Chapter 2

MAXIMUM LIKELIHOOD AND MULTIPLE IMPUTATION MISSING DATA HANDLING: HOW THEY WORK, AND HOW TO MAKE THEM WORK IN PRACTICE

Timothy Hayes and Craig K. Enders

The goal of this chapter is to provide an overview of maximum likelihood estimation and multiple imputation, two major missing data handling strategies with strong support from the methodological literature. The theoretical and computational underpinnings of these methods were mostly developed in the 1970s and 1980s (Dempster et al., 1977; Rubin, 1976, 1987) and both became a practical reality in the 1990s when multiple imputation methods came online (Schafer, 1997) and structural equation modeling software packages began implementing maximum likelihood missing data estimators (Arbuckle, 1996). Both approaches have developed since then, and several important developments have appeared in the methodology literature since the first edition of this handbook.

Many missing data methods have been proposed and investigated in the literature, and there is now broad awareness that older approaches like deleting incomplete data records or filling in missing data with a single prediction (e.g., mean imputation, regression imputation) are seriously flawed. For example, the American Psychological Association's Task Force on Statistical Inference characterized deletion as "among the worst methods available for practical applications" (Wilkinson & Task Force on Statistical Inference, American Psychological Association, Science Directorate, 1999, p. 598). Because descriptions of these older methods abound in the literature, we focus strictly on the two major missing data handling frameworks that have broad theoretical and empirical support: maximum likelihood estimation and multiple imputation.

To set the stage for the material that follows, imagine a researcher who has collected data for a particular study and decided in advance on an analysis model (or set of models) that they intend to run—say, a linear regression analysis—but, upon sitting down to analyze the data, finds themselves confronted with missing values on one or more variables in their intended model. If the researcher is aware that discarding cases will lead to inaccurate model results (e.g., estimated regression coefficients that may be too large or too small, potentially inflated Type I or Type II error rates), they must now think carefully about what should be done instead. At the broadest level, our goal in this chapter is to help empower researchers facing this prototypical dilemma to make informed choices about how to use and configure modern missing data handling approaches in a manner tailored to the unique features of their intended analyses. A major theme of this chapter is that because each analysis one intends to run may involve different sets of variables containing missing data potentially

https://doi.org/10.1037/0000320-002
APA Handbook of Research Methods in Psychology, Second Edition: Vol. 3. Data Analysis and Research Publication, H. Cooper (Editor-in-Chief)

attributable to different causes and because each analysis may, further, contain unique features (e.g., interactions, categorical predictors, outcomes) that require specialized approaches to estimation, a one-size-fits-all approach to missing data handling is rarely appropriate; instead, missing data handling must generally be customized to the needs of each specific analysis.

To help researchers understand how to choose missing data handling methods appropriate for their intended analyses, we begin with a detailed review of the mechanisms that might generate missing data on the variables in a given model. We then provide overviews of missing data handling in the maximum likelihood and multiple imputation frameworks, describe the foundations of these methods as well as more recent extensions, and apply both methods to an illustrative example using simulated data based on a study of $N = 300$ chronic pain patients. Following this, we provide a brief overview of models for data that are missing not at random and close with a set of recommendations for reporting the results of one's missing data analyses.

MISSING DATA MECHANISMS

The manner in which missing data affect the accuracy and precision of one's model estimates depends on the reason why the data are missing. Broadly, scores on a given variable could be missing for purely haphazard reasons unrelated to the data, they could be missing systematically due to scores that are observed in the data, or they could be missing systematically due to the unseen values themselves (Little & Rubin, 1987; Rubin, 1976, 1987). Each of these *missing data mechanisms* (Rubin, 1976) carries distinct implications for how best to handle missing data as well as the potential consequences of mishandling it.

To build an intuition for these concepts, consider a researcher interested in understanding the relationship between levels of self-reported chronic pain (measured with a trichotomous indicator coded: -1 = low pain, 0 = moderate pain, +1 = severe pain) and levels of psychosocial disability due to pain (a construct capturing pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, and so on) in a sample of chronic pain patients. Now, imagine that not all chronic pain patients in the sample choose to report their psychosocial disability scores. As summarized in Table 2.1, the missing data mechanism that applies in this case depends on whether the probability of missing data on disability (y) is systematically related to the patients' chronic pain levels (x) and whether, within each level of chronic pain, the probability of missing data is related to the unseen values of disability (y). The combinations of these two scenarios in the four cells define three missing data processes or mechanisms: missing completely at random, missing (conditionally) at random, and missing not at random (with focused and diffuse subtypes).

Correspondingly, the four panels of Figure 2.1 use simulated data to illustrate how the conditional distributions of the disability scores within each level of chronic pain might appear under each missing data mechanism from Table 2.1. Observed scores are shown as circles and missing (unseen) values are asterisks. Before proceeding, it is important to emphasize that in Figure 2.1 we present the distributions of missing values alongside the observed values for purely pedagogical reasons: to illustrate the underlying missing data theory by providing a "God's-eye view" of how

TABLE 2.1

Rubin's (1976) Missing Data Mechanisms for a Simple Regression of y on x Classified by Whether or Not the Missing Data Are Observed at Random and Missing at Random

Question 2: Is the probability of missing data on *y* equal for all possible values of *y* after conditioning on *x*?	Question 1: Is the probability of missing data on *y* equal for all possible observed values of *x*?	
	Yes	No
Yes	MCAR	MAR
No	Focused MNAR	Diffuse MNAR

Note. MCAR = missing completely at random, MAR = missing at random, MNAR = missing not at random.

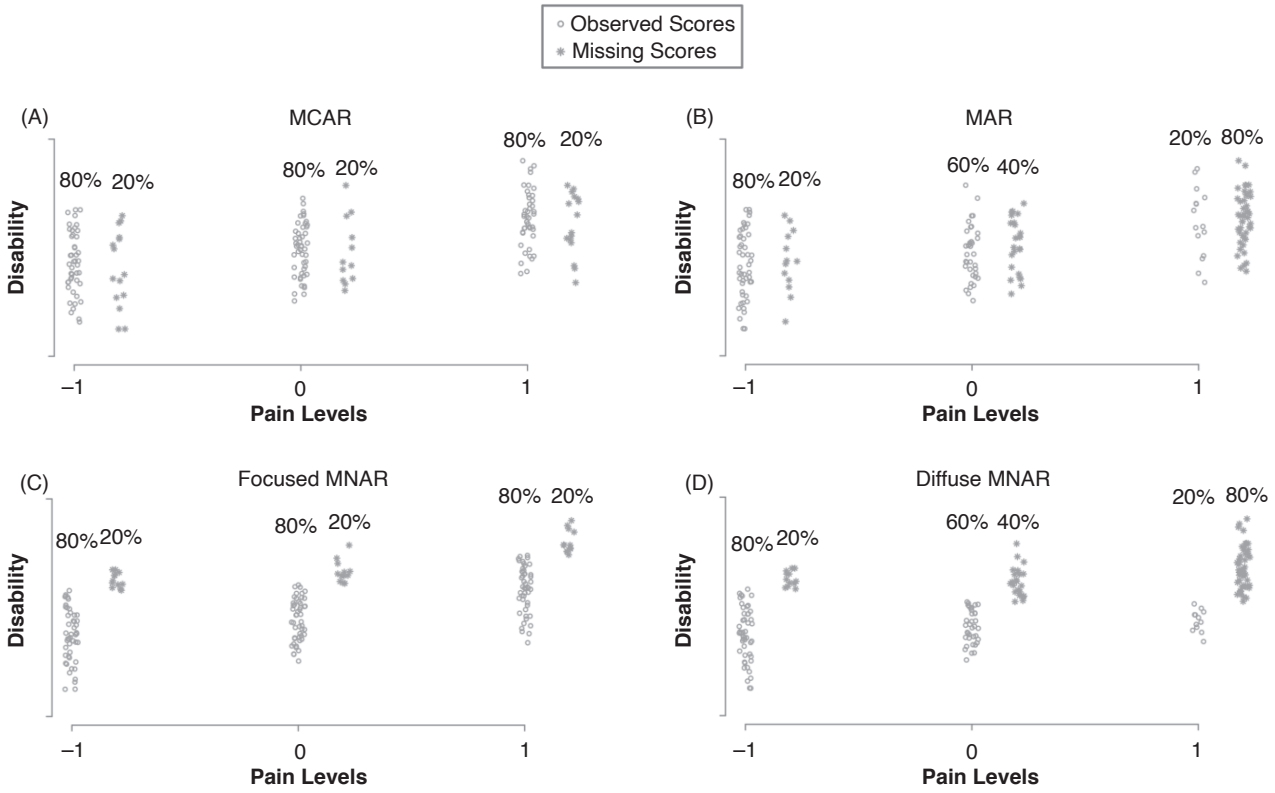

FIGURE 2.1. Missing data mechanisms illustrated using simulated data on chronic pain and psychosocial disability.

the distributions of unseen values might appear under each missing data mechanism if they could be observed. With real data, however, one would only have access to the observed values (circles) and would never know the distributions of the would-be scores (asterisks) that participants might have reported if their data were not missing.

