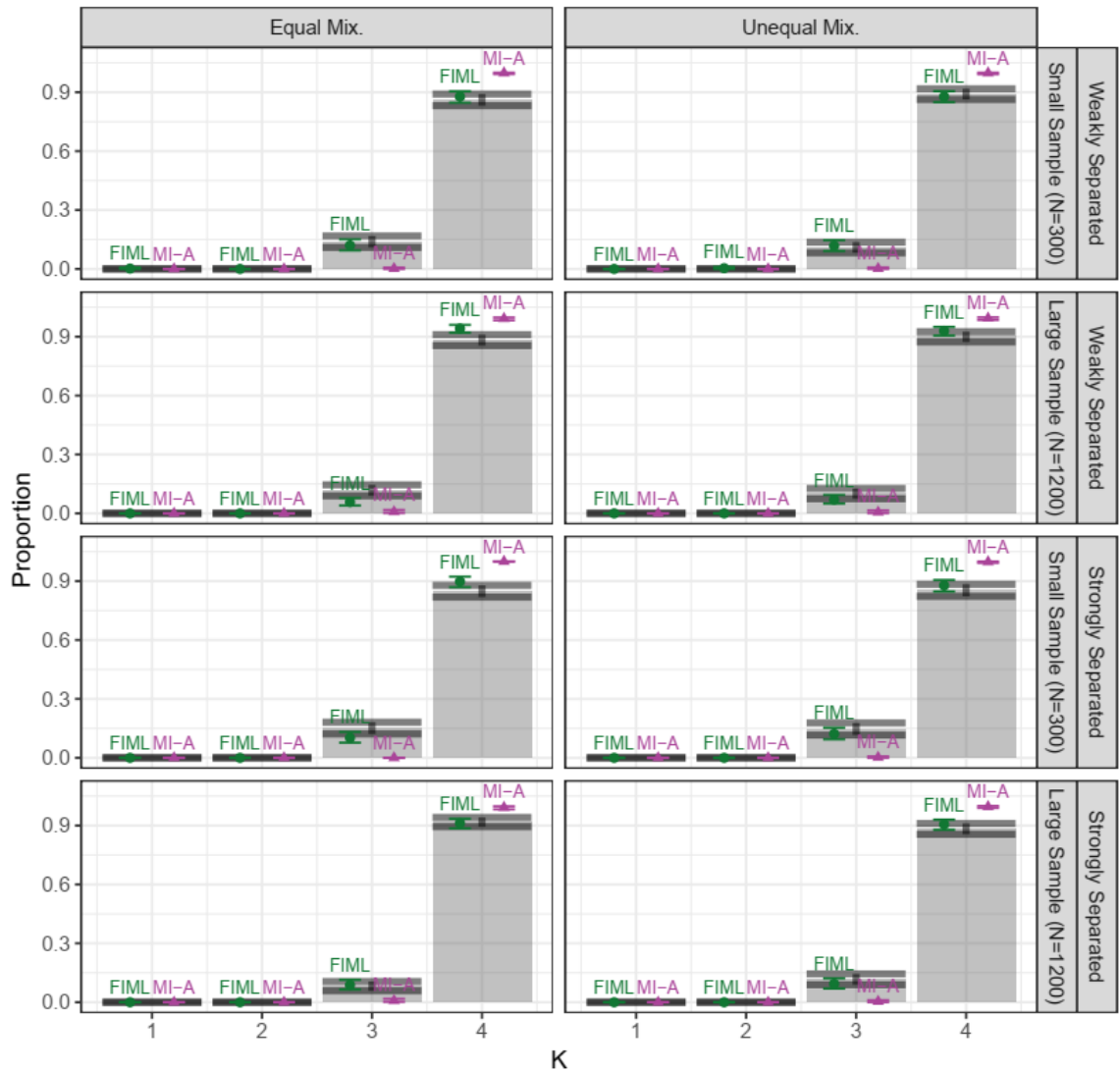
**Figure 4.3**

*Analytic Model*



**Figure 4.4**

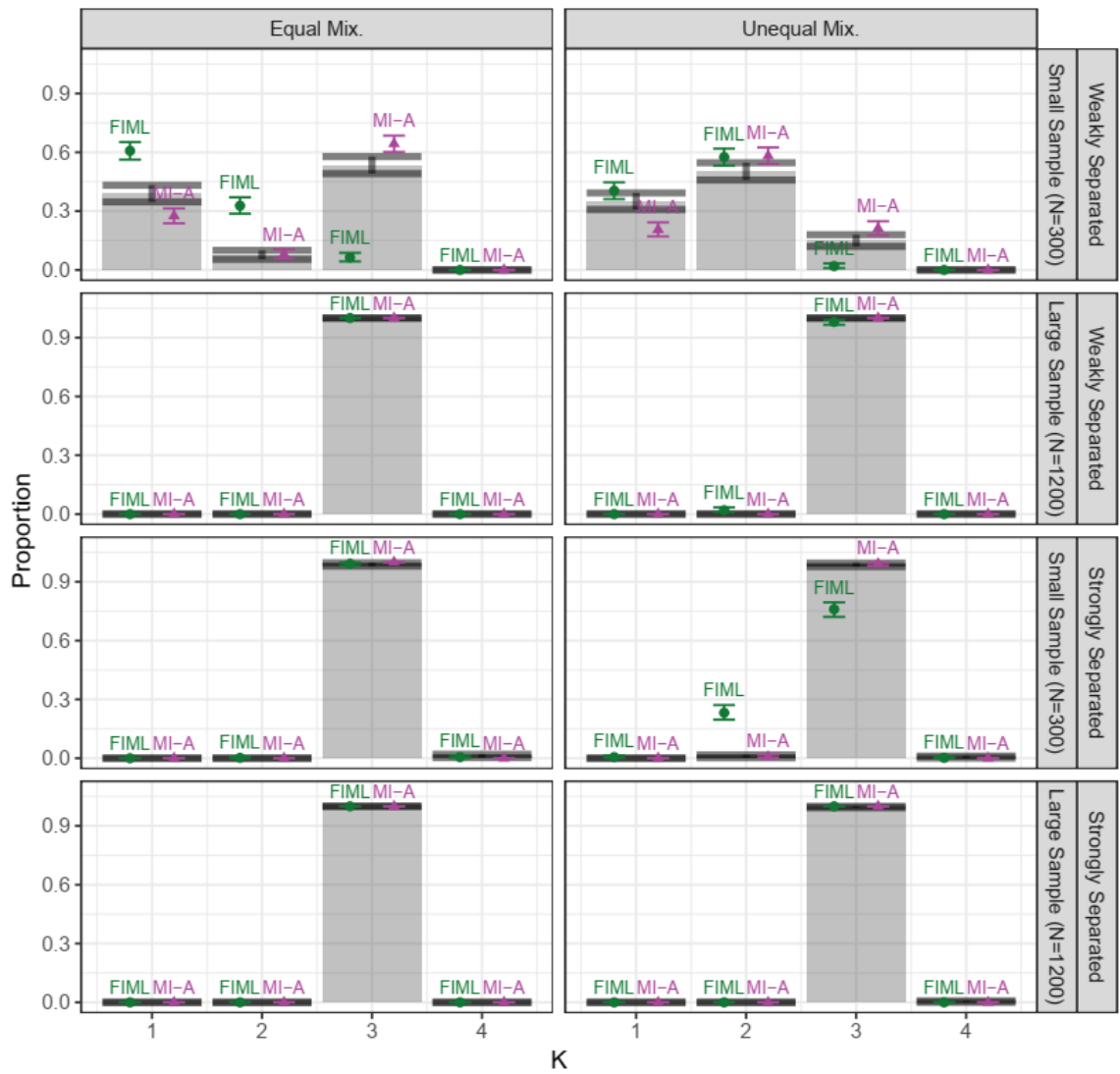
*AIC Model Selection*



*Notes.* AIC model selection decisions using common approaches to treat the missing data. Complete data results shown in grey. A FIML strategy with the AIC calculated using (4.14) is shown in green. Multiple imputation selection decisions obtained by pooling by averaging are shown in purple.

**Figure 4.5**

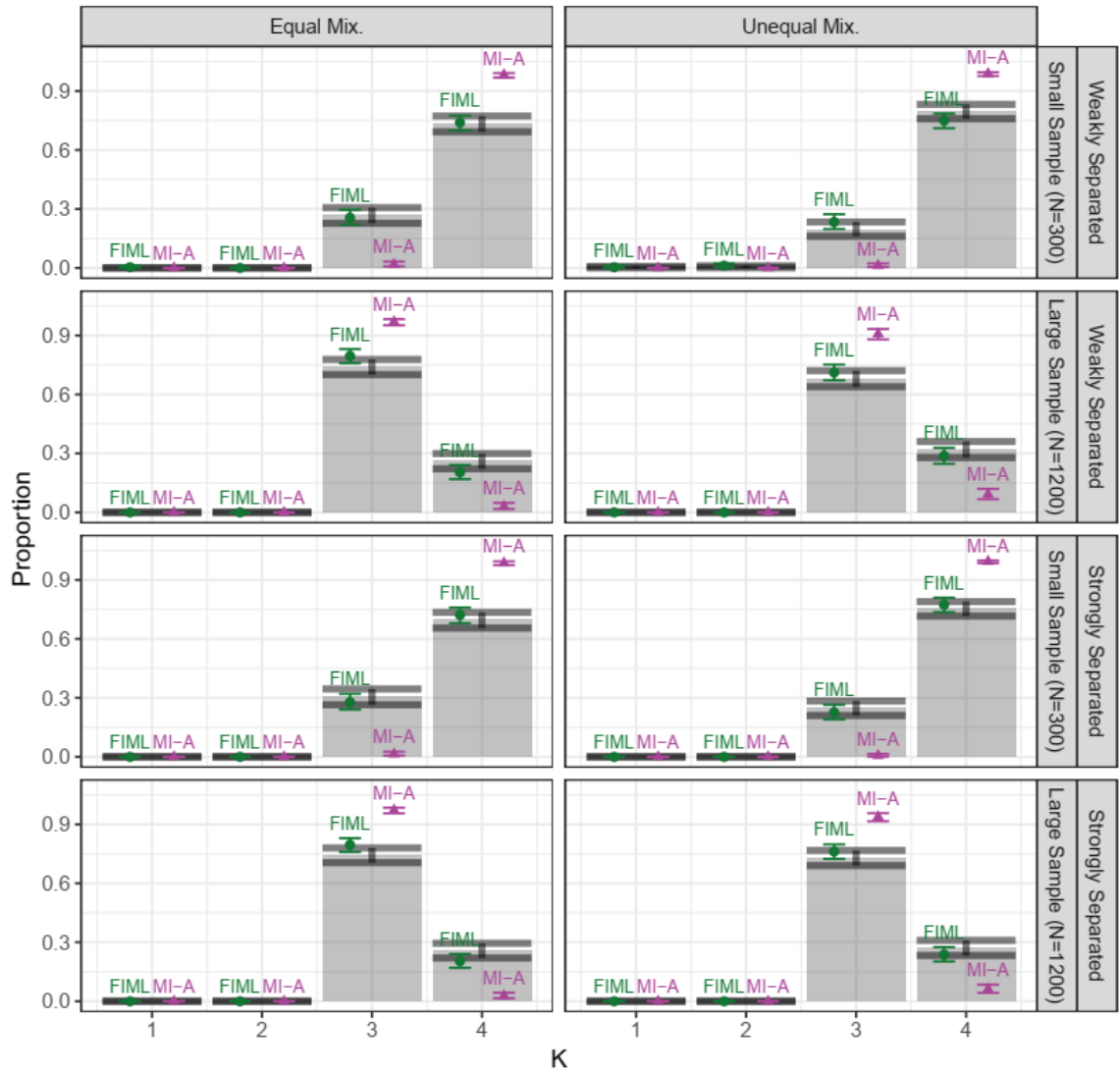
*BIC Model Selection*



*Notes.* BIC model selection decisions using common approaches to treat the missing data. Complete data results shown in grey. A FIML strategy with the BIC calculated using (4.15) is shown in green. Multiple imputation selection decisions obtained by pooling by averaging are shown in purple.

