Journal of Clinical Child & Adolescent Psychology, 45(2), 167–175, 2016
Copyright © Taylor & Francis Group, LLC
ISSN: 1537-4416 print/1537-4424 online
DOI: 10.1080/15374416.2014.958839

Routledge
Taylor & Francis Group

UPDATE ON QUANTITATIVE METHODS

Cautions on the Use of Multiple Imputation When Selecting Between Latent Categorical versus Continuous Models for Psychological Constructs

Sonya K. Sterba
Department of Psychology and Human Development, Vanderbilt University

Clinical psychology researchers studying adolescents and young adults long have been interested in characterizing the latent categorical (classes/profiles) versus continuous (factors) nature of psychological syndromes. To inform this debate, researchers sometimes compare the fit of finite mixture versus factor analysis models to symptom data. This study explains and evaluates how missing data handling methods can impact results of this important model fit comparison. Via simulation, we assess three missing data-handling methods previously recommended to researchers fitting these models: multiple imputation using a saturated multivariate normal imputation model, multiple imputation using a hypothesized model, or full information maximum likelihood using the EM algorithm (FIML-EM). Results show that, under certain conditions, the method used to handle missing data can interfere with clinical psychologists' ability to accurately discriminate latent classes from continua. For instance, certain imputation methods increase the chance of selecting latent continua when latent classes truly exist. FIML-EM performed best overall. Recommendations for practice are discussed.

There have been long-standing debates regarding the underlying categorical (profiles/classes) versus continuous (dimensional) structure of many psychological syndromes and behavioral constructs (see Sterba, 2014, for review). Increasingly, statistical analyses have been used to inform these debates (see Brown & Barlow, 2005; Helzer, van den Brink, & Guth, 2006; Kraemer, Shrout, & Rubio-Stipec, 2007; Krueger, Markon, Patrick, & Iacono, 2005; Trull & Durrett, 2005; Widiger & Samuel, 2006). Specifically, researchers may compare the fit of alternative discrete or continuous latent variable models to data. Finding that a particular model is best fitting is one piece of evidence consistent with that model representing the data-generating process, though it should be considered in the context of power, potential model misspecifications, and evidence of convergent and discriminant validity (e.g., Bauer & Curran, 2004; Lubke, 2012). Such studies, often on adolescents or young adults, commonly pertain to externalizing behavior (e.g., Clark et al., 2013; Krueger et al., 2008; Walton, Ormel, & Krueger, 2011), substance use (Gillespie, Neale, Legrand, Iacono, & McGue, 2011; Muthén, 2006; Witkiewitz et al., 2013), attention-deficit/hyperactivity problems (e.g., Hudziak et al., 1998; Lubke et al., 2007), and borderline personality (e.g., Conway, Hammen, & Brennan, 2012; Hallquist & Pilkonis, 2012). This topic has received additional attention in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-V*; American Psychiatric Association, 2013).

In particular, an increasingly common model comparison is that between a finite mixture model (e.g., latent profile/class model)—implying a categorical latent syndrome—and a factor analysis model—implying a dimensional latent syndrome (for reviews, see Lubke & Neale, 2006, 2008; Markon & Krueger, 2006). In addition, developmental psychopathology studies have extended statistical comparisons of the categorical versus dimensional nature of latent constructs to the longitudinal context (e.g., Hirsh-Pasek & Burchinal, 2006). Such studies

Correspondence should be addressed to Sonya K. Sterba, Quantitative Methods Program, Department of Psychology and Human Development, Vanderbilt University, Peabody #552, 230 Appleton Place, Nashville, TN 37203. E-mail: Sonya.Sterba@vanderbilt.edu

may compare latent class growth models—implying discrete class-specific patterns of change—versus latent factor growth models—implying continuous individual differences in change in, say, early adolescent conduct problems (Kreuter & Muthén, 2008) or substance use (Feldman, Masyn, & Conger, 2009).

Several methodological aspects of such model comparisons have been addressed (see Bauer & Curran, 2004), including how often a true (population-generating) model for the symptoms is selected when it is one of those being compared (e.g., Lubke, 2012; Lubke & Neale, 2006, 2008). One important aspect of such model comparisons that has not been addressed is how they are affected by the handling of missing data—despite the fact that missing data are ubiquitous in empirical research. This article concerns the handling of missing data in studies where the goal is to compare the dimensional versus categorical nature of a latent syndrome or construct. More specifically, this article concerns missingness on outcomes (*y*s); often, when selecting between the dimensional versus categorical nature of a latent syndrome, predictors have not been included in the model (Lubke & Muthén, 2007). Under missing-at-random (or missing-completely-at-random) assumptions,[1] two alternative procedures have been recommended. One alternative is multiple imputation (MI)—most commonly implemented using a saturated multivariate normal (MVN) imputation model (e.g., Barker et al., 2010; Biggs et al., 2010; Ingoldsby et al., 2006; Jonkmann, Trautwein, & Ludke, 2009; Missall, Mercer, Martinez, & Casebeer, 2012; West, Hill, Hewison, Knapp, & House, 2010).[2] Another alternative is full information maximum likelihood estimation using the iterative expectation-maximization algorithm (FIML-EM). When fitting mixture models, FIML-EM often is used for a different purpose—to handle individuals' unknown latent class memberships (McLachlan & Peel, 2000).