Returning to the chronic pain scenario, what if the patients who opt not to report their disability scores are a random subsample of patients with respect to both their pain levels and their disability levels—as would occur, for example, if a glitch in the survey software used to collect data caused the software to randomly crash when assessing disability levels? Figure 2.1A shows the conditional distributions of both observed (circles) and missing (asterisks) disability scores within each level of chronic pain in this type of scenario. Here, we see that the percentage of patients opting not report their disability scores is equal for all observed pain levels (missing values are random with respect to *x*), and the distributions of the observed and unseen values are roughly the same (i.e., have the same center and spread) within each level of chronic pain. Because in this situation the probability of missing data is random with respect to both the observed and missing values, the pattern of missing data is, in essence, as random as it could ever possibly be—hence, the designation *missing completely at random* (MCAR, e.g., Little & Rubin, 1987; Rubin, 1987). This condition is reflected in the upper left of Table 2.1.

By contrast, what if individuals experiencing higher levels of chronic pain were systematically less likely to report their levels of disability (e.g., because their pain levels interfered with their ability to participate)? Figure 2.1B depicts such a scenario. Here, the percentages of missing scores on disability (presented above each distribution of asterisks) are systematically higher at higher levels of chronic pain (i.e., missing values are no

longer random with respect to *x*), but the distributions of the observed and unseen values are once again roughly the same within each level of chronic pain (i.e., missing *y* values are randomly distributed after conditioning on *x*). Although such a mechanism can no longer be said to be completely random, the missing data are said to be *missing at random* (MAR) because the unseen (missing) *y*-values (disability scores) remain randomly distributed within each level of *x* (chronic pain). This condition is designated in the upper right of Table 2.1.

Typical applications of the maximum likelihood and multiple imputation methods described below operate under the assumption that all missing data are MAR. Importantly, under a MAR mechanism, the observed data distributions (circles) act as reasonable proxies for what the complete data distributions (circles + asterisks) would have been, and inferences based on the observed data should yield accurate estimates. In line with this idea, maximum likelihood missing data handling utilizes all of the observed data to help identify the optimal parameter estimates, and multiple imputation uses the conditional distributions of the observed data as the basis for filling in the missing values. We describe these methods in detail in later sections.

Finally, consider next what would happen if patients with higher levels of disability (*y*) were less likely to report their disability scores than those with lower levels of disability, as depicted in Figure 2.1C and D. As the bottom row of Table 2.1 implies, any time that the probability of missing data is unequal across values of *y*, the data are considered *missing not at random* (MNAR; see Little & Rubin, 1987; Rubin, 1976). Following Gomer and Yuan (2021), the exact type of MNAR mechanism depends on whether the probability of missingness is also related to the observed values of *x*. The scenario depicted in Figure 2.1C, in which the probability of missing data is systematically related only to the unseen scores on *y* (disability) but not to the observed values of *x* (pain), is called a *focused MNAR mechanism*, whereas the scenario depicted in Figure 2.1D, in which the probability of missing data is systematically related to both the unseen scores on *y* (disability) and the observed values of *x* (pain,) is called a *diffuse MNAR mechanism*. Whether focused or diffuse, under an MNAR mechanism the observed data distributions (circles) do not act as reasonable proxies for what the complete data distributions (circles + asterisks) would have been, and inferences based exclusively on these observed distributions are no longer be guaranteed to yield accurate results. We briefly discuss approaches to handling MNAR missing data later in the chapter.

Auxiliary Variables

In the previous section, we followed an example in which a researcher was interested in the relationship between chronic pain (*x*) and disability (*y*), and we considered the consequences that might result when missing disability scores (*y*) were generated by a completely random process, a systematic process related to the observed values of chronic pain (*x*), or a systematic process related to the unseen values of disability (*y*) itself (or a combination of both). But what would happen if missing data were caused by a variable *other* than pain or disability—that is, a measured variable in the data that isn't part of the main analysis plan?

Continuing with the bivariate example depicted in Figure 2.1, what if stress was correlated with both pain and disability and was also the cause of missing data, such that individuals with higher levels of stress were less likely to report their disability scores than individuals with moderate or low levels of stress? In such a scenario, the probability of reporting one's disability score would be completely random within levels of stress (that is, the missing data would be MAR, if one conditioned on stress), but it would not necessarily be so within levels of chronic pain. Although pain and stress may be correlated, because they are not perfectly correlated, the conditional distributions of disability within levels of stress are not identical to the conditional distributions within levels of pain depicted in Figure 2.1B. For this reason, the probability of missing data on disability within each level of chronic pain does not necessarily remain

equal across all values of disability, as required by a MAR process (i.e., Question 2 in Table 2.1).

Omitting an important determinant of missingness that is also correlated with the main analysis variables leads to what may be termed an *MNAR-by-omission process* (see Collins et al., 2001, for a detailed discussion of this topic; see also Enders, 2021, Chapter 1, for this terminology). Applied to our bivariate example, omitting the stress scores from the analysis requires the relation between pain severity and disability to absorb the entire influence of the stress scores on missingness. Depending on the magnitude of the correlations, the net result is that the analysis partially rather than fully conditions on the determinants of missingness. Because the distinction between MAR and MNAR data is defined in terms of the probability of observing each value of y after conditioning on x, it follows that this *MNAR-by-omission* mechanism is more severe to the extent that the missing data cause (e.g., stress) is highly correlated with the residuals of y (disability) after conditioning on x (chronic pain; see Collins et al., 2001; Raykov & West, 2016)—in effect, there is more information about the distribution of missing values being omitted from the analysis, leading to greater misspecification and nonresponse bias.

In order to avoid preventable MNAR-by-omission mechanisms and increase the plausibility that the data are MAR, researchers must decide which *auxiliary variables* (e.g., demographic variables, participants' responses to additional psychological questionnaires) from outside of the substantive model of interest should ultimately be included to aid missing data estimation, raising an important question as to how one can best approach this task (Collins et al., 2001). Ideally, plausible determinants of missing data (e.g., stress in the bivariate example) could be identified through a combination of substantive theory, practical experience, and data exploration. As noted previously, the goal is to identify variables that both predict missingness and have salient correlations (or more accurately, semipartial or residual correlations) with the incomplete analysis variables.

Although a variety of statistical approaches might potentially be applied to this task, one particularly useful search strategy is to include auxiliary variables that exhibit at least moderate (e.g., $|r| = .3$) correlations with the residuals of each incomplete variable (i.e., external variables with moderate semipartial correlations). To help accomplish this, Raykov and West (2016) developed a latent variable approach to estimating these correlations within the *structural equation modeling* (SEM) framework. As a simpler alternative to this method, users could first estimate the matrix of bivariate correlations between the substantive model variables and candidate auxiliary variables using a modern approach to missing data handling such as the maximum likelihood or multiple imputation procedures described below and then scan these bivariate correlations for entries greater than .3 in absolute value. Screening based on correlations is an effective strategy for identifying auxiliary variables because strong predictors of missingness (i.e., variables on which participants with and without missing values differ) are only capable of introducing nonresponse bias if they are also correlated with the analysis variables.

MAXIMUM LIKELIHOOD ESTIMATION

The goal of *maximum likelihood* (ML) *estimation* is to identify the model parameter values most likely responsible for producing the data. The missing data handling aspect of maximum likelihood happens behind the scenes as a part of the same process. Importantly, ML estimation does not discard incomplete data records nor does it impute them. Rather, when confronted with missing values, maximum likelihood uses the normal curve to deduce the missing parts of the data as an optimization algorithm iterates to a solution.[1] The resulting parameter values are those with maximum support from (or best fit to) the observed data. To understand how this procedure accommodates missing data, it is first necessary to understand a bit about how ML estimation

[1]Technically, the estimator marginalizes over the missing values.

works in the context of complete data, beginning with what exactly is meant by the concept of a *likelihood*, a quantity that essentially functions as a measure of fit between a person's data and a set of model parameters.

Likelihoods are closely related to probabilities. One way of understanding probabilities is as the relative frequencies with which one would expect to observe a set of events or score values over many repeated trials (e.g., Hoel, 1984, p. 8). For example, if one flipped a fair coin many times, one would expect roughly half of those flips to come up heads, assuming that the true population proportion was .50. The idea of a likelihood inverts this logic by asking what population parameters could most plausibly have generated the observed data (i.e., given an observed sample of 50 heads and 50 tails from 100 flips, what population proportion of heads is most likely to have generated it?).