**Figure 4.6**

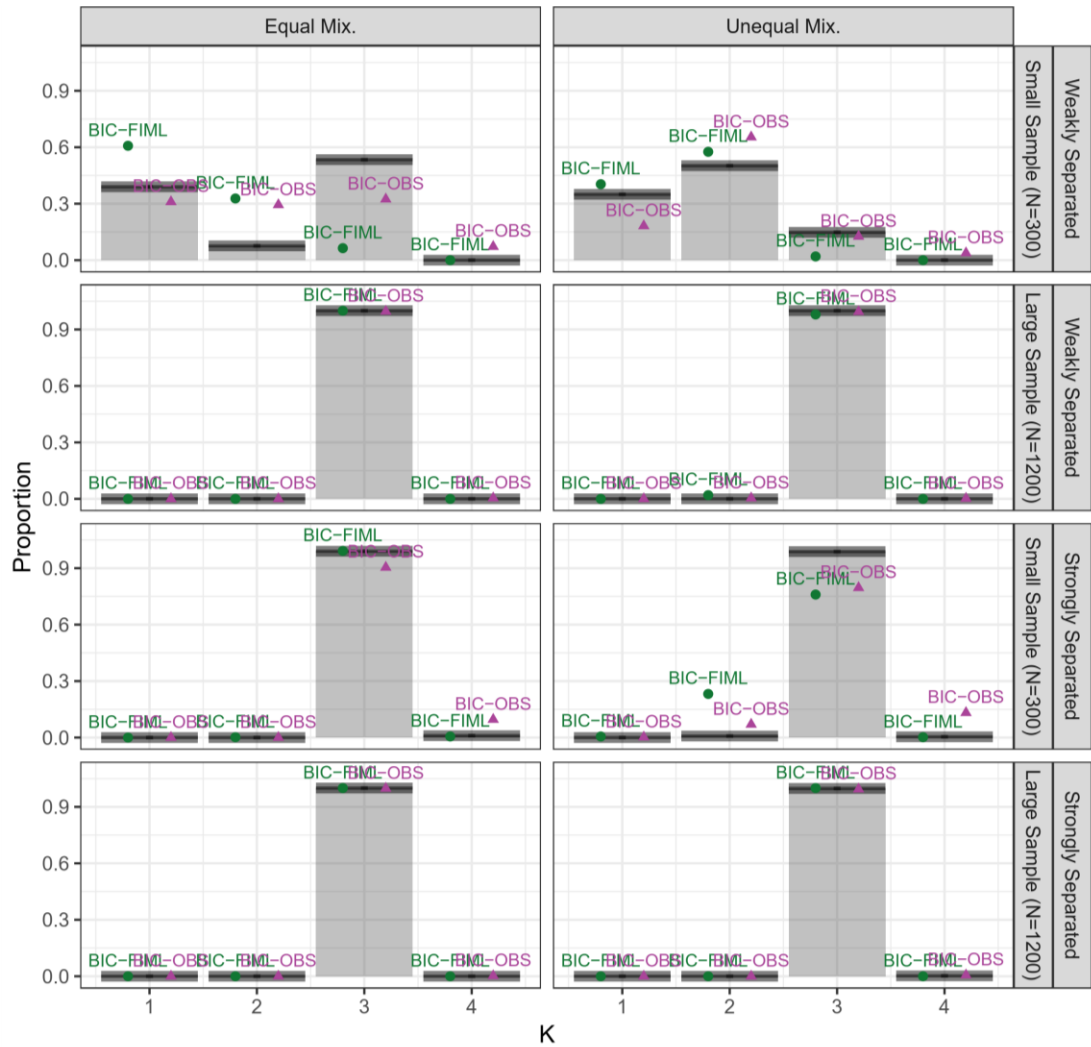
*aBIC Model Selection*



*Notes.* aBIC model selection decisions using common approaches to treat the missing data. Complete data results shown in grey. A FIML strategy with the aBIC calculated using (4.16) is shown in green. Multiple imputation selection decisions obtained by pooling by averaging are shown in purple.

**Figure 4.7**

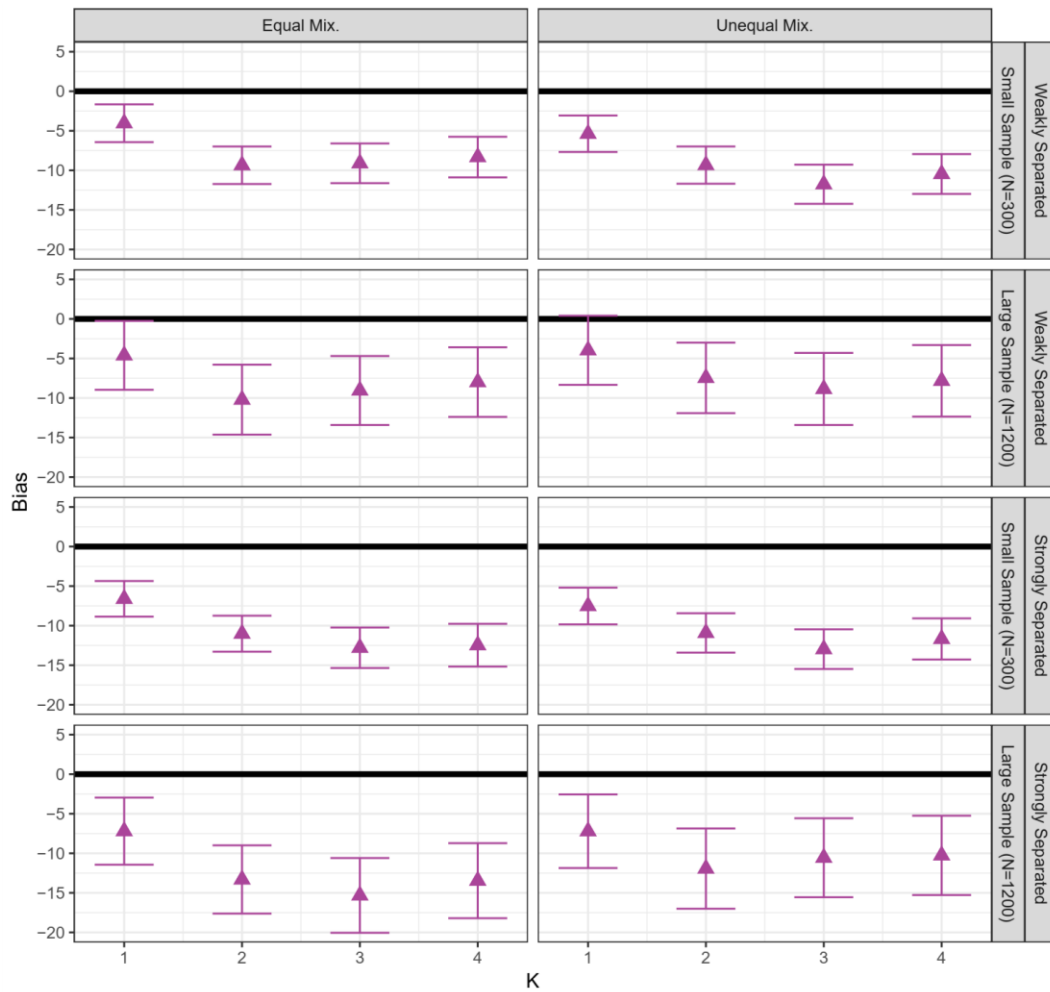
*Model Selection Decisions after Correcting the BIC for Evidence Loss*



*Notes.* Comparing BIC model selection results between the traditional FIML procedure that uses the observed-data likelihood ( $BIC^{FIML}$ ) and a proposed sample-size correction procedure ( $BIC^{OBS}$ ) that adjust the sample size to account for the fact that  $N$  complete observations are not observed so that there is a loss in model evidence.