Some prior recommendations have included MI as an option for missing data handling with finite mixtures (e.g., Asparouhov & Muthén, 2010; Collins & Lanza, 2010; Lanza, Coffman, & Xu, 2013). On the other hand, others have discouraged it (Enders, 2010) based on the potential for MI to interfere specifically with recovery of covariate effects that differ across latent classes (Enders & Gottschall, 2011). Enders and Gottschall's (2011) investigation concerned fitting a model with the true number of classes and did not investigate how imputation affects the class enumeration. The consequences of MI for model selection involving mixtures have not been clarified for applied researchers. Clinical applications have stated uncertainty about employing MI in this context given the lack of methodological research on this topic (Colder et al., 2001; Costello, Dierker, Jones, & Rose, 2008).

The remainder of this article proceeds as follows. First, we motivate three hypotheses regarding the consequences of alternative missing data handling methods for model selection between latent continua versus categories. Next, to evaluate these hypotheses, a simulation is described in which missingness occurs on the *y*s and the true, population-generating model is a categorical latent variable model (a latent profile analysis [LPA]). An example of such a population-generating LPA would be if distinct etiological processes gave rise to typical and expressive latent temperament classes in toddlers, which phenotypically manifested in different patterns of social fear, anger proneness, and activity level. To test the three hypotheses, the fit of a true LPA model is compared to the fit of a continuous latent variable model (factor analysis [FA]) under alternative missing data handling approaches. These hypotheses are of interest to researchers who have missing outcomes and yet are interested in comparing the fit of categorical versus dimensional models for their construct of interest (e.g., toddler temperament).

HYPOTHESES

H1: When latent classes exist, imputing missing *y*s from a MVN saturated model can decrease the chance of correctly selecting latent classes (LPA) over continua (FA).

In general (nonmixture) modeling contexts, it is thought quite benign to impute even strongly nonnormal data (like psychiatric symptoms) under a MVN imputation model (Demirtas, Freels, & Yuncel, 2008; Graham & Schafer, 1999; Rubin & Shenker, 1986; Schafer, 1997). However, the key to being able to recover classes/profiles, when they truly exist, is the preservation of higher order moments (e.g., skew, kurtosis) in the symptom data. Only LPA, but not FA, makes use of these higher order moments (Molenaar & Von Eye, 1994). Hence, when a generating LPA implies greater nonnormality in the marginal (across-class) distributions of the *y*s (say, due to greater separation between class means), LPA will tend to fit the data better than FA (Lubke & Neale, 2006, 2008). For instance, all else equal, greater class mean separation of latent toddler temperament profiles would correspond with means on social fear, anger proneness, and activity-level items that are more distinct across profiles. As visually depicted in Figure 1

[1] The *missing-completely-at-random assumption* is that missingness depends neither on observed variables in the model nor on unobserved variables predictive of the outcome(s) and associated with model variables. The *missing-at-random assumption* is that missingness may depend on observed variables in the model but not on such unobserved variables.

[2] Some mixture applications report using a single imputation due to lengthy computational times.