These same principles can be applied to continuous statistical distributions such as the normal curve displayed in Figure 2.2A. The relative probability of obtaining a particular score from a normal distribution with a known mean and standard deviation can be calculated using the *univariate normal density function*. More formally, the probability density, p_i, that individual i's score in a data set, y_i, was obtained from a univariate normal distribution with mean μ and standard deviation σ, may be written:

$$p_i(y_i|\mu,\sigma^2) = \text{constant} \times \exp\left(-.5 \times \frac{(y_i-\mu)^2}{\sigma^2}\right)$$
$$= \text{constant} \times \exp(-.5(z_i^2)) \quad (1)$$

where exp( ) indicates the exponential function and "constant" represents a collection of terms that ensure that the area of the normal density curve sums (integrates) to 1. Because these constant terms do not change with new input to the function, they can be ignored in order to simplify our present discussion. Importantly the vertical pipe in the notation $p_i(y_i|\mu, \sigma^2)$ makes clear that Equation 1 returns the relative probability of a given score, y_i, conditional or dependent on a known mean and variance, μ and σ^2. Visually, the probability density, p_i is the height of the

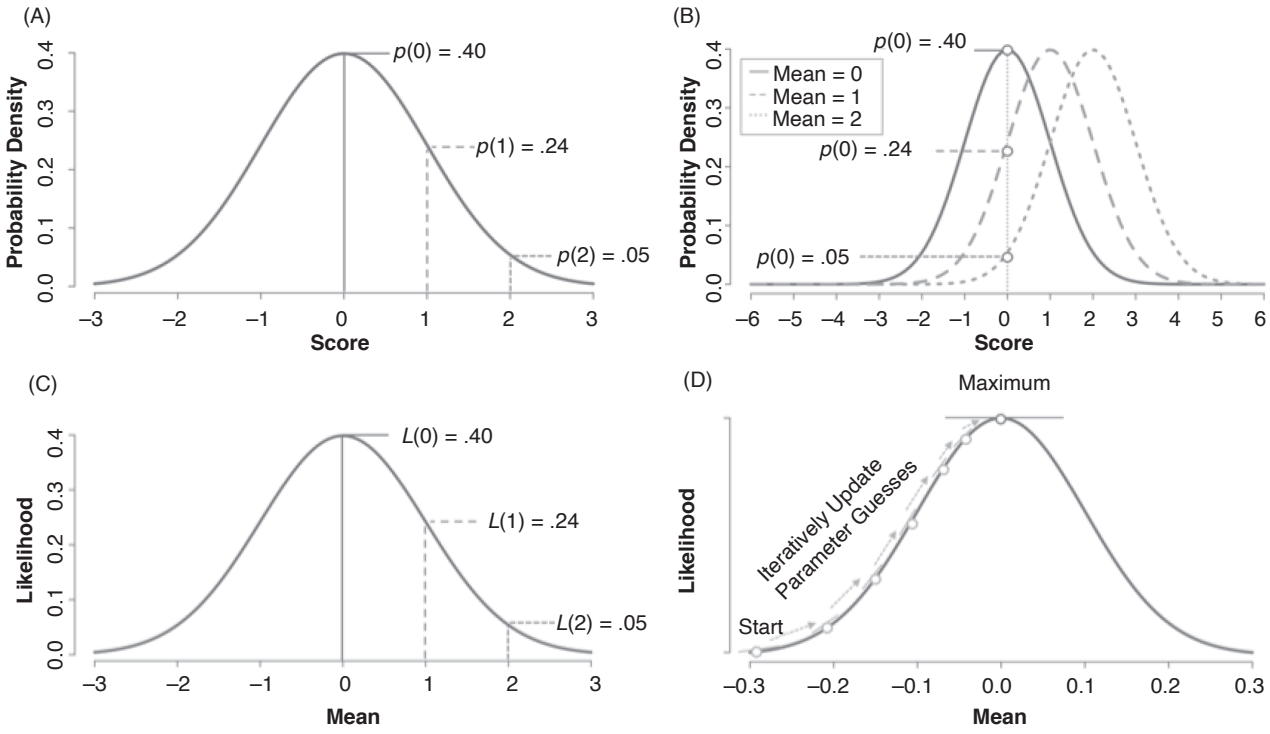

FIGURE 2.2. Illustrations of univariate normal density curves and univariate normal likelihood functions.

normal distribution at a particular value of y_i. The bottom row of the expression makes clear that the term in the exponent contains a squared *z*-score. In Equation 1, smaller squared *z*-scores associated with scores closer to the mean result in larger probability densities than those of larger squared *z*-scores further from the center.

To illustrate, Figure 2.2A graphs the density function from Equation 1 applied to scores from a standard normal distribution. Panel A shows the probability densities associated with three different scores in a standard normal distribution: a score at the mean (0), one standard deviation (*SD*) above the mean (1), and two standard deviations above the mean (2). As the figure shows, in the context of a standard normal distribution, the relative probability of obtaining a score at the mean, $p(0) = .40$, is roughly twice as large (i.e., twice as high in vertical elevation) as the relative probability associated with obtaining a score one *SD* above the mean, $p(1) = .24$, and it is eight times larger than the relative probability associated with obtaining a score two *SD*s above the mean, $p(2) = .05$.

To move from probability densities toward the idea of likelihoods, Figure 2.2B shows the probability densities associated with drawing a score of approximately 0 from three hypothetical normal curves with the same variance ($\sigma^2 = 1$) but different means. Panel B highlights that a score of 0 has a high relative probability of being drawn from a distribution with mean 0, $p(0) = .40$, a comparatively lower relative probability of being drawn from a distribution with mean 1, $p(0) = .24$, and a very low relative probability of being drawn from a distribution with mean 2, $p(0) = .05$. Although these statements concern relative probabilities rather than likelihoods, we might utilize the knowledge that drawing a score approximately equal to 0 is a far more probable outcome for some population means than for others: as noted previously, a likelihood reverses this logic and asks, what population mean is most likely to have produced the data on hand (i.e., a particular score of 0). Based on this single observation, we can infer that $\mu = 0$ is most likely responsible for the datum.

A univariate normal likelihood has the same formula as Equation 1. The key difference is that after obtaining a sample of data, the *y* values become known, and the parameters become unknowns. Symbolically, we can write this as $L_i(\mu, \sigma^2|y_i)$. To illustrate, Figure 2.2C graphs the likelihood of different parameter values based on a single observed data point, $y_i = 0$. Because the parameters and data have switched roles, the horizontal axis now lists the unknown parameter values instead of hypothetical score values. Importantly, the meaning of the vertical coordinates has also changed; the likelihood is no longer a relative probability but an index of support for different parameter values. Panel C shows that a score of $y_i = 0$ has the most support for a population mean of 0 and decreasing support for a mean of 1 and 2.

Shifting from individual scores to an entire data set, the overall likelihood, *L*, for a sample of *N* individuals can be calculated as the product of the individual likelihoods. The resulting graph would look like Figure 2.2C, but the vertical coordinates would represent the entire sample's support for different unknown parameter values. Note that, in practice, ML optimization algorithms work with the natural logarithm of the overall likelihood in order to turn the product of probabilities into a more mathematically tractable sum. The purpose of our discussion of ML estimation here is to provide readers with a broad conceptual overview of the logic of the method, however, so we omit that detail here (for further information on the details of ML, readers are encouraged to consult Eliason, 1993).

In practice, ML estimation is typically implemented using *iterative optimization algorithms*. Iterative optimization algorithms begin with a blunt guess—a *starting value*—for a proposed parameter estimate and iteratively improve upon this guess until finding the estimate(s) that maximize the likelihood of producing the observed data (i.e., the parameter values that have the most support from the data). This process is visualized in Figure 2.2D for a sample of $N = 100$ standard normal observations, with circles representing sequential iterative updates and tangent lines

representing the slope of the curve at each guess. The flat tangent line at the peak of the likelihood function is a mathematical indication that the optimizer has found the optimal estimate for the data.

Building up to a bivariate analysis example, it turns out that the univariate normal density function of Equation 1 is all that is needed to estimate a linear regression model with complete data. To understand how this is so, examine Figure 2.3, which depicts a hypothetical bivariate scatterplot from the regression of disability on chronic pain levels. The normal curves convey the distributions of the disability scores at the three different pain levels. The regression line cuts through these distributions at their (conditional) means, which are just the predicted scores shown as black dots. As the annotations in the figure imply, one way of understanding the regression of disability on pain levels is to view the analysis as attempting to estimate a single regression intercept, single regression slope, and single estimated residual variance that best capture the conditional means ($\hat{y}$ values) and constant (residual) variance of the conditional distributions of disability for individuals with each level of self-reported pain. With this in mind, we can use ML estimation to estimate a bivariate regression model by setting $\mu = \hat{y}_i$ and $\sigma^2 = \sigma^2_{\hat{e}}$ in Equation 1 as follows:

$$\begin{aligned} L_i &= \text{constant} \times \exp\left(-.5 \times \frac{[y_i - (b_0 + b_1 X_1)]^2}{\sigma^2_{\hat{e}}} \right) \\ &= \text{constant} \times \exp\left(-.5 \times \frac{(y_i - \hat{y}_i)^2}{\sigma^2_{\hat{e}}} \right) \\ &= \text{constant} \times \exp(-.5(z_i^2)). \end{aligned} \quad (2)$$

Following the earlier example, the likelihood represents a single score's support for the three unknown parameter values in the equation (b_0, b_1, and $\sigma^2_{\hat{e}}$), and the entire sample's support for different parameter values combines N individual likelihoods. Although there are now multiple parameters, the iterative optimization algorithm updates each parameter one at a time following the same process depicted in Figure 2.2D. Note that, although both x_i and y_i appear in Equation 2, it is only the univariate distribution of y_i that determines the likelihood; the predictor values, x_i, are simply considered to be fixed constants used to define the conditional mean of each y_i. Applied to Figure 2.3, the analysis makes no assumptions about the distribution of pain severity nor does it make any attempt to estimate the parameters of such a distribution.

