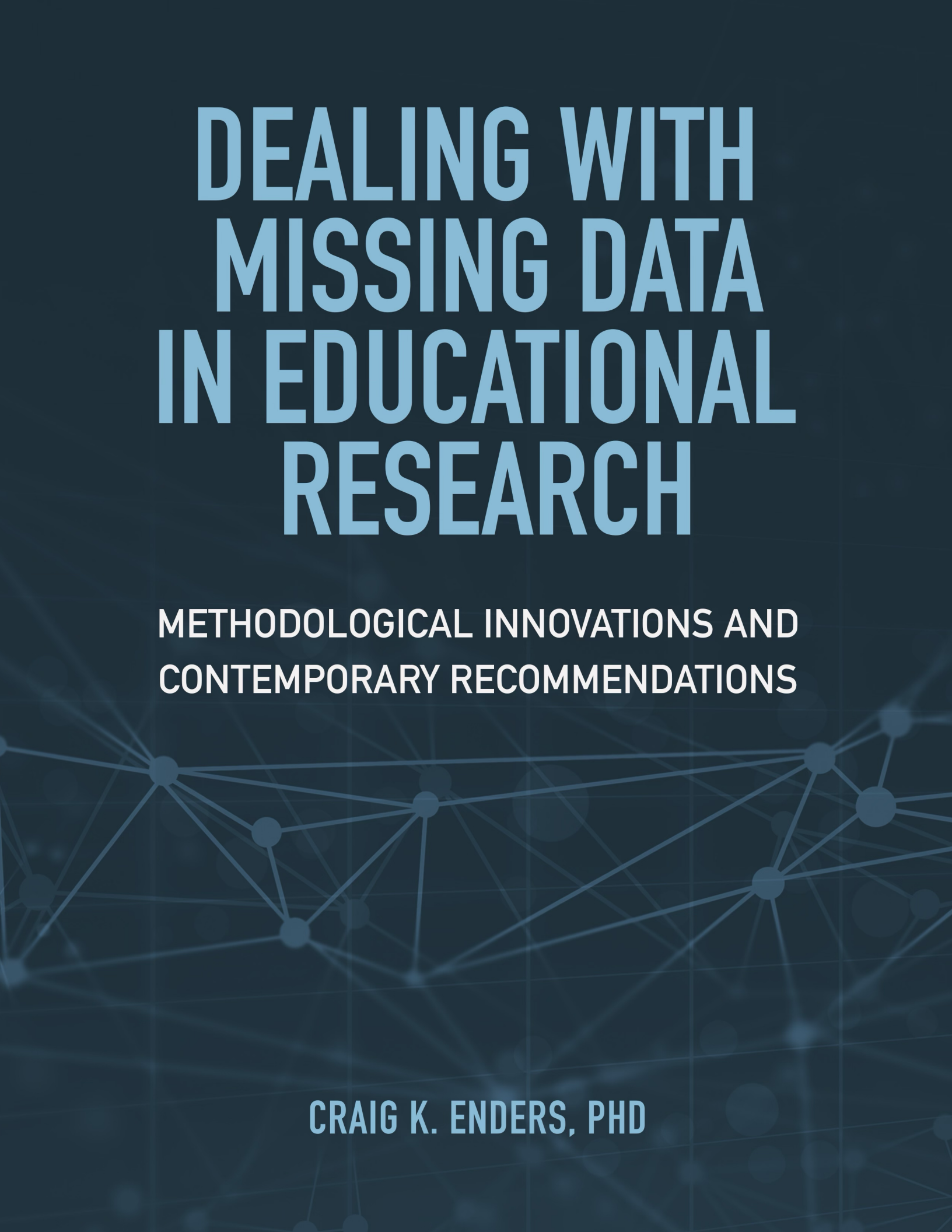
DEALING WITH MISSING DATA IN EDUCATIONAL RESEARCH

METHODOLOGICAL INNOVATIONS AND CONTEMPORARY RECOMMENDATIONS



CRAIG K. ENDERS, PHD

DEALING WITH MISSING DATA IN EDUCATIONAL RESEARCH

Methodological Innovations and Contemporary Recommendations

Craig K. Enders

This review paper was developed as part of a missing data toolkit supported by Institute of Educational Sciences award R305D22000

Preface

This review paper was developed as part of a missing data toolkit supported by Institute of Educational Sciences award R305D22000. The toolkit includes two additional components: a Software Tutorials document with annotated analysis scripts and outputs from several software programs, and a series of YouTube videos demonstrating missing data analyses with the Blimp application (www.appliedmissingdata.com/videos). Blimp was developed with support from Institute awards R305D150056 and R305D190002. The software is available as a free download for MacOS, Windows, and Linux (www.appliedmissingdata.com/blimp).

This review paper catalogs and describes a collection of missing data procedures that represent the current state of the methodological art. The goal is to provide educational researchers with a springboard for accessing the most up-to-date missing data handling methodologies. The paper primarily focusses on three approaches that have gained broad support in the missing data literature: maximum likelihood estimation, Bayesian estimation, and multiple imputation. I henceforth refer to this basket of methods as the "Big Three". Classic incarnations of the Big Three have been available in software programs for more than 20 years. Not surprisingly, missing data methodologies have evolved and improved, and contemporary variants of these methods readily pair with most analytic procedures that enjoy widespread use in educational research applications.

Procedurally, the Big Three *appear to be* very different. Maximum likelihood integrates missing data handling into its estimation machinery, deducing optimal parameter estimates directly from an incomplete data set without ever filling in the missing values. Bayesian estimation similarly identifies optimal parameter estimates from the data, but it does so with an iterative imputation scheme. Multiple imputation requires a preliminary analytic step that creates a collection of filled-in data sets for the subsequent analyses. Procedural differences aside, the Big Three usually produce equivalent numeric estimates, given the same data and assumptions. All things being equal, choosing among competing methods is often a matter of personal preference. Ultimately, the composition of the focal analysis model—in particular, whether it includes nonlinear effects such as interactions, curvilinear terms, or random coefficients—usually determines the type of missing data strategy that works best. This important take-home message appears repeatedly throughout the document.

The structure of the review paper is as follows. Section 1 provides a brief summary of missing data theory, as described by Rubin and colleagues (Little & Rubin, 2020; Rubin, 1976). This section introduces the concept of missing data auxiliary variables, and it describes simple steps to prepare for a missing data analysis. The second section describes an emerging modeling framework called factored regression specification. Unlike classic versions of the Big Three, which often adopt a

multivariate normal distribution for incomplete variables, factored specifications invoke distributional assumptions on a variable-by-variable basis. This flexibility addresses longstanding limitations of classic missing data methods. The next three sections describe the Big Three missing data methods: maximum likelihood, Bayesian estimation, and multiple imputation. These sections describe classic approaches that are widely available in software packages, and they highlight how the frameworks have evolved in the past two decades. Section 6 describes missing data handling for multilevel models that are ubiquitous in educational research. Examples include data where students are nested in schools and data with repeated measurements nested in students. The seventh section describes analyses for a particularly challenging missingness process where the propensity for missing data relates to the unseen score value itself (called missing not at random). The final section provides a summary of some current software options. Most sections include analysis examples, and the Software Tutorials document includes annotated analysis scripts and outputs for 20 common analyses (see www.appliedmissingdata.com/videos).

Contents

1

Introduction to Missing Data

As outlined in the Preface, this review paper focusses on three approaches that have gained broad support in the missing data literature: maximum likelihood, Bayesian estimation, and multiple imputation. I henceforth refer to this basket of methods as the "Big Three". Relative to older approaches, the Big Three offer substantial advantages, including greater power and unbiased estimates under a broader range of applications. However, it is important to understand when they work and what assumptions are required to reap these benefits. To that end, this section summarizes Rubin and colleagues' theoretical framework for missing data problems (Little & Rubin, 2020; Mealli & Rubin, 2016; Rubin, 1976). This classification system defines three mechanisms or processes by which the probability of nonresponse relates to the data: missing completely at random, missing at random, and missing not at random. In practical terms, these foundational concepts function as assumptions for a missing data analysis. Although these assumptions are mostly untestable, there are steps we can take to make their key propositions more plausible.

1.1 Older Missing Data Handling Methods

Of course, the missing data literature outlines numerous other approaches beyond the Big Three. Some have enjoyed widespread use, and others are little more than a historical footnote. Broadly speaking, these methods deal with missing data either by removing incomplete data records or replacing the missing scores with imputations. Deletion options include removing cases with any missing values (listwise deletion or complete-case analysis) or removing cases on an analysis-by-analysis basis (pairwise deletion). Outdated imputation approaches include arithmetic mean imputation, regression imputation, person-mean imputation for questionnaire items, and last observation forward imputation for repeated measures data, among others.

Deletion methods have two important limitations: they reduce power and require a completely unsystematic nonresponse process where missingness is unrelated to the data. For these reasons, the American Psychological Association's Taskforce on Statistical Inference characterized deletion as "among the worst methods available for practical applications" (Wilkinson and Task Force on Statistical Inference, 1999, p. 598). In truth, there are a few situations where deletion produces optimal estimates under the same assumptions as the Big Three (Vach, 1994; von Hippel, 2007; White & Carlin, 2010). However, these occur too infrequently to consider this strategy as a viable option for most missing data problems.

Single imputation methods like mean imputation and regression imputation fill in the data with a *one* set of replacement scores. These approaches also have substantial limitations. Many single imputation methods introduce bias, even when missingness is purely unsystematic. Virtually all distort standard errors and significance tests, even if they sometimes produce unbiased estimates. A problematic version of regression imputation labeled "EM imputation" in the popular SPSS programs deserves a brief discussion. The procedure is potentially misleading because it first uses maximum likelihood estimation[1] to estimate the means, variances, and covariances. Next, the procedure converts these summary statistics to regression equations, then it uses these models to compute predicted values that replace the missing scores. Regression imputation introduces bias because it attenuates variation, and using maximum likelihood to estimate the initial summary statistics does not mitigate this problem. von Hippel (2004) provides a thorough discussion of "EM imputation" in SPSS.

1.2 Missing Data Processes (Mechanisms)

Rubin and colleagues (Little & Rubin, 2020; Mealli & Rubin, 2016; Rubin, 1976) outlined a classification system for missing data problems that describes three ways nonresponse can relate to the data. These so-called missing data mechanisms or processes are missing completely at random, missing at random, and missing not at random. These terms can be confusing because the word "random" implies a probabilistic rather than haphazard process. For example, a missing at random process (the default assumption for most Big Three applications) describes a *systematic* relation between the data and nonresponse.

Rubin's theoretical framework conceptualizes a data set as consisting of observed and unseen (missing) score values. To illustrate, Table 1 shows a data excerpt for 10 observations and three

[1] The EM abbreviation references the expectation-maximization or EM optimization algorithm used to get the maximum likelihood estimates.

variables. The values in the leftmost set of columns comprise the observed data, and the middle set of columns contain the would-be values of the missing data. I refer to these partitions as $\mathbf{Y}_{(obs)}$ and $\mathbf{Y}_{(mis)}$, respectively. The rightmost set of columns in Table 1 display indicators that recode the observed data as 0 = observed and 1 = missing. I collectively refer to the missing data indicators as **M** below.

TABLE 1. Observed and Missing Data Partition

Observed			Missing			Indicators		
Y_1	Y_2	Y_3	Y_1	Y_2	Y_3	M_1	M_2	M_3
--	91	--	63	--	59	1	0	1
76	82	82	--	--	--	0	0	0
97	109	81	--	--	--	0	0	0
66	--	81	--	77	--	0	1	0
63	69	--	--	--	91	0	0	1
83	89	78	--	--	--	0	0	0
73	--	77	--	67	--	0	1	0
63	72	76	--	--	--	0	0	0
--	69	74	--	--	--	1	0	0
68	85	--	--	--	98	0	0	1

Formally, Rubin's mechanisms are defined by statements that indicate whether the observed or missing scores relate to binary missing data indicators. This mechanism is what researchers think of as purely unsystematic missingness. The formal definition of MCAR is as follows.

$$\Pr\left(\mathbf{M} = 1 | \mathbf{Y}_{(obs)}, \mathbf{Y}_{(mis)}\right) = \Pr(\mathbf{M} = 1) \tag{1}$$

The left side of the expression says that the probability of a missing score (i.e., a missing data indicator equal to one) *could* depend on both the observed and missing parts of the data to the right of the vertical pipe. The right side of the equality is a simplified function where nonresponse is unrelated to both parts of the data. In practical terms, the right side of Equation 1 says that all participants have the same chance of missing data. As alluded to previously, most deletion

applications invoke an MCAR mechanism, and estimates are prone to bias when this restriction assumption is violated.

The missing at random (MAR) mechanism states that nonresponse depends on the observed data *but not the unseen scores*. The right side of the equality encodes this definition.

$$\Pr\left(\mathbf{M} = 1|\mathbf{Y}_{(obs)}, \mathbf{Y}_{(mis)}\right) = \Pr\left(\mathbf{M} = 1|\mathbf{Y}_{(obs)}\right) \quad (2)$$

Counterintuitively, the MAR mechanism defines a systematic process where missingness varies across observed score profiles. That is, two participants with identical observed score profiles share the same chance of missing values, whereas two participants with different observed scores have different missingness rates. To clarify that missingness is random after conditioning on or controlling for the observed data, Graham (2009) referred to this process as *conditionally* missing at random (CMAR). I adopt his terminology to enhance clarity. The key advantage of the Big Three is that they produce unbiased estimates when the process is CMAR (or MCAR)[2]. Unfortunately, we cannot demonstrate that missingness is unrelated to the unseen scores because doing so would require knowing their values. Consequently, this default assumption for most Big Three applications does not have testable propositions.

Finally, a missing not at random (MNAR) process states that nonresponse depends on the unseen scores, even after accounting for one's observed score profile. The formal definition below says that the missing data indicators can depend on either the observed or missing parts of the data.

$$\Pr\left(\mathbf{M} = 1|\mathbf{Y}_{(obs)}, \mathbf{Y}_{(mis)}\right) \quad (3)$$

Under such a process, two participants with identical observed score profiles no longer have the same likelihood of missing data, as the would-be scores carry important information not contained in the observed data. Section 7 describes modeling frameworks that pair the Big Three with missing not at random processes.

To illustrate some possible realizations of Rubin's missing data mechanisms, consider a cluster-randomized trial that assigns a sample of middle schools to receive either the district-standard math curriculum or a new cognitive strategy instructional intervention designed to enhance math problem-solving (Montague et al., 2014). The researchers collected seven monthly assessments of

[2] Technically, estimates are "consistent", meaning that bias decreases to zero when the sample size gets large enough. Of course, the definition of "large enough" varies by model. In many cases, estimates are approximately unbiased with sample sizes that are typical for educational research.

an IRT-calibrated math problem-solving test during the school year. As is typical for longitudinal designs, the proportion of missing test scores increased over time.

To begin, a subset of participants had incomplete data because researchers used a planned missing data design (Graham et al., 2001; Graham et al., 2006; Mistler & Enders, 2011) where they intentionally administered only four of the seven monthly assessments. These values are missing completely at random because nonresponse was determined by the researchers and not the data. Other scores were missing for reasons beyond the researchers' control. Housing mobility was a potential determinant of missingness because students from low-income households were more likely to move out of the participating district. This source of missingness would be conditionally missing at random *if nonresponse is unrelated to a student's unseen problem-solving score.* Finally, a missing not at random mechanism could produce item-level missingness if students skipped problem-solving questions because they did not possess adequate knowledge to formulate a response (Finch, 2008; Mislevy & Wu, 1996). To reiterate, the observed data do not contain the information needed to identify, confirm, or differentiate CMAR and MNAR processes because both involve propositions about the *unseen* score values. Unfortunately, the default CMAR assumption invoked by most Big Three applications is inherently untestable. Only expert judgment and substantive knowledge can rule out an MNAR mechanism. When in doubt, sensitivity analyses involving the selection or pattern mixture modeling procedures described in Chapter 7 can be a useful adjunct to CMAR-based analyses.

1.3 Missing Data Auxiliary Variables

Practically speaking, the conditionally missing at random (CMAR) assumption for a Big Three application requires that all important determinants of missingness are contained within a model's observed data. Failing to satisfy this assumption can produce biased estimates. However, some of these determinants may be variables that would not have been considered had the data been complete. Returning to the randomized control trial scenario, consider an ANCOVA model that evaluates intervention group differences with baseline scores as a student-level covariate. If researchers believed that socioeconomic status is a potential determinant of missingness, then this variable could be important for missing data handling even though it was not part of the analytic plan.

Extraneous variables like socioeconomic status are known as auxiliary variables. Methodologists routinely recommend a so-called inclusive analysis strategy that incorporates these additional variables into a missing data handling procedure (Collins et al., 2001; Rubin, 1996; Schafer & Graham, 2002). Leveraging auxiliary variables can fine-tune a missing data analysis in

two important ways. Expanding the observed data can reduce nonresponse bias by making the CMAR assumption more plausible, and introducing additional sources of sources of correlation can improve power by recouping lost information. However, not all auxiliary variables are created equal. Collins et al. (2001) provide a hierarchy that helps prioritize their selection. To illustrate, Figure 1 depicts a bivariate analysis involving X and Y and single auxiliary variable A. The outcome variable Y has missing values, and M_Y is its missing data indicator. The solid arrow represents the focal analysis parameter, and dashed arrows are auxiliary variable assocations.

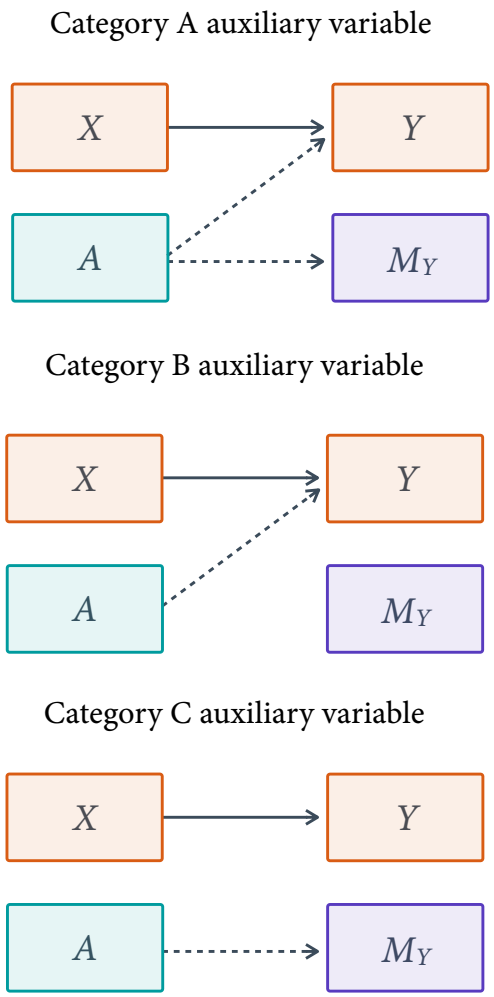


FIGURE 1. Auxiliary variable classification system from Collins, Schafer, and Kim (2001). Dashed arrows depict possible associations between an extraneous variable A, a missingness indicator M_Y, and an incomplete variable Y.

To begin, the top panel of the Figure 1 depicts a category A auxiliary variable. The arrow from *A* to *Y* conveys that the auxiliary variable contains unique covariation not contained in *X*, and the arrow from *A* to M_Y indicates that it also predicts missingness. Incorporating category A variables into a Big Three application is a priority because doing so can reduce nonresponse bias. Next, the middle panel shows a category B auxiliary variable that uniquely correlates with *Y* but does not predict its missingness. Although category B auxiliary variables do not mitigate bias, they can improve power by leveraging additional sources of correlation. Finally, the bottom panel depicts a category C auxiliary variable that predicts missingness but has no unique relation with *Y*. Although type C variables are useful for certain types of MNAR analyses, they are not otherwise beneficial as auxiliary variables.

As a practical matter, the number of additional variables in many data sets is often so large that an inclusive strategy can be daunting to implement. One possibility is to reduce a candidate set of auxiliary variables into one or two principal components and use the component scores as auxiliary variables (Howard et al., 2015). However, software packages that create component scores usually require complete data, thus precluding the use of incomplete auxiliary variables. Fortunately, a targeted approach that selects one or two salient auxiliary variables often works just as well as casting a broad net. Unused repeated measures variables are often excellent auxiliary variables because of their beneficial collinearity.

1.4 Preparing for a Missing Data Analysis

Returning to Figure 1, important category A auxiliary variables have two features: they are correlates of missingness and provide unique sources of covariation beyond that contained in a model's observed data. This section describes some simple strategies for identifying such patterns. These methods are imperfect, but selecting auxiliary variables need not be an exact science.

To begin, researchers routinely use a pattern mean difference approach to identify potential correlates of nonresponse. To implement this strategy, you first create a binary missing data indicator for each incomplete analysis variable with substantial missing data[3]. Treating the indicator as a grouping variable, you then examine whether the incomplete cases exhibit mean differences on other variables. For example, the next section illustrates a multiple regression where

[3] Although "substantial" is a clearly a subjective term, applying this method to a variable with a very small amount of missing data will produce a near-constant indicator comprised mostly of zeros. Unless the sample size is very large, such an indicator would produce noisy pattern mean difference estimates. Moreover, auxiliary variables would not be beneficial for such a variable because there is little missing information to recover.

I create a missing data indicator by recoding the dependent variable as 0 = observed and 1 = missing. I then examined whether the two groups exhibited mean differences on a set of auxiliary variables. There is no need to examine mean differences for variables already in the analysis (e.g., the predictors) because the Big Three automatically leverage on this information.

Most often, significance tests are not very useful for evaluating mean differences because unbalanced group sizes (e.g., an indicator with a preponderance of zeros) decrease power. Instead, standardized mean difference effect sizes or bivariate correlations can provide a practical metric for evaluating pattern differences. There is no harm in adopting a conservative approach that flags candidate auxiliary variables that produce a small effect size (e.g., $|d|> 0.20$ or $|r| > .10$). I adopt this approach in the subsequent example.

Logistic regression is another common approach for identifying correlates of missingness. This strategy reverses the role of the variables, treating the missingness indicator as the outcome and candidate auxiliary variables as predictors. Logistic regression can yield different conclusions from the pattern mean difference approach because it considers a multivariate system where each coefficient reflects a partial effect that controls for other predictors. As such, selecting extraneous variables based on their odds ratios or statistical significance will likely identify a smaller auxiliary set than the pattern mean difference approach, which is inherently bivariate.

Regardless of method, it important to emphasize that a salient association between a missing data indicator and auxiliary variable *does not* mean that the extraneous variable is a potential source of nonresponse bias—that conclusion also requires a residual correlation between the incomplete variable and auxiliary variable (see Figure 1a). However, identifying a correlate of missingness *does* rule out a purely haphazard missing completely at random process. In fact, the Little's (1988) MCAR test available in some software packages is simply a multivariate version of the pattern mean difference approach. Although such a test may seem appealing, evaluating the null hypothesis that missingness is unsystematic has limited practical utility because the Big Three do not require this strict assumption. Moreover, finding evidence for or against haphazard missingness does not change the recommendation to use the Big Three. Raykov (2011, p. 428) emphasized this point, stating that "the desirability of the MCAR condition has been frequently overrated in empirical social and behavioral research".

As mentioned previously, the presence or absence of *residual* or *semipartial* correlation between an auxiliary variable and an incomplete analysis variable largely determines the extra variable's worth. Conceptually, these correlations are computed by first regressing an incomplete variable on all other variables in the analysis, then correlating its residuals with a set of auxiliary variables. Statistical packages readily compute these semipartial correlations, although they generally use listwise or pairwise deletion to do so. Noisier estimates that require unsystematic

missingness are not necessarily a problem at the initial data screening stage. Alternatively, Raykov and West (2015) outlined a latent variable modeling approach to estimating semipartial correlations that readily pairs with maximum likelihood and Bayesian estimation. These approaches use all available data and accommodate a systematic conditionally missing at random process.

A remarkable semipartial correlation signals that an auxiliary variable contains unique information about the missing values not contained in a model's observed data. How large does a residual correlation need to be? Computer simulations from Collins et al. (2001) suggest $|r| > .30$ as an approximate rule of thumb. Their study showed that ignoring category A variables with weaker correlations generally produced parameter estimates with little to no bias, whereas adjusting for sources of stronger correlation reduced or eliminated nonresponse bias. We are often taught that collinearity is detrimental, but it is beneficial in this context—the stronger the semipartial correlations, the more missing information there is to recover. Practical experience suggests that one or two "good" auxiliary variables is usually sufficient. In truth, it is often difficult to find beneficial variables because our models usually capture most salient covariation in the data. Finally, incomplete auxiliary variables can still be beneficial if their scores are mostly observed in the data records where the analysis variables are missing (Enders, 2008). Auxiliary variables that are missing in conjunction with the analysis variables have no information to contribute.

1.5 Analysis Example

I use multiple regression analyses to illustrate Big Three applications throughout the report. This section demonstrates the process of evaluating candidate auxiliary variables for these analyses. The illustration uses the `behaviorachievement.dat` data set from a longitudinal study that followed 138 students from primary through middle school. The file includes three annual assessments of broad reading and math achievement beginning in the first grade, seventh grade standardized achievement test scores taken from a statewide assessment, and a final measure of broad reading and math obtained in ninth grade. The data also contain teacher ratings of behavioral symptoms and learning problems were also obtained in the first grade. The Software Tutorials document provides additional information about this data set. Table 2 below shows the specific variables for this analysis example.

TABLE 2. Variables for the Multiple Regression Analysis

Name	Definition	Missing	Scale
	Focal Analysis Variables		
$BEHSYMP_1$	1st grade emotional/behavior problems	2.2%	Numeric (17 to 92)
$LRNPROB_1$	1st grade learning problems	2.2%	Numeric (31 to 88)
$READ_1$	1st grade broad reading	6.5%	Numeric (39 to 153)
$READ_9$	9th grade broad reading	17.4%	Numeric (41 to 123)
	Auxiliary Variables		
$READ_2$	2nd grade broad reading	9.4%	Numeric (20 to 150)
$READ_3$	3rd grade broad reading	14.5%	Numeric (46 to 138)
$STANREAD_7$	7th grade standardized reading	19.6%	Numeric (100 to 399)

The analysis model features ninth grade broad reading scores regressed on three academic and behavioral measures collected in first grade: the broad reading composite, teacher-rated learning problems, and teacher-rated behavioral problems.

$$READ_9 = \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) + \beta_3(BEHPROB_1) + \varepsilon \quad (4)$$

To reiterate, the conditionally missing at random assumption requires that missingness is fully determined by the model's observed data—unseen score values carry no additional information. Introducing missing data auxiliary variables into a Big Three analysis broadens the observed data, potentially making this assumption more plausible. Selecting auxiliary variables for ninth grade reading scores is the priority because the outcome has substantial missing data. Conversely, the learning and behavioral problems measures would not benefit from extraneous variables because they have very little missing information.

Researchers usually have large data sets with dozens of possible auxiliary variables. Substantive knowledge usually allows researchers to identify a smaller number of variables that potentially correlate with analysis variables. To keep the illustration simple, I consider three candidates: second and third grading broad reading scores and seventh grade standardized reading scores. Returning to Table 2, the auxiliary variables themselves have substantial missing data. However, these variables are still viable because there are relatively few missing data patterns where an auxiliary variable is concurrently missing with the outcome (a pattern where the auxiliary variable

carries no additional information). For example, only five students are missing both the standardized reading test and ninth grade broad reading scores.

To implement the pattern mean difference approach, I created a binary missing data indicator for the ninth grade test scores, and I then examined whether the two groups exhibited mean differences on other variables. The first column of Table 3 shows Cohen's (1988) standardized mean difference effect sizes from this step[4]. First, notice that the incomplete cases exhibited moderate to large mean differences on the learning and behavioral problems measures. The positive difference implies that elevated academic problems are associated with missingness. These differences do not require additional action because the Big Three automatically adjust for systematic missingness among the analysis variables. The incomplete cases also had small, negative mean differences on the auxiliary variables, such that lower reading scores were associated with higher rates of missingness.

The pattern mean difference approach is inherently univariate and does not account for shared covariation among the variables. To examine whether the auxiliary variables uniquely predicted missingness after controlling for the analysis variables, I conducted a binary logistic regression analysis with the missing data indicator as the outcome. The middle panel of Table 3 shows the resulting regression slopes, standard errors, and significance tests. With the important caveat that the analysis likely suffers from low power, the regression results suggested that learning problems ratings—a variable already in the focal model—was the most salient correlate of missingness. There was little or no evidence that the auxiliary variables uniquely determined missingness after controlling for the predictor variables from the focal model. This does not mean that the auxiliary variables are unimportant, however. Their semipartial (residual) correlations determine that.

The rightmost column of Table 3 shows the semipartial correlations between ninth grade reading scores and the auxiliary variables. Conceptually, these correlations link the auxiliary variables to the unexplained part of ninth grade reading scores that remains after residualizing on the predictors. As a reminder, computer simulations from Collins et al. (2001) suggest $|r| > .30$ as a rough rule of thumb. Two of the correlations exceeded the cutoff, suggesting that they possess unique information about the missing values not contained in the model's observed data. Like the pattern mean difference approach, the residual correlations do not account for shared covariation among the variables. Because the auxiliary variables are very highly correlated (all rs > .70), there is probably little benefit to including them all in a Big Three analysis. For this reason, I use second

[4] Cohen (1988) characterized a small effect size as $|d| > .20$, medium as $|d| > .50$, and a large effect as $|d| > .80$.

grade broad reading scores and seventh grade standardized test scores as auxiliary variables in all subsequent analysis examples.