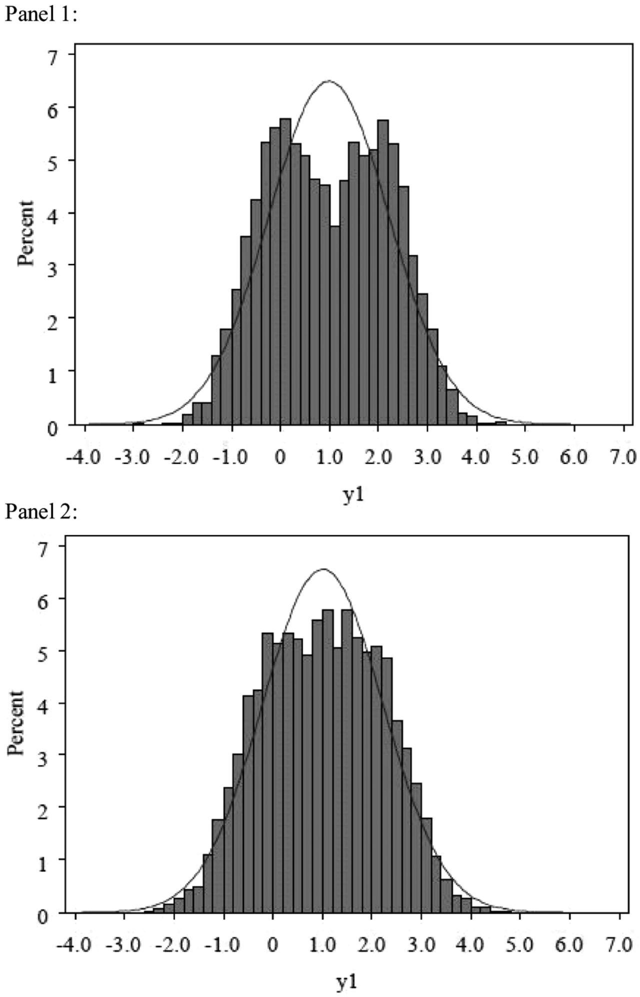

FIGURE 1 Histogram of the marginal (across-class) distribution of an item from a mixture model before multiple imputation (Panel 1) and after multiple imputation (Panel 2). *Note*. Data for Figure 1 were generated from a two-class latent profile model, and this plot pertains to the first item. 50% missingness was induced for the first item under a missing completely at random mechanism. This data set was used only for the purpose of this pedagogical visualization. The solid line is a superimposed normal distribution. In Panel 1, skew = 0, kurtosis = −.87; in Panel 2, skew = 0, kurtosis = −.62.

for a single symptom item generated from a two-class LPA, imputing from an MVN distribution reduces nonnormality in the filled-in complete data—to the extent that, under certain circumstances, an LPA may no longer fit better than an FA, even when latent classes truly exist. We hypothesize that these circumstances are (1a) when the proportion of missing data is larger, *and* (1b) when class mean separation in the generating LPA is moderate. Regarding (1a), when the missingess proportion is larger, misspecifications inherent in the imputation model should have a larger influence on the filled-in complete-case results (Meng, 1994; Schafer, 1997). Regarding (1b), when class mean separation is *moderate*, model selection results should be more sensitive to the use of MVN MI than when separation is large or small. For large class separation, profound nonnormality, favoring LPA, remains despite MVN MI. For small class separation (e.g., similar means across profiles on each toddler temperament item), trivial nonnormality exists anyway, which—even in the case of no missingness—makes correct recovery of a generating LPA nearly impossible (see Tueller & Lubke, 2010).

H2: Imputing missing *y*s from a researcher's hypothesized model (here, LPA or FA) instead of a saturated MVN model also will be problematic. Recovery of the true, generating model in the data analysis will be either helped or hindered depending on whether the imputation model is correct.

Alternatives to saturated MVN MI have been suggested, such as imputing from the researcher's hypothesized model (Asparouhov & Muthén, 2010; Merkle, 2011). According to Asparouhov and Muthén (2010), "ground breaking opportunities arise, such as, imputation from LCA models and factor analysis models," which could mitigate convergence problems that may be encountered with saturated MVN MI, and can be considered "a viable alternative as long as the estimated model for the imputation fits the data well" (p. 3). Relatedly, Merkle (2011) asked, "If you believe that a factor analysis model best describes the data, why use a saturated multivariate normal model to impute the data?" (p. 461). However, the correctness of a researcher's hypothesized model would not be known in advance. When the missingness proportion is larger, imputing from an incorrect model (here, FA) should *decrease* the probability of correctly selecting LPA, whereas imputing from a correct model (here, LPA) should *increase* this probability.

H3: FIML-EM can outperform MI for distinguishing latent classes from continua in the context of missing *y*s.

Prior research in other modeling contexts has shown that when an MI model is misspecified in a manner that makes it more restrictive than a model fitted with FIML-EM ("uncongeniality"), parameter estimates have greater bias under MI (Collins, Schafer, & Kam, 2001; Meng, 1994; Schafer, 2003). A MVN saturated MI model, or for that matter a factor analysis MI model, implies oversimplified marginal (across-class) distributions for *y*s. In contrast, LPA fitted with FIML-EM allows a more flexible marginal distribution for *y*s. For this reason, FIML-EM should, on average, outperform the MI approaches—in particular, the saturated MVN MI and factor analysis MI approaches—in recovering the generating model through model selection.

METHODS

To investigate Hypotheses 1 to 3, we generated 500 samples of data from a two-class LPA model in a fully crossed design defined by two y-missingness proportions and two class separations. The two y-missingness proportions used, .15 and .35, have been considered realistic for practice (Enders & Bandalos, 2001; Merkle, 2011; Wothke, 2000). These y-missingness proportions refer to the (.15 or .35) probability that person i is missing the jth item. Within a condition of the simulation design, the missingness proportion is the same for all persons and for all items. The two class separations used (moderate vs. large) have been operationalized in prior research as a 2.0 versus 3.0 Mahalanobis distance (MD) between profiles (Lubke & Neale, 2006, 2008). Mahalanobis distance is a conventional measure of class separation used in the mixture literature that is similar to a standardized mean difference. For two classes: $MD = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ where $\boldsymbol{\Sigma}$ is the within-class residual covariance matrix and $\boldsymbol{\mu}_k$ is a vector of means specific to class k. Hypotheses 1 to 3 could be demonstrated with either missing at random or missing completely at random y-missingness mechanisms. We chose the simplest missingness mechanism, missing completely at random, to highlight the serious consequences that can arise simply due to misspecifying the MI model, even when missing values are "benign" (see Enders & Gottschall, 2011, for a similar rationale). Results would generalize to circumstances where outcomes are missing at random such that, for example, the probability of missingness on y_2 (e.g., anger proneness item) depends on the observed score of y_1 (e.g., social fear item).