Moving to missing data, how can we compute Equation 2 if either x_i or y_i is missing? The fundamental idea behind *full information maximum likelihood* (FIML, Arbuckle, 1996) estimation procedures is to utilize all available data on all

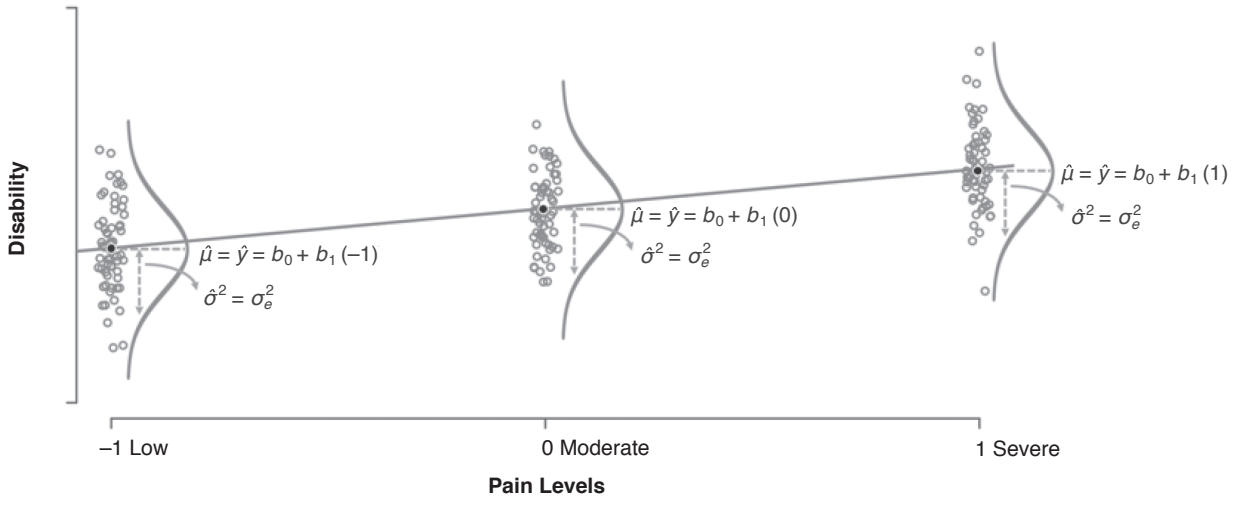

FIGURE 2.3. Regression of psychosocial disability on chronic pain with superimposed conditional normal distributions.

model variables to inform the final model estimates. However, to do so requires switching from a univariate estimation procedure whose only goal is to estimate the parameters associated with y_i to a *multivariate* estimation procedure that also considers the parameters of x_i as estimated quantities in the analysis (i.e., instead of treating the predictor values as fixed, x is also assigned a distribution). We can accomplish this by first incorporating the observed values of both the predictors and outcome(s) into a multivariate normal density function, and then quantifying their support for a set of proposed parameters by rewriting this density as a *multivariate normal likelihood function*:

$$\begin{aligned} &L_i(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{y}_i) \\ &\quad = \text{constant} \times \exp\left[-.5 \times (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu})\right] \\ &\quad = \text{constant} \times \exp(-.5(z_i^2)). \end{aligned} \tag{3}$$

where "constant" again indicates a collection of terms that can be ignored for our present purposes, $\mathbf{y}_i$ is a vector of scores on k multivariate outcomes for person i, $\boldsymbol{\mu}$ is a vector of k model-implied means, and $\boldsymbol{\Sigma}$ is a $k \times k$ model-implied variance–covariance matrix corresponding to the same multivariate outcomes. In the bottom expression, the z_i^2 is a shorthand notation that now represents the *Mahalanobis distance* for case i—a multivariate analog of a squared z-score quantifying the standardized distance of an individual's scores on the outcomes in $\mathbf{y}_i$ from the center of the proposed multivariate normal distribution. Conceptually, the equation works exactly the same as before. That is, each likelihood is essentially a vertical coordinate that measures the scores' support for a particular combination of unknown parameter values. The goal of estimation is to find the parameter values that maximize fit to the data.

The simple regression model can be readily extended to a multivariate estimation framework by setting $\mathbf{y}_i' = [x_i \quad y_i]'$ and by populating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with predicted parameter values derived from the regression model—obtained using algebraic mean and covariance expectations (e.g., Bollen, 1989, pp. 21–36)—such that the squared z-score from Equation 3, the Mahalanobis distance, becomes

$$\begin{aligned} z_i^2 &= (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \\ &= \left(\begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \mu_x \\ b_0 + b_1 \mu_x \end{bmatrix} \right)' \begin{bmatrix} \sigma_x^2 & b_1 \sigma_x^2 \\ b_1 \sigma_x^2 & b_1^2 \sigma_x^2 + \sigma_e^2 \end{bmatrix}^{-1} \\ &\quad \left(\begin{bmatrix} x_i \\ y_i \end{bmatrix} - \begin{bmatrix} \mu_x \\ b_0 + b_1 \mu_x \end{bmatrix} \right). \end{aligned} \tag{4}$$

In the context of this model, an iterative optimization procedure would search for estimates of the unknown regression parameters, b_0, b_1, σ_e^2, μ_x, and σ_x^2 that maximize the overall sample likelihood or fit to the data.

To extend such a multivariate model to handle missing data, the FIML function (Arbuckle, 1996) strategically alters Equation 3 as follows:

$$\begin{aligned} &L_i^{FIML}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i | \mathbf{y}_i) \\ &\quad = \text{constant} \times \exp\left[-.5 \times (\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right]. \end{aligned} \tag{5}$$

The crucial elements of Equation 5 are the i subscripts added to $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\Sigma}}$, implying that the likelihood for the ith individual is calculated with a mean vector and covariance matrix subsetted to contain elements corresponding to only those variables on which individual i has complete data. For example, if an individual has observed data on y but not x, Equation 5 calculates the individual's likelihood by ignoring all quantities related to x and setting $\mathbf{y}_i = [y_i]$, $\boldsymbol{\mu}_i = [b_0 + b_1 \mu_x]$, and $\boldsymbol{\Sigma}_i = [b_1^2 \sigma_x^2 + \sigma_e^2]$. Conversely, if an individual has observed data on x but not y, Equation 7 calculates the individual's likelihood by ignoring all quantities related to y and setting $\mathbf{y}_i = [x_i]$, $\boldsymbol{\mu}_i = [\mu_x]$, and $\boldsymbol{\Sigma}_i = [\sigma_x^2]$. Finally, individuals with complete data on both x and y have matrices $\mathbf{y}_i$, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ defined as in Equation 4. In this way, FIML incorporates all observed values on every variable into the estimation process.

Importantly, FIML estimation does not discard incomplete data records nor does it impute them. Although it isn't obvious from the equations,

maximum likelihood uses the normal curve to deduce the missing parts of the data as the optimization algorithm iterates to a solution. To illustrate, reconsider Figure 2.3. Intuitively, knowing an individual's pain level provides important information about disability, as an individual with high pain is much more likely to have a higher rather than lower disability score. In a similar vein, knowing an individual's disability score also carries information about their pain. Although the missing values are never filled in, maximum likelihood can nevertheless be viewed as an "implicit imputation" routine (Widaman, 2006) in the sense that it intuits the missing parts of the data based on the observed scores.

Incorporating Auxiliary Variables in an FIML Analysis

Revisiting earlier ideas, recall that an MNAR-by-omission process occurs when the analysis or imputation procedure fails to condition on a correlate of missingness that also correlates with the residuals of the analysis variables. Because FIML estimates may become biased when missing data on an outcome or predictor variable are MNAR, researchers are well-advised to guard against this possibility by introducing auxiliary variables pertinent to both predictors and outcomes with missing data. Note that auxiliary variables can themselves have missing values (see Enders, 2008), although their utility diminishes if their values are missing with the analysis variables.

Because FIML handles missing data directly as a part of the model estimation process, auxiliary variables that lie outside of one's substantive analysis model of interest must be incorporated into this model in some way. One intuitive possibility would be to specify all auxiliary variables as additional exogenous covariates in the substantive model—for example, adding stress as an explicit predictor in the regression of disability on pain. However, doing so would convert this bivariate regression into a multiple regression, changing the meaning of the coefficients. In this new model, the *partial* regression coefficient for pain would reflect the influence of pain on disability after removing overlapping variance between stress and pain rather than the intended bivariate regression coefficient reflecting the total influence of pain on disability.