TABLE 3. Auxiliary Variable Screening Results

		Logistic Regression				
Variable	Cohen's *d*	Est.	*SE*	*z*	*p*	Residual *r*
		Analysis Variables				
$BEHSYMP_1$	0.65	–0.01	0.03	–0.37	.71	--
$LRNPROB_1$	0.86	0.10	0.04	2.29	.02	--
$READ_1$	-0.23	–0.04	0.04	–0.92	.36	--
		Auxiliary Variables				
$READ_2$	–0.29	0.01	0.04	0.28	.78	.30
$READ_3$	–0.19	0.04	0.04	0.86	.39	.27
$STANREAD_7$	–0.26	–0.01	0.005	–0.75	.45	.38

To summarize, the pattern mean difference approach and logistic regression analysis revealed a systematic process where students with missing ninth grade scores differed from those with observed data. These results provide compelling evidence against a purely haphazard missing completely at random process. Importantly, we cannot conclude that these systematic patterns are unrelated to the *unseen* score values—a vital component of the conditionally missing at random assumption for a Big Three analysis. Returning to the Collins et al. (2001) hierarchy in Figure 1, the auxiliary variables most closely resemble category B variables because they do not uniquely predict missingness. In practice, it is not important to precisely categorize auxiliary variables because a salient residual correlation justifies their inclusion.

2

Factored Regression Specifications

Missing data handling is most straightforward in the (perhaps unusual) situation where missing values are relegated to the outcome variable. The situation becomes more complex when predictor variables have missing data because the explanatory variables require their own models and statistical assumptions. Classic versions of the Big Three assume that incomplete variables share a common distribution, typically multivariate normal. Adopting a multivariate distribution for missing data handling can be restrictive because it assumes all variables share the same metric. As an example, this assumption precludes an analysis with an incomplete multicategorical and numeric variables. Moreover, applying multivariate normal missing data methods to models with interactive or nonlinear effects can produce substantial bias.

Factored regression specifications express a multivariate distribution as a sequence of simpler distributions, the collection of which is equivalent to the original joint function. Beyond acknowledging that a multivariate distribution exists, factorization makes no assumptions about its shape or form. Rather, distributional assumptions enter on a variable-by-variable or model-by-model basis. The theory for this framework dates to work by Ibrahim and colleagues (Huang et al., 2005; Ibrahim, 1990; Ibrahim et al., 2002; Ibrahim et al., 1999; Lipsitz & Ibrahim, 1996). Factored specifications have received considerable attention in the recent missing data literature because they solve long-standing problems with classic multivariate normal missing data methods that came online in the early 2000s (Enders, 2023).

To illustrate the idea behind a factored regression specification, consider a regression model with two incomplete predictors[5]. Figure 2 depicts multivariate and factored specifications. Colored boxes denote variables with the same distribution assumption. Figure 2a is a multivariate

[5] The term "factored regression" does not imply that the analysis model must be a regression. These specifications can be applied to a broad range of generalized linear models, measurement models, and structural equation models.

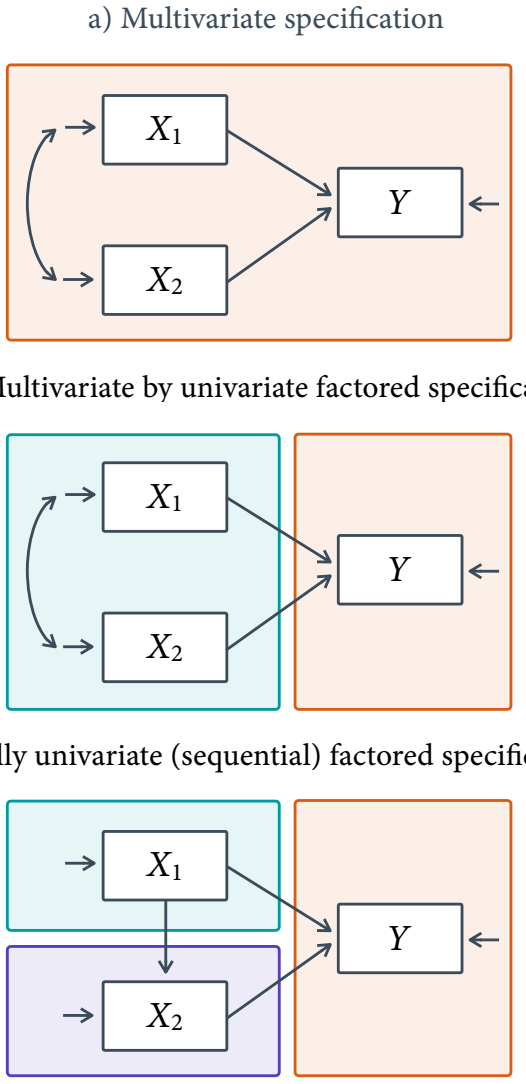


FIGURE 2. Colored boxes enclose variables that share the same distribution. Panel (a) depicts a multivariate specification where all variables have the same metric. Panel (b) is a factored specification that assigns a bivariate distribution for the predictors and a distinct distribution to the outcome. Panel (c) is a specification where each variable has its own distribution.

specification where all variables share a common distribution (e.g., multivariate normal). Figure 2b separates the original trivariate distribution into two simpler distributions: a bivariate distribution for the predictors, and a univariate distribution for the outcome. Finally, Figure 2c

further separates the predictors into distinct models, resulting in a fully univariate specification (Erler et al., 2016; Lüdtke et al., 2020b).

It is useful to establish some shorthand notation to reference the specifications in Figure 2. In the following equation, each f generically references a distribution comprised of the variables in parentheses[6].

$$f(Y, X_1, X_2) = f(Y|X_1, X_2) \times f(X_1, X_2) = f(Y|X_1, X_2) \times f(X_1|X_2) \times f(X_2) \quad (5)$$

Each term with a vertical pipe represents the univariate conditional distribution of an outcome variable in a regression equation, and the variables to the right of the pipe are its regressors. At a high level, the equation says we have three equivalent ways to configure the distributions of the three incomplete variables in Figure 2. The leftmost term is a trivariate distribution that corresponds to Figure 2a. The middle pair of terms separate the multivariate distribution into two simpler distributions: the univariate distribution of *Y* given the *X*s (the focal regression analysis) and a bivariate distribution for the predictors. These terms align with Figure 2b. Finally, the rightmost trio of terms corresponds to the fully univariate specification in Figure 2c.

Although the factorizations are symbolically equivalent, they aren't necessarily the same in practice. For example, Equation 4 uses academic and behavioral measures collected in first grade to predict ninth grade reading achievement. Because all variables are numeric with reasonably symmetric distributions, it makes no difference how the model is specified. Instead, suppose the analysis is a logistic regression where the outcome is a binary measure of whether the student has attained average or higher reading proficiency in ninth grade (see Examples 2 and 7 in the Software Tutorials document). In this case, we do not have an off-the-shelf multivariate distribution that describes the cooccurrence of binary and numeric variables. Adopting the multivariate specification on the left side of Equation 5 (or Figure 2a) is not an option. In contrast, either factored specification is appropriate. The univariate-by-multivariate factorization (the two middle terms in Equation 5 or Figure 2b) would pair a logistic focal model (Bernoulli outcome distribution) with a bivariate normal model for the predictors. Finally, the fully univariate specification (the three rightmost terms in Equation 5 or Figure 2c) would pair the logistic model with a pair of linear regressions with normal errors.

Analyses with incomplete nonlinear terms are another important use case for factored specifications. This includes models with interactions, curvilinear effects, and random coefficients, among others. Such models are incompatible with a multivariate normal distribution (Bartlett et

[6] The factorization is derived by applying the probability chain rule to the multivariate distribution in the leftmost term.

al., 2015; Liu et al., 2014), meaning that they require features that are mathematically impossible with multivariate normal data (e.g., heteroscedasticity). Numerous methodology studies have demonstrated that applying normal distribution assumptions to interactive effects can introduce substantial bias (Cham et al., 2017; Enders et al., 2014; Humberg & Grund, 2022; Seaman et al., 2012; Zhang & Wang, 2017). In contrast, both factored specifications depicted in Figure 2 readily accommodate interactions and nonlinear terms in the focal analysis model, and the fully univariate specification in Figure 2c additionally allows nonlinear associations among predictors (Lüdtke et al., 2020b). For example, X_1 could exert a curvilinear effect on X_2.

If the data and model allow, we may choose a multivariate missing data handling method because doing so is convenient (software options abound), and a wealth of supporting literature has accumulated over the last 25 years. However, the key takeaway from Equation 5 is that we don't *need* to work with a multivariate distribution—whatever the joint function's shape or form, we can always reproduce it by adopting a collection of simpler submodels, each with its own distributional assumption. This flexibility allows us to tailor a missing data handling procedure that preserves important features of the data and analysis model. The Software Tutorials document features numerous analysis examples with factored regression specifications.

The flowchart in Figure 3 depicts a decision tree for missing data analyses. Starting on the left, the first decision point depends on whether the focal analysis features any type of nonlinearity. This includes interactive effects, curvilinear associations, and random coefficients, among other things. If the answer to this first question is yes, then factored regression specification is the only choice. If no, then subsequent steps depend on the variable types. If the analysis is restricted to normal variables, classic multivariate versions of the Big Three are appropriate, as are newer factored specifications. Finally, if the analysis includes a mix of discrete and numeric variables, certain classic multiple imputation methods like fully conditional specification (i.e., the MICE algorithm; van Buuren, 2007; van Buuren et al., 2006; van Buuren & Groothuis–Oudshoorn, 2011) are appropriate, as are factored regression specifications. The flowchart oversimplifies a nuanced issue, but it provides a high-level heuristic for classifying missing data handling options.

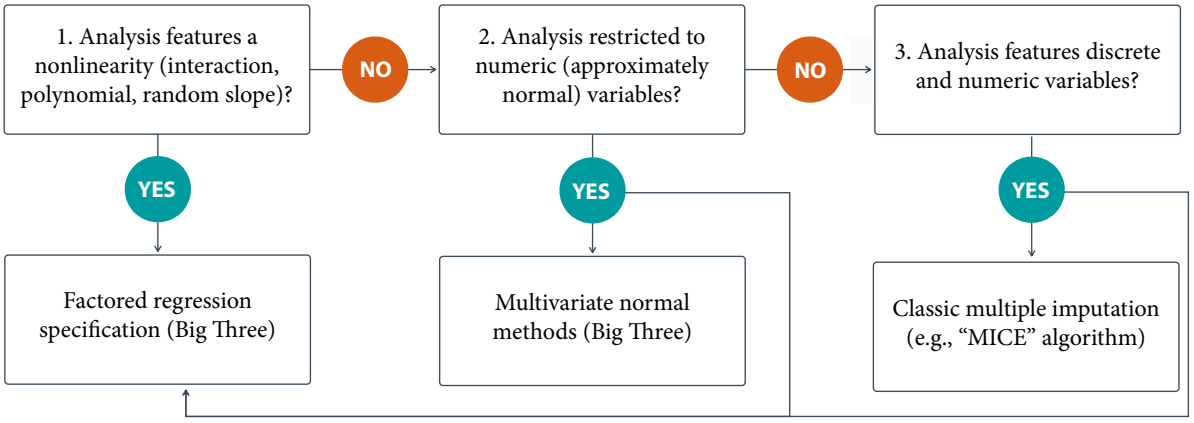


FIGURE 3. Flowchart depicting decision tree for a missing data analysis. Starting on the left, the first decision point depends on whether the focal analysis features any type of nonlinearity. The second decision is whether all variables share the same metric. Although classic methods are not always appropriate, factored specifications are.

3

Maximum Likelihood Estimation

The origins of modern maximum likelihood estimators date back as far as the 1950s (Anderson, 1957; Edgett, 1956; Hartley, 1958; Lord, 1955), but the foundational theoretical and computational advancements mainly occurred in the 1970s (Beale & Little, 1975; Dempster et al., 1977; Finkbeiner, 1979; Hartley & Hocking, 1971; Orchard & Woodbury, 1972). For researchers in education (and the social and behavioral sciences more broadly), maximum likelihood estimation became a practical reality in the 1990s when structural equation modeling software packages started offering so-called full information maximum likelihood (FIML) estimators based on raw data (Arbuckle, 1996). Since then, a substantial amount of methodological work has accumulated that extends FIML's utility and documents its limitations.

Maximum likelihood estimators have evolved considerably, and factored regression specifications that accommodate mixtures of categorical and continuous variables are now widely available (Ibrahim, 1990; Ibrahim et al., 1999; Lüdtke et al., 2020a; Muthén et al., 2016; Pritikin et al., 2018; Rabe-Hesketh et al., 2004; Rockwood & Jeon, 2019). However, this functionality varies across software packages, and not all combinations of metrics are currently available. Support for binary and ordinal variables is common, but missing data handling for other variable types is currently more limited.

3.1 The Classic FIML Estimator

The classic FIML estimator uses an iterative optimization algorithm to identify model parameters that minimize the sum of squared, standardized distances between a model's predictions and the observed data. At a high level, maximum likelihood's goal is the same as ordinary least squares (OLS), which is to find estimates that minimize residuals. The normal distribution function—more accurately, its natural log or log-likelihood—provides a formal metric for quantifying that data–model fit. The observed-data log-likelihood function for a sample of N cases is shown below.

$$\ln(\ell) = \left[-\sum_{i=1}^{N} \frac{V_i}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \ln|\boldsymbol{\Sigma}_i| \right] - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \tag{6}$$

The equation can be understood as follows. First, the multivariate normal distribution function quantifies each participant's data–model fit on a probability-like metric called a likelihood. Taking the natural log of the distribution equation recodes everyone's fit equation to a logarithmic metric where higher (less negative) log-likelihood values represent better fit and smaller (standardized and squared) residuals. Finally, Equation 6 sums the individual log-likelihood (fit) values into a sample-level summary.

At a high level, the terms in square brackets are scaling factors that ensure the area under the normal curve equals one. The terms outside the brackets (the distribution's "kernel") form the key part of the equation that captures data-model fit. Setting aside the scaling terms and focusing on the kernel, $(\mathbf{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$ is the sum of squared, standardized residuals between a person's data and the model's predictions[7]. Because this person-specific sum is a chi-square variable, we can use the following conceptual expression to simplify the log-likelihood equation and emphasize the role of the standardized residuals.

$$\ln(\ell) = -[\text{sum of } N \text{ scaling terms}] - \frac{1}{2} \sum_{i=1}^{N} \chi_i^2 \tag{7}$$

The goal of maximum likelihood is to find the parameter values in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that minimize the sum of the residuals—a goal it shares with least squares estimation.

At a more granular level, $\mathbf{Y}_i$ is a vector that contains an individual's observed data, and the V_i in the scaling part is the number of data points in $\mathbf{Y}_i$. The variables in $\mathbf{Y}_i$ can differ across people due to missing data. Each person's fit (chi-square) is computed using whatever observed data are present in $\mathbf{Y}_i$. Data are neither imputed nor discarded. When $\mathbf{Y}_i$ has missing values, only a subset of the model's parameters contributes to each chi-square calculation. For example, a person missing Y_3 has no information about μ_3 and σ_3^2 (or covariances involving Y_3), so these parameters cannot be used to compute χ_i^2. Equation 6 uses $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ to represent the subset of parameters that link to the observed data in $\mathbf{Y}_i$ (i.e., the parameters that determine each person's fit contribution).

To make the discussion more concrete, reconsider the regression model from Figure 2a. further, suppose that Y is ninth grade academic achievement and X_1 and X_2 are academic and behavioral measures from first grade. For a student with complete data, the standardized residual calculation uses all three scores and every model parameter.

[7] In multivariate statistics, the square root of $(\mathbf{Y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$ is known as Mahalanobis distance.

$$\chi_i^2 = \begin{pmatrix} X_1 - \mu_{X_1} \\ X_2 - \mu_{X_2} \\ Y - \mu_Y \end{pmatrix}' \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} & \sigma_{X_1Y} \\ \sigma_{X_2X_1} & \sigma_{X_2}^2 & \sigma_{X_2Y} \\ \sigma_{YX_1} & \sigma_{YX_2} & \sigma_Y^2 \end{pmatrix}^{-1} \begin{pmatrix} X_1 - \mu_{X_1} \\ X_2 - \mu_{X_2} \\ Y - \mu_Y \end{pmatrix} \tag{8}$$

As a second example, the data–model fit for students with missing *Y* values (e.g., ninth grade achievement) involves just the *X*s and their corresponding parameters.

$$\chi_i^2 = \begin{pmatrix} X_1 - \mu_{X_1} \\ X_2 - \mu_{X_2} \end{pmatrix}' \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} \\ \sigma_{X_2X_1} & \sigma_{X_2}^2 \end{pmatrix}^{-1} \begin{pmatrix} X_1 - \mu_{X_1} \\ X_2 - \mu_{X_2} \end{pmatrix} \tag{9}$$

As a final example, the chi-square fit variable for students with missing *X*s (e.g., first grade measures) reflects only *Y* and its parameters.

$$\chi_i^2 = (Y - \mu_Y)'\left(\sigma_Y^2\right)^{-1}(Y - \mu_Y) = \frac{(Y - \mu_Y)^2}{\sigma_Y^2} \tag{10}$$

As a small clarifying point, the means, variances, and covariances in the previous expressions are themselves predictions that derive from the regression model parameters in Figure 2a (i.e., the elements in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are functions of the regression model parameters). For example, the outcome variable's variance σ_Y^2 reflects explained variation due to the *X*s via their regression slopes plus unexplained residual variance. Similarly, μ_Y is a function of the regression intercept and an adjustment term that depends on predictor means and regression slopes. These algebraic mappings from the regression model to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ follow from viewing the path diagram in Figure 2a as a structural equation model.

The previous equations demonstrate that the classic FIML estimator uses all available data, but they don't necessarily convey *how* the partial data records improve the final estimates. To illustrate, consider a bivariate scenario involving first grade learning problems and ninth grade reading scores. Further, suppose that the ninth grade reading scores are predominantly missing for students in the upper tail of the learning problems distribution. This situation is similar to the one from the analysis example in Section 1.5. As an aside, this process conforms to a conditionally missing at random mechanism because missingness in ninth grade is explained by first grade learning problems ratings.

Figure 4a shows a scatterplot of an artificial data set that depicts the hypothetically-complete data from the bivariate scenario. The orange circles correspond to students with missing ninth grade reading scores, and the green circles are students with complete data on both measures. The black dot represents the "true" means that would have resulted from analyzing a complete data set.

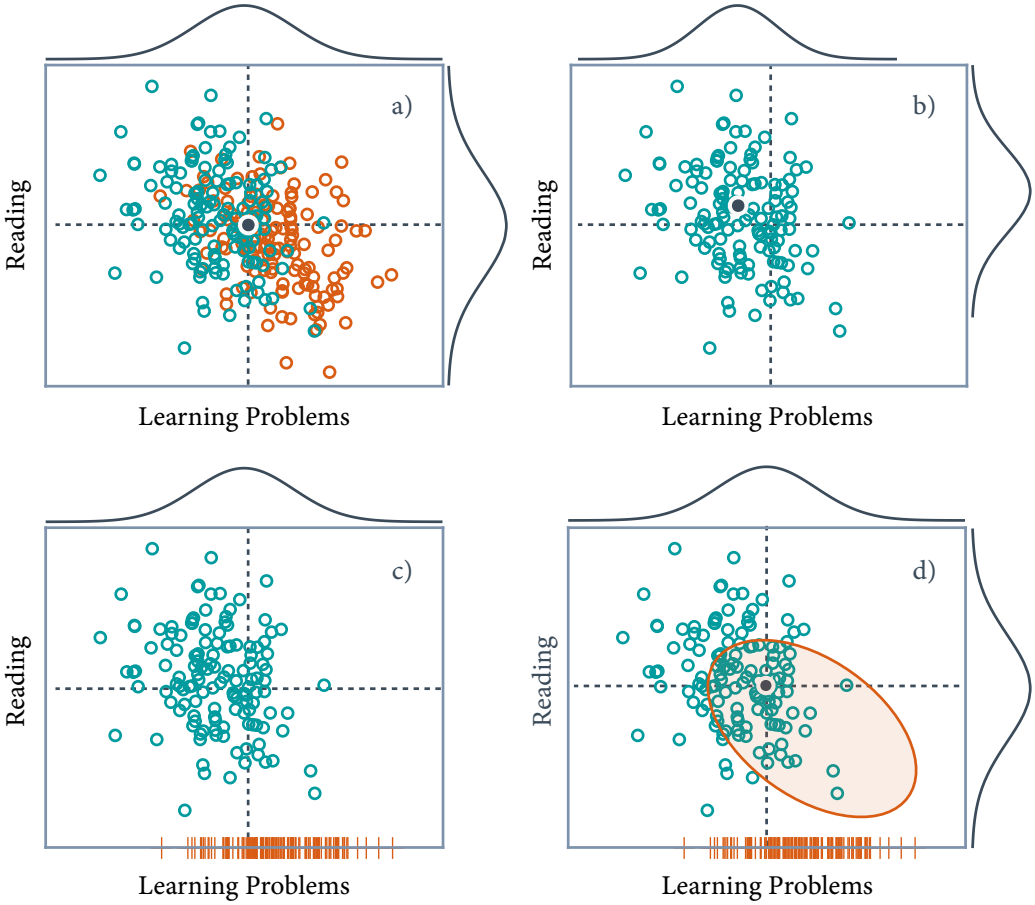


FIGURE 4. Panel (a) depicts the hypothetically-complete data from a bivariate analysis. The orange circles correspond to students with missing ninth grade reading scores, and the green circles are students with complete data on both measures. Panel (b) shows the sample after deletion, and panel (c) depicts partial data records as hashmarks on the horizontal axis. Panel (d) shows the distributions from FIML based on incomplete data records.

Figure 4b is a scatterplot of just the complete data records. Excluding students with incomplete data truncates the upper tail of the learning problems distribution and the lower tail of the reading distribution. As a consequence, the distributions in the marginals are distorted—the variances are too small, and the means are biased—because the complete cases are not representative of the original sample. The FIML estimator incorporates the partial data records from the upper tail of the learning problems distribution. Figure 4c shows the partial data as hashmarks along the horizontal axis. Importantly, the additional scores restore the variation in the learning problems

variable and recenter the variable's mean at the correct location on the horizontal axis. In a bivariate normal distribution with a negative correlation, the high learning problems scores only make sense if they are paired with low reading scores. Although FIML does not impute the missing reading scores, this property of the normal distribution implies that the unobserved scores must fall in the neighborhood of the elipse in Figure 4d. Essentially, the partial data records allow the estimator to intuit the location of the missing scores, thereby adjusting the mean and variance of the incomplete variable to match a hypothetical complete-data analysis.

3.2 Fine-Tuning FIML: Auxiliary Variables and Nonnormal Data

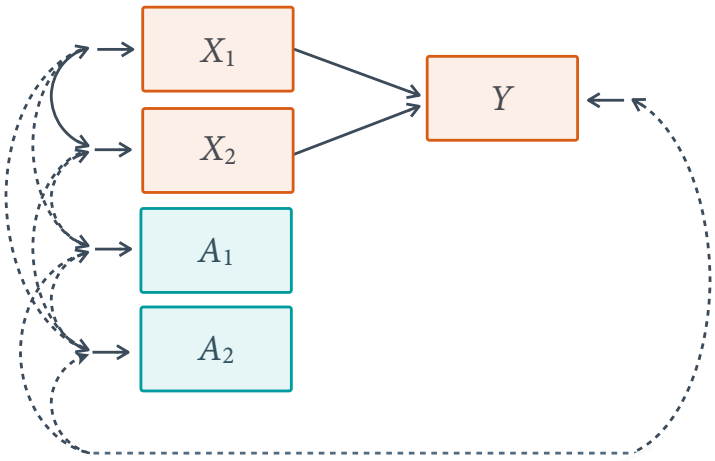
The conditionally missing at random assumption described in Section 1.2 requires that the unseen score values contain no unique information about missingness beyond that contained in the observed data. Practically speaking, this assumption requires that all important determinants of missingness are contained within a model's observed data. In the multiple imputation literature, the long-standing advice is to include additional auxiliary variables when treating missing data because doing so minimizes the risk of nonresponse bias (Collins et al., 2001; Rubin, 1996). Porting this recommendation to maximum likelihood requires care because we need to add the extra variables in a way that does not affect the interpretation of our focal model's parameters (the estimates and standard errors could change due to the influx of additional information). For example, adding extra auxiliary variables as covariates is undesirable because doing so changes the composition of the analysis model. Instead, we want to incorporate the additional variables while maintaining the structure of the model that we would have fit *had there been no missing data*.

Figure 5 shows two auxiliary variable model specifications for a multiple regression model. Figure 5a is Graham's (2003) saturated correlates model for FIML applications with structural equation modeling software. The curved, dashed arrows are correlations that connect the auxiliary variables to the residuals of all analysis variables as well as to each other. This specification is automated in M*plus* (Muthén & Muthén, 1998–2017), EQS (Bentler, 2000-2008), and the R package `semTools` (Jorgensen et al., 2022). The saturated correlates approach has two noteworthy limitations: It is restricted to normally distributed variables, and it is known to produce convergence failures, especially as the number of auxiliary variables increases (Savalei & Bentler, 2009). Convergence issues can be mitigated by using a small number of salient auxiliary variables or by performing a preliminary data reduction step that reduces a large set of additional variables into one or two principal components (Howard et al., 2015).

Alternatively, auxiliary variables can enter a model as additional outcomes that are predicted by the analysis variables and by each other (Lüdtke et al., 2020b). Figure 5b shows this specification,

with dashed lines depicting the auxiliary variable regression slopes. With this approach, the focal analysis model is embedded in larger network of variables. Importantly, the additional regressions do not align with substantive theory and need not reflect a logical causal order. Rather, the additional regressions are simply a tool for linking the analysis variables to the auxiliary variables. Importantly, pairing the additional outcome approach with a factored regression specification readily accommodates discrete auxiliary variables.

a) Saturated Correlates Model



b) Extra Outcome Model

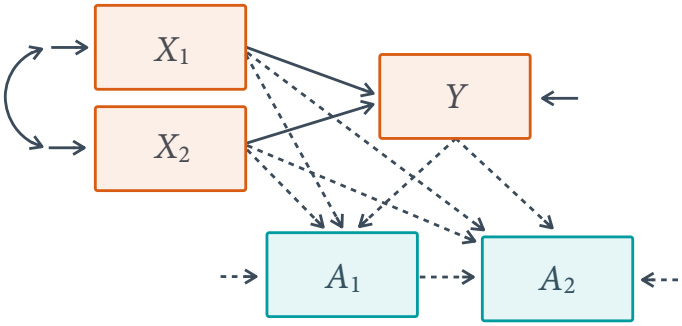


FIGURE 5. Solid lines denote the focal regression model parameters, and dashed lines are auxiliary variable parameters. Panel (a) depicts a saturated correlates model that connects the analysis variables to the auxiliary variables via correlated residuals. Panel (b) is an alternate specification where auxiliary variables function as additional outcomes.

The classic maximum likelihood estimator that enjoys widespread use also assumes multivariate normality. Of course, real data rarely conform to this ideal. A good deal of analytic and simulation work has clarified that maximum likelihood point estimates are generally unbiased in large samples (i.e., consistent) when the mechanism is conditionally missing at random. However, the same is not true for standard errors and test statistics (Savalei, 2010a, 2010b; Takai & Kano, 2013; Yuan, 2009; Yuan & Bentler, 2010; Yuan & Savalei, 2014; Yuan et al., 2012). Fortunately, robust (sandwich estimator) standard errors and test statistics have long been

available for missing data analyses (Arminger & Sobel, 1990; Yuan & Bentler, 2000), and numerous software packages offer these corrections. Savalei and Rosseel (2022) provide a comprehensive summary of these technical innovations. Bootstrap standard errors and test statistics are an alternative corrective procedure for nonnormal missing data (Enders, 2002; Savalei & Yuan, 2009).

Models that feature both discrete and continuous variables are a case where correctives for nonnormal data may or may not be useful. The complete-data literature suggests that treating ordinal outcomes as normal is not problematic if the discrete distribution is symmetric and has at least five response options (Rhemtulla et al., 2012). There is every reason to expect this finding to hold with missing data. Perhaps surprisingly, computer simulation results suggest that applying normal distribution assumptions to incomplete *binary predictors* may not introduce bias in single-level regression models (Muthén et al., 2016). However, this conclusion does not appear to extend to multilevel models with discrete level-2 predictors (Grund et al., 2018). Applying normal distribution assumptions to multicategorical nominal predictors (or their dummy codes) always produces nonsensical results.

3.3 Factored Regression Specifications

Maximum likelihood has evolved considerably since the classic FIML estimator came online in the late 1990s. Factored regression estimators (Ibrahim, 1990; Ibrahim et al., 1999) are a particularly important innovation and a focus of ongoing methodology research. An earlier section explained that factored regression expresses a multivariate distribution as a collection of submodels, each with distinct distributional assumptions. In lieu of a multivariate log-likelihood like the one from Equation 6, factored specifications adopt a chain of distinct log-likelihood functions that follow the pattern from Equation 5 (in this context, each f represents a likelihood function). These multi-function expressions are a challenging computational problem. Iterative optimization routines use numerical or Monte Carlo integration schemes that fill in the missing parts of the data in an imputation-esque manner. Lüdtke et al. (2020a) provide an accessible introduction to maximum likelihood estimation with factored specifications.

Analyses with combinations of discrete and continuous variables are an important use case for factored regression specifications. Estimators that accommodate some combinations of categorical and continuous variables are widely available in software packages (Grund et al., 2021a; Lüdtke et al., 2020a; Muthén et al., 2016; Pritikin et al., 2018; Rabe-Hesketh et al., 2004; Rockwood & Jeon, 2019). Functionality varies substantially, however, and not all combinations of metrics are currently available. Software programs routinely support binary and ordinal variables, but missing

data handling for other discrete metrics is more limited. Bayesian MCMC algorithms discussed in Section 4 are far more capable, at least for now.

Analyses with incomplete nonlinear terms (interactions, curvilinear effects, random coefficients) are another important use case for factored specifications. Until recently, researchers have been forced to use a so-called just-another-variable approach that treats product and polynomial terms as unique, normally distributed variables (Enders et al., 2014; von Hippel, 2009). Analytic and computer simulation work uniformly demonstrates this strategy's propensity for bias (Cham et al., 2017; Enders et al., 2014; Humberg & Grund, 2022; Seaman et al., 2012; Zhang & Wang, 2017). Factored regression estimators instead treat product and curvilinear effects as deterministic functions of their component variables. The limited research to date suggests that a factored specifications are uniformly superior to normal-theory FIML estimation with the just-another-variable strategy (Humberg & Grund, 2022; Lüdtke et al., 2020a). The Software Tutorials document features FIML analysis examples with factored regression specifications.