The generating LPA for persons $i = 1 \ldots N$, items $j = 1 \ldots J$, and classes $k = 1 \ldots K$ of a latent classification variable c is

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} p(c_i = k) \prod_{j=1}^{J} f(y_{ij}|c_i = k)$$

where $\mathbf{y}_i$ is a $J \times 1$ response vector for person i. $p(c_i = k)$ is the probability of membership in class k. $f(y_{ij}|c_i = k)$ is a univariate normal probability density function of y_{ij} in the kth class. Within class k, y_{ij} is normally distributed, $y_{ij}|c_i = k \sim N(\mu_j^{(k)}, \sigma_j^{2(k)})$, with class-specific mean $\mu_j^{(k)}$ and variance $\sigma_j^{2(k)}$. Parameters for the generating LPA with moderate class-separation were taken from two classes of a temperament LPA application in young children (van den Akker, Dekovic, Prinzie, & Asscher, 2010) with three symptoms (social fear, anger proneness, activity level). Temperament is a developmental precursor of personality. In class 1, labeled 'typical', parameters were as follows. Item means: $\mu_1^{(1)} = -.35$; $\mu_2^{(1)} = -.48; \mu_3^{(1)} = -.41$; item variances: $\sigma_j^{2(1)} = 1$; class proportion: $p(c_i = 1) = .8$. The other class, labeled "expressive," had parameters: $\mu_1^{(2)} = -.08; \mu_2^{(2)} = .9$; $\mu_3^{(2)} = 1; p(c_i = 2) = .2; \sigma_j^{2(2)} = 1$. That is, the expressive class was distinguished from the typical class predominantly by its higher activity level and higher anger proneness. Item means in the larger class separation condition were multiplied by 2.5 to achieve an MD = 3. This corresponds with toddler temperament profiles which have increased between-profile differences relative to their within-profile variation. All generated samples had $N = 500$, which was found to be a typical sample size used in social science mixture applications by Sterba, Baldasaro, and Bauer (2012).

Missingness was handled with four alternate approaches: FIML-EM, an MI saturated imputation model, an MI two-class LPA imputation model with equal residual variances, or an MI one-factor confirmatory FA imputation model with equal residual variances. For a given MI model, 100 imputations were drawn per sample. MI used the MCMC algorithm with a Gibbs sampler.[3] Each generated data set was fit with four alternative models: three LPAs (one, two, or three classes, each with equal residual variances) and a one-factor confirmatory FA (also with equal residual variances). Data were generated, imputed (where relevant), and fitted with FIML-EM[4] in M*plus* 6.12 (Muthén & Muthén, 1998–2012).

RESULTS

The outcome of interest was which model was found best fitting. Fitted models were compared within sample, paralleling analysis procedures that would be used in an empirical application on toddler temperament, where only one sample is available. The frequency with which each model was selected as best fitting across samples within cell of the simulation design was recorded. The Bayesian Information Criterion (BIC) was used for model selection. Of the information criteria, BIC has typically performed best in previous mixture simulations

[3] Two independent chains were used after 50 identical iterations from one chain with a maximum of 50,000 iterations. Selected sensitivity analyses continuing with one chain yielded the same pattern of results. Convergence for MI was monitored using the Gelman-Rubin approach (with Potential Scale Reduction Factor $\leq$ 1.025 for any single parameter). Selected samples were inspected for adequate mixing and lack of class label-switching. Procedures recommended by Cho, Cohen, and Kim (2011) and Asparouhov and Muthén (2010) were employed until no label switching was observed across chain (during MI), across imputation within sample, or across repeated fitted LPAs within a cell of the simulation design, in empirical and graphical checks.

[4] In other words, in all cells FIML-EM was used for model fitting. In only one cell, FIML-EM was also used for handling missing ys (all other cells already had imputed-y data by the analysis stage).