**Figure 4.8**

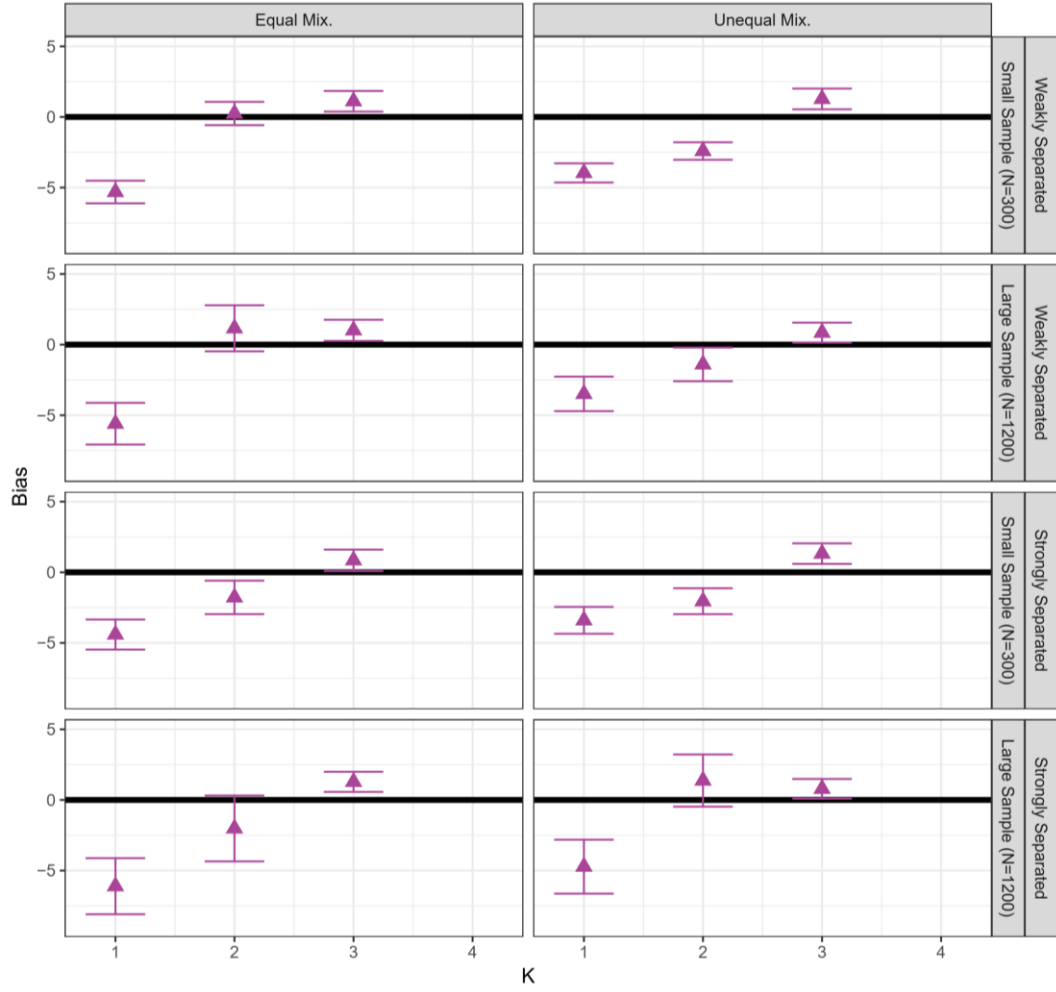
*Pooled Model Deviances by Averaging*



*Notes.* Bias in the model deviance between the deviance calculated by averaging and the deviance that would have been calculated had the data been complete. Standard errors calculated by bootstrapping across replications.

**Figure 4.9**

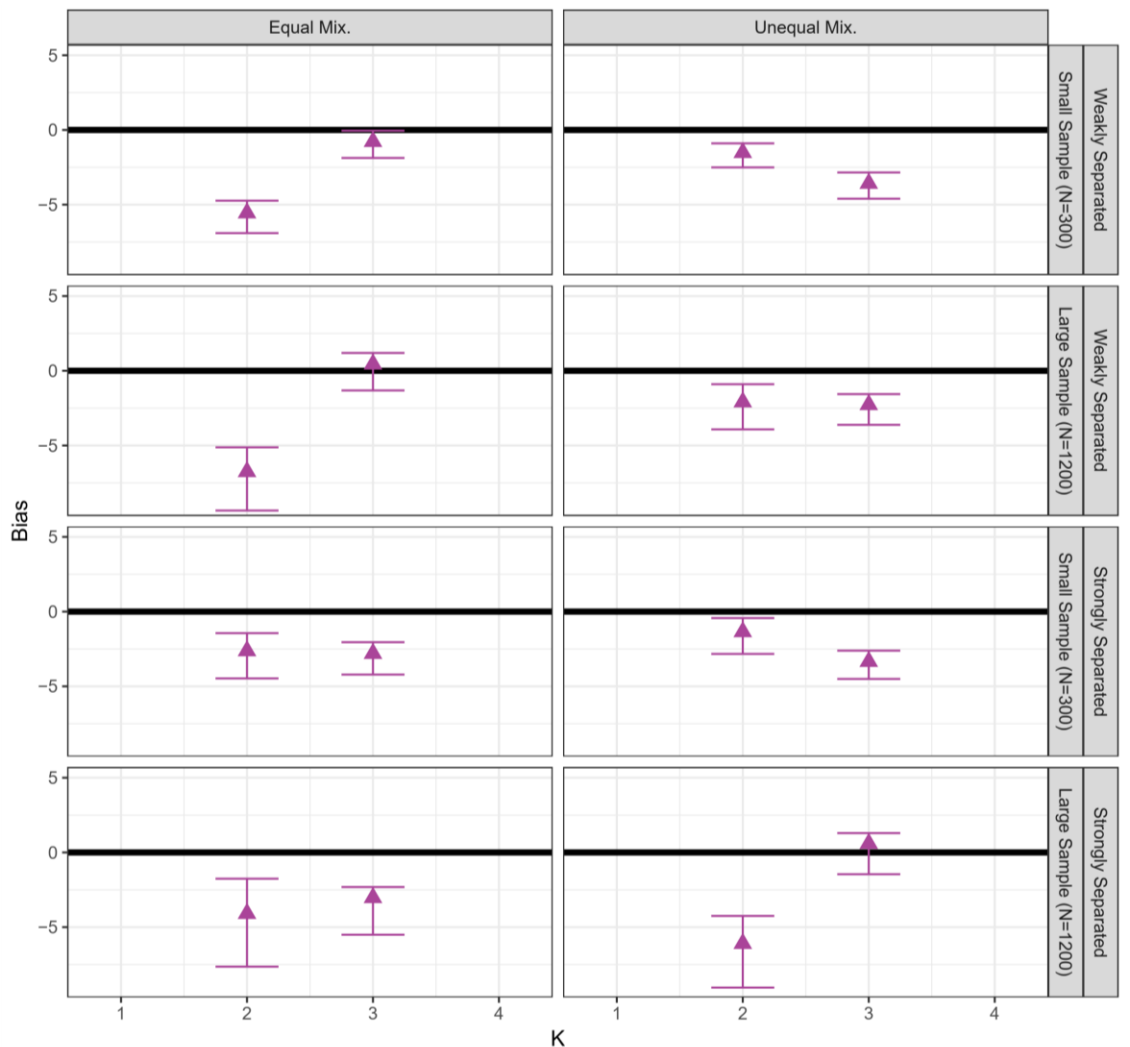
*First Forward Differences of the Pooled Model Deviances by Averaging*



*Notes.* Bias in the rate of change of the model deviance between model  $\mathcal{M}_{K+1}$  and  $\mathcal{M}_K$ . Standard errors calculated by bootstrapping across replications.

Figure 4.10

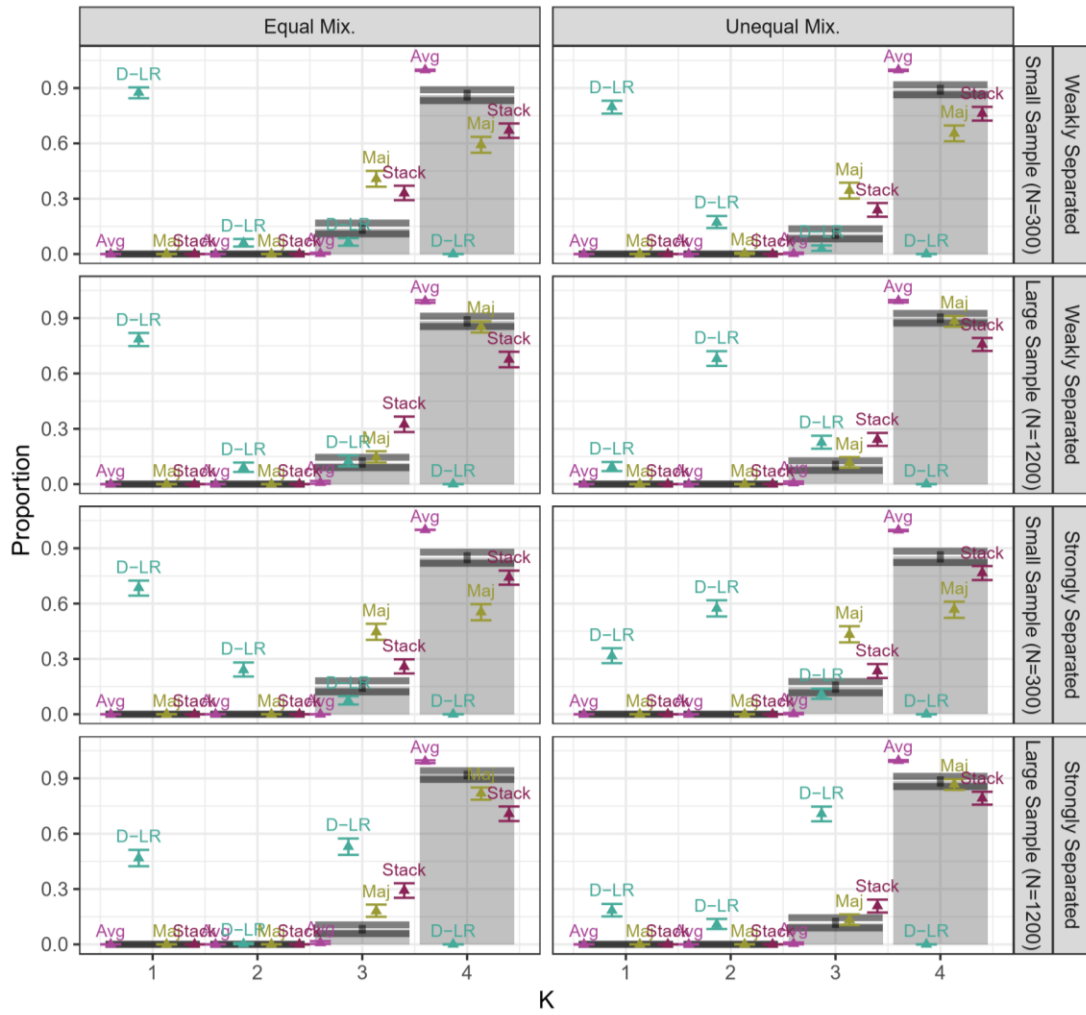
*Second Central Differences of the Pooled Model Deviances by Averaging*



*Notes.* Bias in the second-order rate of change of the model deviance evaluated at  $\mathcal{M}_K$ . Standard errors calculated by bootstrapping across replications.