To avoid this undesirable side effect, Graham (2003) proposed that auxiliary variables in a FIML-estimated SEM could be specified as *saturated correlates*—free-floating variables in one's model specified to covary with (a) all predictors in the model, (b) the residuals of all outcomes, and (c) each other. Figure 2.4A shows a path diagram[2] of a substantive model involving two predictor variables (x_1 and x_2) and two outcome variables (y_1 and y_2), with the focal model parameters drawn in light gray, and two auxiliary variables (a_1 and a_2) specified as saturated correlates. Because the auxiliary variables in a saturated correlates model are related to all other model variables via two-headed arrows (i.e., correlations or residual correlations, depicted using solid black lines), they are able to assist missing data estimation in all parts of the model without altering the meaning of the partial regression coefficients.

Alternatively, Graham (2003) suggested that one might specify all auxiliary variables as *extra dependent variables* (extra DVs) —additional outcomes regressed on the predictors, with their residuals covarying with one another and with the residuals of the outcome variables. Mirroring Figure 2.4A, Figure 2.4B shows a path diagram of the same model with the auxiliary variables reconfigured as extra DVs. The diagrammatic conventions of Figure 2.4B are the same as those of Figure 2.4A, with the exception that black one-headed arrows are now used to indicate the regressions of the auxiliary variables on the predictors. The logic of this method is analogous to the logic of the saturated correlates approach: the extra variables pass their information to

[2]In conventional path diagrammatic notation, rectangles indicate observed (manifest) variables, circles indicate unobserved (latent) residuals, two-headed arrows indicate variances when attached to a single rectangle or circle and covariances when connecting two rectangles or circles, and one-headed arrows connect regression predictors to outcomes.

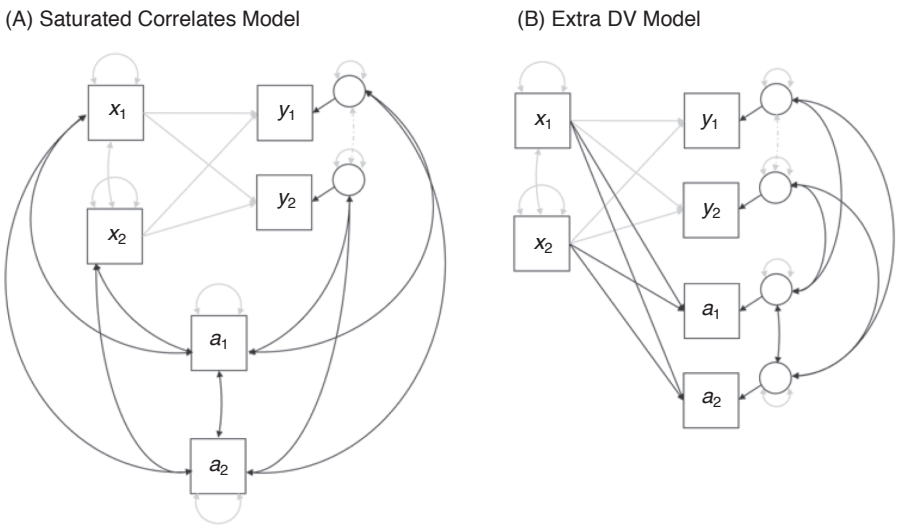

FIGURE 2.4. Graham's (2003) models for incorporating auxiliary variables into a FIML analysis.

every other variable but, because the auxiliary variables are modeled as additional outcomes rather than covariates, the meaning of all partial regression coefficients associated with the focal analysis (the gray arrows in Figure 2.4) will once again remain intact.

In sum, both of Graham's (2003) models provide conceptually straightforward strategies for incorporating auxiliary variables into one's model when using FIML estimation with missing data. The saturated correlates and extra DV approaches do come with some limitations, however. Importantly, including more than a few auxiliary variables in these models can result in matrices with an improper structure and, thus, an increased probability of convergence failures (Savalei & Bentler, 2009). One possible solution is to include one or two auxiliary variables with the highest correlation (or residual correlation) with the incomplete variables. In practice, it is often difficult to find more than one or two extra variables that meaningfully increase explained variance, so there is often little benefit to including large numbers of auxiliary variables. A second solution, proposed by Howard et al. (2015), is to first select a large number auxiliary variables for inclusion, then utilize principal components analysis to extract a smaller subset of components that, in turn, function as auxiliary variables. Their simulations show that even a single principal component score is an effective surrogate for a large number of auxiliary variables.

Maximum Likelihood: Recent Developments

Maximum likelihood analyses have evolved considerably in recent years. The estimators that were widely available when the first edition of this handbook was published were generally limited to multivariate normal data (e.g., Equation 3). This is still a common (and very reasonable) assumption for missing data analyses, but flexible estimation routines that accommodate mixtures of categorical and continuous variables are now widely available (Lüdtke et al., 2020; Muthén et al., 2016).

Estimators for mixed response types generally deploy a so-called *factored regression strategy* that breaks the overall likelihood function into a set of component likelihoods (Ibrahim, 1990). To illustrate its simplest incarnation, reconsider the bivariate normal example from Equation 3. Factored regression models use the probability chain rule to convert the bivariate distribution into the product of two or more univariate distributions, each of which corresponds to

a regression model. Using generic notation, the factorization for the bivariate regression analysis is

$$f(Y, X) = f(Y|X) \times f(X) \tag{6}$$

Where each "*f* of something" represents a probability distribution (or likelihood) induced by a regression model; the leftmost term represents the multivariate distribution of the variables from Equation 3, the first term after the equals sign corresponds to the focal regression model (the linear regression model depicted in Figure 2.3), and the rightmost term is a supporting (empty) regression for the incomplete predictor. By avoiding the multivariate distribution on the left and working with the univariate distributions on the right, we can mix and match distributions that honor the metrics of the data. For example, reconsider the linear regression depicted in Figure 2.3. Suppose instead that the three pain levels represented qualitative rather than quantitative differences among participants. Mixtures of categorical and numeric variables are at odds with a multivariate normal distribution, but the factored approach readily accommodates this by specifying $f(X)$ as a logistic regression.

Factorizing a multivariate distribution into component univariate models also paves the way for estimating interactions and nonlinear effects with missing data (Lüdtke et al., 2020; Robitzsch & Luedtke, 2021). For example, consider the following moderated multiple regression analysis where the influence of depression varies by sex (0 = female, 1 = male) to influence psychosocial disability:

$$\begin{aligned} DISABILITY_i = {} & b_0 + b_1(DEPRESSION_i) \\ & + b_2(SEVERE\ PAIN_i) + b_3(MALE_i) \\ & + b_4(DEPRESSION_i)(MALE_i) + e_i. \end{aligned} \tag{7}$$

Note that, in line with the example data presented later in the chapter, *SEVERE PAIN* refers to a binary indicator where 0 = no/little/moderate pain and where 1 = severe pain. Using generic notation, the factored regression specification for the model is

$$\begin{aligned} & f(DISABILITY|DEPRESSION, SEVERE\ PAIN, \\ & \quad MALE, DEPRESSION \times MALE) \\ & \times f(DEPRESSION|SEVERE\ PAIN, MALE) \\ & \times f(SEVERE\ PAIN|MALE) \times f(MALE) \end{aligned} \tag{8}$$

where the first term corresponds to the moderated regression analysis, and the remaining terms are supporting regression models for the predictors. Two points are worth noting. First, two of the predictors are binary, and their supporting models would be logistic regressions, as previously noted. Second, the focal analysis involves the product of two variables, but the product term itself is not a unique variable (i.e., it does not get predicted by other regressors). Factored regression models are an important recent innovation, as classic estimators based on multivariate normality are known to introduce bias when applied to models with interactions and other types nonlinear effects (e.g., so-called *just-another-variable approaches*; Cham et al., 2017). The analysis examples illustrate the factor regression approach.

MULTIPLE IMPUTATION

To review, maximum likelihood is a direct estimator that incorporates missing data handling into each analysis in a single step. Because maximum likelihood addresses both the missing data model and the substantive analysis model simultaneously, users must choose the maximum likelihood estimator that best captures the important features of both models, for example, relying on the multivariate normal likelihood function of Equation (5) to handle simple scenarios in which all analysis variables and auxiliary variables are continuous and all relationships in the model are linear (no-moderated) but switching to the factored likelihood of Equation (8) to handle more complex scenarios in which key model variables are categorical, key relationships in the model are nonlinear (e.g., moderated), or both. In this way, maximum likelihood estimation

proceeds on an analysis-by-analysis basis, tailoring the estimation approach to the specific features of each model. Importantly, because maximum likelihood integrates missing data handling into the estimation of the substantive model, auxiliary variables must be actively incorporated into the substantive analysis model as saturated correlates, extra DVs, or additional terms in a factored likelihood.