3.4 Maximum Likelihood Analysis Example 1

The first analysis example illustrates a multiple regression model comprised of incomplete continuous variables. The analysis uses the `behaviorachievement.dat` data set from a longitudinal study that followed 138 students from primary through middle school. The file includes three annual assessments of broad reading and math achievement beginning in the first grade, seventh grade standardized achievement test scores taken from a statewide assessment, and a final measure of broad reading and math obtained in ninth grade. The data also contain teacher ratings of behavioral symptoms and learning problems were also obtained in the first grade. The Software Tutorials document provides additional information about this data set. Table 2 from Section 1.5 shows the specific variables for this analysis example.

The analysis model featured ninth grade broad reading scores regressed on three academic and behavioral measures collected in first grade: the broad reading composite, teacher-rated learning problems, and teacher-rated behavioral problems.

$$READ_9 = \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) + \beta_3(BEHSYMP_1^*) + \varepsilon \tag{11}$$

The analysis also included second grade broad reading scores and seventh grade standardized test scores as auxiliary variables. These additional variables were included because they exhibited significant residual covariation with the analysis variables (see Section 1.5). Applying the flowchart from Figure 3, either a multivariate or factored regression specification is appropriate. I use the classic FIML estimator for multivariate normal data along with a saturated correlates specification

for the auxiliary variables. Figure 6 shows a path diagram of the analysis model. The dashed arrows are the residual correlations connecting the auxiliary variables to each other and to the analysis variables.

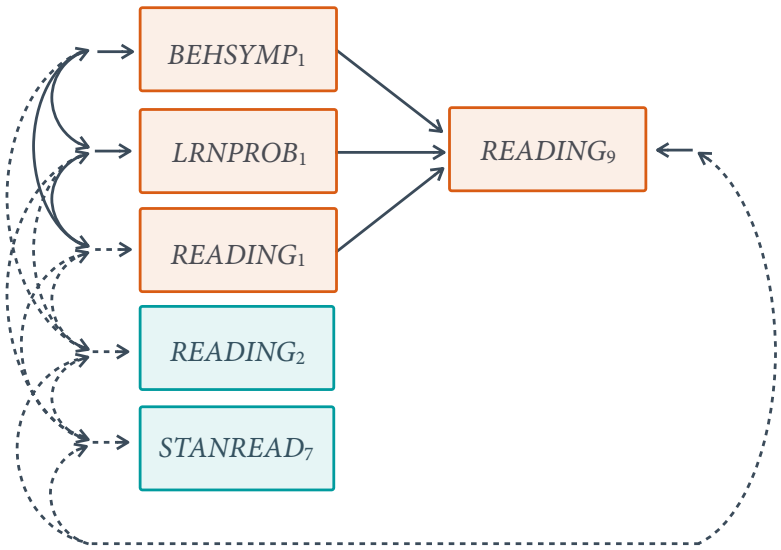


FIGURE 6. Solid lines denote the focal regression model parameters, and dashed lines are auxiliary variable parameters. The path diagram depicts a saturated correlates model that connects the analysis variables to the auxiliary variables via correlated residuals.

Example 1 from the Software Tutorials document provides annotated syntax and output files for the analysis. The top panel of Table 4 shows the parameter estimates, standard errors, and significance tests from the analysis. The results are interpreted in the same way as a complete-data regression analysis. For example, consider the first grade reading slope. The model predicts that two individuals who differ by one point in first grade should differ by 0.50 points on the outcome, holding constant teacher-rated learning and behavioral problems. The corresponding test statistic indicates that the slope coefficient is statistically different from zero ($z = 12.00$, $p < .001$). Because they are not of substantive interest, the table omits the auxiliary variable parameters. As a comparison, the bottom panel also shows the estimates from a fully univariate factored regression specification. The flowchart in Figure 3 indicates that either specification is appropriate, and the close correspondence of the two sets of results is consistent with this conclusion.

TABLE 4. Maximum Likelihood Estimates From a Multiple Regression

Effect	Est.	Std. Error	*z*	*p*	2.5% LCL	97.5% UCL
Multivariate Normal Model						
Intercept	66.03	5.85	10.35	< .001	54.57	77.49
$READ_1$	0.50	0.04	12.00	< .001	0.42	0.59
$LRNPROB_1$	–0.25	0.12	–2.14	.03	–0.47	–0.02
$BEHSYMP_1$	–0.18	0.10	–1.79	.07	–0.38	0.02
R-square	.59	--	--	--	--	--
Factored Regression Specification						
Intercept	65.72	5.89	11.16	< .001	54.18	77.26
$READ_1$	0.51	0.04	12.11	< .001	0.43	0.59
$LRNPROB_1$	–0.26	0.12	–2.22	.03	–0.49	–0.03
$BEHSYMP_1$	–0.18	0.10	–1.73	.08	–0.38	0.02
R-square	.59	--	--	--	--	--

3.5 Maximum Likelihood Analysis Example 2

The second analysis example illustrates a moderated regression model with an incomplete interaction effect and an incomplete binary covariate. The analysis uses the same data as the first example. The Software Tutorials document provides additional information about this data set, and Table 5 below shows the specific variables for this analysis example.

The analysis model featured ninth grade broad reading scores regressed on academic and behavioral measures collected in first grade: reading achievement, teacher-rated learning problems, the product of first grade reading scores and learning problems, and a dummy code indicating whether a student is considered at risk for developing an emotional or behavioral disorder.

$$READ_9 = \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) + \beta_3(READ_1)(LRNPROB_1) + \beta_4(ATRISK) + \varepsilon \quad (12)$$

TABLE 5. Variables for the Moderated Regression Analysis

Name	Definition	Missing	Scale
	Focal Analysis Variables		
ATRISK	Emotion/behavior disorder risk	2.2%	0 = Low risk, 1 = At risk
$LRNPROB_1$	1st grade learning problems	2.2%	Numeric (31 to 88)
$READ_1$	1st grade broad reading	6.5%	Numeric (39 to 153)
$READ_9$	9th grade broad reading	17.4%	Numeric (41 to 123)
	Auxiliary Variables		
$READ_2$	2nd grade broad reading	9.4%	Numeric (20 to 150)
$STANREAD_7$	7th grade standardized reading	19.6%	Numeric (100 to 399)

The analysis also included second grade broad reading scores and seventh grade standardized test scores as auxiliary variables. These additional variables were included because they exhibited significant residual covariation with the analysis variables (see Section 1.5).

Applying the flowchart from Figure 3, models with interactive effects require a factored regression specification that expresses the multivariate distribution into a series of simpler distributions. This example uses a fully univariate specification where each variable is the outcome in its own regression model. Using established symbolic notation, the factorization consists of six regression models, each with its own distribution assumption.

$$\begin{gathered} f(STANREAD_7|READ_2, READ_9, READ_1, LRNPROB_1, ATRISK) \times \\ f(READ_2|READ_9, READ_1, LRNPROB_1, ATRISK) \times \\ f(READ_9|READ_1, LRNPROB_1, ATRISK) \times \\ f(READ_1|LRNPROB_1, ATRISK) \times f(LRNPROB_1|ATRISK) \times f(ATRISK) \end{gathered} \tag{13}$$

The first two lines are the regression models for auxiliary variables, and the third line corresponds to the focal regression from Equation 12. The terms on the last line are regressions linking the predictors to one another.

Figure 7 shows a path diagram of the analysis model. The solid lines highlight the focal model regression slopes, with the arrow pointing from learning problems to the first grade reading test slope denoting the interaction coefficient. The equation and path diagram illustrate a cascading pattern where the binary risk indicator predicts first grade learning problems, both variables predict first grade broad reading performance, and all three regressors predict ninth grade reading

achievement (the focal outcome). The auxiliary variables enter the model as additional outcomes that are predicted by the analysis variables and by each other. The binary risk indicator required an empty logistic model, and all other models were linear regressions with normal errors.

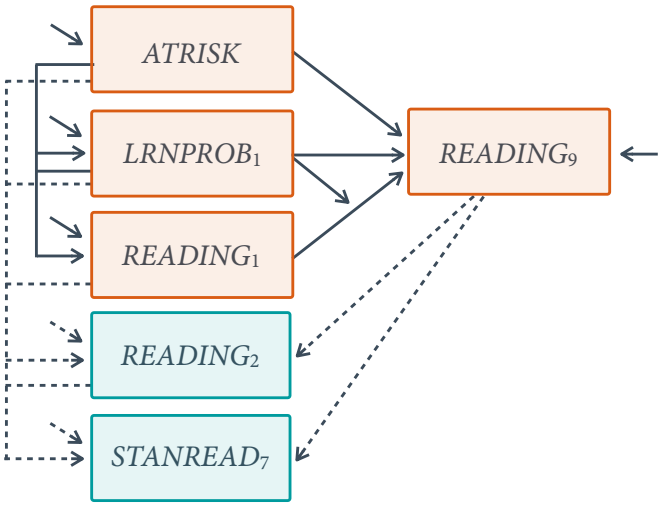


FIGURE 7. Solid lines denote the focal regression model parameters, and dashed lines are auxiliary variable parameters. The path diagram depicts a model where auxiliary variables function as additional outcomes. All analysis variables predict the auxiliary variables, and second grading reading also predicts seventh grade standardized test scores.

Example 4 from the Software Tutorials document provides annotated syntax and output files for the analysis. Table 6 displays the parameter estimates, standard errors, and significance tests from the analysis. The lower-order terms in a moderated regression are conditional effects that depend on scaling or centering. To facilitate interpretation, the analysis centered the interacting variables at the maximum likelihood estimates of their grand means. Centering defines the first grade reading slope ($\hat{\beta}_1 = 0.51$) as a conditional effect at the mean of the learning problems distribution, and the learning problems slope ($\hat{\beta}_2 = -0.38$) similarly reflects a conditional effect at the reading achievement average. The interaction slope captures the change in the first grade reading slope for each one-unit increase in learning problems (and vice versa). The positive coefficient ($\hat{\beta}_3 = 0.013$) indicates that the association between first and ninth grade reading scores becomes stronger (i.e., more positive) for students with elevated learning problems.

TABLE 6. Maximum Likelihood Estimates From a Moderated Regression

Effect	Est.	Std. Error	*t*	*p*	2.5% LCL	97.5% UCL
Intercept	89.05	1.42	62.76	< .001	86.27	91.83
$READ_1$	0.51	0.04	11.55	< .001	0.42	0.59
$LRNPROB_1$	–0.38	0.08	–4.54	< .001	–0.54	–0.21
$READ_1 \times LRNPROB_1$	0.013	0.005	2.83	.005	0.003	0.02
ATRISK	–1.91	1.79	–1.06	.29	–5.43	1.61
R-square	.62	--	--	--	--	--

Note. First grade reading scores and learning problems ratings are centered at the grand mean.

4

Bayesian MCMC Estimation

Bayesian Markov chain Monte Carlo (MCMC) estimation is the second member of the Big Three. Maximum likelihood and Bayesian estimation are similar in the sense that a researcher fits their desired model to the incomplete data, and the software returns parameter estimates and measures of uncertainty that assume a conditionally missing at random process. When confronted with missing values, maximum likelihood uses the normal distribution to deduce the missing parts of the data as it iterates to a solution (more precisely, the estimator marginalizes or averages over a distribution of plausible scores for each person). In contrast, Bayesian estimation algorithms iteratively impute missing values as they update the parameters. The resulting parameter summaries average over numerous realizations of the missing values. Like maximum likelihood, missing data handling is integrated into estimation, and the primary goal is to estimate model parameters.

Bayesian missing data handling routines for multivariate normal data have a long history in the literature (Schafer, 1997, 2001; Schafer & Graham, 2002; Schafer & Olsen, 1998). The multivariate normal model has since evolved into a powerful data analytic framework that mimics the classic FIML estimator in scope (Muthén & Asparouhov, 2012). Long-established methods for factored regression specifications (Huang et al., 2005; Ibrahim et al., 2002) have similarly evolved. Among other things, Bayesian factored regression specifications extend to moderated and curvilinear regression models (Asparouhov & Muthén, 2021a; Keller & Enders, 2023; Lee et al., 2007; Lüdtke et al., 2020b; Zhang & Wang, 2017), models with discrete and nonnormal variables (Asparouhov & Muthén, 2021b; Lee & Mitra, 2016; Lüdtke et al., 2020b), latent variable models (Aroian et al., 1978; Keller & Enders, 2021; Lee & Shi, 2000; Lüdtke et al., 2020b; Merkle & Rosseel, 2018; Palomo et al., 2007), models for missing not at random processes (Du et al., 2021), models with auxiliary variables (Daniels et al., 2014; Lüdtke et al., 2020b), multilevel models (Enders et al., 2020; Erler et al., 2019; Erler et al., 2016; Goldstein et al., 2014; Grund et al., 2021a), and models featuring scale scores with missing item responses (Alacam et al., 2023). The development of

Bayesian missing data handling procedures has arguably outpaced that of maximum likelihood, as these applications exceed what is currently possible with likelihood-based estimation. The ensuing discussion focuses strictly on factored regression specifications.

4.1 The Bayesian Paradigm

Maximum likelihood estimation is fundamentally situated in a frequentist paradigm that defines model parameters as fixed quantities in the population. In this framework, each unique sample of a given size yields a different estimate of an unknown parameter, and a sampling distribution describes the hypothetical behavior of estimates from different random samples. The standard error quantifies the expected difference between an estimate and its true parameter (i.e., the standard deviation of the *estimates*).

In contrast, the Bayesian paradigm defines a parameter as an unknown variable rather than a constant. Through this lens, some realizations of a parameter are more likely to have produced our data than others. Probability distributions called posteriors encode knowledge about a model's parameters after analyzing the data. For example, after fitting a regression model, a normal curve characterizes slope parameters that are consistent with the data. The center of this distribution (the posterior mean or median) quantifies the most likely parameter given the data, and its spread (the posterior standard deviation) quantifies uncertainty. These summary statistics are often numerically equivalent to maximum likelihood point estimates and standard errors, respectively, but their interpretations do not reference hypothetical estimates from other samples.

Notice that the frequentist and Bayesian frameworks reverse what varies and what is fixed. In the frequentist paradigm, estimates from different samples vary around a fixed parameter. In the Bayesian framework, evidence about different parameter values varies and the sample data are fixed. Bayes' theorem is the mathematical device that produces this important reversal. The theorem's composition is as follows.

$$(\text{varying parameter} \mid \text{fixed data}) = \frac{\text{prior} \times (\text{varying data} \mid \text{fixed parameter})}{\text{scaling factor}} \tag{14}$$

The left term is the posterior distribution, and the numerator of the right side is the product of a prior distribution and a likelihood function. The prior is a probability distribution that conveys expectations about a parameter before analyzing the data. This information could come from a pilot study or meta-analysis, but most Bayesian applications use off-the-shelf diffuse (noninformative) priors that contain as little information as possible. Finally, the denominator on the right side is an unnecessary scaling term that does not depend on the parameter.

Adopting diffuse prior distributions that convey little or no information is straightforward for estimands like means and coefficients. For example, the normal curve is a common prior for regression coefficients. Centering the curve at zero and specifying a massive standard deviation effectively produces a flat function over all slope values that could have produced the data. Invoking a prior distribution where every parameter value is equally likely effectively nullifies the influence of the prior, such that the data alone determine the posterior's shape. In contrast, off-the-shelf priors for variance and covariance parameters usually impart information that *could* influence the results. For example, common prior distributions for variance parameters tend to assign higher a priori weights to values close to zero. Fortunately, information from the data overwhelms the prior as the sample size increases, so the impact of the prior distribution is often negligible in practice.

From a practical perspective, researchers should be most concerned about prior distributions when estimating variance and covariance parameters in small samples. Multilevel analyses are a context where prior distributions can matter. In education applications, multilevel data sets often feature many students at level-1 nested in a small number of schools at level-2. In such applications, the choice of prior distribution for the variance–covariance matrix of the school-level random effects is potentially impactful (Gelman, 2006). In practice, there is no way to know which prior is best for a given situation. When in doubt, you can perform a sensitivity analysis that examines whether the choice of prior meaningfully impacts the results. The ensuing data analysis examples demonstrate this idea.

4.2 Markov Chain Monte Carlo (MCMC) Algorithms

In most cases, the posterior distribution of the parameters on the left side of Equation 14 is a complicated multivariate function. Bayesian estimation leverages iterative MCMC algorithms that break a complex problem involving multiple parameters and missing values into separate, simpler steps. Each step estimates one unknown at a time, treating the current values of all other quantities as constants. To illustrate, consider a simplified version of an earlier analysis example where ninth grade broad reading scores are regressed on first grade reading. A factored regression specification consists of two submodels: one for the focal regression and a secondary model for the incomplete predictor. Figure 8 shows a schematic of the MCMC algorithm for the analysis. A single MCMC iteration consists of two broad steps. Using the filled-in data from the previous iteration, the algorithm first estimates the parameters of each submodel. Parameters can be estimated individually or in blocks of like terms (e.g., each model's coefficients can be estimated in a single

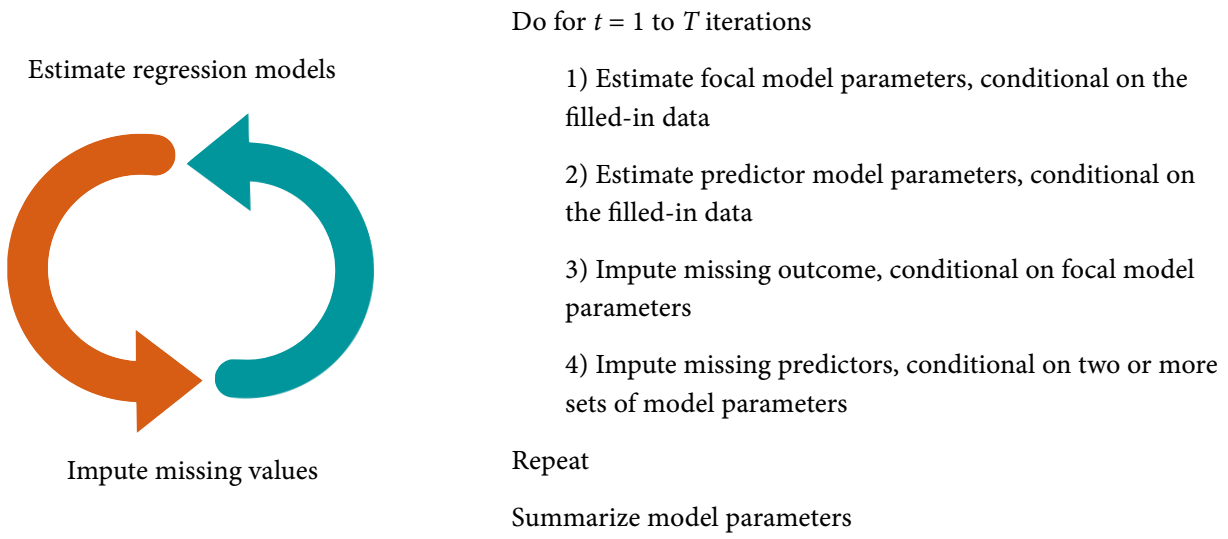


FIGURE 8. Schematic of the MCMC algorithm for the analysis. The algorithm first estimates the parameters of each submodel using the filled-in data from the previous iteration. Having updated the parameter values, the algorithm then draws imputations from distributions based on the current parameters.

step). Having updated the parameter values, the algorithm then samples replacement scores (imputations) from distributions based on the current parameters.

Maximum likelihood applications deploy iterative algorithms that successively and deterministically adjust parameters with the goal of minimizing the sum of squared, standardized residuals. The algorithm converges when estimates no longer change from one iteration to the next, at which point the resulting estimates maximize data–model fit. In contrast, MCMC uses Monte Carlo computer simulation (the second "MC" in MCMC) to sample plausible parameter values at random from a distribution. For example, a normal distribution usually generates regression coefficients, and a right-skewed inverse gamma distribution often generates variance estimates. Conceptually, the process of simulating new parameter values can be understood as computing predicted parameters from the filled-in data then adding random noise (residuals) to each. For example, to update the intercept and slope from a linear regression, the MCMC algorithm uses OLS point estimates and standard errors to define the predicted values and spread of the random noise terms, respectively. Adding computer-generated normal residuals to the predicted coefficients gives updated parameter values.

To illustrate MCMC estimation more concretely, Figure 9 shows line graphs (trace plots) of the bivariate regression model parameters from 50 MCMC cycles. Notice that the parameters continually oscillate in a random pattern from one iteration to the next, and they never converge

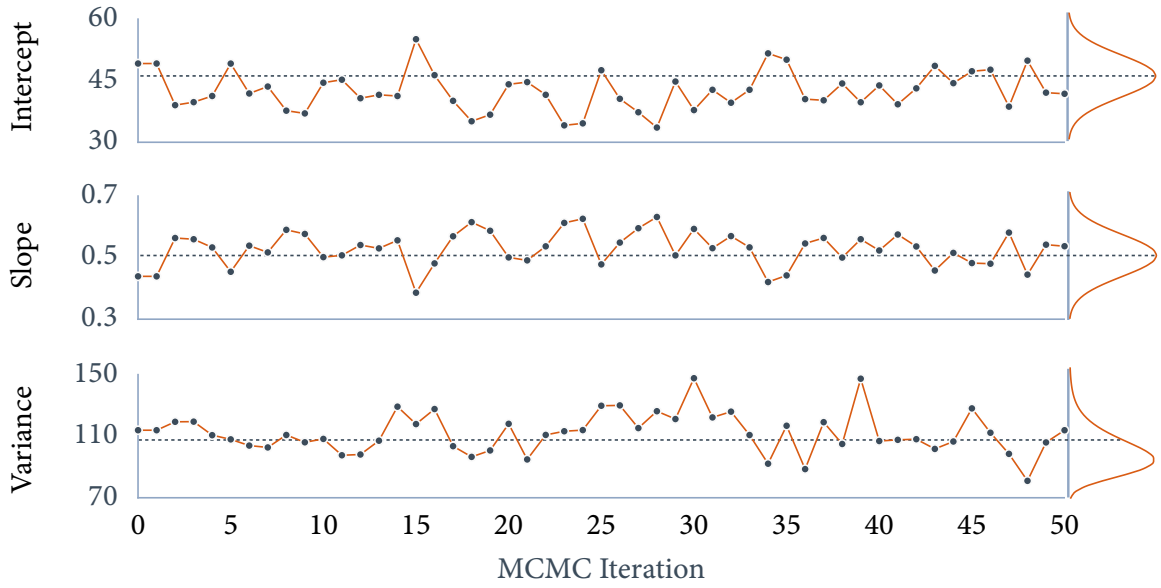


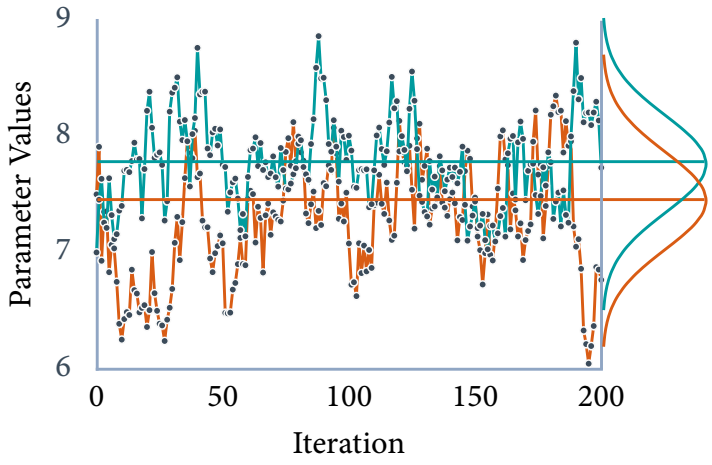
FIGURE 9. Trace plots of bivariate regression model parameters from 50 MCMC cycles. Because they are sampled at random from distributions, parameter values continually oscillate in a random pattern from one iteration to the next. The posterior distributions on the right side of the graph depict the accumulation of parameter values across many computational cycles.

on a single, fixed value. The distributions on the right side of the graph depict the accumulation of parameter values across many computational cycles (the posterior distributions), and the dashed horizontal lines are the Bayesian point estimates (posterior means or medians).

Whereas maximum likelihood algorithms converge when estimates no longer change across successive iterations, MCMC estimation converges when the iterative algorithm generates parameter values that form a stable distribution. Practically speaking, convergence is achieved when accumulating additional iterations does not change the mean and variance of the simulated parameter values. The potential scale reduction factor (PSRF; Gelman & Rubin, 1992) is a popular index that uses ANOVA mean squares expressions to compare parameter distributions generated from two unique MCMC processes (chains). To illustrate, the trace plots in Figure 10 show 200 parameter estimates (the small black dots) from a pair of MCMC chains with different random starting values. Figure 10a is consistent with an MCMC process that has not converged because two independent chains produce parameter values with different means. In contrast, the trace plot in Figure 10b is consistent with a process that has converged because the distributions have similar center and spread. Numerically, PSRF values are scaled such that values closer to one (e.g., PSRF < 1.05) reflect convergence. Keller and Enders (2021, Chapter 3) provide a detailed illustration of

MCMC convergence diagnostics, and the Software Tutorials document similarly uses the PSRF for this purpose.

a) MCMC has not converged (PSRF > 1.05)



b) MCMC has converged (PSRF < 1.05)

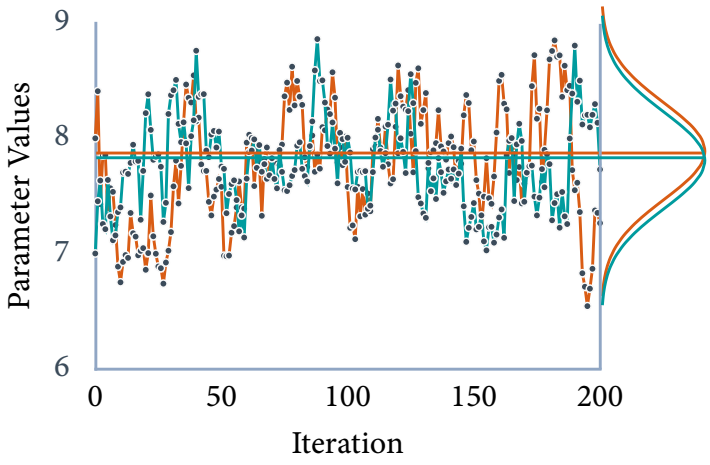


FIGURE 10. Panel (a) depicts an MCMC algorithm that has not converged because the mean parameter values from two independent processes are very different. Panel (b) has converged because the means are very similar.

4.3 Missing Data Imputation

Returning to Figure 8, each MCMC cycle samples plausible replacement scores from distributions based on the current parameter values. The missing data imputation step warrants discussion because it is common to Bayesian analyses and to the multiple imputation procedures discussed in Section 5. As you will see, the imputation phase of a multiple imputation analysis coopts MCMC estimation to create and save a small collection filled-in data sets that are subsequently reanalyzed using frequentist inference. In a Bayesian analysis, the imputations play a supporting role behind the scenes, as the goal is to obtain parameter summaries that average over thousands of realizations of the missing data.

To illustrate MCMC's imputation step, consider a simple regression model where first grade reading predicts ninth grade performance. Both variables have missing data. MCMC uses a factored specification that expresses the bivariate distribution as a pair of univariate distributions. Both are normal in this example, but they need not be. The symbolic expression for the factorization is shown below with the corresponding regression models.

$$\begin{aligned} f(READ_9, READ_1) &= f(READ_9|READ_1) \times f(READ_1) \\ READ_9 &= \beta_0 + \beta_1(READ_1) + \varepsilon \\ READ_1 &= \gamma_0 + \epsilon \end{aligned} \tag{15}$$

Symbolically, $f(READ_9|READ_1)$ says that the conditional distribution of ninth grade scores depends on first grade reading (i.e., ninth grade scores are normally distributed around predicted values), and $f(READ_1)$ is the marginal distribution for the first grade scores alone. The factorization is central to imputation because it dictates which models contribute to a variable's distribution of missing values. The focal regression solely determines ninth grade imputations because the incomplete outcome appears in only one model. In contrast, first grade scores appear in two equations, so both models contribute to imputation.

To illustrate imputation, Figure 11a shows the distribution of plausible ninth grade reading imputations at three levels of first grade performance. The solid black dots on the regression line are predicted ninth grade scores, and the white circles are plausible imputations. By assumption, scores are normally distributed around the regression line, and the residual standard deviation from the regression model (estimated in the first part of each MCMC cycle) dictates the spread of the three normal curves. Candidate imputations fall exactly on vertical lines, but I added horizontal jitter to enhance visual clarity. The MCMC algorithm generates an imputation for each person by randomly selecting a value from their candidate distribution (technically, the algorithm draws from a full distribution of replacement scores, not just those displayed in the graph). Figure 11b

shows three imputations drawn at random from these distributions. The straight arrows pointing from the predicted values to imputed scores are residuals, which MCMC "samples" by simulating random numbers from a normal curve. Conceptually, each imputation is a predicted value plus a random noise term.