TABLE 1
Best-Fitting Model From the Latent Profile Analysis versus Factor Analysis Comparison When *y*-Missingness Is Handled With FIML-EM versus MI Using Each of Three Alternative Imputation Models

		Fitted Model Selected			
Class Separation	*% Missing*	*1-Factor CFA*	*1-Class LPA*	*2-Class LPA (Generating)*	*3-Class LPA*
Missing Data Handling: EM-Algorithm (FIML)					
Smaller	15	17.80%	5.20%	**77.20%**	0.00%
Larger	15	0.00%	0.00%	**100.00%**	0.00%
Smaller	35	23.40%	23.60%	**53.00%**	0.00%
Larger	35	0.20%	0.00%	**99.80%**	0.00%
Missing Data Handling: Multiple Imputation From Saturated Model					
Smaller	15	34.00%	1.60%	**64.40%**	0.00%
Larger	15	0.00%	0.00%	**100.00%**	0.00%
Smaller	35	**76.40%**	3.80%	19.80%	0.00%
Larger	35	24.00%	0.00%	**75.00%**	1.00%
Missing Data Handling: Multiple Imputation From 2-Class LPA					
Smaller	15	15.90%	1.41%	**82.70%**	0.00%
Larger	15	0.00%	0.00%	**100.00%**	0.00%
Smaller	35	20.75%	2.73%	**76.52%**	0.00%
Larger	35	0.00%	0.00%	**100.00%**	0.00%
Missing Data Handling: Multiple Imputation From 1-Factor CFA					
Smaller	15	33.60%	2.00%	**64.40%**	0.00%
Larger	15	0.00%	0.00%	**100.00%**	0.00%
Smaller	35	**73.49%**	6.22%	20.28%	0.00%
Larger	35	22.40%	0.00%	**76.80%**	0.80%

Note. The boldface model was selecting as better fitting than competitors in the highest percentage of samples within cell, according to the Bayesian Information Criterion. Hence the boldface model is labeled the best-fitting model. FIML-EM = full information maximum likelihood using the Expectation-Maximization algorithm; MI = multiple imputation; CFA = confirmatory factor analysis model; LPA = latent profile analysis model.

(e.g., Nylund, Asparouhov, & Muthén, 2007) and has been used for discriminating classes from continua (e.g., Lubke & Neale, 2006, 2008). Further, a recent review found it to be the most common, and often the *only*, selection index used in certain mixture applications (Sterba et al., 2012). For MI, the average of the BIC point estimate across imputed data sets was used (also the M*plus* default). Five times more imputations than the usual recommendation (i.e., 100) were employed to maximize precision of this result (Graham, Olchowski, & Gilreath, 2007). Applications using mixtures with MI most often have relied solely on the across-imputation-average value of BIC for model selection (e.g., Barker et al., 2010; Biggs et al., 2010; Ingoldsby et al., 2006; Missall et al., 2012; Vaughn, Shook, & McMillin, 2008; West et al., 2010).[5]

Convergence ranged from 95% to 100% across cells. Table 1 shows the percentage of samples per cell best fit by each fitted model. The most frequently selected model per cell is in boldface.

The boldface results for FIML-EM versus saturated MI in Table 1 are consistent with hypotheses. Specifically, in line with Hypotheses 1 and 3, under the conditions of moderate class separation *and* larger missingness proportions, imputing from an MVN saturated model can make one less likely to find evidence of unobserved heterogeneity compared with using FIML-EM to handle *y*-missingness. That is, saturated MVN MI can make one less likely to find evidence of latent classes over latent continua, for a psychological construct such as temperament, when latent classes (e.g., expressive and typical classes) truly exist. This means that an applied researcher would be expected to incorrectly select a dimensional representation of a truly categorical temperament construct greater than 76% of the time when imputing from the saturated MVN model, under these conditions.

In addition, the boldface results in Table 1 are consistent with Hypotheses 2 and 3. Specifically, imputing from a model other than the saturated MVN helps (or not) depending on whether this alternative imputation model

[5] The conventional likelihood ratio test (LRT) developed for MI is not suited for comparing models with different numbers of classes due to the violation of regularity conditions. Conversely, adjusted versions of the LRT (Lo-Mendell-Rubin-LRT) suited for comparing models with different numbers of classes have not been adapted for MI. The bootstrap LRT (and the lesser used selection index in Markon & Krueger, 2006) also have not been adapted for MI and would be computationally prohibitive in the present simulation.

is correctly (or incorrectly) specified—something that an empirical researcher interested in statistically investigating the latent structure of temperament or psychopathology presumably would not know in advance. That is, imputing from the correct two-class LPA improved model selection accuracy over using FIML-EM to handle *y*-missingness. But imputing from the incorrect one-factor FA was worse than FIML-EM and about the same as imputing from the saturated MVN model.

DISCUSSION

Results of this study suggest exercising caution regarding the use of saturated MVN MI when the goal is to discriminate the latent continuous (factor) versus categorical (classes) nature of a psychological syndrome. This goal has become increasingly relevant to debates surrounding changes from *DSM-IV* to *DSM-V* (Regier, Kuhn, & Kupfer, 2013). Such caution is needed, particularly in the situation when there is limited available information to recover classes—for instance, moderate class separation and considerable missingness. In other situations examined, model selection results were on average robust to the use of saturated MVN MI. Whereas imputing from a hypothesized model can help select the correct model when the imputed model is correct (consistent with Merkle, 2011), from this illustration FIML-EM seems a less risky strategy overall.