In contrast to maximum likelihood, multiple imputation is a two-stage procedure that separates missing data handling from the analysis. The first stage creates multiple copies of the data (e.g., at least 20), each containing different estimates of the missing values. A typical imputation routine uses the estimated model parameters to compute predicted values of the missing data that are augmented with random noise to preserve variation. Having created a set of filled-in data sets, the second stage consists of analyzing each data set and using "Rubin's rules" (Little & Rubin, 1987; Rubin, 1987) to combine estimates and standard errors into a single package of results. These pooled point estimates and standard errors average over many plausible values for the missing data.

Separating missing data imputation from data analysis can be a benefit or an Achilles heel. On the one hand, because numerous auxiliary variables—both continuous or categorical—can be easily included as predictors in the initial imputation stage, these variables are no longer required to be incorporated into the analysis model as somewhat awkward saturated correlates or extra DVs. On the other hand, this separation also creates the possibility that the model charged with constructing the imputations contradicts or is somehow incompatible with the second-stage analysis model (Bartlett et al., 2015; Meng, 1994). For example, this could happen because the first-stage imputation model omits one of the analysis variables (all analysis model variables must be included in the imputation model, regardless of whether they are complete or incomplete, in order to preserve their correlational structure in the resulting imputations), incorrectly specifies an incomplete variable's distribution, or fails to preserve a structural feature of the data such as clustering or grouping. The importance of these issues cannot be overstated: an incorrectly specified imputation model that does not preserve the key features of an intended analysis can actually inject bias into one's results, potentially compounding existing bias caused by missing data or even creating bias where none previously existed.

For this reason, we find it useful to distinguish imputation methods according to the similarity between the imputation and analysis models. When the analysis model(s) one intends to run are relatively straightforward, it is sometimes possible to create a single, general set of multiple imputations that serve a variety of different analytic goals. For example, a researcher could use a multivariate regression model (or a series of univariate regressions) for imputation and then fit any number of subsequent linear regression models to the filled-in data sets, so long as these models do not feature interactions or polynomial terms that add complexity to the analysis. Although this option does not have a common name, Enders (2021) used the phrase "agnostic imputation" to convey that imputations aren't tailored to one specific analytic goal. Common examples of agnostic imputation approaches include the popular joint model (Schafer, 1997) and fully conditional specification frameworks (Raghunathan et al., 2001; van Buuren et al., 2006).

By contrast, when the analysis model(s) one intends to estimate are more complex, it is best to apply the same models at both stages in order to tailor the imputations to the unique features of each specific analysis. For example, a researcher could use a moderated regression approach to create imputations that accurately reflect the nonlinearity caused by the product term(s), and the second stage analysis would be an identical moderated regression model. The literature has described this option using numerous labels, including the sequential specification, model-based imputation, fully Bayesian imputation, and substantive model-compatible multiple imputation (Bartlett et al., 2015; Enders et al., 2020; Lüdtke et al., 2020; Zhang & Wang, 2017).

By classifying procedures according to the alignment between the imputation and analysis

models, our goal is to emphasize that an analysis model's composition—in particular, whether it includes nonlinear effects such as interactions, polynomial terms, or random effects—determines the type of imputation strategy that works best. A tailored, model-based approach is preferable for analyses that feature these types of nonlinearities, whereas agnostic imputation schemes are well suited for descriptive summaries and additive models that do not include such terms. It is also perfectly acceptable to mix and match these two approaches within a given project as needed, and it is similarly acceptable to use some combination of maximum likelihood estimation and multiple imputation.

Stage 1. Creating Imputed Data Sets

Digging a bit deeper into the mechanics, most multiple imputation procedures use Bayesian estimation and Markov chain Monte Carlo (MCMC) algorithms for the initial imputation stage. These algorithms iterate between two steps: update the parameter estimates conditional on the filled-in data, then update the missing values conditional on the new parameter estimates. To illustrate, suppose it is of interest to estimate a bivariate association between pain levels and psychosocial disability. To simplify the notation, we generically refer to these variables as y and x, respectively. For now, assume that y (disability) has missing values and x (pain) is complete.

Agnostic imputation is appropriate for this model because the analysis does not involve interactive or nonlinear effects. A typical agnostic imputation scheme would use linear regression such as the one depicted in Figure 2.3 as the first-stage imputation model, and the supporting MCMC algorithm would repeatedly alternate between estimating the regression model parameters and imputing the data. Focusing on the imputation step, each missing data point is replaced by an estimate that equals the sum of a predicted value and a normally distributed random noise term. Using generic notation, the following equation defines the imputations

$$y_{i(\text{Imp})} = b_0 + b_1 x_i + \dot{e}_i = \hat{y}_{i(\text{Imp})} + \dot{e}_i \qquad (9)$$

where $y_{i(\text{Imp})}$ is the imputation generated for the ith observation on incomplete variable y_i (e.g., disability), x_i is a complete variable (e.g., pain), b_0 and b_1 are regression coefficients updated to reflect the current iteration of the MCMC algorithm, $y_{i(\text{Imp})}$ is the predicted value of y_i given an individual's observed x_i score, and $\dot{e}_i$ is a synthetic residual term (i.e., a random number) sampled from a normal distribution, the spread of which depends on the estimated residual variance.

To illustrate, Figure 2.3 can be reconstrued as depicting the distributions of plausible disability imputations at three values of pain. The black dots on the regression line are the predicted values and the spread of the normal curves reflects the estimated residual variation (i.e., the variation of the $\dot{e}_i$ terms). Candidate imputations fall exactly on vertical lines, but Figure 2.3 uses horizontal jitter to improve visibility of the circles in high-density portions of the distribution.

The MCMC algorithm imputes missing values by selecting circles from each distribution at random, depending on one's pain level. This agnostic imputation procedure can easily be extended to multivariate missing data (e.g., estimating the correlations among three incomplete variables, such as pain, disability, and depression) by using a round robin sequence of linear regression models, each of which features an incomplete variable regressed on all other variables (complete or previously imputed; see van Buuren et al., 2006, for a description of this *fully conditional specification* [FCS] procedure).

As a second example, suppose the analysis model is the moderated regression from Equation 7. An agnostic imputation scheme is no longer appropriate for this analysis (Bartlett et al., 2015), as the presence of a product term requires a model-based approach that tailors imputation to this specific analysis. In fact, model-based imputation invokes the same factored regression specification shown in Equation 8; in lieu of the round robin scheme described in the previous paragraph, imputation uses a collection of equations that consists of the focal analysis variable and supporting regression models for each incomplete predictor. Although the basic idea is still the

same—imputation equals a predicted value plus noise—model-based imputation is more complex because the distribution of imputations can depend on more than one regression equation.

The MCMC algorithm repeats the two-step estimation procedure (update the parameters, then update the imputations) for many computational cycles. A typical application saves a relatively small number of complete data sets—a common recommendation is to use $M = 20$ data sets (Graham et al., 2007)—from a much longer MCMC process consisting of hundreds or even thousands of computational cycles. One way to do this is to save each imputed data set from the final iteration of a unique MCMC process or chain. In order to ensure that the MCMC algorithm produces accurate, representative imputations, it is critical to determine an appropriate total number of iterations T (also called the *burn-in period*), as the MCMC algorithm must iterate long enough to escape its dependence on random starting values and converge to a steady state. To determine this number, T, researchers must assess one or more MCMC *convergence diagnostics*. Although graphical displays such as trace plots (Schafer, 1997) might be used for this purpose, the potential scale reduction factor (a measure capturing the similarity of MCMC chains initiated from different random starting values; Gelman & Rubin, 1992) is especially useful because simple rules of thumb nearly always produce acceptable results (e.g., determine the number of iterations required for the index to drop below 1.05, then set T to that value).

Stage 2. Analyzing Imputations and Pooling Estimates

The product of the first stage is a set of M filled-in data sets. Although it might seem reasonable to do so, averaging the filled-in values themselves is inappropriate; the correct procedure is to analyze each filled-in data set separately and combine multiple sets of estimates and standard errors into one package of results. Repeating an analysis many times sounds tedious, but most software packages have built-in routines that automate this process.

Rubin (1987) provided the rules or equations for pooling estimates and standard errors. The multiple imputation point estimate is simply the arithmetic average of the M estimates

$$\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{10}$$

where $\hat{\theta}_m$ is a parameter value from data set m, and $\bar{\theta}$ is the average estimate. Pooling standard errors is a bit more complicated because a simple average of the complete-data standard errors would overstate precision. The correct pooling expression is

$$\begin{aligned} SE_{\bar{\theta}} &= \sqrt{\text{mean}\left(SE^2\right) + \text{var}\left(\hat{\theta}_m\right) + \frac{\text{var}\left(\hat{\theta}_m\right)}{M}} \\ &= \sqrt{V_W + V_B + SE_{\bar{\theta}}^2} \end{aligned} \tag{11}$$

where the first term under the radical represents the average squared standard error (the within-imputation sampling variance, V_W), the second term depends on the variance of the M parameter estimates around their average (the between-imputation variance, V_B), and the final term represents the squared standard error of the pooled estimate ($SE_{\bar{\theta}}^2$). Conceptually, the first term estimates the sampling error of a complete-data analysis, and the next two terms are essentially correction factors that inflate the standard error to compensate for uncertainty due to the imputations–that is, additional uncertainty (sampling variability) in the parameter estimates caused by missing data. We note that this additional missing data uncertainty is also reflected in FIML standard errors, though their calculation is not as straightforward.