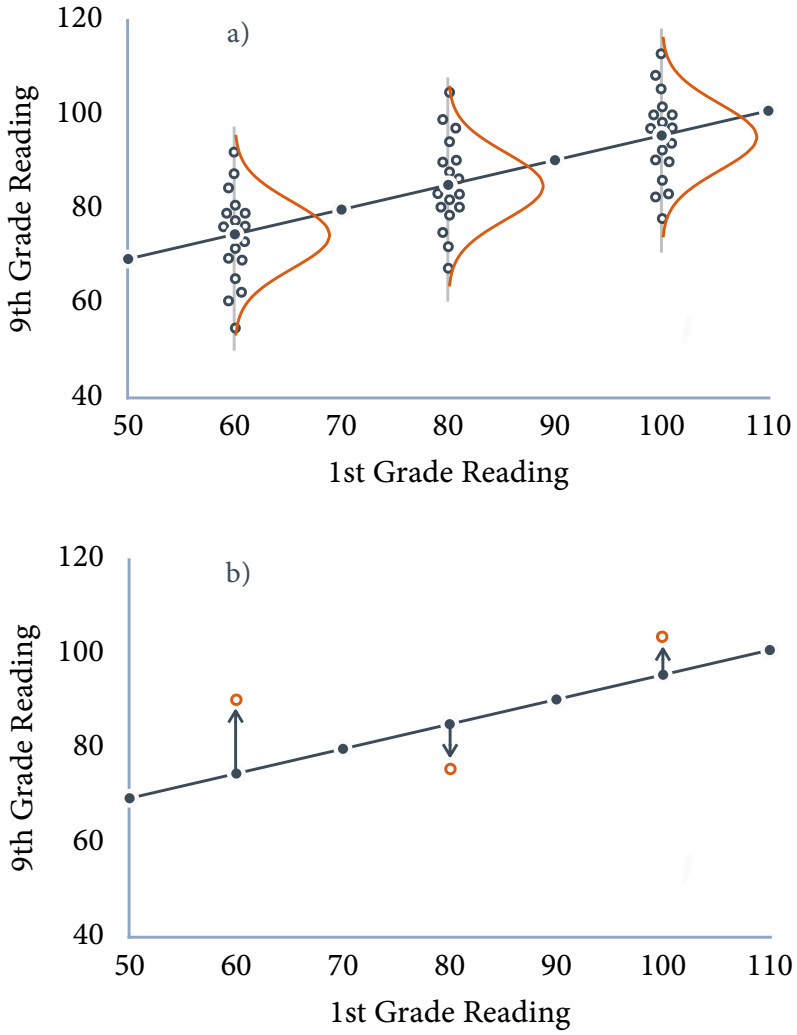


FIGURE 11. Panel (a) shows distributions of ninth grade imputations at three levels of first grade reading. Panel (b) shows three imputations drawn at random from each distribution. Each imputation can be viewed as the sum of a predicted value (the black dots) and a normal residual (arrows).

Returning to Equation 15, the first grade reading variable appears as a predictor in the focal regression and an outcome in its own supporting model. MCMC again samples replacement scores at random from a distribution of plausible values. To illustrate, Figure 12a shows the distribution of first grade reading imputations at three levels of ninth grade performance. The solid black dots represent predicted values, and the white circles are plausible imputations. The predicted scores no longer fall on the regression line, and the distributions of imputations are located on horizontal

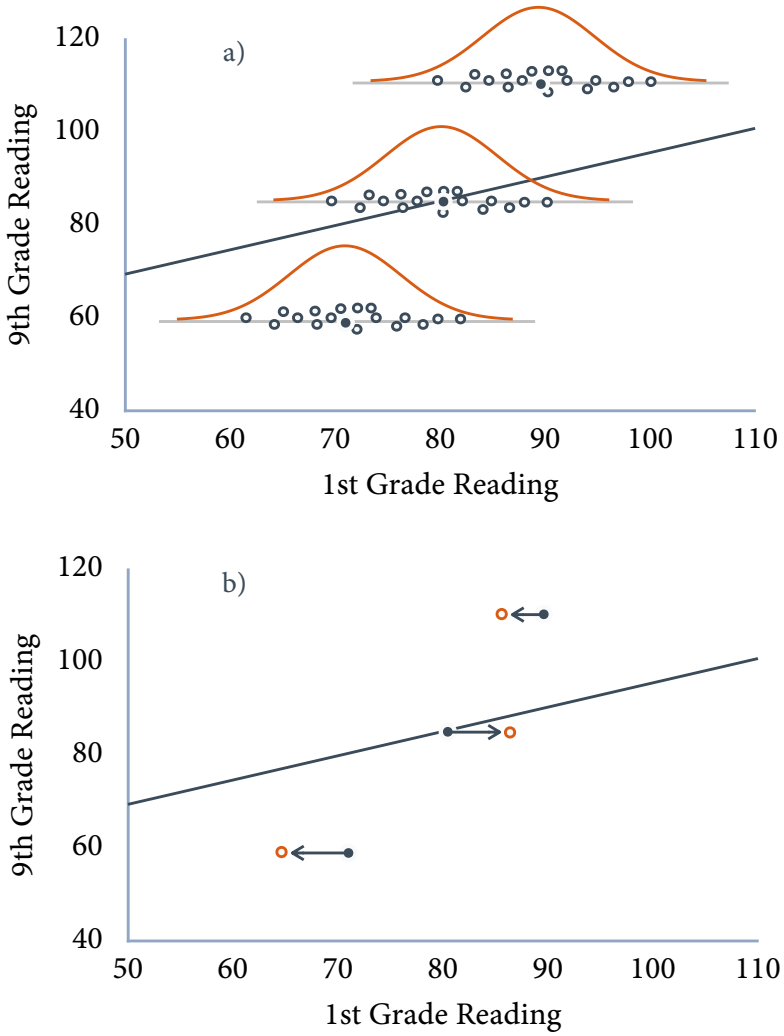


FIGURE 12. Panel (a) shows distributions of first grade imputations at three levels of ninth grade reading. Panel (b) shows three imputations drawn at random from each distribution. Each imputation can be viewed as the sum of a predicted value (the black dots) and a normal residual (arrows).

slices in the bivariate coordinate system. Because the first grade scores appear in two models, the predicted values and spread of the distributions are now complex functions of two sets of model parameters (Enders, 2022; Eq. 5.12). This technical detail aside, the composition of the filled-in data points is the same—each imputation is the sum of a predicted value and random noise term. To emphasize this point, Figure 12b shows three imputations drawn at random. As before, the straight arrows pointing from the predicted values to imputed data points are residuals, which MCMC "samples" by simulating random numbers from a normal curve. After imputing all missing data points, the MCMC algorithm forwards the filled-in data to the next iteration, where the model parameters are estimated again from the new data.

4.4 Fine-Tuning MCMC: Auxiliary Variables and Nonnormal Data

The conditionally missing at random assumption described in Section 1.2 requires that the unseen score values contain no unique information about missingness beyond that contained in the observed data. Practically speaking, this assumption requires that all important determinants of missingness are contained within a model's observed data. Adding extraneous auxiliary variables can help satisfy this assumption, minimizing the risk of nonresponse bias (Collins et al., 2001; Rubin, 1996). Like a maximum likelihood analysis, the goal is incorporate the additional variables while maintaining the structure of the model that we would have fit *had there been no missing data*.

A straightforward way to leverage auxiliary variables is to treat them as additional outcomes that are predicted by the analysis variables and by each other. Figure 5b depicts this specification, with dashed lines depicting the auxiliary variable regression slopes. With this approach, the focal analysis model is embedded in larger network of variables. The additional regressions do not align with substantive theory and need not reflect a logical causal order. Rather, they are simply a tool for linking the analysis variables to the auxiliary variables. This strategy is simple to implement, and it readily accommodates mixtures of numeric and discrete auxiliary variables. The ensuing analysis examples illustrate this additional outcome approach, as do the examples in the Software Tutorials document.

Turning to nonnormal data, Bayesian estimation readily accommodates a variety of variable metrics, and incomplete variables need not be continuous or normal. Estimators for discrete variables are broadly available in software programs, and these facilities include support for binary, ordinal, multicategorical, and count variables (Albert & Chib, 1993; Asparouhov & Muthén, 2021b; Keller & Enders, 2021; Polson et al., 2013). For continuous variables, the statistical justification for robust (sandwich estimator) standard errors and test statistics does not extend naturally to the Bayesian framework because it presupposes a misspecified model (Bayesian

estimation assumes that the fitted model and its corresponding distribution is correctly specified). Some authors argue that Bayesian analyses are robust in the sense that the posterior distributions of the model parameters naturally reflect features of the data. For example, 95% credible intervals do not invoke large-sample (asymptotic) arguments, and the interval limits may be asymmetric around a point estimate.

In lieu of robust corrections, researchers routinely apply normalizing transformations like taking the natural logarithm of a variable. One of the challenges with transformations is selecting one that is appropriate for the data's shape. The Yeo–Johnson normalizing transformation (Yeo & Johnson, 2000) is a promising option that is gaining traction in the literature (Keller & Enders, 2023; Lüdtke et al., 2020b). Briefly, the procedure envisions a nonnormal variable Y that links to a normalized variable Y^* via a shape parameter. In the context of a regression model, the normalized variable serves as the outcome, and the MCMC algorithm iteratively estimates a shape parameter that is appropriate for the observed data's shape. The Yeo–Johnson procedure is attractive because it subsumes several common functions, including inverse, logarithmic, square root, and Box–Cox transformations. Moreover, the procedure accommodates negative score values and positively or negatively skewed distributions. The transformation has shown promise for imputing nonnormal missing data (Lüdtke et al., 2020b), and it is readily available in statistical software (Keller & Enders, 2021; Robitzsch & Lüdke, 2023).

4.5 Bayesian Analysis Example 1

The first Bayesian analysis example illustrates a multiple regression model comprised of incomplete continuous variables. The analysis uses the `behaviorachievement.dat` data set from a longitudinal study that followed 138 students from primary through middle school. The file includes three annual assessments of broad reading and math achievement beginning in the first grade, seventh grade standardized achievement test scores taken from a statewide assessment, and a final measure of broad reading and math obtained in ninth grade. The data also contain teacher ratings of behavioral symptoms and learning problems were also obtained in the first grade. The Software Tutorials document provides additional information about this data set. Table 2 from Section 1.5 shows the specific variables for this analysis example.

The analysis model featured ninth grade broad reading scores regressed on three academic and behavioral measures collected in first grade: the broad reading composite, teacher-rated learning problems, and teacher-rated behavioral problems.

$$READ_9 = \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) + \beta_3(BEHSYMP_1^*) + \varepsilon \tag{16}$$

The analysis also included second grade broad reading scores and seventh grade standardized test scores as auxiliary variables. These additional variables were included because they exhibited significant residual covariation with the analysis variables (see Section 1.5).

Applying the flowchart from Figure 3, either a multivariate or factored regression specification is appropriate. I strictly use the latter for the Bayesian analyses. Following Figure 2b, the focal variables were represented by two submodels, each with distinct distributional assumptions. For completeness, the symbolic representation for this two-part factorization is as follows.

$$f(READ_9|READ_1, LRNPROB_1, BEHPROB_1) \times f(READ_1, LRNPROB_1, BEHPROB_1) \quad (17)$$

The first term corresponds to the univariate regression model from Equation 16, and the second term represents a multivariate normal distribution for the predictors. The Blimp application used in the Software Tutorials document automatically configures the predictor distributions; the user simply needs to specify the focal regression model. Finally, auxiliary variables enter the model as additional outcomes that are predicted by the analysis variables and by each other. This additional

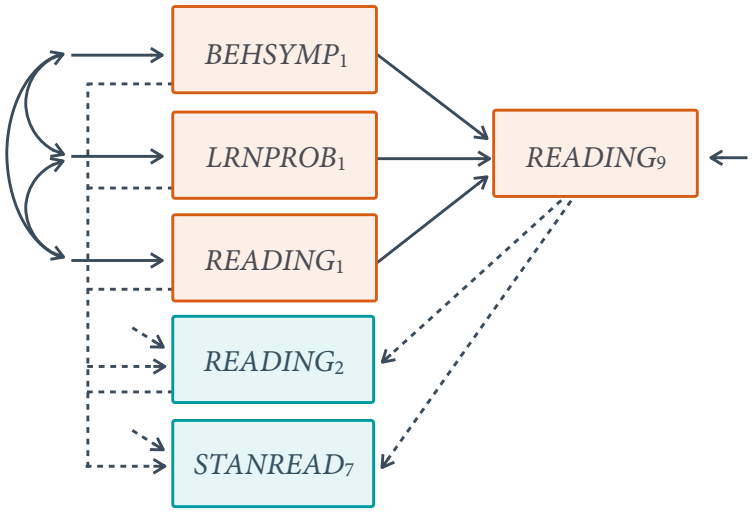


FIGURE 13. Solid lines denote the focal regression model parameters, and dashed lines are auxiliary variable parameters. The path diagram depicts a model where auxiliary variables function as additional outcomes. All analysis variables predict the auxiliary variables, and second grading reading also predicts seventh grade standardized test scores.

part of the factorization consists of two linear regression models, the symbolic notation for which is below.

$$f(STANREAD_7|READ_2, READ_1, LRNPROB_1, BEHPROB_1) \times$$
$$f(READ_2|READ_9, READ_1, LRNPROB_1, BEHPROB_1) \quad (18)$$

Figure 13 shows a path diagram of the analysis model, with dashed lines denoting the auxiliary variable regressions.

Example 6 from the Software Tutorials document provides annotated syntax and output files for the analysis. Table 7 shows the Bayesian parameter summaries. The estimate (posterior median) and standard deviation columns describe the center and spread of the posterior distributions. Although they make no reference to drawing repeated samples, these quantities are analogous to frequentist point estimates and standard errors. The 95% credible interval columns give ranges that capture 95% of each parameter's distribution. These are akin to confidence intervals, but they describe a range of likely parameter values rather than the long-run behavior of intervals from different random samples. Importantly, the numeric summaries in Table 7 are effectively identical to the corresponding maximum likelihood estimates from Table 4. This is exactly what you would expect when applying Big Three approaches that use the same data and invoke the same assumptions.

TABLE 7. Bayesian Parameter Summaries From a Multiple Regression

Effect	Est.	Std. Dev.	2.5% LCL	97.5% UCL	χ^2	*p*
Intercept	66.01	6.05	53.94	77.80	118.94	< .001
$READ_1$	0.50	0.04	0.42	0.59	134.29	< .001
$LRNPROB_1$	–0.25	0.12	–0.48	–0.01	4.23	.04
$BEHSYMP_1$	–0.18	0.11	–0.39	0.03	3.00	.08
R-square	.60	.05	.49	.68	--	--

To illustrate the interpretations, Figure 14 shows the posterior distributions of the parameter values from 10,000 MCMC estimation cycles. The solid vertical lines are the medians, and the shaded regions contain values inside the 95% credible interval limits. To reiterate, the posterior distributions characterize plausible *parameter values* that could have produced these data. From a practical perspective, the posterior summaries have interpretations that parallel any multiple

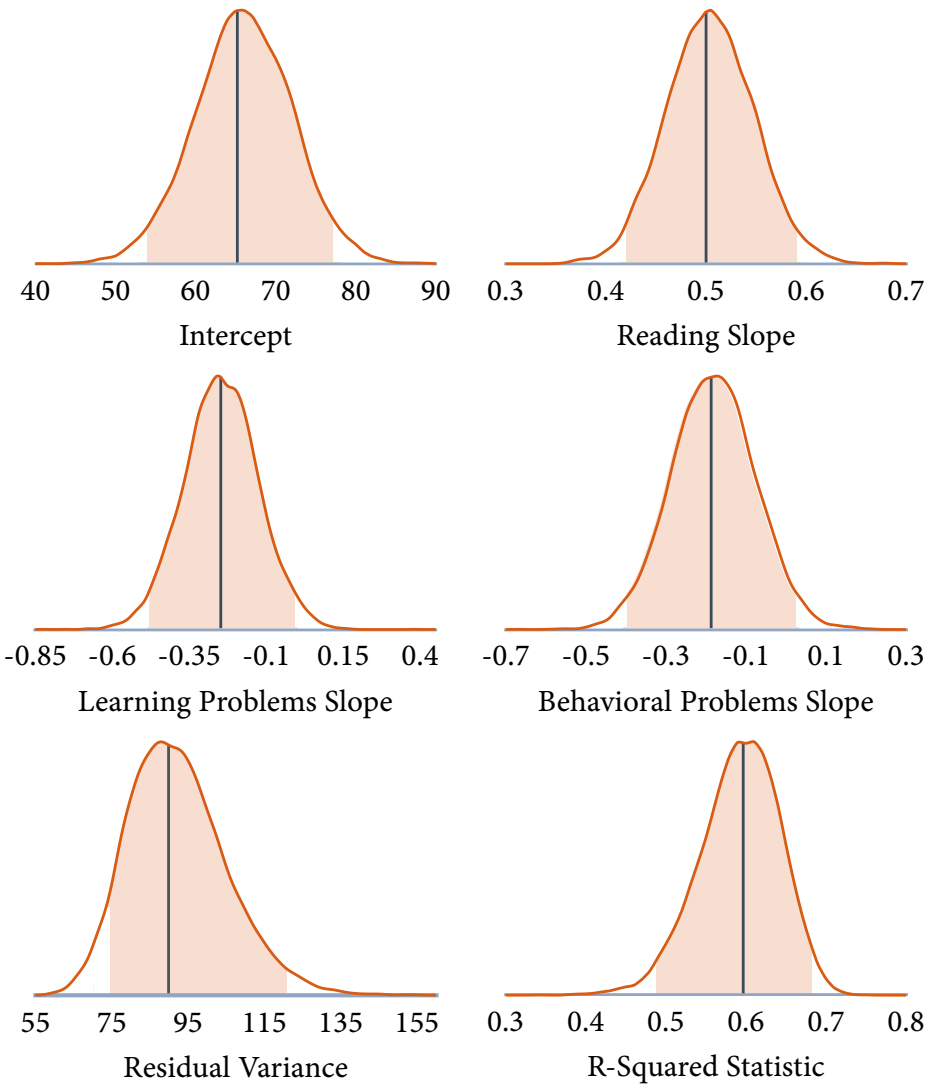


FIGURE 14. Posterior distributions of parameter values from 10,000 MCMC iterations. The solid vertical lines are the medians, and the shaded regions contain values inside the 95% credible interval limits.

regression analysis. For example, consider the first grade reading slope. The coefficient predicts that two individuals who differ by one point in first grade should differ by 0.50 points in ninth grade, holding constant teacher-rated learning and behavioral problems. The posterior standard deviation (*SD* = .04) quantifies uncertainty, much like the corresponding standard error from the maximum likelihood analysis. Finally, the credible interval conveys that there is a .95 probability that the parameter falls between 0.42 to 0.59. To reiterate, the Bayesian summary of the reading

slope is numerically equivalent to its maximum likelihood counterpart from Table 4 ($\hat{\beta}_1 = 0.50$, $SE = .04$, and CI[0.42, 0.59]).

The Bayesian summaries also lend themselves to familiar hypothesis testing logic. Returning to the reading slope, the credible interval limits spanning 0.42 to 0.59 included 95% of the area under the coefficient's posterior distribution. From this, we can conclude that the parameter is statistically different from zero ($p < .05$) because the null value is well outside the 95% interval. That is, the probability that the parameter is less than 0.42 or greater than 0.59 is .05. Alternately, (Asparouhov & Muthén, 2021) proposed a Bayesian Wald chi-square statistic that can be used to evaluate a broad range of hypotheses. For researchers who prefer a frequentist-like test statistic, the rightmost columns of Table 7 show the chi-square test statistics and p-values for the coefficients. These quantities lead to the same conclusions as the 95% credible intervals (e.g., the first grade reading slope is significant with $p < .001$). Levy and McNeish (2023) use the phrase "computational frequentism" to describe applications that use MCMC estimation to approximate frequentist point estimates and test statistics.

The Blimp application invokes flat priors for the coefficients and common off-the-shelf priors for variances and covariances (Keller & Enders, 2021). The prior distributions influence variance parameters by altering the number of independent data points and/or the residual sums of squares at each MCMC cycle. Roughly speaking, alternate prior distributions induce differences that are analogous to those between restricted and full information maximum likelihood (McNeish, 2017; D. McNeish & L. M. Stapleton, 2016) or least squares and full information maximum likelihood (Enders, 2022). The choice of prior distribution is potentially impactful in small samples (McNeish, 2016). To explore the sensitivity of the results to this issue, I fit the model with two alternate prior distributions (also off-the-shelf options that are invoked with a single keyword). Changes to the results were generally inconsequential. At a high level, the R-square summaries were stable across priors (.595 to .604), and differences in the point estimates were generally in the third decimal. Finally, inferences were similarly unaffected by the prior distribution. For example, the first grade reading slope's lower credible limit varied between only 0.421 and 0.423, and the upper limit ranged from 0.589 to 0.591. Considered as a whole, I conclude that the choice of prior distribution did not meaningfully impact on the MCMC results.

4.6 Bayesian Analysis Example 2

The second analysis example illustrates a moderated regression model with an incomplete interaction effect and an incomplete binary covariate. The corresponding maximum likelihood analysis appears in Section 3.5. To refresh, the analysis model features ninth grade broad reading

scores regressed on academic and behavioral measures collected in first grade: reading achievement, teacher-rated learning problems, the product of first grade reading scores and learning problems, and the at-risk dummy code.

$$\begin{aligned} READ_9 = \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) \\ + \beta_3(READ_1)(LRNPROB_1) + \beta_4(ATRISK) + \varepsilon \end{aligned} \tag{19}$$

The analysis also included second grade broad reading scores and seventh grade standardized test scores as auxiliary variables.

Applying the flowchart from Figure 3, models with interactive effects require a factored regression specification that expresses the multivariate distribution into a series of simpler distributions. Following Figure 2b, the focal variables were represented by two submodels, each with distinct distributional assumptions. For completeness, the symbolic representation for this two-part factorization is as follows.

$$f(READ_9|READ_1, LRNPROB_1, ATRISK) \times f(READ_1, LRNPROB_1, ATRISK) \tag{20}$$

The first term corresponds to the univariate regression model from Equation 19, and the second term represents a multivariate normal distribution for the predictors. The Blimp application used in the Software Tutorials document automatically configures the predictor distributions; the user simply needs to specify the focal regression model. In the multivariate distribution, the binary risk indicator appears as a normally distributed latent response variable, consistent with the probit regression framework (Albert & Chib, 1993). The categorical variable model does not impact the variable's role in the focal model, where it appears as a dummy code. Finally, auxiliary variables enter the model as additional outcomes that are predicted by the analysis variables and by each other. This additional part of the factorization consists of two linear regression models, the symbolic notation for which is below.

$$\begin{aligned} f(STANREAD_7|READ_2, READ_1, LRNPROB_1, ATRISK) \times \\ f(READ_2|READ_9, READ_1, LRNPROB_1, ATRISK) \end{aligned} \tag{21}$$

Figure 15 shows a path diagram of the analysis model. The focal and auxiliary variables are color coded, and dashed lines denote the auxiliary variable regressions. The arrow pointing from learning problems to the first grade reading test slope is the interaction coefficient.

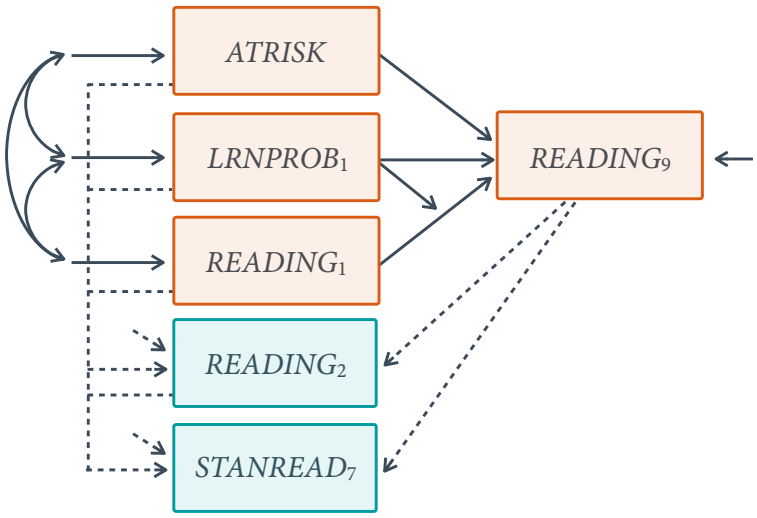


FIGURE 15. Solid lines denote the focal regression model parameters, and dashed lines are auxiliary variable parameters. The path diagram depicts a model where auxiliary variables function as additional outcomes. All analysis variables predict the auxiliary variables, and second grading reading also predicts seventh grade standardized test scores.

Example 10 from the Software Tutorials document provides annotated syntax and output files for the analysis. Table 8 shows the Bayesian parameter summaries. To reiterate, the posterior distributions characterize plausible *parameter values* that could have produced these data. The median of each distribution is the MCMC point estimate, and the standard deviation is the "Bayesian standard error", albeit with no reference to repeated sampling. From a practical perspective, the posterior summaries have interpretations that parallel any moderated regression analysis. To begin, the lower-order terms in a moderated regression are conditional effects that depend on scaling or centering. To facilitate interpretation, the analysis centered the interacting variables at their iteratively-estimated grand means. Centering defines the first grade reading slope ($\beta_1 = 0.50$) as a conditional effect at the mean of the learning problems distribution, and the learning problems slope ($\beta_2 = -0.37$) similarly reflects a conditional effect at the reading achievement average. The interaction slope captures the change in the first grade reading slope for each one-unit increase in learning problems (and vice versa). The positive coefficient ($\beta_3 = 0.012$) indicates that the association between first and ninth grade reading scores becomes stronger (i.e., more positive) for students with elevated learning problems.

TABLE 8. Bayesian Parameter Summaries From a Moderated Regression

Effect	Est.	Std. Dev.	2.5% LCL	97.5% UCL	χ^2	p
Intercept	87.85	1.32	85.19	90.38	4419.07	< .001
$READ_1$	0.50	0.05	0.41	0.59	114.66	< .001
$LRNPROB_1$	–0.37	0.09	–0.55	–0.20	17.09	< .001
$READ_1 \times LRNPROB_1$	0.012	0.005	0.003	0.022	1.08	< .01
ATRISK	–1.93	1.87	–5.67	1.63	6.86	.32
R-square	.61	.05	.50	.70	--	--

Note. First grade reading scores and learning problems ratings are centered at their grand means.

As explained previously, the Bayesian summaries also lend themselves to familiar hypothesis testing logic. Returning to the interaction coefficient, the credible interval limits spanning 0.003 to 0.021 included 95% of the area under the coefficient's posterior distribution. From this, we can conclude that the parameter is statistically different from zero ($p < .05$) because the null value is outside the 95% interval. One could also adopt a computational frequentism perspective that views the MCMC summaries as surrogates for frequentist point estimates and standard errors (Levy & McNeish, 2023). To this end, the rightmost pair of columns in Table 8 show Wald chi-square statistics (squared *z*-tests) and frequentist probability values (Asparouhov & Muthén, 2021).

The Blimp application invokes flat priors for the coefficients and common off-the-shelf priors for variances and covariances (Keller & Enders, 2021). Roughly speaking, alternate prior distributions induce differences that are analogous to those between restricted and full information maximum likelihood (McNeish, 2017; D. McNeish & L. M. Stapleton, 2016) or least squares and full information maximum likelihood (Enders, 2022). To explore the sensitivity of the results to this issue, I fit the model with two alternate prior distributions (also off-the-shelf options that are invoked with a single keyword). Changes to the results were generally inconsequential. At a high level, the *R*-square summaries were stable across priors (.610 to .619), and differences in the point estimates were generally in the third decimal. The binary risk indicator's coefficient exhibited the most noticeable changes, ranging from –1.88 to –1.93. However, these variations are trivial when compared to the parameter's standard deviation or 95% confidence limits. Finally, inferences were similarly unaffected by the prior distribution. For example, the interaction's lower credible limit was consistently 0.003, and the upper limit ranged between 0.021 and 0.022. Considered as a whole, I conclude that the choice of prior distribution did not meaningfully impact on the MCMC results.

5

Multiple Imputation

Maximum likelihood and Bayesian estimation are similar in the sense that a researcher fits a desired model to the incomplete data, and the software returns estimates and measures of uncertainty that assume a conditionally missing at random process. When confronted with missing values, maximum likelihood uses the normal curve to deduce the missing parts of the data as it iterates to a solution, whereas Bayesian MCMC repeatedly fills in the missing data. In both cases, missing data handling happens behind the scenes, and the primary goal is to estimate model parameters. In contrast, multiple imputation prioritizes the imputations, and the goal is to create filled-in data sets for later analysis.

Multiple imputation's history traces 1977, when Donald Rubin proposed the procedure to the Social Security Administration and Census Bureau as a solution for missing survey data[8]. Rubin published his seminal multiple imputation book (Rubin, 1987) a decade later, and Joe Schafer's subsequent text (Schafer, 1997) fully generalized the methodology. In the 40 years or so since its inception, multiple imputation has developed into a large collection of diverse methods; modern texts catalog these various incarnations (Carpenter et al., 2023; van Buuren, 2018). Published applications of multiple imputation abound, and virtually every general-use software program has imputation facilities.

A typical multiple imputation application consists of three steps. In the imputation and analysis phases, the researcher creates multiple copies of the data with different imputations, after which they fit one or more analysis models to each filled-in data set. Analyzing each data separately produces multiple sets of estimates and standard errors. The final pooling phase uses "Rubin's rules" (Little & Rubin, 2020; Rubin, 1987) to combine the imputation-specific results into a single set of frequentist point estimates, standard errors, and test statistics. Figure 16 depicts these three steps.

[8] Rubin's 1977 report is available in a 2004 volume of the *American Statistician*.