**Figure 4.11**

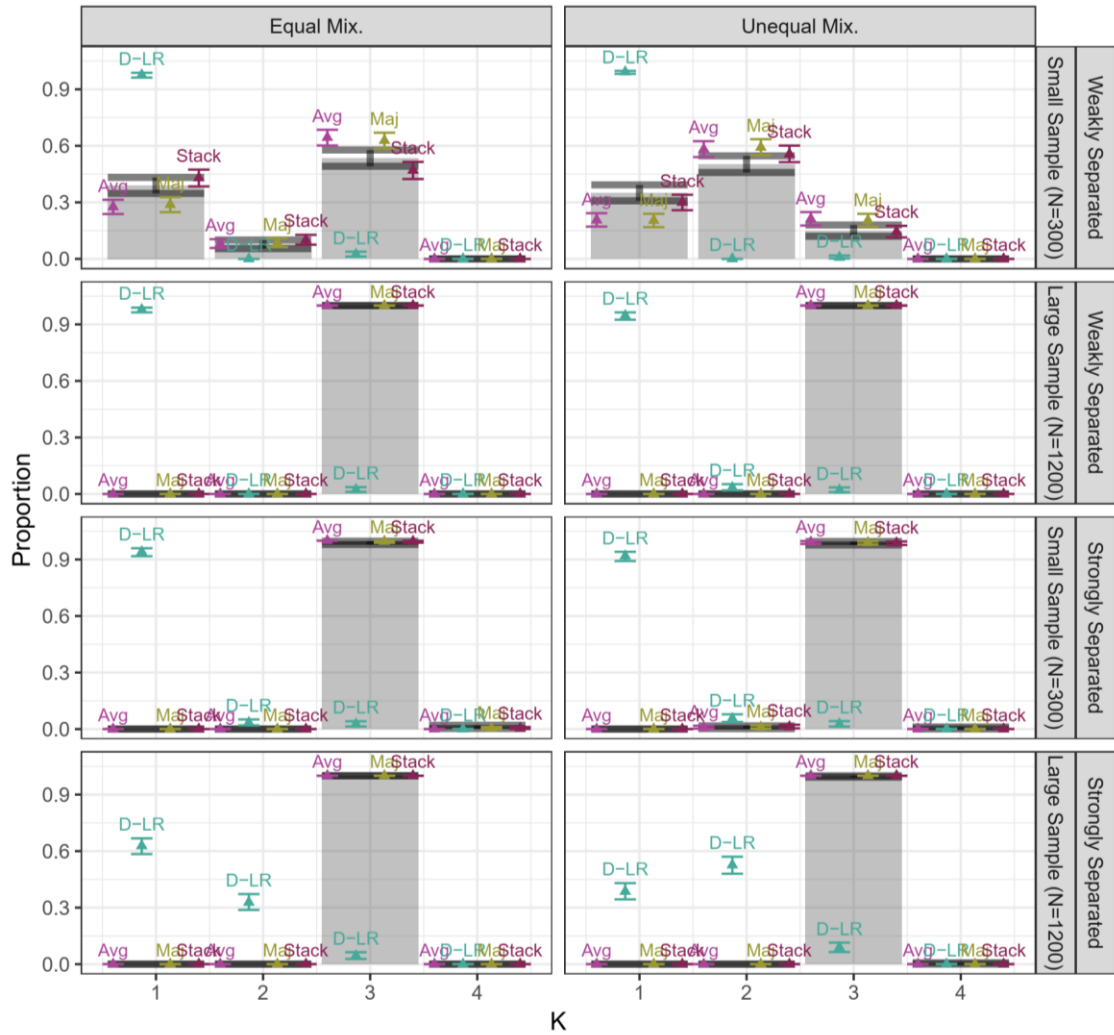
*AIC: Multiple Imputation Model Selection Decisions by Pooling Strategy*



*Notes.* Avg – pooling by averaging; D-LR – pooling by the  $D_{LR}$  statistic; Maj – pooling by majority vote; Stack – pooling by stacking.

**Figure 4.12**

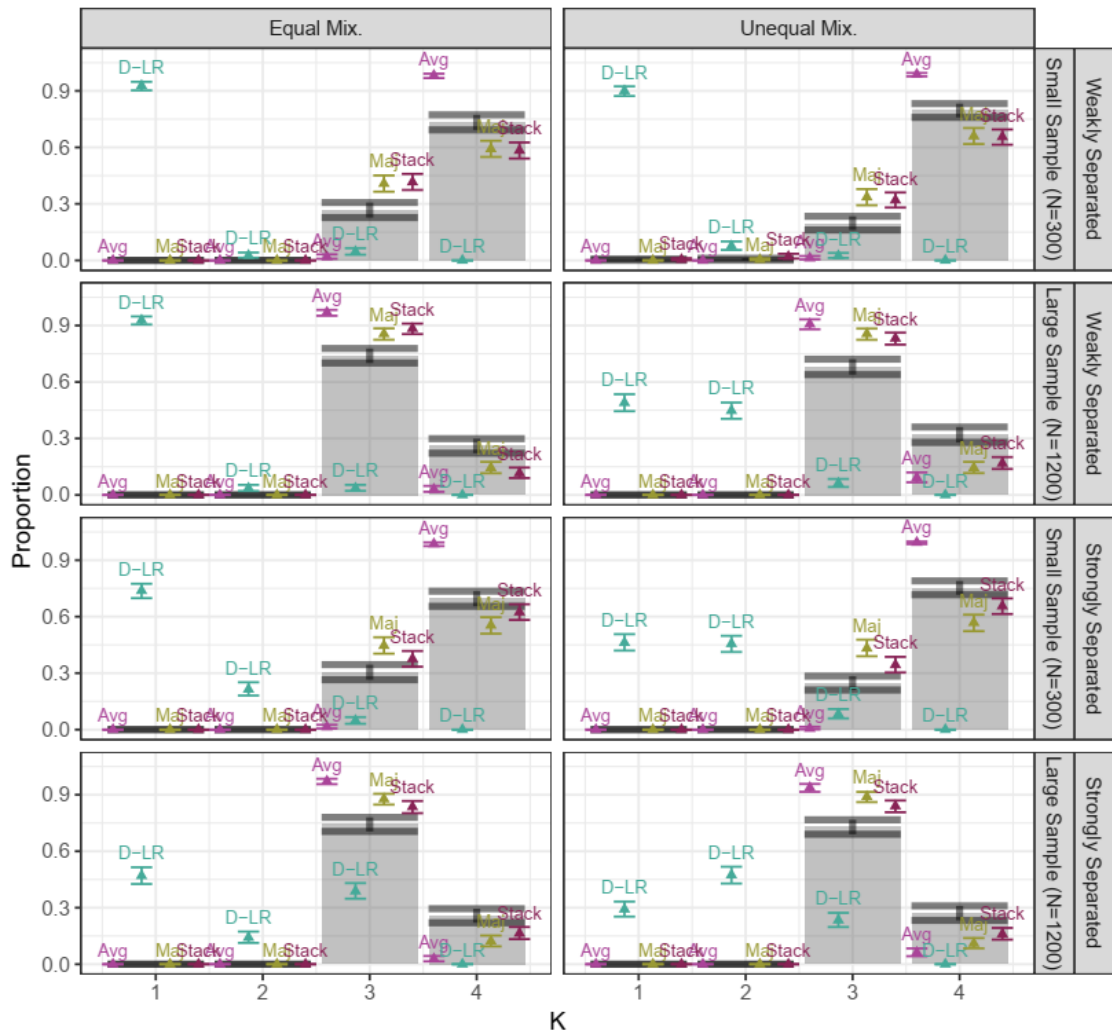
*BIC: Multiple Imputation Model Selection Decisions by Pooling Strategy*



*Notes.* Avg – pooling by averaging; D-LR – pooling by the  $D_{LR}$  statistic; Maj – pooling by majority vote; Stack – pooling by stacking.

**Figure 4.13**

*BIC: Multiple Imputation Model Selection Decisions by Pooling Strategy*



*Notes.* Avg – pooling by averaging; D-LR – pooling by the  $D_{LR}$  statistic; Maj – pooling by majority vote; Stack – pooling by stacking.

## References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Bowers, A. J., & Sprott, R. (2012). Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *Journal of Educational Research*, 105(3), 176–195. <https://doi.org/10.1080/00220671.2011.552075>
- Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. [PhD thesis, Erasmus University, Rotterdam]. RePub. Retrieved from <https://repub.eur.nl/pub/19790>
- Celeux, G., Fruewirth-Schnatter, S., & Robert, C. P. (2018). Model selection for mixture models - perspectives and strategies. In S. Frühwirth-Schnatter, G. Celeux, & C. P. Robert (Eds.), *Handbook of mixture analysis* (pp. 117–154). New York, NY: CRC Press.
- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2), 195–212. <https://doi.org/10.1007/BF01246098>
- Chaurasia, A., & Harel, O. (2012). Using AIC in multiple linear regression framework with multiply imputed data. *Health Services and Outcomes Research Methodology*, 12(2–3), 219–233.
- Chernick, M. R. (2011). *An introduction to bootstrap methods with applications to R*. (R. A. LaBudde, Ed.). Hoboken, New Jersey: Wiley.
- Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351. <https://doi.org/10.1037/1082-989X.6.4.330>
- Consentino, F., & Claeskens, G. (2010). Order selection tests with multiply imputed data. *Computational Statistics & Data Analysis*, 54(10), 2284–2295.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.
- Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2), 323–380.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.
- Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, 18(1), 35–54.
- Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4), 1386–1403.

- Freedman, D. A. (1965). On the asymptotic behavior of Bayes estimates in the discrete case II. *The Annals of Mathematical Statistics*, 36(2), 454–456.
- Frühwirth-Schnatter, S., Celeux, G., & Robert, C. P. (2019). *Handbook of mixture analysis*. Boca Raton: CRC Press, Taylor & Francis Group.
- Grimm, K. J., Mazza, G. L., & Davoudzadeh, P. (2017). Model selection in finite mixture models: A k-fold cross-validation approach. *Structural Equation Modeling*, 24(2), 246–256. <https://doi.org/10.1080/10705511.2016.1250638>
- Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621–638.
- He, J., & Fan, X. (2019). Evaluating the performance of the k-fold cross-validation approach for model selection in growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1), 66–79. <https://doi.org/10.1080/10705511.2018.1500140>
- Kaplan, D. (2014). *Bayesian Statistics for the Social Sciences*. New York, NY: The Guilford Press.
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), 49–69.
- Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. New York, NY: Springer New York.
- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, N.J.: John Wiley & Sons.
- Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, 39(2), 151–162. <https://doi.org/10.1093/jpepsy/jst048>
- Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767–778.
- Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The oxford handbook of quantitative methods* (pp. 551–611). New York, NY: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199934898.013.0025>
- McCoy, D. C., Jones, S., Roy, A., & Raver, C. (2017). Classifying trajectories of social-emotional difficulties through elementary school: Impacts of the Chicago School Readiness Project. *Developmental Psychology*, e-pub ahead of print.
- McLachlan, G. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, 36(3), 318. <https://doi.org/10.2307/2347790>
- McLachlan, G., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.).