Our study thus identifies a separate and additional limitation of applying MVN MI in the context of mixtures, beyond that identified by Enders and Gottschall (2011). When exclusively fitting the generating mixture, Enders and Gottschall noted that the application of the popular saturated MVN MI interfered with recovery of moderated (class-varying) effects of observed covariates—because the imputation model oversimplified relationships between outcomes and covariates. The present article focused instead on demonstrating that saturated MVN MIs oversimplified distributional assumptions about *y*s lead to reduced nonnormality, which in turn can interfere with the ability to correctly discriminate latent classes and continua. In different ways, both articles underscore the need to ensure the MI model is not more restrictive than candidate data-generating processes (and missingness mechanisms). Although our simulation used models for cross-sectional data, results should generalize to the parallel context of discriminating continuous versus categorical variability in change, in developmental psychopathology studies.

Limitations

Several limitations should be noted. First, the simulation demonstrations held total sample size constant at 500; for mixture models, even smaller sample sizes together with substantial proportions of missing data could lead to high rates of estimation problems and empirical underidentification of small classes (McLachlan & Peel, 2000), which were not a focus here. Second, we considered saturated MVN MI in particular because of its widespread use. In theory, nonparametric (e.g., approximate Bayesian bootstrap) MI approaches are an alternative; however, they suffer from several key practical limitations (see Molenberghs & Kenward, 2007). Third, if missingness proportions had been trivial, the choice between missing data methods considered here likely would be of little consequence.

Future Directions

Although our demonstrations showed that imputing from a factor model could decrease the probability of finding classes when classes exist, we anticipate that the opposite problem could arise. For example, if a factor model actually generated responses on social fear, anger proneness, and activity level items, imputing from an LPA might decrease the probability of correctly selecting a dimensional representation of the temperament construct. In addition, although our demonstrations involved a representative but relatively simple mixture model, related issues would be expected to apply when hybrid models (combining both latent classes and continua; e.g., factor mixture models) are compared to models that specify either latent classes or continua. Expanding the focus of investigation to involve alternative generating models and fitted models could be useful for future research. Relatedly, because this study was intended to isolate the impact of missing data handling methods on class versus continua model selection results, additional conditions with known potential to affect such results were not simultaneously introduced. For instance, the within-class model could have been misspecified or *y*s could have been ordinal but treated as continuous, as is common in psychopathology applications. Future research could consider the effects of a larger combination of conditions on model selection results. Finally, note that this study did not consider taxometric methods because they tend to perform less well than mixture models under general conditions (Lubke & Tueller, 2010) and cannot accommodate missing data (e.g., Ruscio, 2006).

Conclusions and Recommendations

Understanding the latent structure of psychological syndromes and constructs is a topic of great interest to clinical psychology researchers. In recent years, statistical comparisons of the fit of alternative latent variable models implying categorical versus dimensional syndromes have become more common, particularly in adolescent samples (e.g., Conway et al., 2012; Gillespie et al., 2011). In this light, the present study compared two missing data methods that could be used in such studies (saturated MVN MI and FIML-EM) and considered some newer MI approaches. Although MI has in the past been described as an option for handling missing data in mixture/latent class analyses (e.g., Asparouhov & Muthén, 2010; Collins & Lanza, 2010), its consequences had not been previously considered. We showed that the manner in which *y*-missingness is handled can under certain circumstances interfere with the ability to statistically discriminate between the latent categorical (classes) versus continuous (factors) nature of a psychological construct. Our results suggest the use of FIML-EM during model selection in future studies with this objective, because it performed well under the broadest range of circumstances.

In general, if FIML estimation and the chosen MI model impose the exact same assumptions using the same data and analysis, they should asymptotically lead to the same inferences. However, there are contexts in which conventional implementations of FIML and MI models imply meaningfully different assumptions. In this study's context, assumptions of conventional FIML-EM were more appropriate than those of saturated MVN MI. Other times, in nonmixture modeling contexts, conventional MI implementations may be more realistic; for instance, they routinely incorporate auxiliary variables predictive of missingness, unlike typical FIML implementations (but see Graham, 2003). It is important to move beyond recommendations that simply restate benefits of both FIML and MI over, say, listwise deletion (e.g., Croy & Novins, 2005; Jelicic, Phelps, & Lerner, 2009), to focus on choosing *among alternative* implementations of FIML and/or MI. In this regard, a researcher can even conduct a sensitivity analysis to see if his or her own model selection results are sensitive to alternative missing data methods and assumptions. Finally, researchers should report on the form and assumptions of the MI (or FIML) model so others can, for instance, "judge if the imputation model can be misleading for a particular intended analysis" (Meng, 1994, p. 554).

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: American Psychiatric Press.