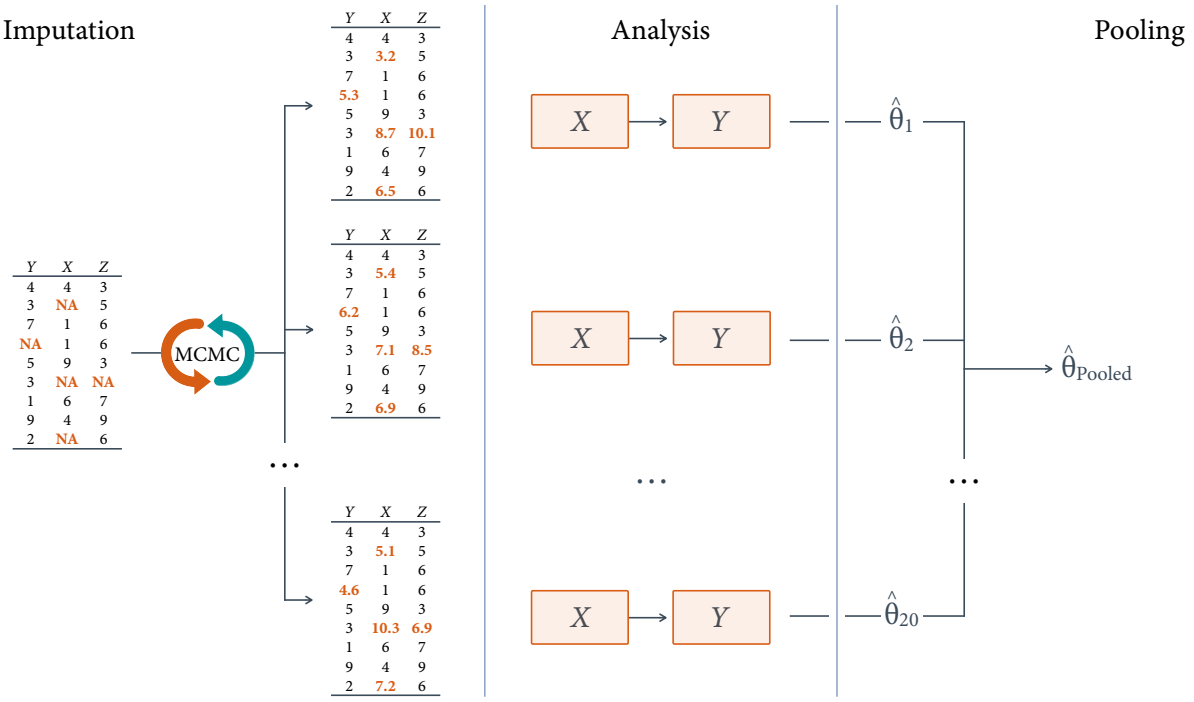


FIGURE 16. The imputation phase creates multiple filled-in data sets that the researcher subsequently analyzes in the analysis phase. The imputation-specific estimates from the second step are subsequently pooled in a single set of frequentist results.

The initial imputation phase coopts Bayesian MCMC estimation described in Section 4.2. Recall that a Bayesian analysis produces parameter summaries that average over thousands of filled-in data sets. In a multiple imputation application, the Bayesian parameter summaries are not the scientific focus. Rather, the main goal is to save a small number of the imputed data sets that MCMC produces (e.g., 20 is an oft-cited recommendation; Graham et al., 2007). These data sets are the inputs for obtaining frequentist points estimates, standard errors, and significance tests.

Multiple imputation involves two rounds of model fitting. The researcher first deploys MCMC to fit an imputation model (usually some type of regression), after which they reanalyze the filled-in data to obtain frequentist inferences. It is useful to classify multiple imputation procedures according to whether the initial MCMC analysis is the same or different from the subsequent frequentist analysis. Borrowing terminology from Enders (2022), a model-agnostic approach deploys an imputation model that differs from the analysis phase model, whereas a model-based procedure uses the same model for imputation and reanalysis. Model-agnostic methods include classic multivariate imputation approaches (Schafer, 1997; Schafer & Olsen, 1998) and fully conditional specification (van Buuren, 2007, 2018), among others. Newer model-based methods include substantive model-compatible imputation (Bartlett et al., 2015), fully Bayesian estimation

(Enders et al., 2020; Zhang & Wang, 2017), and the sequential specification (Erler et al., 2016; Lüdtke et al., 2020b).

To illustrate the key difference between agnostic and model-based imputation strategies, consider a journal paper that requires multiple imputations for *t*-tests and a small number of univariate regression models. A typical application of model-agnostic imputation would create common imputations for all these analyses. In that scenario, the MCMC imputation model must be broad enough to incorporate and preserve *all* associations examined in the paper. Such an imputation model would necessarily differ from the individual analysis models because it includes numerous additional parameters and variables. The imputation model usually represents associations differently too. For example, Schafer's (1997) classic joint model imputation uses a multivariate regression model with complete variables predicting incomplete variables. Alternatively, van Buuren's (2006) fully conditional specification approach (also called the MICE algorithm, for Multiple Imputation by Chained Equations) uses a round robin sequence of univariate regression models where each incomplete variable is predicted by all other variables. Notice that a variable's status as an outcome or predictor in the MCMC imputation model depends on whether it is incomplete or complete. The imputation phase is effectively agnostic about a variable's role in the subsequent analysis phase.

In contrast, a researcher adopting a model-based imputation strategy uses the same model to impute the data as they do to analyze the data. Although the imputation phase may include additional auxiliary variables, the imputation and analysis phase models are otherwise congruent. Crafting imputations to align with a specific analytic model implies that each frequentist analysis requires its own filled-in data sets. Returning to the example, researchers would create unique imputations for each *t*-test and regression analysis in the paper. The MCMC imputation phase for each *t*-test would deploy a regression model with a dummy code predictor, and the imputation phase for each regression analysis would deploy a matching regression model. Note that model-based imputation is analogous to FIML estimation, which also integrates missing data handling on an analysis-by-analysis basis.

How does one decide between model-agnostic and model-based imputation? The composition of the focal analysis model—in particular, whether it includes nonlinear effects such as interactions, polynomial terms, or random coefficients—determines the type of imputation strategy that works best. The flowchart in Figure 17 depicts a decision tree for selecting a multiple imputation procedure. Starting on the left, the first decision point depends on whether the focal analysis features a nonlinear effect. If the answer to this first question is yes, then model-based multiple imputation based on a factored regression specification is the only choice. If no, then various model-agnostic procedures are appropriate depending on the variable metrics. The

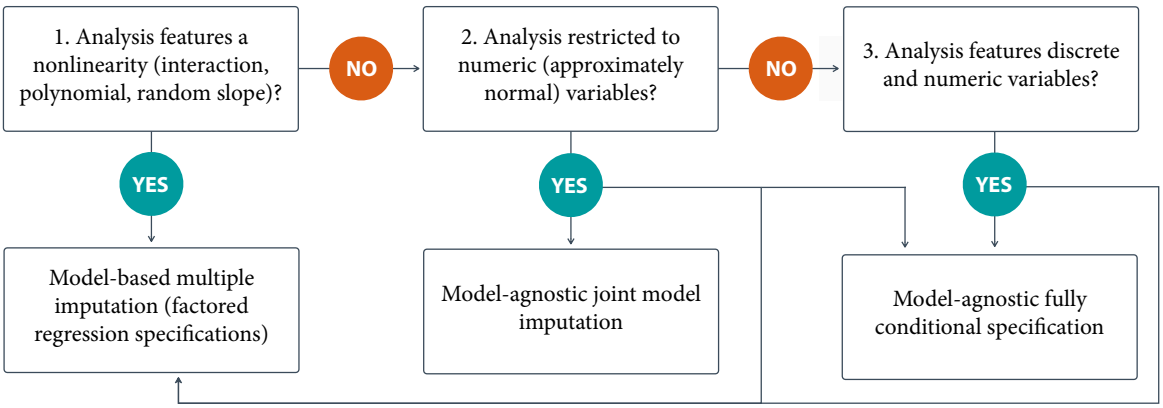


FIGURE 17. Flowchart depicting decision tree for choosing agnostic or model-based multiple imputation. Starting on the left, the first decision point depends on whether the focal analysis features any type of nonlinearity. The second decision is whether all variables share the same metric. Although classic methods are not always appropriate, factored specifications are.

flowchart oversimplifies a nuanced issue, but it provides a heuristic for classifying multiple imputation options. The flowchart also parallels the high-level decision tree for missing data analyses in Figure 3. Finally, it is perfectly fine to use a combination of model-based and model-agnostic procedures within the same project or paper.

Joseph Schafer's classic textbook (Schafer, 1997) popularized agnostic imputation based on the multivariate normal distribution. The MCMC algorithm for his approach repeatedly updates the model parameters—a mean vector and covariance matrix—conditional on the current filled-in data, after which it draws new imputations from normal distributions based on the parameter values. Saving a small number of imputed data sets from a long iterative sequence and performing analyses on the filled-in data sets gives estimates that average over different realizations of the missing values. Accessible descriptions of this classic procedure are widely available in the literature (Schafer, 1999; Schafer & Graham, 2002; Schafer & Olsen, 1998). This paper instead focuses on model-agnostic fully conditional specification because it is more broadly available in computer software programs. Finally, a discussion of model-based imputation is relegated to the analysis example in Section 5.6 because the imputation phase exactly coopts the MCMC algorithm described earlier in Section 4.2—the only new details involve saving and reanalyzing imputations.

5.1 Model-Agnostic Fully Conditional Specification

Fully conditional specification (van Buuren, 2007; van Buuren et al., 2006; van Buuren & Groothuis–Oudshoorn, 2011) imputes variables one at a time using a sequence of univariate

regression models. The composition of these models follows a round robin scheme where each incomplete variable is predicted by all other variables, complete or previously imputed. Importantly, each regression in the sequence is tailored to the incomplete variable's metric, allowing for a potentially diverse collection of generalized linear imputation models. This procedure is also known as the MICE algorithm (Multiple Imputation by Chained Equations) after Stef van Buuren's popular R program (van Buuren & Groothuis–Oudshoorn, 2011). Variations of the procedure are available for imputing latent response variables (Enders, 2022; Grund et al., 2021b; Keller & Enders, 2021), incomplete covariates in moderated regression models (Bartlett & Morris, 2015; Bartlett et al., 2015), multilevel data structures (Enders et al., 2018; van Buuren, 2011), count and zero-inflated variables (Kleinke & Reinecke, 2013), classification and regression trees (Doove et al., 2014; Shah et al., 2014), and regularized regression (Deng et al., 2016; Zhao & Long, 2016), among others. Van Buuren's multiple imputation book (van Buuren, 2018) details this framework and some of its extensions (https://stefvanbuuren.name/fimd/).

To illustrate the basic idea behind fully conditional specification, suppose a researcher wants to obtain descriptive statistics and correlations for three incomplete variables, Y_1 to Y_3, and one complete variable, Z. Recall from an earlier section that MCMC estimation iteratively repeats two broad steps: estimate model parameters using the current filled-in data, then create imputations using the updated parameter values (imputation = predicted value + random noise). The MCMC algorithm for fully conditional specification applies this estimation–imputation sequence to each incomplete variable in turn.

For the trivariate imputation problem, each computational cycle fits three regression models, one per incomplete variable. The imputation regression equations are shown below, and Figure 18 displays these imputation regression models as path diagrams.

$$\begin{aligned} Y_1 &= \gamma_{10} + \gamma_{11}(Y_2^{(t-1)}) + \gamma_{12}(Y_3^{(t-1)}) + \gamma_{13}(Z) + \epsilon_1 \\ Y_2 &= \gamma_{20} + \gamma_{21}(Y_3^{(t-1)}) + \gamma_{22}(Y_1^{(t)}) + \gamma_{23}(Z) + \epsilon_2 \\ Y_3 &= \gamma_{30} + \gamma_{31}(Y_1^{(t)}) + \gamma_{32}(Y_2^{(t)}) + \gamma_{33}(Z) + \epsilon_3 \end{aligned} \tag{22}$$

Notice that the imputation regression equations follow a round robin scheme where each incomplete variable takes a turn as the to-be-imputed outcome, after which it functions as a fully complete predictor in all other equations. Each model invokes estimation and imputation steps that exactly follow Sections 4.2 and 4.3. The only difference is that each MCMC iteration applies these updating steps to multiple models in a sequence. The t and $t - 1$ superscripts indicate whether the filled-in values on the right side of each equation originate from the current or previous iteration, respectively. For simplicity, the imputation models are all linear regressions, but the

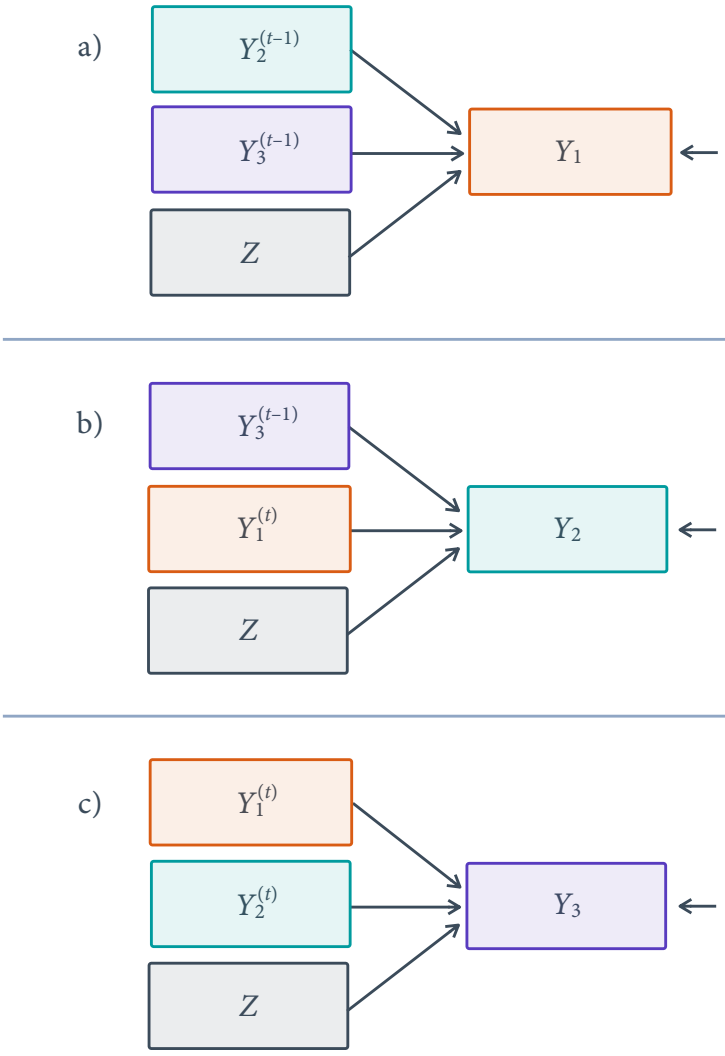


FIGURE 18. Fully conditional specification imputation models with three incomplete variables. Each computational cycle fits three regression models, one per incomplete variable. Panel (a) is Y_1's imputation model, panel (b) is Y_2's regression model, and panel (c) is Y_3's imputation model.

procedure readily accommodates discrete metrics as well. For example, if Y_1 was binary, the first equation would be a logistic or probit regression that spawns dichotomous imputations. Similarly, if Y_1 was multicategorical, the final equation in the sequence would be a multinomial logistic regression.

Section 4.3 described MCMC's imputation procedure, with carries over to fully conditional specification with no modification. For numeric variables, the algorithm uses Monte Carlo computer simulation to "sample" replacement scores from a person-specific normal distribution. A predicted value from a regression equation defines the center of each normal curve, and the model's residual standard deviation defines spread. For categorical variables, logistic or probit imputation models yield predicted probabilities for each discrete response, and MCMC samples replacement scores from a probability mass function that looks like a bar graph. Figure 11 illustrates the imputation = predicted value + noise concept for an incomplete numeric variable. Predictive mean matching is a nonparametric alternative that instead draws imputations from a pool of observed scores taken from people whose predicted values are similar to that of the person with missing data (Kleinke, 2017; Lee & Carlin, 2017; van Buuren, 2018; Vink et al., 2014). Morris et al. (2014) catalog variations of predictive mean matching in software programs, and they provide recommendations about the optimal matching criterion and donor pool size.

Notice that the imputation regression models all differ from the intended frequentist analyses (descriptive statistics and correlations). This is the defining feature of agnostic imputation. Although the imputation and analysis models need not match, it is vital that imputation includes all to-be-analyzed variables and preserves any special features of the frequentist analyses (Collins et al., 2001; Schafer, 2003). In this example, the three regression equations do not invoke assumptions or impose restrictions that contradict the intended frequentist analysis; the imputation phase includes all analysis variables, and the imputation and analysis models both assume linear associations. Suppose the data are hierarchical with students nested in schools. The imputation models from Equation 22 would distort the frequentist point estimates and inference because the fail to incorporate the multilevel structure. Similarly, the imputations would be inappropriate for a moderated regression analysis because the models assume that all interactions equal zero.

Two early strategies that integrated nonlinear terms into a fully conditional specification scheme—just-another-variable imputation and passive imputation—warrant brief discussion due to their propensity for bias. The so-called just-another-variable approach treats product and polynomial terms as standalone, normally distributed variables (von Hippel, 2009). Returning to the variables from Equation 22, suppose the intended frequentist analysis is a moderated regression where Y_1 and Y_2 interact to predict Y_3. The product term—which is incomplete whenever Y_1 or Y_2 is missing—would simply appear as an additional incomplete variable in the imputation scheme. In contrast, passive imputation (also known as impute-then-transform; von Hippel, 2009) would adopt the imputation models from Equation 22, updating the product term by multiplying the imputed Y_1 and Y_2 scores (Seaman et al., 2012). Analytic and computer simulation studies

repeatedly demonstrate that both methods are prone to substantial bias (Cham et al., 2017; Enders et al., 2014; Humberg & Grund, 2022; Seaman et al., 2012; Zhang & Wang, 2017). A newer variation of fully conditional specification called substantive model-compatible imputation addresses this shortcoming (Bartlett et al., 2022; Bartlett & Morris, 2015; Bartlett et al., 2015; Goldstein et al., 2014). However, I classify this extension as a model-based imputation procedure because the appropriate sequence of univariate imputation models is inherently built around one analysis with a specific configuration of interactive or nonlinear effects.

5.2 Saving Imputed Data Sets

A typical MCMC process generates thousands of imputed data sets, one per computational cycle. The multiple imputation analysis phase requires just a few of these. Classic resources recommend three to five imputed data sets for frequentist analyses (Rubin, 1987; Schafer, 1997). Graham et al. (2007) used computer simulations to demonstrate that analyzing 20 or more imputations can produce nontrivial power gains. Other studies suggest that 100 or more imputations may be required to achieve precise estimates of quantities like confidence interval half-widths and probability values (Bodner, 2008; Harel, 2007; von Hippel, 2020). The examples in the Software Tutorials document use 20 imputed data sets.

After deciding on the number of imputations, we need to extract the data sets from a much longer MCMC process. Consecutive iterations of the MCMC algorithm produce correlated estimates and imputations, and this autocorrelation can last for many computational cycles. Analyzing data sets with correlated imputations should be avoided because doing so can attenuate frequentist standard errors. A simple way to eliminate autocorrelation is to specify a unique MCMC process for each data set, then save the filled-in data from the final iteration of each "chain". To illustrate, suppose a researcher wants to save 20 imputed data sets. After conducting a preliminary MCMC run to evaluate convergence (see Section 4.2 and Figure 10), they determine that the algorithm becomes sufficiently stable within 1,000 iterations. To create multiple imputations, they would specify 20 parallel MCMC processes and save the filled-in data from the 1,000th iteration of each unique chain. The Software Tutorials document uniformly adopts this strategy because the resulting imputations are automatically uncorrelated.

5.3 Analyzing Imputations and Pooling Results

The product of the initial imputation phase is a collection of filled-in data sets. Although intuition might suggest averaging the imputations into a single data set, doing so would attenuate standard errors. Instead, the correct approach is to analyze each data set individually then average the estimates and standard errors. Fortunately, most major software platforms have facilities for automating this process. To illustrate, consider an example where the goal is to estimate the mean change between the first- and ninth grade broad reading assessments. The analysis and pooling phases involve the following steps: (a) compute difference scores in each imputed data set by

TABLE 9. Imputation-Specific Estimates

Imputation	Estimate	Std. Error	(Std. Error)^2
1	6.652	0.541	0.293
2	6.509	0.526	0.277
3	6.734	0.533	0.284
4	6.376	0.541	0.292
5	6.318	0.549	0.301
6	6.267	0.533	0.284
7	6.698	0.537	0.289
8	6.400	0.534	0.285
9	6.321	0.544	0.296
10	6.563	0.544	0.296
11	6.440	0.554	0.307
12	6.433	0.527	0.278
13	6.374	0.535	0.286
14	6.416	0.536	0.287
15	5.900	0.549	0.301
16	6.659	0.527	0.277
17	6.532	0.543	0.294
18	6.092	0.539	0.291
19	6.388	0.553	0.305
20	6.289	0.543	0.295
Mean	6.418		0.291
Variance	0.041		

subtracting first grade scores from the ninth grade scores, (b) compute the mean difference score and its standard error in each data set, and (c) pool estimates and standard errors. This section demonstrates the various components of the pooling process. Example 12 from the Software Tutorials document provides annotated syntax and output files for model-agnostic fully conditional specification.

Table 9 shows the mean difference estimates, standard errors, and squared standard errors (i.e., sampling variances) from the 20 imputed data sets. These quantities naturally vary because each data set contains different predictions about the missing values. The estimates and squared standard errors are the inputs to Rubin's pooling rules in Table 10 (Little & Rubin, 2020; Rubin, 1987). The pooling formulas use M to denote the number of data sets, in this case $M = 20$. Focusing on the mean change score, the multiple imputation point estimate is the arithmetic average of the imputation-specific estimates. The summary section near the bottom of Table 9 reports the mean estimate as 6.42. On average, broad reading scores improved by more than six points. Although it averages over multiple realizations of the missing data, the point estimate has the same interpretation as one from a complete-data analysis.

TABLE 10. Summary of Rubin's Pooling Rules

Quantity	Description	Formula
Pooled point estimate	Arithmetic average of the point estimates	$\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$
Within-imputation variance	Arithmetic average of the squared standard errors (complete-data sampling error)	$mean(SE^2) = \frac{1}{M}\sum_{m=1}^{M}SE_m^2$
Between-imputation variance	Variance of parameters across imputed data sets (missing data uncertainty)	$var(\hat{\theta}) = \frac{1}{(M-1)}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta})^2$
Pooled standard error	Complete-data sampling error plus additional noise due to missing values	$SE = \sqrt{mean(SE^2) + var(\hat{\theta}) + \frac{var(\hat{\theta})}{M}}$
Test statistic	Difference between the pooled estimate and null value relative to the pooled standard error	$t = \frac{\hat{\theta} - \theta_{H_0}}{SE}$

Rubin's pooling rules for standard errors are more complex because they incorporate two sources of uncertainty. To begin, each of the squared standard errors in the rightmost column of

Table 9 estimates the expected amount of random sampling error from a *complete* data set. The average of the squared standard errors provides a more precise estimate of this quantity. The second equation in Table 10 shows this component of the pooling expression, which is commonly called the within-imputation variance. The summary section near the bottom of Table 9 reports the average squared standard error as 0.29. Of course, the data are not complete, so taking the square root of this value would produce an attenuated standard error that fails to acknowledge missingness. Returning to the point estimates, the mean difference estimates vary for only one reason—the data sets that spawned the estimates contain different imputations. Leveraging this fact, computing the variance of the point estimates around their average quantifies additional uncertainty due to missing data. The third equation in Table 10 shows this component of the pooling expression, which is often called between-imputation variance. The summary section near the bottom of Table 7 reports the between-imputation variance of the point estimates as 0.41.

The pooled standard error equation in Table 10 combines the within- and between-imputation variance into a single measure of uncertainty. Conceptually, the standard error expression starts with the attenuated estimate of complete-data sampling error, then it uses between-imputation variance as a correction factor that inflates the standard error according to the amounts and patterns of missing data. Substituting the within- and between-imputation variance estimates into the standard error formula gives $SE = 0.58$. This composite standard error is noticeably larger than the complete-data standard errors from Table 9 because it appropriately reflects missingness-induced imprecision. Finally, although their composition is very different, it is important to note that the multiple imputation standard errors are usually numerically equivalent to those of maximum likelihood and to Bayesian posterior standard deviations. The subsequent analysis examples highlight this point.

Numerous procedures for conducting frequentist significance tests are available for multiply imputed data. The familiar t-statistic for evaluating single-parameter hypotheses is a ratio that compares the discrepancy between the point estimate and null parameter value relative to the standard error. The final equation in Table 10 shows the test statistic. Returning to the reading achievement analysis, substituting the pooled mean difference and standard error into the expression gives $t = 11.11$. Statisticians have developed different degrees of freedom expressions for the t-statistic. Barnard and Rubin's (1999) expression provides good performance with small samples and is widely available in statistical software. Multiple imputation versions of the Wald (Li et al., 1991; Rubin, 1987) and likelihood ratio test statistics (Chan & Meng, 2021, December 30; Meng & Rubin, 1992) are also available for multiparameter hypotheses (e.g., omnibus and nested model tests). The R packages `mitml` (Grund et al., 2023) and `semTools` (Jorgensen et al., 2022)

offer a complete compendium of multiple imputation significance tests that are not widely available in commercial software.

5.4 Fine-Tuning Multiple Imputation: Auxiliary Variables and Nonnormal Data

The conditionally missing at random assumption described in Section 1.2 requires that the unseen score values contain no unique information about missingness beyond that contained in the observed data. Practically speaking, this assumption requires that all important determinants of missingness are contained within the imputation model's observed data. Introducing missing data auxiliary variables broadens the observed data, potentially making this assumption more plausible (Collins et al., 2001). The ease of introducing auxiliary variables is an oft-cited advantage of multiple imputation; additional variables are included during imputation and ignored during analysis. This contrasts maximum likelihood and Bayesian missing data handling approaches that integrate auxiliary variables on an analysis-by-analysis basis.

In practice, the number of variables that can be included in the imputation model is a complex function of the sample size and missing data patterns. In the hypothetical limit, regression models that provide the backbone of MCMC imputation routines require more observations than variables. Not surprisingly, missing values can dramatically restrict the size of the initial imputation model. In my experience, a pair of categorical variables with low or zero cell counts in a two-way contingency table is the most common cause of MCMC convergence failures. Such patterns imply that the data do not contain sufficient information to estimate bivariate associations. Collapsing categories, combining multiple variables into one variable, or excluding problematic variables are possible solutions. Previous studies have also documented difficulties applying model-agnostic multiple imputation to large numbers of ordered categorical variables (e.g., test or questionnaire items; Alacam et al., 2023). For many researchers, these practical issues place a relatively low upper limit on the number of variables multiple imputation can support. For example, a sample size of 300 participants may support fewer than 20 or 30 variables, especially if several are categorical .

The classic multivariate multiple imputation approach popularized by Joseph Schafer (Schafer, 1997) assumes multivariate normality. Contemporary variants of this strategy use a latent response variable (i.e., probit regression) framework to accommodate incomplete binary, ordinal, and multicategorical nominal variables (Asparouhov & Muthén, 2010; Carpenter & Kenward, 2013; Demirtas, 2017; Goldstein et al., 2009; Quartagno & Carpenter, 2019). For example, binary and ordinal responses can be modeled as a normally distributed latent variable, the distribution of which is carved into discrete regions by one or more threshold parameters (Johnson & Albert,

1999). Fully conditional specification imputation also accommodates a range of discrete variable types (Kleinke & Reinecke, 2013; van Buuren, 2007; van Buuren, 2010; van Buuren, 2018). Returning to the hypothetical imputation scheme in Equation 22, the individual regression equations would simply become generalized linear models, and MCMC would sample imputations from probability mass functions that look like bar graphs.

Turning to numeric variables, several methodology studies have examined the impact of pairing normally distributed imputations with nonnormal observed data (Demirtas et al., 2008; Lee & Carlin, 2017; von Hippel, 2013; Yuan et al., 2012). Although this approach can produce imputes outside the range of the data, a common finding is that distributional misspecifications do not substantially impact means and regression coefficients; it can distort variances and estimands that quantify the distribution's tails, however. Related work describes a transform-then-imputation strategy that first applies normalizing transformations to incomplete variables prior to imputation then back-transforms the variables prior to analysis (Goldstein et al., 2009; Lee & Carlin, 2017; Schafer & Olsen, 1998; Su et al., 2011; van Buuren, 2018; von Hippel, 2013). Studies caution against this strategy because it can exacerbate rather than mitigate bias due to nonnormal data (Lee & Carlin, 2017; von Hippel, 2013).

An alternate strategy is to sample imputations from a nonnormal distribution. As mentioned previously, predictive mean matching is a nonparametric approach that draws imputations from a pool of observed scores taken from people whose predicted values are similar to that of the person with missing data (Kleinke, 2017; Lee & Carlin, 2017; van Buuren, 2018; Vink et al., 2014). If the observed data are nonnormal, the resulting imputations will be as well. Morris et al. (2014) catalog variations of predictive mean matching in software programs, and they provide recommendations about the optimal matching criterion and donor pool size. The literature also describes various parametric methods for creating nonnormal imputations (de Jong et al., 2016; Demirtas, 2017; Demirtas & Hedeker, 2008a, 2008b; He & Raghunathan, 2009), although most of these approaches are either limited in scope or do not have software implementations. The Yeo–Johnson transformation (Yeo & Johnson, 2000) is a promising approach that readily pairs with model-based multiple imputation (Lüdtke et al., 2020b). The Yeo–Johnson procedure invokes a normally distributed transformed variable and a shape parameter that maps the normalized scores to a nonnormal raw score distribution. The MCMC algorithm iteratively estimates the shape parameter to match the observed-data distribution's shape. This approach, which is available in the Blimp application (Keller & Enders, 2021) and R package `mdmb` (Grund et al., 2021a), subsumes common transformations such as the inverse, logarithmic, square root, and Box–Cox.

5.5 Multiple Imputation Analysis Example 1

The first multiple imputation example illustrates a multiple regression analysis comprised of incomplete continuous variables. The analysis uses the `behaviorachievement.dat` data set from a longitudinal study that followed 138 students from primary through middle school. The file includes three annual assessments of broad reading and math achievement beginning in the first grade, seventh grade standardized achievement test scores taken from a statewide assessment, and a final measure of broad reading and math obtained in ninth grade. The data also contain teacher ratings of behavioral symptoms and learning problems were also obtained in the first grade. The Software Tutorials document provides additional information about this data set.

Given the same data and assumptions, multiple imputation usually produces estimates and inferences that are equivalent to maximum likelihood and Bayesian estimation. To underscore this point, I repeat the same analysis from Sections 3.4 and 4.5. Table 2 from Section 1.5 describes the specific variables for the multiple regression. To refresh, the analysis model featured ninth grade broad reading scores regressed on three academic and behavioral measures collected in first grade: the broad reading composite, teacher-rated learning problems, and teacher-rated behavioral problems.

$$READ_9 = \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) + \beta_3(BEHPROB_1) + \varepsilon \tag{23}$$

The analysis also included second grade broad reading scores and seventh grade standardized test scores as auxiliary variables.