- Hoboken, N.J.: Wiley-Interscience.
- McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York, NY: John Wiley & Sons.
- McNeish, D., & Harring, J. R. (2017). Correcting model fit Criteria for small sample latent growth models with incomplete data. *Educational and Psychological Measurement*, 77(6), 990–1018. <https://doi.org/10.1177/0013164416661824>
- Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79(1), 103–111.
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, California: Muthén & Muthén. Retrieved from [https://www.statmodel.com/html\\_ug.shtml](https://www.statmodel.com/html_ug.shtml)
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14(4), 535–569.
- Ostrander, R., Herman, K., Sikorski, J., Mascendaro, P., & Lambert, S. (2008). Patterns of psychopathology in children with ADHD: A latent profile analysis. *Journal of Clinical Child & Adolescent Psychology*, 37(4), 833–847. <https://doi.org/10.1080/15374410802359668>
- R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from <https://www.r-project.org/>
- Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development*. <https://doi.org/10.1177/0165025409343765>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9(1), 130–134. Retrieved from <http://www.jstor.org/stable/2240875>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons. <https://doi.org/10.1002/9780470316696>
- Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366.
- Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(3), 477–494. <https://doi.org/10.1080/10705511.2012.687669>
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art.

- Psychological Methods*, 7(2), 147.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Sclove, S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3), 333–343.  
<https://doi.org/10.1007/BF02294360>
- Sterba, S. K. (2016). Cautions on the use of multiple imputation when selecting between latent categorical versus continuous models for psychological constructs. *Journal of Clinical Child & Adolescent Psychology*, 45(2), 167–175.
- Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Charlotte, NC: Information Age Pub.
- van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC Press.
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC Press.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.
- Vermunt, J. K., & Magdison, J. (2016). Upgrade Manual for Latent GOLD 5.1. Retrieved from <https://www.statisticalinnovations.com/wp-content/uploads/UpgradeManual5.1.pdf>
- Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., Andries, L., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38(1), 369–397.
- Vidotto, D., Vermunt, J. K., & Kaptein, M. C. (2015). Multiple imputation of missing categorical data using latent class models: State of the art. *Psychological Test and Assessment Modeling*, 57(4), 542–576.
- Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *The Public Opinion Quarterly*, 74(2), 223–243. Retrieved from <http://www.jstor.org/stable/40660641>
- Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27(17), 3227–3246.

## **Chapter 5: Epilogue**

In this chapter, I situate the contributions of my dissertation within the historical context of missing data as a source of methodological inquiry. I then provide three tangible recommendations to applied researchers in the field based on the lessons learned from the studies in this dissertation. I close by offering how a follow-up study can be designed to provide more precise guidance to applied researchers.

### **Contextualizing the Contributions in this Dissertation**

Over the past fifty years, statisticians and methodologists developed an extensive methodological toolkit to tackle the problems associated with missing data. Two gold-standard approaches derived from this toolkit include full information maximum likelihood (FIML) and multiple imputation. Because procedures result in valid inference so long as the missing data patterns depend only on observables (i.e., the data are missing at random [MAR]), influential methodologists have labeled FIML and multiple imputation as both “state of the art” (Schafer & Graham, 2002) methods. Twenty years ago, both strategies were generally accepted as equally viable for effectively treating missing data in the analytic strategies employed at that time.

However, applied researchers in education and the behavioral sciences have rapidly begun adopting more modern statistical techniques than the traditional variable-centered approaches. Specifically, researchers are now asking research questions that seek more nuance in explaining how individuals in a population may differ from one another on some set of outcomes or trajectories. Answering these research questions requires employing person-centered approaches, such as latent profile analysis (LPA). Enders (2010) along with Enders and Gottschall (2011) were the first to warn that

multiple imputation, as currently implemented by applied researchers, may not result in valid inference because the imputation models available in software at that time do not reflect the multiple-group structure in the data. Simulation studies by Sterba (2016) subsequently confirmed that multiple imputation conducted using single-class models does not result in valid inferences.

The three studies produced by Enders (2010), Enders and Gottschall (2011), and Sterba (2016) have had profound consequences on the missing data practices adopted by applied researchers. Although each of these researchers were careful to note that the source of the issue was the single-class models that generated the imputations, this nuance has been lost in translating and explaining their findings and generating recommendations to applied researchers. In effect, many applied researchers are under the impression that multiple imputation is fundamentally inviable as a missing data approach in person-centered analysis. Combined with the fact that FIML is the default estimator in software, the result has been that the expansive missing data toolkit available in variable-centered approaches (i.e, multiple imputation and FIML) has been restricted to FIML in person-centered analyses.

A decade ago, the complete reliance on FIML by applied researchers to treat missing data in person-centered analyses was understandable. After all, software was limited in the options available to generate imputations, and all of the available options inappropriately assumed a single-class structure in the data. Complicating matters further, editors and reviewers of journals appear to remain content with the manner in which FIML is implemented in practice, without due consideration to possible threats to validity when enumerating the classes, conducting model selection, and defining the classes. In

particular, it is common for reviewers and editors to be satisfied with an analysis in which the researcher assumes that the MAR condition is tenable given the observed profile indicator values only. Although missingness at random is not fully testable because the missingness may depend on unobservables, the tenability of the MAR assumption can easily be ruled out if other variables available in the dataset predict the missing data patterns in the profile indicators. It stands to reason that because such an investigation has not been a requirement for publication, applied researchers are operating under the assumption that fitting a mixture model to the profile indicators using FIML is sufficient to guard against possible threats to valid inference.

This dissertation makes it clear that such an operating assumption is misguided. In fact, in real-world settings where the missing data mechanism is not under the direct control of the researcher, it is far more likely that the MAR assumption is violated. Recognizing this problem, methodologists have long implored researchers to adopt an inclusive strategy by incorporating auxiliary variables into the missing data strategy. This applies in any analysis, regardless of whether a variable-centered or person-centered approach is taken, as it is well-established that violations to the MAR assumption are known to result in nonresponse bias. Indeed, I showed in Chapter 2 that key parameters of substantive interest are biased if implementing current missing data practices (i.e., FIML) when conducting an LPA in the more realistic situation that MAR is only tenable conditional on a set of variables ancillary to the main analysis. This includes bias in class-specific means which threatens the validity of conclusions regarding the definitions of the classes. It also includes bias in the marginal class probabilities, suggesting that individuals are being inappropriately assigned to the wrong subpopulation.

Additionally, current missing data practices also threaten the validity of model selection decisions during enumeration. In fact, I showed that missing data threatens model selection decisions if a FIML approach is taken. This is true even in the case where the MAR assumption is tenable and results because the penalty term in the BIC assumes all  $N$  observations contain complete information. In particular, the BIC obtained when estimating using FIML tends to lead to model selection decisions where the researcher under extracts the true number of classes. Consequently, current practices risk obfuscating the important sources of heterogeneity in the data that motivated the analysis in the first place.

Taken together, I have made it clear in this dissertation that education and behavioral science researchers should not be satisfied with current practices for treating missing data in person-centered analysis. Naturally, the next question is how to address the limitations I have identified. Although there are no easy solutions to address all of the problems missing data poses in a person-centered analysis like LPA, the studies contained in this dissertation offer several important starting points.

First, I demonstrated that FIML is fundamentally not amenable to an inclusive missing data strategy because auxiliary variables cannot be incorporated without unduly influencing the class definitions and sacrificing the definitions of the classes (Chapter 2). Thus, addressing limitations of current practice requires looking outside of a FIML strategy entirely.

My underlying premise in this dissertation was that the field has been too quick to close the door on multiple imputation as a viable strategy that can overcome the extant limitations. Although the limited options available in imputation software made it

understandable for researchers to not consider multiple imputation a decade ago, several modern imputation approaches are now available. Among these are the recursive partitioning algorithms available in software which do not impose a single-class structure when constructing the imputations. I showed that in some data conditions experienced by applied researchers (i.e., sample-sizes greater than  $N = 1,200$  with classes exhibiting strong separation so that entropy values averaged .88), recursive partitioning imputation—and classification and regression tree (CART) imputation, in particular—mitigate nonresponse bias associated with FIML when the MAR condition requires auxiliary variables (see Chapter 2). Still, there are many common situations where CART imputation failed to adequately attenuate nonresponse bias, namely in small samples of  $N = 300$  or even in large-sample ( $N = 1,200$ ) settings where class separation is weak (i.e., entropies averaged near .74) and a small class is identified (i.e., a class representing 10% or less of the sample). Although recursive partitioning is not a panacea to treating missing data in LPA, the importance of this study lies in the fact that it opens a door that was previously shut, paving the way for future methodological research to ensure that all options are available in the missing data toolkit. Afterall, I demonstrated that if the imputation model is sufficiently congenial with the true data generating mechanism, then multiple imputation can easily incorporate an inclusive missing data strategy—a strategy that cannot be incorporated in FIML. In other words, Chapter 2 provided a proof of concept for the benefits of multiple imputation over FIML and provided a rationale for why future methodological work should focus on multiple imputation in person-centered analysis.

Clearly, multiple imputation holds strong potential for more effectively treating missing data than FIML in person-centered analysis like LPA, but several technical challenges must be acknowledged and resolved before multiple imputation can become mainstream in treating missing data in a person-centered analysis. First, greater congeniality than what is offered by recursive partition imputation is required for multiple imputation to be effective in small-sample settings (e.g.,  $N = 300$ ) or when class separation is weak (entropy near .74) and a small class (one that represents 10% or less of the population) is present. Although perfect congeniality is impossible because the number of subpopulations in a person-centered analysis is unknown, the initial study in Chapter 3 explored whether a proposed hybrid imputation procedure that utilizes Bayesian model averaging can provide the requisite congeniality and improve inferences. Surprisingly, the proposed procedure poorly replicated complete data results using an empirical example, indicating that the procedure will need to be fine-tuned before it can be recommended to applied researchers.

The pooling phase of model fit information provided by information criteria is the final technical challenge that will need to be explored for multiple imputation to become mainstream. This will ensure the robustness of inferences regarding model selection. The default in software is to pool the information criteria by averaging. I demonstrated that such a technique poorly replicates the decisions that would have been made had the data been complete and that a better strategy is to obtain a pooled information criteria value by stacking the imputed datasets (Chapter 4).

In summary, it remains an open question how best to treat missing data in a person-centered analysis appropriately, and it is clear that future methodological work is

needed to fully resolve the issues with current practices. Even though there are not simple answers to these issues, the simulation studies that formed this dissertation provide valuable insights into some areas where it is clear that best practices can be updated. With this in mind, I now discuss my recommendations to provide guidance to applied researchers conducting LPA in the presence of missing data.

### **Guidance to Applied Researchers**

As methodologists become increasingly aware of the issues regarding how missing data is currently treated, best practices are likely to evolve rapidly in the coming years with the advent of new methodologies. In the following, I provide some recommendations based on the lessons learned from the simulations in this study. As there are no quick fixes to these issues, I strive to incrementally improve current practices with the key findings from the studies in this dissertation.