Because different variables in an analysis (e.g., the predictors and outcome in the regression of Equation 7) may be more or less affected by missing data, it follows that their associated standard errors may be correspondingly influenced by different degrees of missing data uncertainty. To quantify the degree of this influence, it is useful to divide the second and third terms under the radical in Equation 11 by the entire quantity under

the radical to calculate a useful R^2-like quantity known as the *fraction of missing information* (FMI, Rubin, 1987), which gives proportional impact of missing data on the squared standard error. For example, FMI = .20 means that variation due to missing data accounts for 20% of the squared standard error.

In addition to using the FMI to quantify missing data uncertainty in each standard error in a single model, it can also be useful to examine changes in the FMI values that result from adding one or more auxiliary variables to an initial model. To understand why, consider that if an auxiliary variable correlates highly with a certain variable in the substantive analysis model and also contains comparatively more complete data than that variable (e.g., because the auxiliary variable was collected at baseline, before later measurements were affected by participant dropout), the complete values of this auxiliary variable, to some extent, act as proxies for the unseen values of the analysis variable, contributing information to the analysis that might serve to reduce missing data uncertainty (Collins et al., 2001). The success of an auxiliary variable (or set of auxiliary variables) in reducing missing data uncertainty in parameter standard errors is naturally captured by decreased FMI values.[3] Although there is no firm rule regarding how steep a decrease in FMI values must be in order to be considered meaningful, reporting information about which auxiliary variables seem useful for repairing parameter standard errors and recovering lost power can help alert researchers working in the same substantive area to the possible benefits of including these auxiliary variables in their future research. We note that the FMI can also be calculated using FIML estimation, although this option is not available in all software packages (for details, see Savalei & Rhemtulla, 2012). We report these diagnostics in the upcoming analysis examples, where available.

FULL INFORMATION MAXIMUM LIKELIHOOD AND MULTIPLE IMPUTATION DATA ANALYSIS EXAMPLES

To illustrate maximum likelihood and multiple imputation, we present illustrative analyses using synthetic data based on a real study of 300 chronic pain sufferers. The data set includes a number of psychological correlates of chronic pain. The focal variables for the analyses are a gender dummy code (0 = female, 1 = male), a binary severe pain indicator (0 = no, little, or moderate pain, 1 = severe pain), a depression composite, and a scale measuring psychosocial disability. We also considered four self-report auxiliary variables: perceived control over pain, pain interference with daily life activities, anxiety, and stress.

All data and analysis scripts are available for download at https://case.fiu.edu/about/directory/profiles/hayes-timothy.html. We provide scripts for all multiple imputation examples using Blimp software (Enders & Keller, 2021) for the imputation step and Mplus 8 (Muthén & Muthén, 2017) for the analysis and pooling steps as well as R scripts (R Core Team, 2021) for all FIML examples using the lavaan package (Rosseel, 2012) for analyses requiring multivariate normal likelihood functions and the mdmb package (Robitzsch & Luedtke, 2021) for analyses requiring factored likelihood functions. This list of software packages is far from exhaustive, and the software choices used in our supplemental materials represent only one configuration among a dizzying number possible. Because the software landscape is ever-shifting, with new software packages constantly being developed and existing packages frequently releasing updated functionality and syntax, we do not focus on detailed descriptions of software here, referring readers instead to the supplemental materials (see also Grund et al., 2021; Hayes, 2019; Lüdtke et al., 2020; Rosseel, 2012).

Table 2.2 displays descriptive statistics for these variables, computed by analyzing and pooling

[3]We note, however, that it may require pooling the results of many imputed data sets to achieve stable enough FMI values to compare across analyses (e.g., von Hippel, 2018, for a primer on the number of imputations needed in multiple imputation analyses).

TABLE 2.2

Pooled Descriptive Statistics by Sex

			Bivariate correlations (lower triangle) and covariance coverage (upper triangle)						
	M	*SD*	1	2	3	4	5	6	7
					Males				
1. Depression	15.29	6.59	91%	77%	65%	91%	91%	91%	91%
2. Severe pain	0.40	0.49	.36	86%	62%	86%	86%	86%	86%
3. Disability	22.35	4.73	.22	.17	72%	72%	72%	72%	72%
4. Anxiety	12.06	4.82	**.57**	.18	**.34**	100%	100%	100%	100%
5. Stress	4.14	1.76	**.54**	.21	.26	.69	100%	100%	100%
6. Control	20.33	5.58	**−.35**	−.21	**−.48**	−.30	−.16	100%	100%
7. Interfere	27.92	8.53	**.30**	**.36**	**.40**	.28	.20	−.45	100%
					Females				
1. Depression	13.85	5.66	88%	77%	63%	88%	88%	88%	88%
2. Severe pain	0.19	0.39	.07	85%	60%	85%	85%	85%	85%
3. Disability	21.80	4.97	.46	.28	71%	71%	71%	71%	71%
4. Anxiety	11.19	4.37	**.59**	−.02	.27	100%	100%	100%	100%
5. Stress	3.73	1.82	**.53**	.11	**.30**	.69	100%	100%	100%
6. Control	21.03	5.11	**−.30**	−.03	−.29	−.30	−.36	100%	100%
7. Interfere	27.15	8.75	**.36**	.25	.24	.23	.19	−.36	100%

Note. Descriptive statistics were pooled across $M = 100$ multiply imputed data sets, with imputations generated separately for males and females. For each group, entries in the lower triangle of the correlation matrix represent bivariate correlations, whereas entries on the main diagonal and upper triangle indicate covariance coverage (percentage of data present). Bolded values indicate correlations between auxiliary variables and model variables exceeding values of $|r| = .30$.

100 imputed data sets. The bolded entries in Table 2.2 indicate correlations between auxiliary variables and model variables greater than $|r| = .30$. Because all four auxiliary variables correlated with at least one substantive model variable at or above this level, we included all four auxiliary variables as extra DVs in the FIML analyses and as additional variables in the imputation models. Table 2.2 also reports *covariance coverage* information for all variables: the main diagonal provides the percentage of observed data for each variable, whereas the off-diagonals in the upper triangle provides the percentage of data present for each pair of variables. As described below, in our experience, covariance coverage tables provide a quick and convenient way for readers to assess the prevalence of missing values in a given data set.

The first analysis example compares FIML estimation to multiple imputation in the following linear regression model:

$$\begin{aligned} DISABILITY_i = b_0 &+ b_1\left(DEPRESSION_i\right) \\ &+ b_2\left(SEVERE\ PAIN_i\right) \\ &+ b_3\left(MALE_i\right) + e_i. \end{aligned} \tag{12}$$

Agnostic imputation is appropriate for this analysis because the model does not feature interactive or nonlinear terms. In the first section of Table 2.3 we present a side-by-side comparison of results from this linear regression analysis using (a) standard multivariate normal FIML estimation, (b) the factored FIML estimation procedure that treats each binary predictor more appropriately with its own logistic regression submodel (for more information, see Lüdtke et al., 2020), and (c) the agnostic FCS imputation method described two sections earlier, with dichotomous variables imputed using latent probit imputation methods.

TABLE 2.3

Estimates and Standard Errors From Linear and Moderated Regression Model by Estimation Method

	Linear regression								Moderated regression				
	FIML: Multivariate normal			FIML: Factored		Agnostic multiple imputation			FIML: Factored		Model-based multiple imputation		
Parameter	Est.	SE	FMI	Est.	SE	Est.	SE	FMI	Est.	SE	Est.	SE	FMI
Intercept	17.91***	0.90	.44	17.93***	0.90	17.98***	0.86	.38	21.38***	0.41	21.55***	0.43	.30
Depression	0.25***	0.06	.45	0.25***	0.06	0.26***	0.06	.40	0.40***	0.07	0.41***	0.08	.44
Severe Pain	1.92*	0.76	.40	1.85*	0.75	1.85*	0.82	.45	1.89**	0.73	1.88*	0.74	.40
Male	–0.30	0.62	.27	–0.30	0.62	–0.31	0.63	.27	–0.10	0.61	–0.20	0.63	.30
Depression × Male	—	—	—	—	—	—	—	—	–0.30**	0.11	–0.30**	0.10	.34
Residual Variance	19.3***	2.14	0.46	19.34***	—	19.90***	2.30	.50	18.31***	—	19.10***	2.22	.50
R^2	.16	—	—	.16	—	.16**	0.05	.45	.20	—	.20***	.06	.47

Note. * = $p < .05$, ** =. $p < .01$, *** = $p < .001$. FMI = fraction of missing information. Listwise and multivariate normal FIML estimation were conducted using the lavaan package in R (Rosseel, 2012), which returns estimates of the FMI upon request, facilitating comparison with the FMI values returned by the multiple imputation analysis in Mplus (L. K. Muthén & Muthén, 2017). Factored likelihood estimation was conducted using R package mdmb (Robitzsch & Luedtke, 2021). Multiple imputations were generated using Blimp software (Enders & Keller, 2021) and subsequently analyzed and pooled using Mplus version 8.6 (L. K. Muthén & Muthén, 2017). Both agnostic and model-based imputation methods were used to impute, analyze, and pool $M = 100$ data sets.