Applying the flowchart from Figure 17, both model-agnostic and model-based multiple imputation are appropriate, and the two would yield equivalent estimates. To apply model-based imputation, one would simply fit the regression model from the first Bayesian analysis example (see Section 4.6), saving imputations for reanalysis. This example instead uses model-agnostic fully conditional specification (van Buuren, 2018; van Buuren et al., 2006). Focusing on the initial imputation step, each iteration of the MCMC algorithm fits six regression models, one per incomplete variable. The equations follow the same round-robin schematic described in Section 5.1.

$$\begin{aligned} READ_9 = \gamma_{10} &+ \gamma_{11}\left(READ_1^{(t-1)}\right) + \gamma_{12}\left(LRNPROB_1^{(t-1)}\right) + \gamma_{13}\left(BEHPROB_1^{(t-1)}\right) \\ &+ \gamma_{14}\left(READ_2^{(t-1)}\right) + \gamma_{15}\left(STANREAD_7^{(t-1)}\right) + \epsilon_1 \end{aligned}$$

$$\begin{aligned} READ_1 = \gamma_{20} &+ \gamma_{21}\left(LRNPROB_1^{(t-1)}\right) + \gamma_{22}\left(BEHPROB_1^{(t-1)}\right) + \gamma_{23}\left(READ_2^{(t-1)}\right) \\ &+ \gamma_{24}\left(STANREAD_7^{(t-1)}\right) + \gamma_{25}\left(READ_9^{(t)}\right) + \epsilon_2 \end{aligned}$$

$$LRNPROB_1 = \gamma_{30} + \gamma_{31}\left(BEHPROB_1^{(t-1)}\right) + \gamma_{32}\left(READ_2^{(t-1)}\right) + \gamma_{33}\left(STANREAD_7^{(t-1)}\right) + \gamma_{34}\left(READ_9^{(t)}\right) + \gamma_{35}\left(READ_1^{(t)}\right) + \epsilon_3 \quad (24)$$

$$BEHPROB_1 = \gamma_{40} + \gamma_{41}\left(READ_2^{(t-1)}\right) + \gamma_{42}\left(STANREAD_7^{(t-1)}\right) + \gamma_{43}\left(READ_9^{(t)}\right) + \gamma_{44}\left(READ_1^{(t)}\right) + \gamma_{45}\left(LRNPROB_1^{(t)}\right) + \epsilon_4$$

$$READ_2 = \gamma_{50} + \gamma_{51}\left(STANREAD_7^{(t-1)}\right) + \gamma_{52}\left(READ_9^{(t)}\right) + \gamma_{53}\left(READ_1^{(t)}\right) + \gamma_{54}\left(LRNPROB_1^{(t)}\right) + \gamma_{55}\left(BEHPROB_1^{(t)}\right) + \epsilon_5$$

$$STANREAD_7 = \gamma_{60} + \gamma_{61}\left(READ_9^{(t)}\right) + \gamma_{62}\left(READ_1^{(t)}\right) + \gamma_{63}\left(LRNPROB_1^{(t)}\right) + \gamma_{64}\left(BEHPROB_1^{(t)}\right) + \gamma_{65}\left(READ_2^{(t)}\right) + \epsilon_6$$

The imputation models are all linear regressions in this example, although the algorithm readily accommodates discrete metrics as well (see Example 14 in the Software Tutorials document).

Example 13 from the Software Tutorials document provides annotated syntax and output files for the analysis. To summarize, the fully conditional specification imputation algorithm invoked 20 independent MCMC processes (chains), each consisting of 2,500 iterations. The filled-in data from the final iteration of each chain was saved for reanalysis. Fitting the focal regression model from Equation 23 to each imputed data set produced 20 sets of estimates and standard errors. Finally, applying Rubin's rules gave the pooled results shown in Table 11. The results are interpreted in the same way as a complete-data regression analysis. For example, consider the slope of the first grade reading test. The model predicts that two individuals who differ by one point should differ by 0.51 points on the outcome, holding constant teacher-rated learning and

TABLE 11. Agnostic Multiple Imputation Estimates From a Multiple Regression

Effect	Est.	Std. Error	*t*	*p*	2.5% LCL	97.5% UCL
Intercept	66.19	6.22	10.64	< .001	53.81	78.58
$READ_1$	0.51	0.05	10.96	< .001	0.41	0.60
$LRNPROB_1$	–0.25	0.12	–2.06	.04	–0.49	–0.01
$BEHSYMP_1$	–0.18	0.11	–1.73	.09	–0.40	0.03
R-square	.61	--	--	--	--	--

behavioral problems. The corresponding test statistic indicates that the slope coefficient is statistically different from zero, $t(71.14) = 10.96$, $p < .001$.

I have repeatedly stated that the Big Three missing data methods tend to produce numerically equivalent results, given the same input data and assumptions. To emphasize this point, Table 12 shows the parameter estimates and measures of uncertainty for each procedure. Notwithstanding their philosophical differences about repeated sampling, the Big Three were identical, as expected. Returning to decision tree in Figure 3, the choice of missing data method is largely determined by features of the analysis model. Selecting between two approaches that are equally appropriate is largely a matter of personal preference.

TABLE 12. Comparison of Regression Results From the Big Three

	FIML		Bayes		MI	
Effect	Est.	*SE*	Median	*SD*	Est.	*SE*
Intercept	66.03	5.85	66.01	6.05	66.19	6.22
$READ_1$	0.50	0.04	0.50	0.04	0.51	0.05
$LRNPROB_1$	–0.25	0.12	–0.25	0.12	–0.25	0.12
$BEHSYMP_1$	–0.18	0.10	–0.18	0.11	–0.18	0.11
R-square	.60	--	.60	.05	.61	--

5.6 Multiple Imputation Analysis Example 2

The second analysis example illustrates a moderated regression analysis with an incomplete interaction effect and an incomplete binary covariate. This example repeats the same analysis from Sections 3.5 and 4.6. Table 5 from Section 3.5 describes the specific variables for the regression. To refresh, the analysis model features ninth grade broad reading scores regressed on academic and behavioral measures collected in first grade: reading achievement, teacher-rated learning problems, the product of first grade reading scores and learning problems, and the at-risk dummy code.

$$\begin{aligned} READ_9 = {} & \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) \\ & + \beta_3(READ_1)(LRNPROB_1) + \beta_4(ATRISK) + \varepsilon \end{aligned} \tag{25}$$

The analysis also included second grade broad reading scores and seventh grade standardized test scores as auxiliary variables.

Applying the flowchart from Figure 17, models with interactive effects require a factored regression specification that expresses the multivariate distribution into a series of simpler distributions. Such specifications are necessarily model-based because the correct sequence of regression models is constructed around one focal analysis with a particular constellation of nonlinear terms. Section 4.6 illustrated a Bayesian analysis with an interactive effect and auxiliary variables. This identical analysis can also produce multiple imputations for a frequentist analysis; comparable procedures like substantive model-compatible imputation (Bartlett & Morris, 2015; Bartlett et al., 2015) and the sequential specification (Lüdtke et al., 2020b) are equally appropriate.

Example 10 from the Software Tutorials document provides annotated syntax and output files for the analysis. The MCMC algorithm from the earlier Bayesian analysis was modified to invoke 20 independent computational chains, each consisting of 5,500 iterations. Aside from that small procedural detail, the imputation phase was identical to the earlier example. The filled-in data from the final iteration of each MCMC chain was saved for reanalysis, and the moderated regression model from Equation 22 was then fit to each imputed data set. Finally, using Rubin's rules to combine the 20 sets of estimates and standard errors produced the pooled results.

TABLE 13. Model-Based Multiple Imputation Estimates From a Moderated Regression

Effect	Est.	Std. Error	*t*	*p*	2.5% LCL	97.5% UCL
Intercept	87.92	0.87	100.98	< .001	86.19	89.64
$READ_1$	0.50	0.05	11.04	< .001	0.41	0.59
$LRNPROB_1$	−0.38	0.09	−4.32	< .001	−0.54	−0.20
$READ_1 \times LRNPROB_1$	0.012	0.005	2.55	.03	0.002	0.021
ATRISK	−2.05	1.78	−1.15	.25	−5.57	1.47
R-square	.61	--	--	--	--	--

Note. First grade reading scores and learning problems ratings are centered at the grand mean.

Table 13 shows the multiple imputation analysis results. The lower-order terms in a moderated regression are conditional effects that depend on scaling or centering. To facilitate interpretation, the analysis centered the interacting variables at the pooled estimates of their grand

means. Centering defines the first grade reading slope ($\hat{\beta}_1 = 0.51$) as a conditional effect at the mean of the learning problems distribution, and the learning problems slope ($\hat{\beta}_2 = -0.38$) similarly reflects a conditional effect at the reading achievement average. The interaction slope captures the change in the first grade reading slope for each one-unit increase in learning problems (and vice versa). The positive coefficient ($\hat{\beta}_3 = 0.013$) indicates that the association between first and ninth grade reading scores becomes stronger (i.e., more positive) for students with elevated learning problems.

I have repeatedly stated that the Big Three missing data methods tend to produce numerically equivalent results, given the same input data and assumptions. To further emphasize this point, Table 14 shows the parameter estimates and measures of uncertainty for each procedure. Notwithstanding their philosophical differences about repeated sampling, the Big Three were identical, as expected. Returning to decision tree in Figure 3, the choice of missing data method is largely determined by features of the analysis model. Selecting between two approaches that are equally appropriate is largely a matter of personal preference.

TABLE 14. Comparison of Moderated Regression Estimates From the Big Three

	FIML		Bayes		MI	
Effect	Est.	*SE*	Median	*SD*	Est.	*SE*
Intercept	89.05	1.42	87.85	1.32	87.92	0.87
$READ_1$	0.51	0.04	0.50	0.05	0.50	0.05
$LRNPROB_1$	–0.38	0.08	–0.37	0.09	–0.38	0.09
$READ_1 \times LRNPROB_1$	0.013	0.005	0.012	0.005	0.012	0.005
ATRISK	–1.91	1.79	–1.93	1.87	–2.05	1.78
R-square	.62	--	.61	--	.61	--

6

Multilevel Missing Data

Multilevel data structures are ubiquitous in education research. Two-level examples include repeated measurements at level-1 nested within students at level-2 or students at level-1 nested in schools at level-2. A prototypical three-level data hierarchy combines repeated measurements (level-1) within students (level-2) and students within schools (level-3). This section focuses on multilevel regression models with random effects because they are an exceedingly common data analytic tool in educational research applications. For example, substantial methodological work has focused on hierarchical models for cluster-randomized trials where schools are assigned to experimental conditions (Hedges & Hedberg, 2007; Raudenbush, 1997; Raudenbush & Liu, 2000; Spybrook et al., 2011; Spybrook et al., 2016). A variety of sources provide information about alternate modeling approaches for hierarchically structured data sets (Hamaker & Muthén, 2020; McNeish & Kelley, 2019; McNeish et al., 2017).

The emergence of sophisticated missing data handling methods for multilevel regression models is an important recent development in the methodology literature (Carpenter et al., 2023; Carpenter et al., 2011; Enders et al., 2020; Erler et al., 2019; Erler et al., 2016; Goldstein et al., 2009; Goldstein et al., 2014; Grund et al., 2021a; Quartagno & Carpenter, 2016; Shin, 2013; Shin & Raudenbush, 2023; Shin & Raudenbush, 2007, 2013; van Buuren, 2011; Yucel, 2008, 2011). I devote much of this section to Bayesian estimation and model-based multiple imputation because they currently handle a broader range of multilevel missing data problems than maximum likelihood estimators.

6.1 Multilevel Missing Data Handling Options

The full and restricted maximum likelihood estimators in general-use mixed modeling programs readily accommodate incomplete outcomes. When missing values are relegated to the dependent variable, excluding rows with missing outcome scores yields maximum likelihood estimates with

a conditionally missing at random assumption (von Hippel, 2007). Missingness simply creates an unbalanced data set where the number of level-1 observations varies across level-2 units. However, these programs typically have no capacity for treating incomplete predictors. Methodologists have developed maximum likelihood routines for random intercept models with continuous predictors (Shin & Raudenbush, 2007, 2013; Shin & Raudenbush, 2010), and important work on random slope predictors and interaction effects is ongoing (Rockwood, 2020; Shin & Raudenbush, 2023). The HLM (Raudenbush et al., 2019) and M*plus* programs (Muthén & Muthén, 1998–2017) offer the most sophisticated maximum likelihood missing data handling options.

Prior to the advent of sophisticated techniques for multilevel missing data, researchers could dummy code higher-level units (e.g., schools) and include the code variables as predictors in a single-level imputation scheme. This so-called fixed effect imputation scheme can produce unbiased parameter estimates in some random intercept applications (Lüdtke et al., 2017; Reiter et al., 2006). However, inflated standard errors and distorted confidence intervals are undesirable byproducts (Andridge, 2011; van Buuren, 2011). Although newer approaches are far superior, fixed effect imputation may be useful when the number of higher-level clusters is too small to support random effect estimation.

As described in Section 5, Schafer's (1997) classic joint model imputation uses a multivariate regression model with complete variables predicting incomplete variables. The multilevel extension of this approach uses a multivariate mixed effects model where incomplete level-1 variables are regressed on complete level-1 and level-2 variables (Schafer, 2001; Schafer & Yucel, 2002). Flexible variations of this approach allow for incomplete categorical variables and missing values at any level of the data hierarchy (Asparouhov & Muthén, 2010; Carpenter et al., 2011; Carpenter & Kenward, 2013; Goldstein et al., 2009; Goldstein et al., 2014; Yucel, 2008). Importantly, most joint model imputation schemes are limited to random intercept analyses and have no capacity for preserving random slopes for incomplete predictors (Enders et al., 2016). A notable exception is a joint modeling variant that allows the within-cluster covariance matrix to vary across level-2 clusters (Quartagno & Carpenter, 2020; Quartagno & Carpenter, 2016; Yucel, 2011).

Fully conditional specification imputation also extends to two- and three-level models (Audigier et al., 2018; Enders et al., 2018; Keller, 2015; Resche-Rigon & White, 2018; van Buuren, 2011). Like joint model imputation, fully conditional specification should be reserved for random intercept models, in which case the two agnostic imputation procedures are effectively equivalent (Grund et al., 2017, 2018; Grund et al., 2019; Lüdtke et al., 2017; Mistler & Enders, 2017). Importantly, computer simulation studies show that applying fully conditional specification to incomplete random slope predictors—so-called "reverse random coefficient" imputation—can

introduce substantial bias because it misrepresents variation (Enders et al., 2020; Enders et al., 2016; Grund et al., 2016a, 2018).

Factored regression specifications like those described in Section 2 provide the greatest flexibility for addressing incomplete predictors (Enders et al., 2020; Erler et al., 2019; Erler et al., 2016; Grund et al., 2021a). To refresh, factored regression specifications express a multivariate distribution as a sequence of simpler distributions, the collection of which is equivalent the original joint function. Beyond acknowledging that a multivariate distribution exists, factorization makes no assumptions about its shape or form. Rather, distributional assumptions enter on a variable-by-variable via a collection of regression models. Among other things, this factorization readily accommodates variables with different metrics, random slope predictors, interaction or curvilinear effects, and incomplete variables at any level of the data hierarchy.

6.2 Multilevel Analysis Example 1

The first multilevel analysis example illustrates a two-level regression model with random intercepts. The `problemsolving2level.dat` data set is taken from a cluster-randomized educational intervention where 29 schools (level-2 units) were assigned to an intervention and comparison condition (Montague et al., 2014). The comparison condition (i.e., control schools) implemented the district's standard mathematics curriculum, and the intervention schools implemented a new curriculum designed to enhance math problem-solving skills. The 982 student-level (level-1) records include pretest and posttest math problem-solving and self-efficacy scores, standardized math scores taken from a statewide assessment, and several sociodemographic variables. The dependent variable is an end-of-year math problem-solving assessment with IRT-

TABLE 15. Variables for the Multilevel Random Intercept Model

Name	Definition	Missing	Scale
SCHOOL	School (level-2) identifier	0	Integer index
CONDITION	Experimental condition	0	0 = Control, 1 = Experimental
HISPANIC	Ethnicity/race	9.0	0 = Other, 1 = Hispanic
FRLUNCH	Lunch assistance code	4.7	0 = None, 1 = Free or Reduced Lunch
MPSPRE	Math problem-solving pretest	0	Numeric (37 to 66)
MPSPOST	Math problem-solving posttest	20.5	Numeric (37 to 65)

scaled scores. The Software Tutorials document provides additional information about this data set. Table 15 shows the specific variables for this analysis example.

A random intercept model features school-specific intercept coefficients that capture level-2 mean differences. The goal of the analysis is to determine whether the intervention groups differ on an end-of-year math problem-solving test after controlling for three student-level covariates: math problem-solving pre-test scores, a Hispanic dummy code, and a free or reduced lunch assistance dummy code. Following notation from Raudenbush and Bryk (2002), the within-school model describes score variation among students in the same school.

$$MPSPOST_{ij} = \beta_{0j} + \beta_1\left(MPSPRE_{ij}^{cwc}\right) + \beta_2\left(HISPANIC_{ij}^{cwc}\right) + \beta_3\left(FRLUNCH_{ij}^{cwc}\right) + \varepsilon_{ij} \quad (26)$$

The *i* and *j* subscripts on the student-level variables index students and schools, respectively. The β_{0j} coefficient is a school-specific random intercept that represents the average post-test score in school *j*. The *cwc* superscript on the predictors stands for centering within cluster, also known as group mean centering (Enders & Tofighi, 2007; Kreft et al., 1995). This operation subtracts each student's score from their own school's average, as follows.

$$MPSPRE_{ij}^{cwc} = MPSPRE_{ij} - \mu_{j(MPSPRE)} \quad (27)$$

The purpose of this type of centering is to disaggregate the student-level scores by removing school-level variation. The resulting coefficient is a pure within-school association.

Each coefficient on the right side of the within-school regression is an outcome in a school-level regression, as follows.

$$\begin{gathered}\beta_{0j} = \gamma_{00} + \gamma_{01}\left(\mu_{j(MSPRE)}^{cgm}\right) + \gamma_{02}\left(\mu_{j(HISPANIC)}^{cgm}\right) + \gamma_{03}\left(\mu_{j(FRLUNCH)}^{cgm}\right) + \gamma_{04}\left(CONDITION_j\right) + u_{0j} \\ \beta_1 = \gamma_{10} \\ \beta_2 = \gamma_{20} \\ \beta_3 = \gamma_{30}\end{gathered} \quad (28)$$

The expressions for β_1, β_2, and β_3 indicate that the within-school slopes are constant across schools (i.e., there are no u terms indicating a random slope residual). In contrast, the random intercepts (school-level post-test averages) vary as a function of the covariates and the intervention assignment indicator. To appropriately control for student-level covariates, it is important to include the school means of all level-1 variables as predictors (Rights et al., 2020). The *cgm* superscript indicates that these school-level aggregates are centered at their grand means. Centering a level-2 variable is largely a cosmetic operation that defines γ_{00} as the average post-test

problem solving score among comparison schools. Finally, the u_{0j} term is a random intercept residual that captures the unexplained part of each school's post-test mean.

Replacing each β coefficient in the within-school equation from Equation 26 with the right side of its school-level equation gives the combined regression model below.

$$\begin{aligned} MPSPOST_{ij} = &\left(\gamma_{00} + u_{0j}\right) + \gamma_{10}\left(MPSPRE_{ij}^{cwc}\right) + \gamma_{20}\left(HISPANIC_{ij}^{cwc}\right) \\ &+ \gamma_{30}\left(FRLUNCH_{ij}^{cwc}\right) + \gamma_{01}\left(\mu_{j(MSPRE)}^{cgm}\right) + \gamma_{02}\left(\mu_{j(HISPANIC)}^{cgm}\right) \\ &+ \gamma_{03}\left(\mu_{j(FRLUNCH)}^{cgm}\right) + \gamma_{04}\left(CONDITION_{j}\right) + \varepsilon_{ij} \end{aligned} \tag{29}$$

All coefficients with a leading zero subscript are school-level effects, and all coefficients with non-zero leading subscripts are pure within-school effects. The γ_{04} slope is of particular interest because it captures the intervention effect, controlling for covariates.

The incomplete binary predictors complicate missing data handling because existing maximum likelihood estimators are limited to normally distributed incomplete predictors (Shin & Raudenbush, 2023; Shin & Raudenbush, 2013). There is currently very little research that documents the impact of misspecifying incomplete binary predictors as normal in multilevel analyses (Grund et al., 2018). Applying the flowchart from Figure 3, either fully conditional specification multiple imputation or estimation based on a factored regression specification are appropriate. Importantly, fully conditional specification can only preserve the disaggregated covariate effects in Equation 29 if the school-level averages of the level-1 variables are included in the imputation regression models (Enders et al., 2018; Grund et al., 2017). This example applies a version of fully conditional specification that introduces random intercepts (latent cluster means) for the level-1 variables (Keller & Enders, 2021).

Currently, Bayesian estimation is the only way to deploy a factored regression specification for this analysis. Following Figure 2b, I adopt a factorization comprised of two submodels, each with distinct distributional assumptions. The Blimp application used in the Software Tutorials document automatically configures the distributions, and the user simply needs to specify the focal regression model in Equation 29. For completeness, the symbolic representation for the underlying two-part factorization is as follows.

$$\begin{aligned} &f(MPSPOST|MSPRE, HISPANIC, FRLUNCH, CONDITION) \\ &\times f(MSPRE, HISPANIC, FRLUNCH, CONDITION) \end{aligned} \tag{30}$$

The first term corresponds to the univariate regression model from Equation 29. To accommodate the missing data, this model treats the school-level means as level-2 latent variables or random

intercepts (Lüdtke et al., 2011; Lüdtke et al., 2008; Shin & Raudenbush, 2010). The second model is a multivariate normal distribution for the predictors. This submodel further disaggregates predictors into within-school and between-school parts, such that the level-1 predictors link to the level-2 predictors via latent cluster means (random intercepts). Finally, the factored specification readily accommodates incomplete categorical variables with a latent response variable framework (probit regression). Detailed descriptions of the model are available in the literature (Enders, 2022; Enders et al., 2020).

Example 15 from the Software Tutorials document provides annotated syntax and output files for agnostic fully conditional specification multiple imputation. The algorithm invoked 20 independent MCMC processes (chains), each consisting of 2,500 iterations. Blimp's version of fully conditional specification automatically introduces random intercepts (latent cluster means) for the level-1 variables (Keller & Enders, 2021), so it is appropriate for analysis models that disaggregate level-1 predictors. The filled-in data from the final iteration of each chain was saved for reanalysis. Fitting the focal regression model from Equation 29 to each imputed data set produced 20 sets of estimates and standard errors. Finally, applying Rubin's rules gave the pooled results shown in Table 16. The results are interpreted in the same way as a complete-data multilevel regression analysis. Due to centering, the intercept coefficient ($\hat{\gamma}_{00} = 52.70$) represents the post-test problem-

TABLE 16. Agnostic Multiple Imputation Estimates From a Multilevel Regression

Effect	Est.	Std. Error	*t*	*p*	2.5% LCL	97.5% UCL
Intercept	52.70	0.52	100.87	< .001	51.68	53.73
MPSPRE	0.46	0.04	13.23	< .001	0.39	0.52
HISPANIC	1.02	0.42	2.46	.01	0.20	1.84
FRLUNCH	–0.71	0.49	–1.45	.15	–1.69	0.26
MPSPRE Means	0.63	0.22	2.88	< .01	0.20	1.05
HISPANIC Means	5.01	1.39	3.62	< .001	2.29	7.72
FRLUNCH Means	–2.65	2.51	–1.06	.29	–7.57	2.27
CONDITION	2.37	0.72	3.31	.001	0.97	3.78
Intercept variance	2.29	--	--	--	--	--
Residual variance	20.61	--	--	--	--	--

solving mean among the comparison schools that received the district's standard curriculum. The intervention indicator slope indicates that the intervention schools with the new curriculum scored $\hat{\gamma}_{04}$ = 2.37 points higher than comparison schools, on average, holding constant student-level covariates. corresponding test statistic indicates that the slope coefficient is statistically different from zero, $t(2569.30) = 3.31$, $p = .001$.

Example 16 from the Software Tutorials document provides annotated syntax and output files for the Bayesian analysis. Table 17 shows the Bayesian parameter summaries. The estimate (posterior median) and standard deviation columns describe the center and spread of the posterior distributions; although they make no reference to drawing repeated samples, they are analogous to frequentist point estimates and standard errors. The 95% credible interval columns give ranges that capture 95% of each parameter's distribution. These are akin to confidence intervals, but they describe a range of likely parameter values rather than the long-run behavior of intervals from different random samples.

To reiterate, the posterior distributions characterize plausible *parameter values* that could have produced these data. From a practical perspective, the posterior summaries have interpretations

TABLE 17. Bayesian Parameter Summaries From a Multilevel Regression

Effect	Est.	Std. Dev.	2.5% LCL	97.5% UCL	χ^2	*p*
Intercept	52.42	0.73	50.93	53.86	5125.49	< .001
MPSPRE	0.46	0.04	0.39	0.53	160.39	< .001
HISPANIC	0.94	0.44	0.09	1.80	4.67	.03
FRLUNCH	−0.78	0.47	−1.70	0.14	2.79	.10
MPSPRE means	0.71	0.34	0.06	1.41	4.59	.03
HISPANIC means	5.31	1.71	2.01	8.68	9.69	.002
FRLUNCH means	−1.70	3.52	−8.74	5.16	0.23	.63
CONDITION	2.42	0.79	0.90	4.02	9.42	.002
Intercept variance	2.45	1.20	1.06	5.68	--	--
Residual variance	20.59	1.07	18.62	22.79	--	--
R-square coefficients	.33	.04	.25	.41		
R-square intercepts	.07	.03	.03	.15		

that parallel any multilevel regression analysis. Due to centering, the intercept coefficient (γ_{00} = 52.42) represents the post-test problem-solving mean among the comparison schools that received the district's standard curriculum. The intervention indicator slope indicates that the intervention schools with the new curriculum scored γ_{04} = 2.42 points higher than comparison schools, on average, holding constant student-level covariates. Importantly, the numeric summaries in Table 17 are effectively equivalent to the corresponding multiple imputation estimates from Table 16 (e.g., $\hat{\gamma}_{00}$ = 52.70 and $\hat{\gamma}_{04}$ = 2.37). This is precisely what you would expect when applying Big Three approaches that use the same data and invoke the same assumptions.

As explained previously, the Bayesian summaries also lend themselves to familiar hypothesis testing logic. Returning to the intervention coefficient, the credible interval limits spanning 0.90 to 4.02 included 95% of the area under the coefficient's posterior distribution. From this we can conclude that the parameter is statistically different from zero ($p < .05$) because the null value is outside the interval. That is, the probability that the parameter is less than 0.90 or greater than 4.02 is .05. Alternatively, (Asparouhov & Muthén, 2021) proposed a Bayesian Wald chi-square statistic that can be used to evaluate a broad range of hypotheses. For researchers who prefer a frequentist-like test statistic, the rightmost columns of Table 17 show the chi-square test statistics and *p*-values for the coefficients. These quantities lead to the same conclusions as the 95% credible intervals (e.g., the condition effect is significant with $p = .002$). Levy and McNeish (2023) use the phrase "computational frequentism" to describe applications that use MCMC estimation to approximate frequentist point estimates and test statistics.

The Blimp application invokes flat priors for the coefficients and common off-the-shelf priors for variances and covariances (Keller & Enders, 2021). The prior distributions influence variance parameters by altering the number of independent data points and/or the residual sums of squares at each MCMC cycle. Roughly speaking, alternate prior distributions induce differences that are analogous to those between restricted and full information maximum likelihood (McNeish, 2017; D. McNeish & L. M. Stapleton, 2016; D. M. McNeish & L. M. Stapleton, 2016). The choice of prior distribution is potentially impactful to variance estimates when the number of level-2 units is small, as it is here (there are 29 schools at level-2). To explore the sensitivity of the results to this issue, I fit the model with two alternate prior distributions (also off-the-shelf options that are invoked with a single keyword). Changes to the coefficients were generally inconsequential and in the second decimal place. As one might expect, the choice of prior did impact the random intercept residual variance; both alternatives decreased the variance point estimate from 2.45, producing median values equal to 1.89 and 2.07. To put these changes in perspective, the R^2 value in the bottom row Table 17 reports that residual random intercept variance accounted for approximately 7% of the total variation in the post-test scores (Rights & Sterba, 2019). Invoking alternate prior distributions

changed this effect size by less than 1.5% in both cases (R^2 = .056 and .061). From a practical perspective, these differences seem relatively minor, and I would conclude that the prior distributions did not meaningfully impact on the results.

Finally, to apply model-based multiple imputation, one would simply save imputations from the MCMC algorithm for reanalysis with frequentist inference (Little & Rubin, 2020; Rubin, 1987). Example 16 from the Software Tutorials document provides the analysis scripts. I omit the pooled results because they were effectively equivalent to those in Tables 16 and 17. To reiterate, you would expect this correspondence when applying Big Three approaches that use the same data and invoke the same assumptions.

6.3 Multilevel Analysis Example 2

The second multilevel analysis example illustrates a two-level linear growth curve model with repeated measurements at level-1 nested in students at level-2. Students are also nested in schools, but I ignore the third level to keep the example simple. Example 18 in the Software Tutorials document demonstrates a three-level growth curve model. The `problemsolving3level.dat` data set is taken from a cluster-randomized educational intervention where 29 schools (level-2 units) were assigned to an intervention and comparison condition (Montague et al., 2014). The comparison condition (i.e., control schools) implemented the district's standard mathematics curriculum, and the intervention schools implemented a new curriculum designed to enhance math problem-solving skills. The 6874 within-subjects data records include seven monthly measures of an IRT-scaled math problem-solving assessment. Table 18 shows the specific variables for this analysis example.