#### **Recommendation 1: Test the MAR Assumption**

Consistent with what would be expected from missing data theory, the simulations in this study demonstrated that the validity of inferences are threatened when current practices (i.e., fitting a mixture model to the profile indicators by estimating with FIML) are employed. At a minimum, researchers fitting a mixture model using FIML should test whether the tacit assumption that the data are MAR conditional only on the observed indicator values is tenable or not. This can be accomplished by evaluating whether other variables in the dataset are predictive of any of the missing data patterns. Specifically, the MAR assumption is likely violated (and the data are likely missing not at random) if a variable in the dataset other than the profile indicators predicts the missing data patterns, controlling for the observed indicator values. If a researcher

identifies such a variable, then that should warn the researcher that nonresponse bias is possible. The fact that the data are MNAR should be acknowledged in the limitations section, and journal editors should begin holding researchers accountable for describing possible threats to the validity of conclusions induced by missing data. Some researchers will be in a position to conduct a sensitivity analysis, which I now discuss.

### **Recommendation 2: Conduct a Sensitivity Analysis**

The simulation studies in this dissertation show that in settings where the sample size is sufficiently large, the researcher may be positioned to conduct a sensitivity analysis. Specifically, this can be done by fitting the final model to imputed datasets constructed from CART imputation with auxiliary variables. Ideally, the inferences regarding class-specific means and marginal class probabilities from the imputed dataset will align with the conclusions drawn when fitting the mixture model to the profile indicators using FIML.

Researchers should note that the degree to which CART imputation will mitigate nonresponse bias relative to FIML will depend on the separation of the classes when a small class is found. This is true even if the sample size is large. If results are robust across the FIML approach and the imputed datasets generated with CART imputation and the classes are sufficiently separated, then conclusions drawn from the class-specific means and marginal class probabilities are likely robust to the fact that missingness is known to depend on observables other than the profile indicators. In contrast, the classes are not sufficiently separated, the simulations showed that CART imputation will only attenuate nonresponse bias when the classes are equally mixed so that no small class is

found. In this case, CART imputation will provide no additional valuable information on assessing whether the results are sensitive to missing data.

### **Recommendation 3: Adjust the BIC to Account for Evidence Loss if using FIML**

In both small and large samples, applied researchers should use the adjustment for the BIC provided in Chapter 4 to correct for the fact that the BIC leads to model selection decisions with too few classes if mixture models are fit to the profile indicators using FIML. The simulations demonstrated that the proposed adjustment procedure mitigates the tendency for researchers to under extract the number of classes.

### **Future Research Required for More Narrowly Tailored Guidance**

More work is needed to identify when CART imputation would be effective as a sensitivity analysis in an LPA, or even be preferable to a FIML approach entirely. Specifically, the factorial simulation design is limited in identifying useful threshold values that can offer applied researchers more precise data condition values for when CART imputation is valuable for a sensitivity analysis. My recommendation for employing CART imputation in a sensitivity analysis when samples are above  $N = 1,200$  observations follows directly from the fact that the simulation studies defined small samples as  $N = 300$  and large samples as  $N = 1,200$  across all replications. Although these values correspond to ranges in education and psychology literature ( $N = 300$  to  $1,200$  is the interquartile range for a sample of thirty highly cited LPA studies informing the simulation studies), the simulations do not provide clear guidance for intermediate values. For example, would CART imputation perform well if sample size is slightly reduced to, say,  $N = 1,000$ ? Although one can expect that the positive findings would generalize with this sample size, this expectation needs to be verified through a

simulation study tailored to that research question. For instance, a simulation study could be designed where sample size and class separation values are randomly drawn from a uniform distribution at each replication in the simulation. More sophisticated approaches to analyze the resulting data would then be needed in order to smooth the results and provide information on the expected bias at a given sample size and class separation value. Regression analysis or another curve-fitting technique are two options for effectively smoothing the data and determining appropriate threshold values.

Finally, before CART imputation is used in place of FIML, the strong performance of stacking should be evaluated using simulation studies. This is true even in the large-sample settings where the simulations showed it performed well. Specifically, although it is likely that the positive results regarding stacking will generalize to CART imputation, this was not directly tested in Chapter 4 because imputations were drawn from the true data generating probability distribution. Thus, future work should confirm that the results from Chapter 4 generalize to CART imputation in the data conditions where we concluded that CART imputation mitigates nonresponse bias.

### **Summary**

In conclusion, person-centered analysis like LPA presents unique challenges to treating missing data. In this dissertation, I have scrutinized current practices in treating missing data and found many limitations. I have argued that multiple imputation is better suited to treat missing data than the current practice of estimating mixture models fit to profile indicators with FIML, although several challenges remain. Identifying solutions to these challenges is imperative to minimize threats to inference when analyzing real-world data from education and the behavioral sciences.

## References

- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: The Guilford Press.
- Enders, C. K., & Gottschall, A. C. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, 18(1), 35–54.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2), 147.
- Sterba, S. K. (2016). Cautions on the use of multiple imputation when selecting between latent categorical versus continuous models for psychological constructs. *Journal of Clinical Child & Adolescent Psychology*, 45(2), 167–175.

## Technical Appendix A: Useful Principles, Quantities, and Properties in Missing Data Theory

### The Missing Information Principle

The missing information principle (Orchard & Woodbury, 1972) states that the complete-data information matrix is the sum of the observed-data information matrix and the missing information

$$I_{\text{com}}(\hat{\theta}|Y) = I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}) + I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}}). \quad (\text{A. 1})$$

#### *Proof*

The following proof is taken from Schafer (1997, p. 57). The missing information principal follows from the decomposition of the complete-data loglikelihood as the sum of observed and missing components,

$$\log \Pr(Y|\hat{\theta}) = \log \Pr(Y^{\text{obs}}|\hat{\theta}) + \log \Pr(Y^{\text{mis}}|Y^{\text{obs}}, \hat{\theta}).$$

This can be equivalently written as

$$\ell_{\text{com}}(\hat{\theta}|Y) = \ell_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}) + \log \Pr(Y^{\text{mis}}|Y^{\text{obs}}, \hat{\theta}).$$

Taking the negative Hessian of both sides and applying the distributive property of derivative results in the information matrices in (A. 1).

### Fraction of Missing Information

As defined by Schafer (1997, p. 58), the fraction of missing information  $\text{FMI}(\hat{\theta})$  is given by

$$\text{FMI}(\hat{\theta}) = I_{\text{com}}^{-1}(\hat{\theta}|Y)I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}}) = \mathbb{I} - I_{\text{com}}^{-1}(\hat{\theta}|Y)I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}). \quad (\text{A. 2})$$

The latter expression for the fraction of missing information results from solving for  $I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}})$  in (A. 1) and substituting, i.e.,

$$I_{\text{com}}^{-1}(\hat{\theta}|Y)I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}}) = I_{\text{com}}^{-1}(\hat{\theta}|Y) \left( I_{\text{com}}(\hat{\theta}|Y) - I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}) \right)$$

$$= \mathbb{I} - I_{\text{com}}^{-1}(\hat{\theta}|Y)I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}).$$

### Relative Increase in Variance

The relative increase in variance is defined as the proportional increase in the complete-data asymptotic covariance matrix needed to arrive at the observed-data asymptotic covariance matrix, i.e.

$$V_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}) = \left( \mathbb{I} + \text{RIV}(\hat{\theta}) \right) V_{\text{com}}(\hat{\theta}|Y). \quad (\text{A. 3})$$

Solving for  $\text{RIV}(\hat{\theta})$  in (A. 3) results in an expression for the relative increase in variance

$$\text{RIV}(\hat{\theta}) = V_{\text{obs}}(\hat{\theta}|Y^{\text{obs}})V_{\text{com}}^{-1}(\hat{\theta}|Y) - \mathbb{I} = I_{\text{obs}}^{-1}(\hat{\theta}|Y^{\text{obs}})I_{\text{com}}(\hat{\theta}|Y) - \mathbb{I}. \quad (\text{A. 4})$$

### Property 1

The determinant of the identity matrix plus the relative increase in variance matrix is strictly greater than or equal to one:  $|\mathbb{I} + \text{RIV}(\hat{\theta})| \geq 1$ .

### *Proof*

Substituting  $V_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}) = I_{\text{obs}}^{-1}(\hat{\theta}|Y^{\text{obs}})$  and  $V_{\text{com}}(\hat{\theta}|Y) = I_{\text{com}}^{-1}(\hat{\theta}|Y)$  into (A. 3) and then taking the determinant of both sides and then simplifying the expression leads to the following intermediate expression,

$$\frac{1}{|I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}})|} = \frac{|\mathbb{I} + \text{RIV}(\hat{\theta})|}{|I_{\text{com}}(\hat{\theta}|Y)|}.$$

Substituting for  $I_{\text{com}}(\hat{\theta})$  using (A. 1) and solving for  $|\mathbb{I} + \text{RIV}(\hat{\theta})|$  implies that

$$|\mathbb{I} + \text{RIV}(\hat{\theta})| = \frac{|I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}) + I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}})|}{|I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}})|}. \quad (\text{A. 5})$$

Because  $I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}})$  is positive semidefinite the determinant of the combined observed- and missing-data information matrices (i.e, the complete-data information matrix) is strictly greater than the determinant of the observed-data information matrix alone, i.e.,

$$|I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}}) + I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}})| \geq |I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}})|.$$

Applying this inequality into (A. 5) implies that

$$|\mathbb{I} + \text{RIV}(\hat{\theta})| \geq \frac{|I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}})|}{|I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}})|}.$$

Simplifying the above results in the inequality provided in Property 1.