Asparouhov, T., & Muthén, B. (2010). *Multiple imputation with Mplus*. Retrieved from http://statmodel.com

Barker, E. D., Vitaro, F., Lacourse, E., Fontaine, N. M., Carbonneau, R., & Tremblay, R. E. (2010). Testing the developmental distinctiveness of male proactive and reactive aggression with a nested longitudinal experimental intervention. *Aggressive Behavior*, *36*, 127–140.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, *9*, 3–29.

Biggs, B. K., Vernberg, E., Little, T. D., Dill, E. J., Fonagy, P., & Twemlow, S. W. (2010). Peer victimization trajectories and their association with children's affect in late elementary school. *International Journal of Behavioral Development*, *34*, 136–146.

Brown, T., & Barlow, D. (2005). Dimensional versus categorical classification of mental disorders in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* and beyond: Comment on the special section. *Journal of Abnormal Psychology*, *114*, 551–556.

Cho, S. J., Cohen, A. S., & Kim, S.-H. (2011). Markov chain monte carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *78*, 1–29.

Clark, S., Muthén, B., Kaprio, J., D'Onofrio, M., Viken, R., & Rose, R. (2013). Models and strategies for factor mixture analysis. An example concerning the structure underlying psychological disorders. *Structural Equation Modeling*, *20*, 681–703.

Colder, C., Mehta, P., Balanda, K., Campbell, R. T., Mayhew, K. P., Stanton, W. R., … Flay, B. R. (2001). Identifying trajectories of adolescent smoking: An application of latent growth mixture modeling. *Health Psychology*, *20*, 127–135.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis with applications in the social, behavioral, and health sciences*. Hoboken, NJ: 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*, *4*, 330–351.

Conway, C., Hammen, C., & Brennan, P. (2012). A comparison of latent class, latent trait, and factor mixture models of *DSM–IV* borderline personality disorder criteria in a community setting: Implications for *DSM–V*. *Journal of Personality Disorders*, *26*, 793–803.

Costello, D., Dierker, L., Jones, B., & Rose, J. (2008). Trajectories of smoking from adolescence to early adulthood and their psychosocial risk factors. *Health Psychology*, *27*, 811–818.

Croy, C., & Novins, D. (2005). Methods for addressing missing data in psychiatric and developmental research. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 1230–1240.

Demirtas, H., Freels, S., & Yuncel (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, *78*, 69–84.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

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

Enders, C. K., & Gottschall, A. (2011). Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling*, *18*, 35–54.

Feldman, B. J., Masyn, K. E., & Conger, R. D. (2009). New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Developmental Psychology*, *45*, 652–676.

Gillespie, N. A., Neale, M. C., Legrand, L. N., Iacono, W. G., & McGue, M. (2011). Are the symptoms of cannabis use disorder best accounted for by dimensional, categorical, or factor mixture models? A comparison of male and female young adults. *Psychology of Addictive Behaviors*, *26*, 68–77.

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

Graham, J. W., Olchowski, A., & Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size.

In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Hallquist, M., & Pilkonis, P. (2012). Refining the phenotype of borderline personality disorder: Diagnostic criteria and beyond. *Personality Disorders*, *3*, 228–246.

Helzer, J., van den Brink, W., & Guth, S. (2006). Should there be both categorical and dimensional criteria for the substance use disorders in *DSM–V*? *Addiction*, *101*, 17–22.

Hirsh-Pasek, K., & Burchinal, M. (2006). Mother and caregiver sensitivity over time: Predicting language and academic outcomes with variable- and person-centered approaches. *Merrill Palmer Quarterly*, *52*, 449–485.

Hudziak, J. J., Heath, A. C., Madden, P. F., Reich, W., Bucholz, K. K., Slutske, W., . . . Todd, R. D. (1998). Latent class and factor analysis of *DSM–IV* ADHD: A twin study of female adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, *37*, 848–857.

Ingoldsby, E. M., Shaw, D. S., Winslow, E., Schonberg, M., Gilliom, M., & Criss, M. M. (2006). Neighborhood disadvantage, parent–child conflict, neighborhood peer relationships, and early antisocial behavior problem trajectories. *Journal of Abnormal Child Psychology*, *34*, 303–319.

Jelicic, H., Phelps, E., & Lerner, R. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*, 1195–1199.

Jonkmann, K., Trautwein, U., & Ludke, O. (2009). Social dominance in adolescence: The moderating role of the classroom context and behavioral heterogeneity. *Child Development*, *80*, 338–355.

Kraemer, H. C., Shrout, P. E., & Rubio-Stipec, M. (2007). Developing the diagnostic and statistical manual V: What will "statistical" mean in *DSM–V*? *Social Psychiatry and Psychiatric Epidemiology*, *42*, 259–267.

Kreuter, F., & Muthén, B. O. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology*, *24*, 1–31.

Krueger, R., Markon, K., Patrick, C., & Iacono, W. (2005). Externalizing psychopathology in adulthood: A dimensional spectrum conceptualization and its implications for *DSM–V*. *Journal of Abnormal Psychology*, *114*, 537–550.