TABLE 18. Variables for the Multilevel Growth Curve Model

Name	Definition	Missing	Scale
STUDENT	Student (level-2) identifier	0	Integer index
PROBSOLVE	Math problem-solving	11.4	Numeric (37 to 68)
MONTH	Time scores	0	Integer index (0 to 6)
HISPANIC	Ethnicity/race	9.0	0 = Non-Hispanic, 1 = Hispanic
FRLUNCH	Lunch assistance code	4.7	0 = None, 1 = Lunch assistance
CONDITION	Experimental condition	0	0 = Control, 1 = Experimental

A random coefficient growth model features person-specific intercept and slope coefficients that represent idealized linear change trajectories. Following notation from Raudenbush and Bryk (2002), the within-person model describes score variation among students in the same school.

$$PROBSOLVE_{ij} = \beta_{0j} + \beta_{1j}(MONTH_{ij}) + \varepsilon_{ij} \tag{31}$$

The *i* and *j* subscripts on the student-level variables index repeated measurements and students, respectively. The *MONTH* variable is an integer index that codes the timing of the monthly measurements relative to the baseline assessment. That is, *MONTH* = 0, 1, 2, 3, 4, 5, and 6. These time scores define β_{0j} as person *j*'s expected baseline problem-solving and β_{1j} as their monthly change rate. To keep the example simple, I ignore school-level nesting.

Each coefficient on the right side of the within-school regression is an outcome in a person-level regression. The expressions for β_{0j} and β_{1j} indicate that the random intercepts and slopes vary as a function of the Hispanic, lunch assistance, and intervention assignment dummy codes. The random intercept equation includes all three predictors of baseline performance, and the individual change rates vary as a function of ethnicity and intervention condition. This configuration treats the slope predictors as moderators and lunch assistance as a covariate.

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(HISPANIC_j) + \gamma_{02}(FRLUNCH_j^{cgm}) + \gamma_{03}(CONDITION_j) + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(HISPANIC_j) + \gamma_{12}(CONDITION_j) + u_{1j} \end{aligned} \tag{32}$$

The *cgm* superscript on *FRLUNCH* indicates that this variable is centered at its grand mean (Enders & Tofighi, 2007). Centering a level-2 variable is largely a cosmetic operation that defines γ_{00} as the expected baseline problem solving score among non-Hispanic students in comparison schools. In the random slope model, γ_{10} is the monthly change rate among non-Hispanic students in comparison schools, and γ_{11} and γ_{12} are growth rate differences for Hispanic students and intervention schools, respectively. Finally, the u_{0j} and u_{1j} terms are random intercept and slope residuals that captures the unexplained part of each person's growth trajectory.

Replacing each β coefficient in the within-school equation from Equation 31 with the right side of its student-level model from Equation 32 gives the combined regression model below.

$$\begin{aligned} PROBSOLVE_{ij} &= (\gamma_{00} + u_{0j}) + (\gamma_{10} + u_{1j})(MONTH_{ij}) \\ &+ \gamma_{11}(HISPANIC_j)(MONTH_{ij}) + \gamma_{12}(CONDITION_j)(MONTH_{ij}) \\ &+ \gamma_{01}(HISPANIC_j) + \gamma_{02}(FRLUNCH_{ij}^{cgm}) + \gamma_{03}(CONDITION_j) + \varepsilon_{ij} \end{aligned} \tag{33}$$

All coefficients with a leading zero subscript are determinants of baseline performance, and all coefficients with one as a leading subscript define the monthly change rates. Notice that the

predictors of random slopes form cross-level, group-by-time interactions after substitution; γ_{11} is the degree to which ethnicity moderates the change rates, and γ_{12} captures the moderating effect of the intervention.

The incomplete binary predictors and interaction effects complicate missing data handling because existing maximum likelihood estimators are limited to normally distributed predictors (Shin & Raudenbush, 2023; Shin & Raudenbush, 2013). Applying the flowchart from Figure 3, a factored regression specification is the appropriate option. Currently, Bayesian estimation is the only way to deploy a factored specification for this analysis. Following Figure 2b, I adopt a factorization comprised of two submodels, each with distinct distributional assumptions. The Blimp application used in the Software Tutorials document automatically configures the distributions, and the user simply needs to specify the focal regression model in Equation 33.

For completeness, the symbolic representation for the underlying two-part factorization is as follows.

$$\begin{aligned} &f(PROBSOLVE|MONTH, HISPANIC, FRLUNCH, CONDITION) \\ &\times f(MONTH, HISPANIC, FRLUNCH, CONDITION) \end{aligned} \tag{34}$$

The first term corresponds to the univariate growth model from Equation 33. The second model is a multivariate normal distribution for the predictors. This submodel further disaggregates predictors into within-school and between-school parts, such that the level-1 predictors link to the level-2 predictors via latent cluster means (random intercepts). Finally, the dummy codes appear as normally distributed latent response variables in the predictor submodel (i.e., probit regression). Detailed descriptions of the model are available in the literature (Enders, 2022; Enders et al., 2020).

Example 17 from the Software Tutorials document provides annotated syntax and output files for the Bayesian analysis. Table 19 shows the Bayesian parameter summaries. The estimate (posterior median) and standard deviation columns describe the center and spread of the posterior distributions; although they make no reference to drawing repeated samples, they are analogous to frequentist point estimates and standard errors. The 95% credible interval columns give ranges that capture 95% of each parameter's distribution. These are akin to confidence intervals, but they describe a range of likely parameter values rather than the long-run behavior of intervals from different random samples.

TABLE 19. Bayesian Parameter Summaries From a Multilevel Growth Curve Model

Effect	Est.	Std. Dev.	2.5% LCL	97.5% UCL	χ^2	*p*
Intercept	49.36	0.31	48.77	49.97	26003.06	< .001
MONTH	0.27	0.06	0.16	0.39	21.26	< .001
MONTH × *HISPANIC*	0.37	0.05	0.26	0.47	47.32	< .001
MONTH × *CONDITION*	0.22	0.06	0.11	0.34	14.86	< .001
HISPANIC	1.36	0.30	0.77	1.95	20.53	< .001
FRLUNCH	–0.95	0.31	–1.57	–0.34	9.20	.002
CONDITION	–0.42	0.28	–0.97	0.14	2.25	.13
Intercept variance	11.29	0.81	9.78	12.95	--	--
Slope variance	0.11	0.03	0.06	0.18	--	--
Covariance	0.04	0.12	-0.21	0.25	--	--
Residual variance	12.57	0.27	12.04	13.12	--	--
R-square coefficients	0.11	0.01	0.09	0.13	--	--
R-square intercepts	0.44	0.02	0.41	0.47	--	--
R-square slopes	0.02	0.00	0.01	0.02	--	--

To reiterate, the posterior distributions characterize plausible *parameter values* that could have produced these data. From a practical perspective, the posterior summaries have interpretations that parallel a complete-data growth curve analysis. For example, the intercept coefficient (γ_{00} = 49.36) defines the expected baseline problem-solving score among non-Hispanic students attending comparison schools (i.e., the predicted value for a student with zero values on all regressors). Hispanic students scored higher at baseline (γ_{01} = 1.36), and students who received lunch assistance scored lower (γ_{02} = –0.95). Experimental and control school means were not significantly different at the study's outset. Turning to the growth rates, the *MONTH* coefficient (γ_{10} = 0.27) conveys the monthly change rate for non-Hispanic students attending comparison schools. The group-by-time interaction coefficients were both positive, indicating that Hispanic students and students attending intervention schools improved more rapidly (γ_{11} = 0.37 and γ_{12} = 0.22, respectively).

As explained previously, the Bayesian summaries also lend themselves to familiar hypothesis testing logic. Returning to the intervention coefficient, the credible interval limits spanning –0.97 to 0.14 included 95% of the area under the coefficient's posterior distribution. From this we can conclude that zero is a plausible parameter because the null value falls inside the interval ($p > .05$). As a second example, the condition-by-time interaction was significant ($p < .05$) because zero fell outside the 95% credible interval, which spanned from 0.11 to 0.34. One could also adopt a computational frequentism perspective that views the MCMC summaries as surrogates for frequentist point estimates and standard errors (Levy & McNeish, 2023). To this end, the rightmost pair of columns in Table 19 show Wald chi-square statistics (squared *z*-tests) and frequentist probability values (Asparouhov & Muthén, 2021).

The Blimp application invokes flat priors for the coefficients and common off-the-shelf priors for variances and covariances (Keller & Enders, 2021). The prior distributions influence variance parameters by altering the number of independent data points and/or the residual sums of squares and cross-products at each MCMC cycle. Roughly speaking, alternate prior distributions induce differences that are analogous to those between restricted and full information maximum likelihood (McNeish, 2017; D. McNeish & L. M. Stapleton, 2016; D. M. McNeish & L. M. Stapleton, 2016). The choice of prior distribution is potentially impactful to variance estimates when the number of level-2 units is small. To explore the sensitivity of the results to this issue, I fit the model with two alternate prior distributions (also off-the-shelf options that are invoked with a single keyword). Changes to the coefficients were generally inconsequential and in the second decimal place. The choice of prior produced small but noticeable differences in the random slope variance and intercept–slope covariance (e.g., the largest change shifted the random slope variance from 0.11 to 0.09). To put these changes in perspective, the R^2 values in the bottom row Table 19 express the random intercept and slope variance estimates as proportions of the total outcome variation (Rights & Sterba, 2019). Invoking alternate prior distributions changed these effect size estimates by less than one half percent. From a practical perspective, these differences seem relatively minor, and I would conclude that the prior distributions did not meaningfully impact on the results.

Finally, to apply model-based multiple imputation, one would simply save imputations from the MCMC algorithm for reanalysis with frequentist inference (Little & Rubin, 2020; Rubin, 1987). Example 17 from the Software Tutorials document provides the analysis scripts. I omit the pooled results because they were effectively equivalent to those in Table 19. To reiterate, you would expect this correspondence when applying Big Three approaches that use the same data and invoke the same assumptions.

7

Missing Not at Random Processes

The Big Three applications up to this point have assumed a conditionally missing at random process. This systematic mechanism requires that unseen score values carry no unique information about missingness beyond that contained in a model's observed data. This assumption cannot be tested or verified because it involves propositions about the unseen (latent) data. Assuming a conditionally missing at random process is probably reasonable for a broad range of educational research applications. Certainly, the vast majority of published Big Three applications adopt this assumption. An alternative assumption is that the unseen score values *do* carry unique information about missingness. This process is called missing not at random. Section 1.2 described this mechanism in detail, and this section describes analytic tools that invoke this alternate assumption about missingness.

7.1 Missing Not at Random Modeling Frameworks

Returning to Equation 3, the formal definition of a missing not at random process involves binary missing data indicators ($M = 0$ if a score is observed, and $M = 1$ if missing). Accordingly, the two major frameworks for modeling this process—selection models and pattern mixture models—introduce a submodel that describes the occurrence of missing data. To illustrate, consider a simple regression model where first grade broad reading predicts ninth grade reading proficiency. Further, suppose there is reason to believe that whether a student's ninth grade score is missing could depend on their unobserved reading proficiency score (e.g., students with the lowest proficiency were unable to complete enough exam items to produce a valid score). Figure 20 shows path diagrams depicting a basic selection model and pattern mixture model for this scenario. The selection model in Figure 20a features the missing data indicator as an additional outcome variable in a probit or logistic regression model. The pattern mixture model in Figure 20b instead treats the indicator as a grouping variable that predicts the outcome and moderates the association between

the two reading measures (i.e., subgroups defined by missing data pattern have different model parameters).

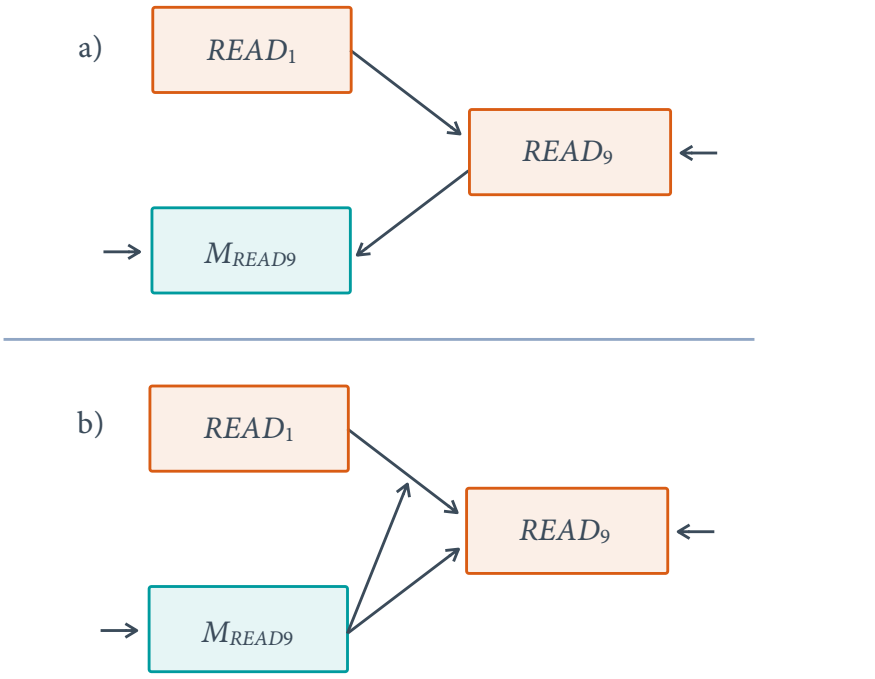


FIGURE 20. Panel (a) is a selection model where the binary missing data indicator is an additional outcome variable in a probit or logistic regression model. Panel (b) is a pattern mixture model where missingness is a grouping variable that predicts the outcome and moderates the focal association.

Selection and pattern mixture models have a long history in the literature, especially in the context of longitudinal data analyses (Diggle & Kenward, 1994; Hedeker & Gibbons, 1997; Little, 1995; Wu & Carroll, 1988). Numerous accessible tutorial papers document these approaches (Albert & Follmann, 2009; Enders, 2011; Feldman & Rabe-Hesketh, 2012; Muthén et al., 2011; Xu & Blozis, 2011), most with real-data software demonstrations[9]. Additionally, a good deal of recent methodological work enhances our understanding of these modeling frameworks, documenting their strengths and weaknesses (Gomer & Yuan, 2021; N. C. Gottfredson et al., 2014; Nisha C Gottfredson et al., 2014; Gottfredson et al., 2017; Sterba & Gottfredson, 2014; Yang & Maxwell, 2014; Yang et al., 2015).

[9] Software scripts for several selection and pattern mixture model applications are available at www.appliedmissingdata.com/analyses.

Selection and pattern mixture models are underutilized analytic tools that are more accessible than ever. Their application nevertheless requires caution. To begin, both modeling frameworks force the researcher to model associations between the data and missingness; selection models treat missing data indicators as outcomes, and pattern mixture models introduce them as predictors. The validity of the resulting estimates presupposes a correctly specified (or approximately so) missingness model. Unfortunately, there is no way to verify the correctness of the model, nor is it possible to formally compare competing missingness models (Molenberghs et al., 2008; Molenberghs et al., 1997; Sterba & Gottfredson, 2014; Verbeke & Molenberghs, 2000). Misspecifying the relations between the data and missingness could exacerbate rather than mitigate nonresponse bias. Although it is not immediately obvious, the observed data do not contain the information needed to estimate the association between an incomplete variable and its missing data indicator. For example, returning to Figure 20b, the arrow connecting the indicator to $READ_9$ is the group mean difference for students without ninth grade data. This path is not estimable because reading scores are completely missing for that subgroup. The corresponding arrow in Figure 20a is also inestimable. Selection models overcome this identification problem by invoking strict distributional assumptions, and pattern mixture models do so with additional input from user. Despite these potential pitfalls, missing not at random models are useful for contexts where missingness is plausibly related to the unseen score values. The subsequent examples demonstrate a sensitivity analysis that examines the stability of a regression model's parameter estimates across different assumptions about the missing data process.

7.2 Missing Not at Random Analysis Example 1: Selection Model

The first missing not at random analysis example illustrates a multiple regression paired with a selection model for the incomplete outcome. The analysis uses the `behaviorachievement.dat` data set from a longitudinal study that followed 138 students from primary through middle school. The file includes three annual assessments of broad reading and math achievement beginning in the first grade, seventh grade standardized achievement test scores taken from a statewide assessment, and a final measure of broad reading and math obtained in ninth grade. The data also contain teacher ratings of behavioral symptoms and learning problems were also obtained in the first grade. The Software Tutorials document provides additional information about this data set. Table 2 from Section 1.5 shows the specific variables for this analysis example.

The analysis model featured ninth grade broad reading scores regressed on three academic and behavioral measures collected in first grade: the broad reading composite, teacher-rated learning problems, and teacher-rated behavioral problems.

$$READ_9 = \beta_0 + \beta_1(READ_1) + \beta_2(LRNPROB_1) + \beta_3(BEHSYMP_1) + \varepsilon \quad (35)$$

The analysis also included second grade broad reading scores and seventh grade standardized test scores as auxiliary variables. These additional variables were included because they exhibited significant residual covariation with the analysis variables (see Section 1.5).

A selection model augments the focal regression with a probit or logistic model where the missing data indicator is the outcome. At a minimum, this model should include the variable hypothesized to follow a missing not at random process, ninth grade reading in this case. Selecting additional variables is difficult because myriad configurations of covariates could appear in the missingness model. Moreover, reducing nonresponse bias requires that the missingness model is approximately correctly specified. I offer a few practical guidelines. First, it is generally less harmful to include unnecessary predictors than to ignore important determinants of missingness. However, indiscriminately adding predictors can introduce substantial noise, reducing precision and power. Next, you should avoid situations where the focal regression and the missingness model share too many of the same variables. Ideally, the selection equation should include predictors of missingness that do not appear in the analysis model. This ideal can be difficult to achieve. Finally, building the missingness model by adding regressors and higher-order terms in a stepwise sequence is often a useful strategy (Ibrahim et al., 2005). Overfitting or grossly misspecifying a selection model often produces some combination of three symptoms: (a) standard errors increase dramatically (e.g., 50% or more) from one model to the next, (b) the missingness model produces an implausibly large pseudo R^2 value (e.g., .70 or larger), and (c) achieving convergence requires unusually long iterative sequences (e.g., MCMC burn-in periods of 100,000 iterations or more). A stepwise model-building procedure can reveal these issues.

Section 1.4 described a pattern mean difference approach to identifying potential correlates of nonresponse. To implement this strategy, you first create a binary missing data indicator *M*, in this case for the outcome variable. Treating the indicator as a grouping variable, you then examine whether the incomplete cases exhibit mean differences on other variables. Applying this strategy revealed three predictors of missingness at ninth grade: first grade learning problems, first grade behavior symptoms, and seventh grade math achievement. To avoid excessive overlap between the focal and missingness models, I used first grade learning problems as an additional regressor in the missingness model. I did not consider a model with interaction or polynomial terms because simple models with main effects can provide an adequate approximation (Gomer & Yuan, 2021; Ibrahim et al., 2005).

$$M^* = \gamma_0 + \gamma_1(READ_9) + \gamma_2(LRNPROB_1) + \epsilon \quad (36)$$

The M^* notation is consistent with probit regression, where the dependent variable is a normally distributed latent response score rather than the binary indicator. In a logistic regression, M^* corresponds to the log odds or logit, and the errors follow a standard logistic rather than standard normal distribution (Johnson & Albert, 1999). Figure 21 shows a path diagram of the analysis model, and it uses an oval to represent the latent M^* variable.

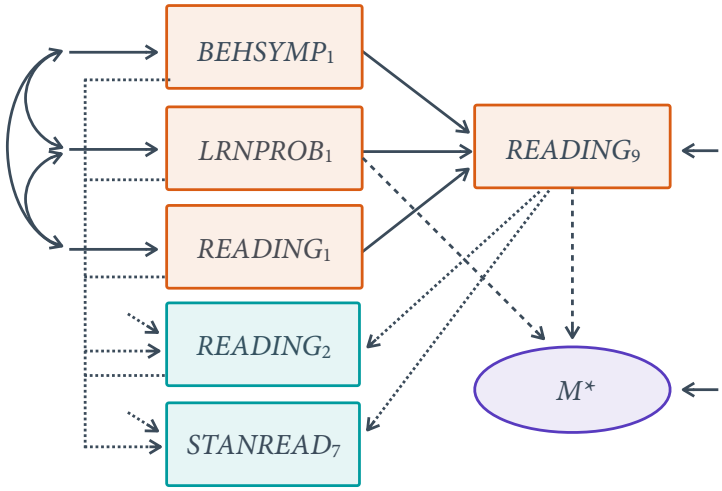


FIGURE 21. Solid lines denote the focal regression model parameters, and dotted lines are auxiliary variable parameters. Dashed lines pointing to M^* are the missingness model parameters.

Example 19 from the Software Tutorials document provides annotated syntax and output files for the analysis. The model did not exhibit the telltale symptoms of overfitting or misspecification. In particular, the MCMC algorithm converged quickly, and the missingness model's pseudo R^2 value of .27 was plausible. The leftmost panel of Table 20 shows the Bayesian analysis results that invoke the conditionally missing at random assumption (see Section 4.5). The middle panel shows the corresponding selection model estimates. The two sets of results were effectively identical. As noted previously, the selection model's validity hinges on a correctly specified missingness model with the right variables and correct functional forms. If we are willing to assume that Equation 36 is reasonably correct, we would conclude that the estimates were not sensitive to the assumptions about the missing data process—at least not this particular missing not at random process. In situations where two models produce discrepant results, there is no way of knowing whether the missing not at random analysis is preferable to a simpler analysis that assumes a conditionally missing at random mechanism. Although the prospect of such discrepancies may seem troubling,

they simply reflect different, plausible assumptions about the missing data process. Both sets of results are defensible and could be included in a research report (e.g., in an online supplement).

TABLE 20. Comparison of Missing Not at Random Regression Models

	CMAR		MNAR (SM)		MNAR (PMM)	
Effect	Est.	*SD*	Median	*SD*	Est.	*SD*
Intercept	66.01	6.05	66.41	5.99	66.57	6.14
$READ_1$	0.50	0.04	0.51	0.04	0.50	0.04
$LRNPROB_1$	–0.25	0.12	–0.25	0.12	–0.24	0.12
$BEHSYMP_1$	–0.18	0.11	–0.18	0.10	–0.18	0.11
Residual var.	91.26	12.80	91.78	12.79	91.69	13.05
R-square	.60	.05	.60	.05	.59	.05

Note. CMAR = conditionally missing at random, MNAR = missing not at random, SM = selection model, PMM = pattern mixture model.

7.3 Missing Not at Random Analysis Example 2: Pattern Mixture Model

Selection models treat missingness as an outcome. In contrast, pattern mixture models treat the missing data indicator as a predictor, such that missing data patterns form qualitatively different subgroups with distinct parameter values. Building on the focal regression model from Equation 35, this section illustrates a pattern mixture model that posits a missing not at random process where students with missing scores have a different reading proficiency mean in ninth grade.

The pattern mixture model below features the missing data indicator as an additional predictor variable.

$$READ_9 = \left[\beta_{0(com)} + \beta_{0(mis)}(M)\right] + \beta_1(READ_1) + \beta_2(LRNPROB_1) + \beta_3(BEHPROB_1) + \varepsilon \quad (37)$$

Focusing on the terms in the square brackets, $\beta_{0(com)}$ is the intercept for students with ninth grade reading scores, and $\beta_{0(mis)}$ is the mean difference for students with missing data. Collectively, these terms define a missing not at random process where the students with missing scores in ninth grade comprise a distinct subpopulation with a unique mean. Figure 22 shows a path diagram of the model, with the purple rectangle representing the missing data indicator. As shown in the Software Tutorials document, the analysis uses a sequential specification for the predictors where

the missingness indicator predicts behavioral symptom ratings, both variables predict learning problems, then all three variables predict first grade reading.

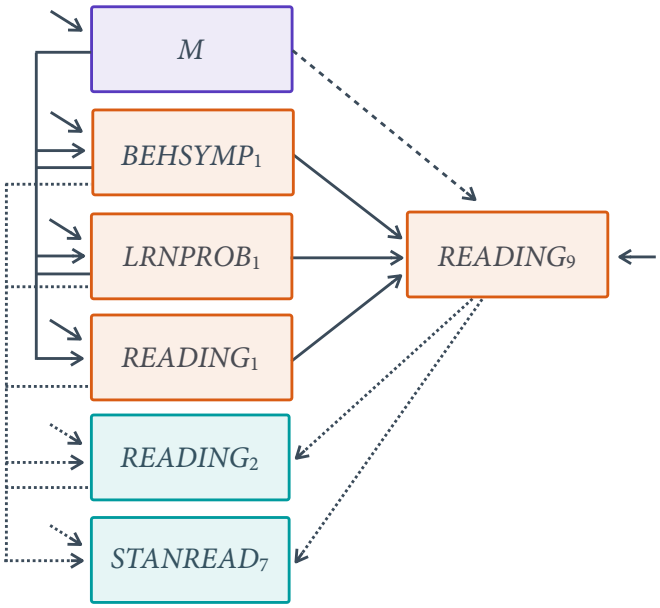


FIGURE 22. Solid lines denote the focal regression model parameters, and dotted lines are auxiliary variable parameters. The dashed line pointing to *M* to the dependent variable is a fixed parameter that captures the mean difference for the cases with missing data.

Equation 37 assumes that all associations—and thus regression slopes—are common to both groups. The model could be expanded to include pattern-specific slopes as well. For example, if there was reason to believe that the association between first and ninth grade scores differed by missingness pattern, and interaction could be added to the model as follows.

$$READ_9 = \left[\beta_{0(com)} + \beta_{0(mis)}(M)\right] + \beta_{1(com)}(READ_1) + \beta_{1(mis)}(M)(READ_1) + \beta_2(LRNPROB_1) + \beta_3(BEHPROB_1) + \varepsilon \quad (38)$$

Returning to Equation 37, the $\beta_{0(com)}$ and $\beta_{0(mis)}$ coefficients are not the parameters of interest because they define subgroup means. The population-level intercept estimate is a weighted average over the missing data patterns

$$\beta_0 = \pi_{(com)}\beta_{0(com)} + \pi_{(mis)}\left(\beta_{0(com)} + \beta_{0(mis)}\right) \quad (39)$$

where $\pi_{(com)}$ and $\pi_{(mis)}$ are weights that capture the subgroup (missing data pattern) proportions. The group-specific slopes from Equation 38 (i.e., $\beta_{1(com)}$ and $\beta_{1(com)}+ \beta_{1(mis)}$) would similarly pool into a population-level estimate of β_1.

Still focusing on Equation 37, the $\beta_{0(mis)}$ mean difference is not estimable because ninth grade reading scores are completely missing within the $M = 1$ subgroup. Similarly, both $\beta_{0(mis)}$ and $\beta_{1(mis)}$ in Equation 38 are inestimable. Pattern mixture models are uniquely challenging because the researcher needs to provide values for inestimable quantities. In this example, specifying a reasonable value for $\beta_{0(mis)}$ is vital because this parameter determines the strength of the missing not at random process. I adopt a simple approach that uses off-the-shelf effect size benchmarks to specify inestimable quantities (Enders, 2022; Section 9.7). Because $\beta_{0(mis)}$ is a mean difference, dividing this coefficient by the outcome variable's standard deviation (or residual standard deviation) gives a standardized mean difference like Cohen's (1988) d effect size. Cohen suggested the following benchmarks: $| d | > 0.20$ (small), 0.50 (medium), and 0.80 (large). This link to d allows us to specify the inestimable mean difference on the standardized metric and solve for $\beta_{0(mis)}$ as follows.

$$\beta_{0(mis)} = d \times \sqrt{\sigma_Y^2} \text{ or } d \times \sqrt{\sigma_\varepsilon^2} \tag{40}$$

The total variation or residual variation (σ_Y^2 and σ_ε^2, respectively) can be estimated directly from the data, so the researcher simply needs to provide a hypothesized (or purely hypothetical) effect size.

To illustrate the process of applying Equation 40, suppose I want to investigate the impact of a missing not at random mechanism where students with missing reading scores have lower ninth grade reading proficiency. Setting $d = -0.30$ conveys that the reading mean for the subpopulation with missing scores is lower by one third of a standard deviation unit. Absent specific knowledge about the size of the effect, one can perform a sensitivity analysis that implements a range of standardized effect sizes. Like the selection model, the pattern mixture model's validity hinges on a correctly specified missingness model. Even with simple rules for deriving inestimable quantities, there is no way to verify whether our choices are accurate. Despite this limitation, some methodologists prefer the pattern mixture approach because it transparently conveys a researcher's assumptions about the presumed missingness process (via their choices about the inestimable parameters).

Example 20 from the Software Tutorials document provides annotated syntax and output files for the analysis model from Equation 37. Following the previous illustration, I specified a standardized mean difference of –0.30. The software automatically estimates the reading standard deviation, fixes $\beta_{0(mis)}$ to the value from Equation 40, and computes the pooled estimate. The

rightmost panel of Table 20 shows the pattern mixture model parameter summaries. Aside from very minor differences in the intercept, the estimates were effectively identical to the conditionally missing at random and selection model results. As noted previously, the model's validity hinges on a correctly specified missingness process—in this case, one where the indicator group has a lower mean but identical associations. If we are willing to assume that Equation 37 is reasonably correct, we could conclude that the estimates were not sensitive to the assumptions about the missing data process. We could further probe the robustness of the estimates by modeling a more complex missingness process where the effect of a key predictor varies by group. Enders (2022; Section 9.7) demonstrates a sensitivity analysis that implements a model like the one from Equation 38.

8

Current Software Landscape

This closing section provides a summary of current software options. The proliferation of specialized R packages and numerous but disparate Big Three applications in commercial software precludes a complete summary. Instead, the goal of this section is to highlight the breadth of missing data handling tools that educational researchers currently have at their disposal. The Software Tutorials document demonstrates a few of these options.

Structural equation modeling software programs offer flexible facilities for implementing maximum likelihood missing data handling. Commercial platforms like SAS (CALIS; SAS Institute Inc., 2011), SPSS (AMOS; Arbuckle, 2019), and Stata (gllamm; Rabe-Hesketh et al., 2004) offer structural equation modeling modules, as do specialized programs like EQS (Bentler, 2000-2008), LISREL (Jöreskog & Sörbom, 2018), and M*plus* (Muthén & Muthén, 1998–2017). Structural equation modeling packages on the R platform include `OpenMx` (Boker et al., 2011), `PLmixed` (Rockwood & Jeon, 2019), and `lavaan` (Rosseel, 2012) with `semTools` (Jorgensen et al., 2022). M*plus*, `OpenMX`, gllamm, and `PLmixed` are notable because they implement factored regression specifications that accommodate mixtures of discrete and normally distributed variables. The R package `mdmb` (Lüdtke et al., 2020a) offers factored regression specifications for single-level regression models with interactive or nonlinear effects, and the moderated latent structural equation model (Kelava et al., 2011) facilities in M*plus* and the R package `nlsem` (Umbach et al., 2017) also accommodate certain types of interactive effects (Cham et al., 2017).