## Property 2

If given the fraction of missing information, the relative increase in variance can be obtained via

$$\text{RIV}(\hat{\theta}) = \text{FMI}(\hat{\theta}) \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1}.$$

## Proof

Solving for  $I_{\text{obs}}(\hat{\theta}|Y^{\text{obs}})$  in (A. 1) and then taking the inverse of the resulting expression leads to the following observed-data information matrix,

$$I_{\text{obs}}^{-1}(\hat{\theta}|Y^{\text{obs}}) = \left( I_{\text{com}}(\hat{\theta}|Y) - I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}}) \right)^{-1} \quad (\text{A. 6})$$

Substituting (A. 6) into (A. 4) leads to the following expression for  $\text{RIV}(\hat{\theta})$ :

$$\text{RIV}(\hat{\theta}) = A(\hat{\theta}) - \mathbb{I}. \quad (\text{A. 7})$$

where

$$A(\hat{\theta}) = \left( I_{\text{com}}(\hat{\theta}|Y) - I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}}) \right)^{-1} I_{\text{com}}(\hat{\theta}|Y). \quad (\text{A. 8})$$

Taking the inverse of both sides in (A. 8),

$$\begin{aligned}
A^{-1}(\hat{\theta}) &= I_{\text{com}}^{-1}(\hat{\theta}|Y) \left( I_{\text{com}}(\hat{\theta}|Y) - I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}}) \right) \\
&= \mathbb{I} - I_{\text{com}}^{-1}(\hat{\theta}|Y) I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}})
\end{aligned} \tag{A.9}$$

Substituting  $\text{FMI}(\hat{\theta}) = I_{\text{com}}^{-1}(\hat{\theta}|Y) I_{\text{mis}}(\hat{\theta}|Y^{\text{mis}})$  from (A.2) into (A.9) and then inverting the left and right hand side results in an expression for  $A(\hat{\theta})$  in terms of  $\text{FMI}(\hat{\theta})$ ,

$$A(\hat{\theta}) = \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1}. \tag{A.10}$$

Substituting (A.10) into (A.7),

$$\text{RIV}(\hat{\theta}) = \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1} - \mathbb{I}. \tag{A.11}$$

Substituting  $\mathbb{I} = \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right) \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1}$  in (A.11),

$$\begin{aligned}
\text{RIV}(\hat{\theta}) &= \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1} - \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right) \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1} \\
&= \left( \mathbb{I} - \mathbb{I} + \text{FMI}(\hat{\theta}) \right) \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1}.
\end{aligned} \tag{A.12}$$

Simplifying (A.12) results in the Property 2.

### Property 3

A useful relationship between the relative increase in variance and the fraction of missing information is as follows:

$$\mathbb{I} + \text{RIV}(\hat{\theta}) = \left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1}.$$

### *Proof*

Solving for  $\left( \mathbb{I} - \text{FMI}(\hat{\theta}) \right)^{-1}$  in (A.11) results in Property 3.

## References

- Orchard, T., & Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proc. Sixth Berkeley Symp. on Math. Statist. And, Prob.*(Vol.1), 697–715.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Raton, FL: CRC Press.

## **Technical Appendix B: Illustrating the Limitations of a Saturated Correlates Approach in Latent Profile Analysis**

To highlight the difficulty in implementing a saturated correlates approach in a latent profile analysis, we utilized data from the Early Childhood Longitudinal Study, Kindergarten (ECLS-K) 1998 cohort. With empirical data we show that implementing Graham's rules (2003) results in individuals switching classes and it leads to undesired changes in class definitions. In other words, while Graham's rules (2003) preserves the interpretation of parameters in regression models, structural equation models, or other covariance structure models, these rules do not preserve the interpretation of the parameters that define the classes in latent profile analysis.

The ECLS-K is a nationally representative, longitudinal study following children enrolled in either full-time or part-time Kindergarten in the 1998-1999 school year to the end of eighth grade. The dataset contains a wide range of achievement, behavioral, psychological, and school-environment outcomes being collected up to two times a school year. We build on an introduction to LPA for applied researchers provided by Berlin, Williams, & Parra (2014) investigating distinct subpopulation of Black, non-Hispanic adolescents in eighth grade who differ according to dietary intake, physical activity, and sedentary behaviors.

### **Measures and Inclusion Criteria**

The constructs of interest include weekly physical activity, sedentary behaviors, and dietary intake. Three items measured physical activities, including participation in school sports (three categories), participation in non-school sports (four categories), days exercised in the past seven days (0-7 days; 8 categories), and average days in PE per

week (0-5 days; 6 categories). Six items measured sedentary behaviors by asking the number of hours per day (0-24 hours; 25 categories) spent watching TV, playing videogames, or using the internet. Nine items measured dietary intake by asking the number of days (0-7 days; 8 categories) specified foods (e.g., carrots, potatoes, fruit, fast food, etc.) or drinks (e.g., a glass of milk, a glass of juice, drank soda, etc.) were consumed.

Because LPA relies on the assumption that indicators are continuous, we parceled items to construct three profile indicators that measure overall physical activity, sedentary behaviors, and dietary intake. The ACTIVITY parcel was constructed by standardizing all items before aggregating. The SEDENTARY and DIETARY parcels were constructed using the items in their raw scales because all items were on the same scale (i.e., hours/day or days/week). The auxiliary variable used in this example is the student's predicted BMI in eighth grade. This value was predicted by fitting a latent growth curve model to BMI data collected in seven waves between Kindergarten and eighth grade.

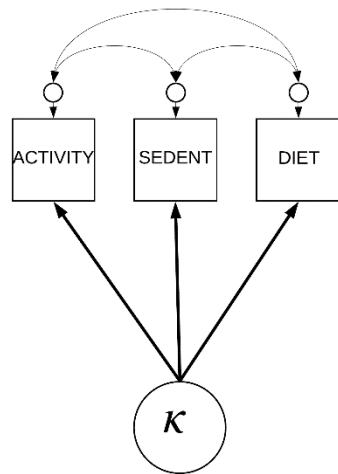
Only data for individuals who identify as Black, non-Hispanic in eighth grade are analyzed in the illustrative example. We required that all class indicator data be present in order to be included in the study. Requiring complete data implies that any difference in the parameter estimates across the specified models is the result of one specification changing the parameter's interpretation, and not differences because missing data are being treated. Applying these inclusion criteria results in a sample size of  $N=608$  girls and boys.

## Analytic Strategy

To identify if Graham's (2003) rules preserves the interpretation of the parameters of the latent classes, we fit two separate mixture models. We specify the first mixture model according to the regular specification. The second mixture model is specified according to Graham's (2003) rules. The total number of classes specified was  $K = 3$ , which was determined by minimizing the BIC.

If Graham's (2003) rules preserves the interpretation of the parameters that define the classes, then the class-specific means should be equivalent across the two fitted models. Similarly, if the two models result in equivalent classifications into the latent classes, then this should be reflected by the marginal class probabilities being equivalent.

### *Regular Specification*



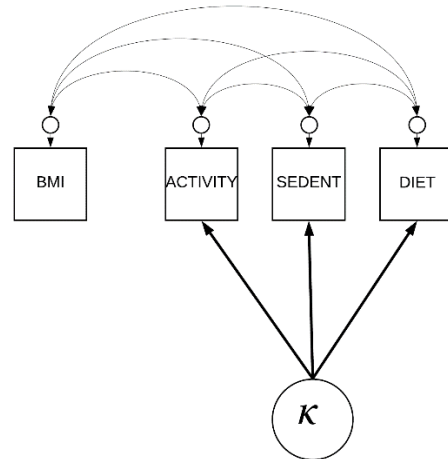
*Notes.* Regular specification of a finite mixture model for latent profile analysis.

Mplus Code:

```
%OVERALL%
[ c#1*2.62697 ];
[ c#2*1.96629 ];

%C#1%
activity WITH sedent*0.02210;
```

### *Graham's (2003) Rules Specification*



*Notes.* Finite mixture model specified according to Graham's (2003) rules.

Mplus Code:

```
%OVERALL%
[ c#1*2.01049 ];
[ c#2*1.56094 ];

%C#1%
```

```

activity WITH diet*0.05300;
sedent WITH diet*-0.05011;

[ activity*-0.02476 ];
[ sedent*2.65952 ];
[ diet*2.26605 ];

activity*0.40506;
sedent*1.21632;
diet*0.26683;

%C#2%

activity WITH sedent*0.28524;
activity WITH diet*0.05147;
sedent WITH diet*-1.00020;

[ activity*0.11883 ];
[ sedent*5.16571 ];
[ diet*3.03186 ];

activity*0.33318;
sedent*5.12648;
diet*0.85182;

%C#3%

activity WITH sedent*1.82689;
activity WITH diet*0.03423;
sedent WITH diet*-0.14302;

[ activity*-0.50638 ];
[ sedent*10.96636 ];
[ diet*3.25351 ];

activity*0.21963;
sedent*28.22388;
diet*1.09756;

```

```

activity WITH sedent*0.02103;
activity WITH diet*0.05051;
activity WITH bmi7hat*-0.43547;
sedent WITH diet*0.04691;
sedent WITH bmi7hat*0.12228;
diet WITH bmi7hat*-0.05482;

[ activity*-0.01788 ];
[ sedent*2.58428 ];
[ diet*2.22146 ];
[ bmi7hat*23.64010 ] (c1);

activity*0.40754;
sedent*1.07994;
diet*0.24012;
bmi7hat*26.41688;

%C#2%

activity WITH sedent*0.22583;
activity WITH diet*0.05782;
activity WITH bmi7hat*-0.09564;
sedent WITH diet*-0.83204;
sedent WITH bmi7hat*-1.07637;
diet WITH bmi7hat*-0.01836;

[ activity*0.07593 ];
[ sedent*4.89019 ];
[ diet*2.90937 ];
[ bmi7hat*23.64010 ] (c1);

activity*0.36193;
sedent*5.17748;
diet*0.66108;
bmi7hat*15.65743;

%C#3%

activity WITH sedent*0.03752;
activity WITH diet*0.05405;
activity WITH bmi7hat*1.35984;
sedent WITH diet*-0.05141;
sedent WITH bmi7hat*-41.36480;
diet WITH bmi7hat*-9.50095;

[ activity*-0.23295 ];
[ sedent*9.79538 ];
[ diet*3.80696 ];
[ bmi7hat*23.64010 ] (c1);

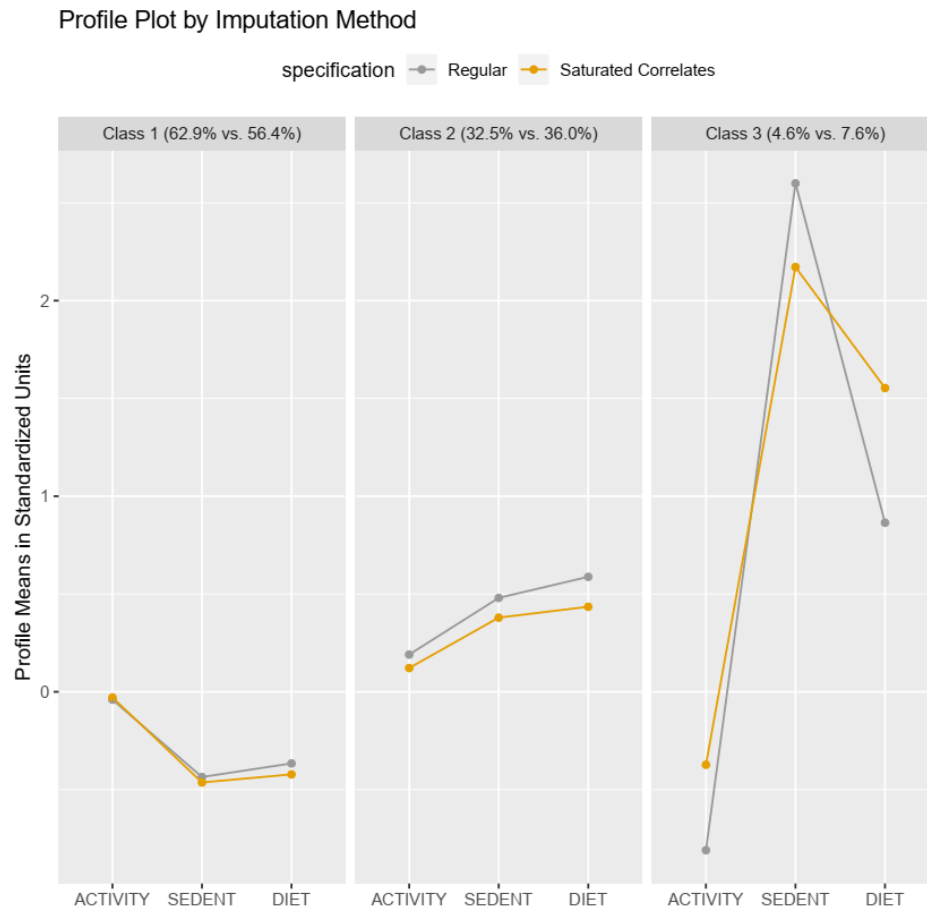
activity*0.32718;
sedent*32.94598;
diet*1.51608;
bmi7hat*188.48395;