Lanza, S., Coffman, D., & Xu, S. (2013). Causal inference in latent class analysis. *Structural Equation Modeling*, *20*, 361–383.

Lubke, G. (2012). Old issues in a new jacket: Power and validation in the context of mixture modeling. *Measurement*, *10*, 212–216.

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

Lubke, G. H., Muthén, B., Moilanen, I. K., McGough, J. J., Loo, S. K., Swanson, J. M., . . . Smalley, S. L. (2007). Subtypes versus severity differences in attention-deficit/hyperactivity disorder in the Northern Finnish Birth Cohort. *American Academy of Child and Adolescent Psychiatry*, *46*, 1584–1593.

Lubke, G., & Neale, M. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, *41*, 499–532.

Lubke, G., & Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: Class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, *43*, 592–620.

Lubke, G., & Tueller, M. (2010). Latent class detection and class assignment: A comparison of the MAXEIG taxonomic procedure and factor mixture modeling approaches. *Structural Equation Modeling*, *17*, 605–628.

Markon, K., & Krueger, R. (2006). Information-theoretic latent distribution modeling: Distinguishing discrete and continuous latent variable models. *Psychological Methods*, *11*, 228–243.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Meng, X. L. (1994). Multiple imputation inference with uncongenial sources of input. *Statistical Science*, *9*, 538–573.

Merkle, E. (2011). A comparison of imputation methods for Bayesian factor analysis models. *Journal of Educational and Behavioral Statistics*, *36*, 257–276.

Missall, K., Mercer, S., Martinez, R., & Casebeer, D. (2012). Concurrent and longitudinal patterns and trends in performance on early numeracy curriculum-based measures in kindergarten through third grade. *Assessment for Effective Intervention*, *37*, 95–106.

Molenaar, P., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 226–242). Thousand Oaks, CA: Sage.

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. West Sussex, UK: Wiley.

Muthén, B. O. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction*, *101*, 6–16.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus users guide*. 6th edition. Los Angeles, CA: Muthén & Muthén.

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*, 535–569.

Regier, D., Kuhl, E., & Kupfer, D. (2013). The *DSM–V*: Classification and criteria changes. *World Psychiatry*, *12*, 92–98.

Rubin, D. B., & Shenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable response. *Journal of the American Statistical Association*, *81*, 366–374.

Ruscio, J. (2006). *Taxometric programs for the R computing environment: Users manual*. Available from http://www.tcnj.edu/~ruscio/TaxProgManual%202012-03-10.pdf

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, UK: Chapman & Hall.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*, 19–35.

Sterba, S. K. (2014). Modeling strategies in developmental psychopathology research: Prediction of individual change. In M. Lewis & K. D. Rudolph (Eds.), *Handbook of developmental psychopathology* (3rd ed., pp. 109–124). New York, NY: Springer.

Sterba, S. K., Baldasaro, R. E., & Bauer, D. J. (2012). Factors affecting the adequacy and preferability of semiparametric groups-based approximations of continuous growth trajectories. *Multivariate Behavioral Research*, *40*, 590–634.

Trull, T., & Durrett, C. (2005). Categorical and dimensional models of personality disorder. *Annual Review of Clinical Psychology*, *1*, 355–380.

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

van den Akker, A., Dekovic, M, Prinzie, P., & Asscher, J. (2010). Toddlers' temperament profiles: Stability and relations to negative and positive parenting. *Journal of Abnormal Child Psychology*, *38*, 485–495.

Vaughn, M. G., Shook, J. J., & McMillin, J. (2008). Aging out of foster care and legal involvement: Toward a typology of risk. *Social Service Review*, *82*, 419–446.

Walton, K., Ormel, J., & Krueger, R. (2011). The dimensional nature of externalizing behaviors in adolescence: Evidence from a direct comparison of categorical, dimensional, and hybrid models. *Journal of Abnormal Child Psychology*, *39*, 553–561.

West, R., Hill, K., Hewison, J., Knapp, P., & House, A. (2010). Psychological disorders after stroke are an important influence on functional outcomes. *Stroke*, *41*, 1723–1727.

Widiger, T., & Samuel, D. (2005). Diagnostic categories or dimensions? A question for the *Diagnostic and Statistical Manual of Mental Disorders—Fifth edition*. *Journal of Abnormal Psychology*, *114*, 494–504.

Witkiewitz, K., King, K., McMahon, R., Wu, J., Luk, J., Bierman, K. L., . . . Conduct Problems Prevention Research Group. (2013). Evidence for a multidimensional latent structural model of externalizing disorders. *Journal of Abnormal Child Psychology*, *41*, 223–237.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. Little, K. Schnabel, & J. Baumert (Eds.), *Testing structural equation models* (pp. 256–293). Mahwah, NJ: Erlbaum.