Turning to Bayesian estimation, M*plus* (Muthén & Asparouhov, 2012) offers a powerful feature set for multivariate normal data, but that multivariate focus carries the same limitations as it does with maximum likelihood. Recent extensions integrate factored specifications that accommodate interaction and nonlinear effects (Asparouhov & Muthén, 2021a). MCMC estimators abound on the R platform, many of which implement factored regression specifications with missing data. Options include `rstan` (Gelman et al., 2015; Guo et al., 2020), `rjags` (Plummer, 2019), `R2OpenBUGS` (Sturtz et al., 2019), `brms` (Bürkner, 2021), `blavaan` (Merkle & Rosseel, 2018),

mdmb (Grund et al., 2021a), nimble (de Valpine et al., 2017), and JointAI (Erler, 2021). The blavaan, brms, mdmb, and JointAI packages are among the most user-friendly options on this list.

The Software Tutorials document solely relies on the Blimp (Keller & Enders, 2021) application for MCMC estimation. Blimp's development was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D150056 and R305D190002 to UCLA. Blimp is an all-purpose data analysis and latent variable modeling program that offers factored regression specifications in a user-friendly application that requires minimal scripting and minimal knowledge about the Bayesian paradigm. The software accommodates missing data handling for normally distributed, binary, ordinal, multicategorical, count, and skewed continuous variables in data sets with up to three levels. The software is freely available for macOS, Windows, and Linux at www.appliedmissingdata.com/blimp. An R version the software called rblimp is also available at github.com/blimp-stats. A User Guide with dozens of analysis examples is available from the application's Help pull-down menu.

Turning to multiple imputation, commercial software packages typically offer agnostic multiple imputation facilities that implement joint model imputation or fully conditional specification. Blimp also offers single-level and multilevel fully conditional specification as well as model-based multiple imputation based on factored regression specifications. Not surprisingly, the R platform offers numerous multiple imputation options. A partial list includes the popular mice package (van Buuren & Groothuis–Oudshoorn, 2011), pan (Grund et al., 2016b; Schafer, 2018), jomo (Quartagno & Carpenter, 2020; Quartagno & Carpenter, 2016), amelia (Honaker et al., 2021), and smcfs (Bartlett et al., 2022; Bartlett et al., 2015). Finally, the R package mitml (Grund et al., 2023) provides a comprehensive toolkit for pooling estimates and conducting significance tests, as does the semTools package.

References

Alacam, E., Enders, C. K., Du, H., & Keller, B. T. (2023). A factored regression model for composite scores with item-level missing data. *Psychological Methods*, *30*(3), 462–481. https://doi.org/https://doi.org/10.1037/met0000584

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. https://doi.org/10.2307/2290350

Albert, P. S., & Follmann, D. A. (2009). Shared-parameter models. In G. Fitzmaurice, M. Davidian, G. Vebeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 433–452). Chapman & Hall.

Anderson, T. W. (1957). Maximum-likelihood estimates for a multivariate normal-distribution when some observations are missing. *Journal of the American Statistical Association*, *52*, 200–203. https://doi.org/10.2307/2280845

Andridge, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal*, *53*(1), 57–74. https://doi.org/10.1002/bimj.201000140

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–278). Lawrence Erlbaum Associates.

Arbuckle, J. L. (2019). *Amos 26.0 User's Guide*. In IBM SPSS.

Arminger, G., & Sobel, M. E. (1990). Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association*, *85*(409), 195–203. https://doi.org/10.2307/2289545

Aroian, L. A., Taneja, V. S., & Cornwell, L. W. (1978). Mathematical forms of the distribution of the product of two normal variables. *Communications in Statistics - Theory and Methods*, *7*(2), 165–172. https://doi.org/10.1080/03610927808827610

Asparouhov, T., & Muthén, B. (2021a). Bayesian estimation of single and multilevel models with latent variable interactions. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 314–328. https://doi.org/10.1080/10705511.2020.1761808

Asparouhov, T., & Muthén, B. (2021b). Expanding the Bayesian structural equation, multilevel and mixture models to logit, negative-binomial and nominal variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 622–637. https://doi.org/doi.org/10.1080/10705511.2021.1878896

Asparouhov, T., & Muthén, B. (2010). Multiple imputation with Mplus.

Asparouhov, T., & Muthén, B. (2021). Advances in bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 1–14.

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., van Buuren, S., & Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, *33*(2), 160–183. https://doi.org/10.1214/18-Sts646

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, *86*(4), 948–955. https://doi.org/DOI 10.1093/biomet/86.4.948

Bartlett, J., Keogh, R., & Bonneville, C. T. (2022). Package 'smcfcs'.

Bartlett, J. W., & Morris, T. P. (2015). Multiple imputation of covariates by substantive-model compatible fully conditional specification. *The Stata Journal*, *15*, 437–456. https://doi.org/10.1177/1536867X1501500206

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487. https://doi.org/10.1177/0962280214521348

Beale, E. M. L., & Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *37*(1), 129–145. https://doi.org/10.1111/j.2517-6161.1975.tb01037.x

Bentler, P. M. (2000-2008). *EQS 6 Structrual Equations Program Manual*. In Multivariate Software, Inc.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 651–675.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., & Bates, T. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306-317. https://doi.org/10.1007/s11336-010-9200-6

Bürkner, P.-C. (2021). Package 'brms'.

Carpenter, J. R., Bartlett, J. W., Morris, T. P., Wood, A. M., Quartagno, M., & Kenward, M. G. (2023). *Multiple Imputation and its Application (2nd. Ed.)*. Wiley.

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software*, *45*(5), 1–14. https://doi.org/10.18637/jss.v045.i05

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. Wiley.

Cham, H., Reshetnyak, E., Rosenfeld, B., & Breitbart, W. (2017). Full information maximum likelihood estimation for latent variable interactions with incomplete indicators. *Multivariate Behavioral Research*, *52*(1), 12–30. https://doi.org/10.1080/00273171.2016.1245600

Chan, K. W., & Meng, X.-L. (2021, December 30). Multiple improvements of multiple imputation likelihood ratio tests.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

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

Daniels, M. J., Wang, C., & Marcus, B. H. (2014). Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Biometrics*, *70*(1), 62–72. https://doi.org/10.1111/biom.12121

de Jong, R., van Buuren, S., & Spiess, M. (2016). Multiple imputation of predictor variables using generalized additive models. *Communications in Statistics-Simulation and Computation*, *45*, 968–985. https://doi.org/10.1080/03610918.2014.911894

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413. https://doi.org/10.1080/10618600.2016.1172487

Demirtas, H. (2017). A multiple imputation framework for massive multivariate data of different variable types: A Monte-Carlo technique. In D. G. Chen & J. D. Chen (Eds.), *Monte-Carlo simulation-based statistical modeling* (pp. 143-162). Springer Nature Singapore.

Demirtas, H., Freels, S. A., & Yucel, R. M. (2008). Plausibility of multivariate normality assumption when multiple imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, *78*, 69–84. https://doi.org/10.1080/10629360600903866

Demirtas, H., & Hedeker, D. (2008a). Imputing continuous data under some non-Gaussian distributions. *Statistica Neerlandica*, *62*(2), 193–205. https://doi.org/10.1111/j.1467-9574.2007.00377.x

Demirtas, H., & Hedeker, D. (2008b). Multiple imputation under power polynomials. *Communications in Statistics—Simulation and Computation*, *37*(8), 1682–1695. https://doi.org/10.1080/03610910802101531

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 (Statistical Methodology)*, *39*, 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific Reports*, *6*(1), 1–10. https://doi.org/10.1038/srep21689

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(1), 49–93. https://doi.org/10.2307/2986113

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. https://doi.org/10.1016/j.csda.2013.10.025

Du, H., Enders, C. K., Keller, B. T., Bradbury, T., & Karney, B. (2021). A Bayesian latent variable selection model for nonignorable missingness. *Multivariate Behavioral Research*, Advance online publication. https://doi.org/10.1080/00273171.2021.1874259

Edgett, G. L. (1956). Multiple regression with missing observations among the independent variables. *Journal of the American Statistical Association*, *51*(273), 122–131. https://doi.org/10.2307/2280808

Enders, C. K. (2002). Applying the Bollen-Stine bootstrap for goodness-of-fit measures to structural equation models with missing data. *Multivariate Behavioral Research*, *37*(3), 359–377. https://doi.org/10.1207/S15327906MBR3703_3

Enders, C. K. (2008). A note on the use of missing auxiliary variables in FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 434–448.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1–16. https://doi.org/10.1037/a0022640.supp

Enders, C. K. (2022). *Applied Missing Data Analysis* (2nd ed.). Guilford Press.

Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*, *Advanced online publication*, 1-18. https://doi.org/dx.doi.org/10.1037/met0000563

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, *19*(1), 39–55. https://doi.org/10.1037/a0035314

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and other nonlinear terms. *Psychological Methods*, *25*, 88–112. https://doi.org/10.1037/met0000228

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods*, *23*(2), 298–317. https://doi.org/10.1037/met0000148

Enders, C. K., Mistler, S. A., & Keller, B. T. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods*, *21*(2), 222–240. https://doi.org/10.1037/met0000063

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Erler, N. S. (2021). Package 'JointAI'.

Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2019). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, *28*, 555–568. https://doi.org/10.1177/0962280217730851

Erler, N. S., Rizopoulos, D., Rosmalen, J., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine*, *35*(17), 2955–2974. https://doi.org/10.1002/sim.6944

Feldman, B. J., & Rabe-Hesketh, S. (2012). Modeling achievement trajectories when attrition is informative. *Journal of Educational and Behavioral Statistics*, *37*(6), 703–736. https://doi.org/10.3102/1076998612458701

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement*, *45*(3), 225–245. https://doi.org/10.1111/j.1745-3984.2008.00062.x

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, *44*, 409–420. https://doi.org/10.1007/BF02296204

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/Doi 10.1214/06-Ba117a

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. https://doi.org/10.3102/1076998615606113

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. https://doi.org/10.1214/ss/1177011136

Goldstein, H., Carpenter, J., Kenward, M. G., & Levin, K. A. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, *9*(3), 173–197. https://doi.org/10.1177/1471082x0800900301

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(2), 553–564. https://doi.org/10.1111/rssa.12022

Gomer, B., & Yuan, K.-H. (2021). Subtypes of the missing not at random missing data mechanism. *Psychological Methods*, *26*, 559–598. https://doi.org/10.1037/met0000377

Gottfredson, N. C., Bauer, D. J., & Baldwin, S. A. (2014). Modeling change in the presence of non-randomly missing data: Evaluating a shared parameter mixture model. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 196–209. https://doi.org/10.1080/10705511.2014.882666

Gottfredson, N. C., Bauer, D. J., Baldwin, S. A., & Okiishi, J. C. (2014). Using a shared parameter mixture model to estimate change during treatment when termination is related to recovery speed. *Journal of Consulting and Clinical Psychology*, *82*(5), 813-827. https://doi.org/10.1037/a0034831

Gottfredson, N. C., Sterba, S. K., & Jackson, K. M. (2017). Explicating the conditions under which multilevel multiple imputation mitigates bias resulting from random coefficient-dependent missing longitudinal data. *Prevention Science*, *18*(1), 12–19. https://doi.org/10.1007/s11121-016-0735-3

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 80–100. https://doi.org/10.1207/S15328007sem1001_4

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev Sci*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing data designs in analysis of change. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). American Psychological Association.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323–343. https://doi.org/10.1037/1082-989X.11.4.323.supp

Grund, S., Lüdke, O., & Robitzsch, A. (2016a). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, *48*(2), 640–649. https://doi.org/10.3758/s13428-015-0590-3

Grund, S., Lüdke, O., & Robitzsch, A. (2016b). Multiple imputation of multilevel missing data: an introduction to the R package pan. *Sage Open*, *6*(4), 1–17. https://doi.org/10.1177/2158244016668220

Grund, S., Lüdke, O., & Robitzsch, A. (2017). Multiple imputation of missing data at level 2: A comparison of fully conditional and joint modeling in multilevel designs. *Journal of Educational and Behavioral Statistics*, *43*, 316–353. https://doi.org/10.3102/1076998617738087

Grund, S., Lüdke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111–149. https://doi.org/10.1177/1094428117703686

Grund, S., Lüdtke, O., & Robitzsch, A. (2019). Missing data in multilevel research. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook for multilevel theory, measurement, and analysis* (pp. 365–386). American Psychological Association.

Grund, S., Lüdtke, O., & Robitzsch, A. (2021a). Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, *53*, 2631–2649. https://doi.org/10.3758/s13428-020-01530-0

Grund, S., Lüdtke, O., & Robitzsch, A. (2021b). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*, *46*(4), 430–465.

Grund, S., Robitzsch, A., & Lüdke, O. (2023). Package 'mitml'. https://cran.r-project.org/web/packages/mitml/mitml.pdf

Guo, J., Gabry, J., Goodrich, B., & Weber, S. (2020). Package 'rstan'.

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*, 365–379.

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, *4*(1), 75–89. https://doi.org/10.1016/j.stamet.2006.03.002

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, *14*(2), 174–194. https://doi.org/10.2307/2527783

Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics*, *27*, 783–823. https://doi.org/10.2307/2528820

He, Y. L., & Raghunathan, T. E. (2009). On the performance of sequential regression multiple imputation methods with non normal error distributions. *Communications in Statistics—Simulation and Computation*, *38*(4), 856–883. https://doi.org/10.1080/03610910802677191

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*(1), 64–78. https://doi.org/10.1037//1082-989x.2.1.64

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Honaker, J., King, G., & Blackwell, M. (2021). Package 'Amelia'.

Howard, W. J., Rhemtulla, M., & Little, T. D. (2015). Using principal components as auxiliary variables in missing data estimation. *Multivariate Behavioral Research*, *50*(3), 285–299. https://doi.org/10.1080/00273171.2014.999267

Huang, L., Chen, M. H., & Ibrahim, J. G. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics*, *61*(3), 767-780. https://doi.org/10.1111/j.1541-0420.2005.00338.x

Humberg, S., & Grund, S. (2022). Response surface analysis with missing data. *Multivariate Behavioral Research*, *57*, 581–602. https://doi.org/10.1080/00273171.2021.1884522

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*(411), 765–769. https://doi.org/10.1080/01621459.1990.10474938

Ibrahim, J. G., Chen, M. H., & Lipsitz, S. R. (2002). Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, *30*(1), 55–78. https://doi.org/10.2307/3315865

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*(469), 332–346. https://doi.org/10.1198/016214504000001844

Ibrahim, J. G., Lipsitz, S. R., & Chen, M. H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(1), 173–190. https://doi.org/10.1111/1467-9868.00170

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. Springer.

Jöreskog, K. G., & Sörbom, D. (2018). *LISREL 10 for Windows [Computer software]*. In Scientific Software International.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). Package 'semTools'. https://cran.r-project.org/web/packages/semTools/semTools.pdf

Kelava, A., Werner, C. S., Schermelleh-Engel, K., Moosbrugger, H., Zapf, D., Ma, Y., Cham, H., Aiken, L. S., & West, S. G. (2011). Advanced nonlinear latent variable modeling: Distribution analytic LMS and QML estimators of interaction and quadratic effects. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 465-491.

Keller, B. T. (2015). *Three-level multiple imputation: A fully conditional specification approach* [Master's thesis, Arizona State University]. Proquest.

Keller, B. T., & Enders, C. K. (2021). Blimp user's guide (Version 3). www.appliedmissingdata.com/blimp

Keller, B. T., & Enders, C. K. (2023). An investigation of factored regression missing data methods for multilevel models with cross-level interactions. *Multivariate Behavioral Research*, *58*(52023), 938–963. https://doi.org/doi.org/10.1080/00273171.2022.2147049

Kleinke, K. (2017). Multiple imputation under violated distributional assumptions: A systematic evaluation of the assumed robustness of predictive mean matching. *Journal of*

Educational and Behavioral Statistics, *42*(4), 371–404. https://doi.org/10.3102/1076998616687084

Kleinke, K., & Reinecke, J. (2013). Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, *67*(3), 311–336. https://doi.org/10.1111/stan.12009

Kreft, I. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1

Lee, K. J., & Carlin, J. B. (2017). Multiple imputation in the presence of non-normal data. *Statistics in Medicine*, *36*(4), 606–617. https://doi.org/10.1002/sim.7173

Lee, M. C., & Mitra, R. (2016). Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. *Computational Statistics & Data Analysis*, *95*, 24–38. https://doi.org/10.1016/j.csda.2015.08.004

Lee, S.-Y., & Shi, J.-Q. (2000). Joint Bayesian analysis of factor scores and structural parameters in the factor analysis model. *Annals of the Institute of Statistical Mathematics*, *52*(4), 722-736. https://doi.org/10.1023/A:1017529427433

Lee, S.-Y., Song, X.-Y., & Tang, N.-S. (2007). Bayesian methods for analyzing structural equation models with covariates, interaction, and quadratic latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 404-434. https://doi.org/10.1080/10705510701301511

Levy, R., & McNeish, D. (2023). Perspectives on Bayesian inference and their implications for data analysis. *Psychological Methods*, *28*(3), 719–739. https://doi.org/doi.org/10.1037/met0000443

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, *86*(416), 1065–1073. https://doi.org/Doi 10.2307/2290525

Lipsitz, S. R., & Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, *83*(4), 916–922. https://doi.org/10.1093/biomet/83.4.916

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/Doi 10.2307/2290157

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*(431), 1112–1121. https://doi.org/Doi 10.2307/2291350

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.

Liu, J. C., Gelman, A., Hill, J., Su, Y. S., & Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika*, *101*(1), 155–173. https://doi.org/10.1093/biomet/ast044

Lord, F. M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association*, *50*(271), 870–876. https://doi.org/10.2307/2281171

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 x 2 taxonomy of multilevel latent contextual models: accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444–467. https://doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 201–229. https://doi.org/10.1037/a0012869

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165. https://doi.org/10.1037/met0000096

Lüdtke, O., Robitzsch, A., & West, S. G. (2020a). Analysis of interactions and nonlinear effects with missing data: A factored regression modeling approach using maximum likelihood estimation. *Multivariate Behavioral Research*, *55*(3), 361–381. https://doi.org/10.1080/00273171.2019.1640104

Lüdtke, O., Robitzsch, A., & West, S. G. (2020b). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods*, *25*, 157–181. https://doi.org/10.1037/met0000233

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *52*(5), 661–670.

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*, 20–35.

McNeish, D., & Stapleton, L. M. (2016). Modeling Clustered Data with Very Few Clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140.

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

Mealli, F., & Rubin, D. B. (2016). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, *103*(2), 491–491. https://doi.org/10.1093/biomet/asw017

Meng, X.-L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, *79*(1), 103–111. https://doi.org/10.2307/2337151

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Mislevy, R. J., & Wu, P.-K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing.*

Mistler, S. A., & Enders, C. K. (2011). An introduction to planned missing data designs for developmental research. In B. Laursen, T. Little, & N. Card (Eds.), *Handbook of developmental research methods* (pp. 742–754). Guilford Press.

Mistler, S. A., & Enders, C. K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics*, *42*(4), 432–466. https://doi.org/10.3102/1076998617690869

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(2), 371–388.

Molenberghs, G., Kenward, M. G., & Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, *84*(1), 33–44.

Montague, M., Krawec, J., Enders, C., & Dietz, S. (2014). The effects of cognitive strategy instruction on math problem solving of middle-school students of varying ability. *Journal of Educational Psychology*, *106*(2), 469–481.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(1), 1–13. https://doi.org/10.1186/1471-2288-14-75

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335.

Muthén, B., Asparouhov, T., Hunter, A. M., & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychological Methods*, *16*(1), 17–33. https://doi.org/10.1037/a0022634

Muthén, B., Muthén, L., & Asparouhov, T. (2016). *Regression and mediation analysis using Mplus*. Muthén & Muthén.

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

Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings from the sixth berkeley symposium on mathematical statistics and probability* (Vol. 1: Theory of Statistics, pp. 697–715). University of California Press.

Palomo, J., Dunson, D. B., & Bollen, K. (2007). Bayesian structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 163–188). Elsevier.

Plummer, M. (2019). Package 'rjags'.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Pritikin, J. N., Brick, T. R., & Neale, M. C. (2018). Multivariate normal maximum likelihood with both ordinal and continuous variables, and data missing at random. *Behavior Research Methods*, *50*(2), 490–500. https://doi.org/10.3758/s13428-017-1011-6

Quartagno, M., & Carpenter, J. (2020). Package 'jomo'.

Quartagno, M., & Carpenter, J. R. (2016). Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, *35*(17), 2938–2954. https://doi.org/10.1002/sim.6837

Quartagno, M., & Carpenter, J. R. (2019). Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical Journal*, *61*, 1003–1019. https://doi.org/10.1002/bimj.201800222

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. https://doi.org/10.1007/Bf02295939

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. https://doi.org/Doi 10.1037//1082-989x.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. (2019). *HLM for Windows [Computer software]*. In Scientific Software International.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037//1082-989x.5.2.199

Raykov, T. (2011). On testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 419–429. https://doi.org/10.1080/10705511.2011.582396

Raykov, T., & West, B. T. (2015). On enhancing plausibility of the missing at random assumption in incomplete data analyses via evaluation of response-auxiliary variable correlations. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 45–53. https://doi.org/10.1080/10705511.2014.937848

Reiter, J. P., Raghunathan, T. E., & Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology*, *32*, 143–150. https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9548-eng.pdf

Resche-Rigon, M., & White, I. R. (2018). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research*, *27*(6), 1634–1649. https://doi.org/10.1177/0962280216666564

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Rights, J. D., Preacher, K. J., & Cole, D. A. (2020). The danger of conflating level-specific effects of control variables when primary interest lies in level-2 effects. *British Journal of Mathematical and Statistical Psychology*, *73*, 194–211. https://doi.org/doi.org/10.1111/bmsp.12194

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*, 309–338. https://doi.org/dx.doi.org/10.1037/met0000184

Robitzsch, A., & Lüdke, O. (2023). Package 'mdmb'. https://cran.r-project.org/web/packages/mdmb/mdmb.pdf

Rockwood, N. J. (2020). Maximum likelihood estimation of multilevel structural equation models with random slopes for latent covariates. *Psychometrika*, *85*(2), 275–300. https://doi.org/10.1007/s11336-020-09702-9

Rockwood, N. J., & Jeon, M. (2019). Estimating complex measurement and growth models using the R package PLmixed. *Multivariate Behavioral Research*, *54*(2), 288-306. https://doi.org/10.1080/00273171.2018.1516541

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

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

SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide*. In SAS Institute Inc.

Savalei, V. (2010a). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods*, *15*(4), 352–367. https://doi.org/10.1037/a0020143

Savalei, V. (2010b). Small Sample Statistics for Incomplete Nonnormal Data: Extensions of Complete Data Formulae and a Monte Carlo Comparison. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 241–264. https://doi.org/10.1080/10705511003659375

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. https://doi.org/10.1080/10705510903008238

Savalei, V., & Rosseel, Y. (2022). Computational options for standard errors and test statistics with incomplete normal and nonnormal data. *Structural Equation Modeling*, *29*, 163–181. https://doi.org/10.1080/10705511.2021.1877548

Savalei, V., & Yuan, K. H. (2009). On the model-based bootstrap with missing data: Obtaining a p-value for a test of exact fit. *Multivariate Behavioral Research*, *44*(6), 741–763. https://doi.org/10.1080/00273170903333590

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

Schafer, J. L. (1999). Multiple imputation: A primer. *Stat Methods Med Res*, *8*(1), 3–15. https://doi.org/10.1177/096228029900800102

Schafer, J. L. (2001). Multiple imputation with PAN. In A. G. Sayer & L. M. Collins (Eds.), *New methods for the analysis of change* (pp. 355-377). American Psychological Association.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*(1), 19–35. https://doi.org/10.1111/1467-9574.00218

Schafer, J. L. (2018). Package 'pan'.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037//1082-989x.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545–571. https://doi.org/10.1207/s15327906mbr3304_5

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*(2), 437–457. https://doi.org/10.1198/106186002760180608

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, *12*, 1–13. https://doi.org/10.1186/1471-2288-12-46

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*, 764–774. https://doi.org/10.1093/aje/kwt312

Shin, Y. (2013). Efficient handling of predictors and outcomes having missing values. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment data analysis: Background, technical issues, and methods of data analysis* (pp. 451–479). Chapman & Hall.

Shin, Y., & Raudenbush, S. (2023). Maximum Likelihood Estimation of Hierarchical Linear Models from Incomplete Data: Random Coefficients, Statistical Interactions, and

Measurement Error. *Journal of Computational and Graphical Statistic*, 1–14. https://doi.org/doi.org/10.1080/10618600.2023.2234414

Shin, Y., & Raudenbush, S. W. (2007). Just-identified versus overidentified two-level hierarchical linear models with missing data. *Biometrics*, *63*(4), 1262–1268. https://doi.org/10.1111/j.1541-0420.2007.00818.x

Shin, Y., & Raudenbush, S. W. (2013). Efficient analysis of Q-level nested hierarchical general linear models given ignorable missing data. *International Journal of Biostatistics*, *9*(1), 109–133. https://doi.org/10.1515/ijb-2012-0048

Shin, Y. Y., & Raudenbush, S. W. (2010). A Latent Cluster-Mean Approach to the Contextual Effects Model With Missing Data. *Journal of Educational and Behavioral Statistics*, *35*(1), 26–53. https://doi.org/10.3102/1076998609345252

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). *Optimal design plus empirical evidence: Documentation for the "Optimal Design" software version 3.0.* In

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two-and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*(6), 605-627.

Sterba, S. K., & Gottfredson, N. C. (2014). Diagnosing global case influence on MAR versus MNAR model comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(2), 294–307. https://doi.org/10.1080/10705511.2014.936082

Sturtz, S., Ligges, U., & Gelman, A. (2019). R2OpenBUGS: A package for running OpenBUGS from R.

Su, Y.-S., Gelman, A. E., Hill, J., & Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, *45*, 1–31. https://doi.org/10.18637/jss.v045.i02

Takai, K., & Kano, Y. (2013). Asymptotic inference with incomplete data. *Communications in Statistics—Theory and Methods*, *42*(17), 3174-3190. https://doi.org/10.1080/03610926.2011.621577

Umbach, N., Naumann, K., Hoppe, D., Brandt, H., Kelava, A., & Schmitz, B. (2017). *Package 'nlsem'*. In https://cran.r-project.org/web/packages/nlsem/nlsem.pdf

Vach, W. (1994). *Logistic regression with missing values in the covariates*. Springer-Verlag.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242. https://doi.org/10.1177/0962280206074463

van Buuren, S. (2010). Item imputation without specifying scale structure. *Methodology*, *6*(1), 31–36. https://doi.org/10.1027/1614-2241/a000004

van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 173–196). Routledge.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall.

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. https://doi.org/10.1080/10629360600810434

van Buuren, S., & Groothuis–Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer-Verlag.

Vink, G., Frank, L. E., Pannekoek, J., & van Buuren, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, *68*(1), 61–90. https://doi.org/10.1111/stan.12023

von Hippel, P. T. (2004). Biases in SPSS 12.0 Missing Value Analysis. *The American Statistician*, *58*, 160–164.

von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*, 83–117. https://doi.org/doi.org/10.1111/j.1467-9531.2007.00180.x

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

von Hippel, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociological Methods and Research*, *42*, 105–138. https://doi.org/10.1177/0049124112464866

von Hippel, P. T. (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods and Research*, *49*, 699–718.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, *29*(28), 2920–2931.

Wilkinson and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/Doi 10.1037//0003-066x.54.8.594

Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*(1), 175–188. https://doi.org/10.2307/2531905

Xu, S., & Blozis, S. A. (2011). Sensitivity analysis of mixed models for incomplete longitudinal data. *Journal of Educational and Behavioral Statistics*, *36*(2), 237–256. https://doi.org/10.3102/1076998610375836

Yang, M., & Maxwell, S. E. (2014). Treatment effects in randomized longitudinal trials with different types of nonignorable dropout. *Psychological Methods*, *19*(2), 188. https://doi.org/10.1037/a0033804

Yang, M., Wang, L., & Maxwell, S. E. (2015). Bias in longitudinal data analysis with missing data using typical linear mixed-effects modelling and pattern-mixture approach: An analytical illustration. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 246–267. https://doi.org/10.1111/bmsp.12043

Yeo, I. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Yuan, K.-H. (2009). Normal distribution based pseudo ML for missing data: With applications to mean and covariance structure analysis. *Journal of Multivariate Analysis*, *100*(9), 1900–1918. https://doi.org/10.1016/j.jmva.2009.05.001

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200. https://doi.org/Doi 10.1111/0081-1750.00078

Yuan, K.-H., & Bentler, P. M. (2010). Consistency of normal distribution based pseudo maximum likelihood estimates when data are missing at random. *American Statistician*, *64*(3), 263–267. https://doi.org/10.1198/tast.2010.09203

Yuan, K.-H., & Savalei, V. (2014). Consistency, bias and efficiency of the normal-distribution-based MLE: The role of auxiliary variables. *Journal of Multivariate Analysis*, *124*, 353–370. https://doi.org/10.1016/j.jmva.2013.11.006

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for missing data with violation of distribution conditions. *Sociological Methods & Research*, *41*(4), 598–629. https://doi.org/10.1177/0049124112460373

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical and Physical Sciences*, *366*(1874), 2389–2403. https://doi.org/10.1098/rsta.2008.0038

Yucel, R. M. (2011). Random-covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling*, *11*(4), 351–370. https://doi.org/10.1177/1471082X1001100404

Zhang, Q., & Wang, L. (2017). Moderation analysis with missing data in the predictors. *Psychological Methods*, *22*(4), 649–666. https://doi.org/10.1037/met0000104

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, *25*, 2021–2035. https://doi.org/10.1177/0962280213511027