```

## Results

The marginal class probabilities and standardized class means across the indicator parcels for both specifications is shown in the figure below. We find that specifying according to Graham's (2003) rules does not result in equivalent parameter estimates in

the class means, suggesting that such a specification risks changing the definitions of the classes. Moreover, the two specifications lead to differences in the marginal class probabilities. For example, the third class was found to comprise only 4.6% of the population according to the regular specification, but the corresponding class comprised 7.6% of the population when the mixture model is specified according to Graham's (2003) rules. Thus, Graham's (2003) rules neither preserves the parameters that define the classes nor results in observations being classified into the same classes.



## References

- Berlin, K. S., Williams, N. A., & Parra, G. R. (2014). An introduction to latent variable mixture modeling (part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology*, 39(2), 174–187. Retrieved from <http://dx.doi.org/10.1093/jpepsy/jst084>
- Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, 10(1), 80–100.

### Technical Appendix C: Derivation of BIC<sup>obs</sup>

To understand why the penalty term  $q_K \log N$  assumes  $N$  complete observations, we first derive the expression for the BIC assuming the data are complete. Then we consider whether the tacit underlying assumptions are appropriate in the presence of missing data.

#### Derivation of the BIC in Complete-Data Settings

The BIC is derived using large sample theory to approximate to the log-integrated likelihood function (also referred to as the model evidence) given by

$$\log \Pr(Y|\mathcal{M}_K) = \int_{\theta_K} \Pr(Y, \theta_K|\mathcal{M}_K) d\theta_K. \quad (\text{C. 1})$$

Making use of Laplace's method (Kass & Raftery, 1995; Tierney & Kadane, 1986), it can be shown that, under certain conditions, (C. 1) can be well approximated as the loglikelihood at the maximum likelihood estimate plus the log determinant of the Fisher information matrix, i.e.,

$$\log \Pr(Y; \mathcal{M}_K) \approx \ell(\hat{\theta}_K|Y) - \frac{1}{2} \log |I(\hat{\theta}_K|Y)| \quad (\text{C. 2})$$

where  $\hat{\theta}_K$  is the MLE and  $I(\hat{\theta}_K|Y)$  is the complete-data Fisher information matrix. The BIC makes use of another approximation: the determinant of  $I(\hat{\theta}_K)$  is approximately given by  $N^{q_K}$ , i.e.,

$$|I(\hat{\theta}_K|Y)| \approx N^{q_K}. \quad (\text{C. 3})$$

The approximation in (C. 3) is justified by the central limit theorem because (a) the standard errors are of order  $N^{-\frac{1}{2}}$  under the regularity conditions, implying that the diagonals of the Fisher information matrix are of order  $N$ , and (b) the determinant is the product of  $q_K$  diagonals of the Fisher information matrix. Substituting the approximation

(C. 3) into (C. 2), simplifying, and then multiplying by -2 results in the familiar expression for the BIC given by Schwarz (1978), i.e.,

$$\text{BIC} = -2\ell(\hat{\theta}_K|Y) + q_K \log N. \quad (\text{C. 4})$$

### **Adjusting the BIC to Account for Evidence Loss due to Missing Data**

In practice, the penalty term in calculating the BIC using only the observed data  $Y^{\text{obs}}$  (i.e.,  $\text{BIC}^{\text{FIML}}$ ) remains the same as in the complete data case. The tacit assumption here is that the determinant of the Fisher information matrix given the observed data is approximately equal the determinant value of the complete-data information matrix, i.e.,

$$|I_{\text{obs}}(\hat{\theta}_K|Y^{\text{obs}})| \approx |I_{\text{com}}(\hat{\theta}_K|Y)| \approx N^{\frac{q_K}{2}}. \quad (\text{C. 5})$$

We assert here that a penalty which approximates the determinant as  $N^{\frac{q_K}{2}}$  over estimates  $|I_{\text{obs}}(\hat{\theta}_K|Y^{\text{obs}})|$  whenever there is nontrivial missing information and especially in small samples. We further assert that model selection based on  $\text{BIC}^{\text{FIML}}$  leads to decisions that favor more parsimonious models than what would have been made had the data been complete.

To justify our assertions, we start with the same foundation of the BIC in the complete data case: Laplace's method. However, in this case we approximate the log-integrated observed-data likelihood using the observed-data likelihood function and the corresponding observed-data Fisher information matrix,

$$\log \Pr(Y^{\text{obs}}; \mathcal{M}_K) \approx \ell(\hat{\theta}_K|Y^{\text{obs}}) + \frac{1}{2} \log |I_{\text{obs}}(\hat{\theta}_K|Y^{\text{obs}})|. \quad (\text{C. 6})$$

Following from the missing information principle in (A.1), we note that the asymptotic covariance matrix if given the observed-data,  $I_{\text{obs}}^{-1}(\hat{\theta}_K|Y^{\text{obs}})$ , is related to the asymptotic complete-data covariance matrix,  $I_{\text{com}}^{-1}(\hat{\theta}_K|Y)$  by the relative increase

invariance matrix,  $\text{RIV}(\hat{\theta}_K)$ . Substituting  $I_{\text{obs}}^{-I}(\hat{\theta}_K|Y^{\text{obs}})$  for  $V_{\text{obs}}(\hat{\theta})$  and  $I_{\text{com}}^{-I}(\hat{\theta}_K|Y)$  for  $V_{\text{com}}(\hat{\theta})$  in (A.4) yields

$$I_{\text{obs}}^{-1}(\hat{\theta}_K|Y^{\text{obs}}) = (\mathbb{I} + \text{RIV}(\hat{\theta}_K)) I_{\text{com}}^{-1}(\hat{\theta}_K|Y). \quad (\text{C. 7})$$

To provide intuition about the relative increase in variance,  $I_{\text{obs}}^{-I}(\hat{\theta}_K|Y^{\text{obs}})$  represents the “total” variance and  $I_{\text{com}}^{-I}(\hat{\theta}_K|Y^{\text{obs}})$  represents the “within-imputation” variance components used in Rubin’s (1987) rules for multiple imputation. Inverting (C. 7) and then taking the determinant implies that the determinant of the observed-data Fisher information matrix is  $N^{q_K}$  scaled by the scalar value  $|\mathbb{I} + \text{RIV}(\hat{\theta}_K)|^{-I}$ ,

$$|I_{\text{obs}}(\hat{\theta}_K|Y^{\text{obs}})| = \frac{|I_{\text{com}}(\hat{\theta}_K|Y)|}{|\mathbb{I}_{q_K} + \text{RIV}(\hat{\theta}_K)|} \approx \frac{N^{q_K}}{|\mathbb{I}_{q_K} + \text{RIV}(\hat{\theta}_K)|}. \quad (\text{C. 8})$$

Substituting the approximation in (C. 8) into (C. 6) and simplifying implies that

$$\log \Pr(Y^{\text{obs}}; \mathcal{M}_K) \approx \ell(\hat{\theta}_K|Y^{\text{obs}}) - \frac{1}{2} (q_K \log N - \log |\mathbb{I}_{q_K} + \text{RIV}(\hat{\theta}_K)|). \quad (\text{C. 9})$$

An expression for the BIC adjusted for evidence loss due to missing data (i.e.,  $\text{BIC}^{\text{obs}}$ ) is then obtained by multiplying the above equation by -2,

$$\text{BIC}^{\text{obs}} = -2\ell(\hat{\theta}_K|Y^{\text{obs}}) + q_K \log N - \log |\mathbb{I}_{q_K} + \text{RIV}(\hat{\theta}_K)| \quad (\text{C. 10})$$

Noting that  $-2\ell(\hat{\theta}_K|Y^{\text{obs}}) + q_K \log N$  is defined as  $\text{BIC}^{\text{FIML}}$ , (C. 10) can be expressed as

$$\text{BIC}^{\text{obs}} = \text{BIC}^{\text{FIML}} - \log |\mathbb{I}_{q_K} + \text{RIV}(\hat{\theta}_K)|. \quad (\text{C. 11})$$

Thus,  $\text{BIC}^{\text{obs}}$  is simply the traditional FIML BIC formulation corrected for the added variability resulting from the missing data, and  $\log |\mathbb{I}_{q_K} + \text{RIV}(\hat{\theta}_K)|$  represents the amount of evidence lost from the model evidence that the BIC approximates.

The relative increase in variance is related to the fraction of missing information,  $\text{FMI}(\hat{\theta}_K)$ , a quantity of theoretical importance because, among other reasons, it governs

the convergence rate of the EM algorithm (Dempster, Laird, & Rubin, 1977). Thus, the evidence loss can also be expressed in terms of the fraction of missing information. In particular, Property 3 in Technical Appendix A states that  $|\mathbb{I}_{q_K} + \text{RIV}(\hat{\theta}_K)| = |\mathbb{I}_{q_K} - \text{FMI}(\hat{\theta}_K)|^{-1}$ .  $\text{BIC}^{\text{obs}}$  can, therefore, be expressed as

$$\text{BIC}^{\text{obs}} = \text{BIC}^{\text{FIML}} + \log |\mathbb{I}_{q_K} - \text{FMI}(\hat{\theta}_K)|. \quad (\text{C.12})$$

## References

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Johnson Wiley & Sons. <https://doi.org/10.1002/9780470316696>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393), 82–86. <https://doi.org/10.2307/2287970>