Note that all three missing data handling methods incorporated the same model variables and auxiliary variables. Additionally, the factored FIML and agnostic multiple imputation methods used comparable procedures to treat all dichotomous variables appropriately. Because it has long been known that FIML and multiple imputation converge on similar solutions when the same input data is used for both (see Collins et al., 2001, pp. 336–338 and Table 1), it comes as no surprise that the results of all three methods are near-identical. In fact, despite ignoring the correct scaling of the dichotomous predictors in the model, the results produced by multivariate normal FIML estimation in this analysis were even comparable to the other methods. This finding mirrors simulation results from Muthén et al. (2016).

The second analysis example compares FIML estimation to multiple imputation in the moderated regression model from Equation 7. Because the analysis model includes an incomplete interaction effect, standard multivariate normal FIML estimation and agnostic imputation routines are no longer appropriate. Instead, a tailored approach is necessary for this situation. Thus, in the second section of Table 2.3, we present a side-by-side comparison of results from this moderated regression model using (a) the factored FIML estimation procedure from Equation 8 and (b) the model-compatible imputation method described above, with dichotomous variables once again treated appropriately using either logistic regression (ML) or latent probit methods (imputation). Because these methods incorporated the same model variables and auxiliary variables (e.g., they employed comparable strategies both to address the scaling of the dichotomous predictors and the presence of the product term in the moderated regression analysis), it is again unsurprising that their resulting estimates, standard errors, and patterns of significance are nearly indistinguishable.

METHODS FOR MNAR MISSING DATA

In contrast to a missing-at-random mechanism, in which the unseen scores are unrelated to the probability of missingness after conditioning on or controlling for the observed data, a MNAR mechanism is one where the unseen scores still carry information about missingness even after

conditioning on the observed data. When this is true, the analysis model must include an additional component that describes the occurrence of missing data. The two major modeling approaches for MNAR mechanisms—selection models and pattern mixture models—do just that, albeit in different ways. A selection model includes an additional regression equation with a binary missing data indicator (0 = observed, 1 = missing) as the dependent variable, whereas a pattern mixture model uses that missing data indicator as a predictor.

To illustrate the two models, reconsider the simple regression model where *x* predicts *y* (see Figure 2.3), and suppose that the unseen values of *y* determine whether the outcome is missing (e.g., the individuals with the highest disability levels opt not to report their scores, as in the MNAR mechanisms of Figure 2.1C and D). We also need a dummy variable M_y that codes whether *Y* is missing. Figure 2.5A shows the selection model as a path diagram. Notice that the composition of the diagram resembles a single-mediator model where the analysis variables predict missingness via direct and indirect effects. In contrast, a pattern mixture model resembles a moderated process where M_y defines qualitatively different subgroups with unique parameter values. Figure 5B shows the path diagram; the arrow connecting M_y to *Y* is an intercept difference, and the dashed line indicates that *x*'s slope differs between groups.

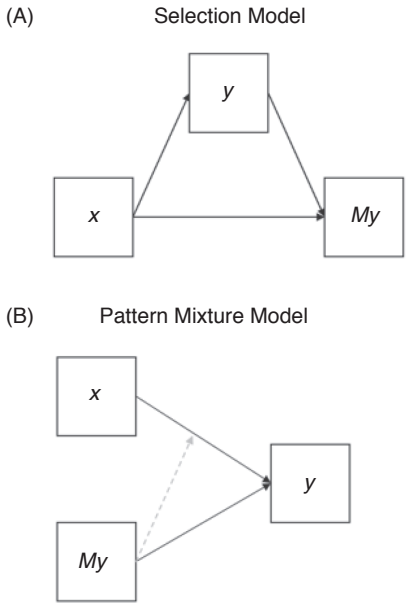

FIGURE 2.5. Conceptual diagrams of MNAR missing data approaches.

MNAR analyses require strict, unverifiable assumptions (e.g., a correct missingness model with the right configuration of effects), and simple misspecifications can produce biased estimates. Consequently, methodologists often suggest using these models as part of a sensitivity analysis that explores different missingness assumptions. A simple example is one where the researcher augments the main MAR analysis with one or more selection or pattern mixture models. An online supplemental document can present side-by-side comparisons of two or more sets of analysis results, with any discrepant estimates noted in the main body of the manuscript.

REPORTING THE RESULTS OF A MISSING DATA ANALYSIS

Although many of the principles summarized in this chapter have been known in the methodological literature for decades, misconceptions about missing data remain prevalent (van Ginkel et al., 2020), and surveys of the published literature have repeatedly found missing data reporting practices to be woefully inadequate (e.g., Jeličić et al., 2009; Nicholson et al., 2017). As a result, researchers aspiring to learn good reporting practices may find themselves without trustworthy examples to refer to, left to simply do their best to describe missing data analyses to an audience of readers (and reviewers) whose knowledge of the topic may be incomplete or even misguided. To help researchers navigate this landscape, in this section we provide several broad recommendations for missing data reporting.

The first thing researchers must report is descriptive statistics for all variables in the sample, including the prevalence of missing data on all model variables. With this in mind, our first recommendation is to use a modern method like FIML or multiple imputation to appropriately

deal with missing data when estimating and reporting descriptive statistics, including all auxiliary variables deemed useful to estimation. We highlight this recommendation because, in our experience, many researchers use modern missing data handling methods only when estimating their main analysis models, often using specialized software, but return to their favorite general software package to estimate descriptive statistics using default deletion settings (e.g., using FIML in Mplus to estimate an SEM but computing descriptive statistics using listwise deletion in SPSS). This is unfortunate, because descriptive statistics based on marginal (unconditional) distributions tend to be even more severely affected by missing data than regression coefficients based on conditional distributions like those shown in Figure 2.1. In this way, modern missing data handling methods are just as crucial (if not more) for estimating descriptive statistics as they are for estimating more complex analysis models.

Our second recommendation is to report the prevalence of missing data by including covariance coverage (a matrix with the proportion of complete data for each variable or variable pair) in the upper triangle of the reported correlation matrix, as we have in Table 2.2. This provides readers with a compact yet comprehensive overview of the pattern of missing data affecting each model variable and pair of variables. If the proportion of missing data for a particular model variable or pair of variables is high, we recommend reminding readers of the general principle that *the greater the percentage of missing data affecting a particular variable or analysis model, the greater the need to address missing data using a modern method.* Citing quantitative simulation studies demonstrating the accuracy of FIML and multiple imputation under a variety of missing data rates (e.g., Collins et al., 2001, who showed across four studies that FIML and multiple imputation provide unbiased—and equivalent—results even under 50% missing data)[4] can help to counter, or to preempt, the commonly held misconception that there is some amount of missing data that is simply too much for modern methods to handle (e.g., "it is dangerous to impute a variable with 50% missing values . . .").

Our next major recommendation is to provide theoretical justification for the missing data mechanism(s) one believes to be affecting each substantive model. This recommendation stems from the reality that, although one could plausibly rule out an MCAR (or focused MNAR) mechanism by finding any evidence at all that the observed variables in one's data set predict the probability of missingness on a given variable (e.g., in a *t*-test of missing vs. nonmissing cases or a logistic regression analysis predicting a missing data indicator), analogous statistical procedures for ruling out a MAR (or MCAR) mechanism by finding evidence that the conditional distributions of observed and unseen values differ systematically from one another (as in Figure 2.1C and D) are undefined without access to the missing scores (the asterisks in Figure 2.1). As such, the MAR mechanism assumed by standard FIML and multiple imputation methods can neither be proven (or shown) nor ruled out using real data but must instead be argued for on the basis of theory.

When initially considering which missing data mechanisms seem theoretically plausible, we recommend starting with the assumption that when participants decline to answer a certain question or decide to skip a particular testing session, they generally do so for a reason, making a pure MCAR mechanism highly unlikely. It is also unlikely that one's intended analysis model already contains all missing data causes necessary on which to condition in order to meet the MAR assumption. Instead, it seems most reasonable to begin with the assumption that the missing data on all model variables are at least MNAR by omission at the outset, requiring the identification of the auxiliary variables required to make the data MAR (e.g., using the bivariate or semipartial correlation approaches described earlier). Reporting information about any auxiliary variables

[4]It might also be useful to point out that planned missing data designs, such as the popular *three form design*, intentionally produce extremely high missing data rates, (e.g., close to 70% between some variable pairs; see Graham et al., 2006).

identified is crucial to helping future researchers know which measures they should consider including in their study designs.

In some lucky cases, a combination of substantive theory and practical judgment can rule out an MNAR mechanism. For example, attrition in a study of infants' spatial skills could only occur if the infants' parents or caregivers declined to return them to the study, making it straightforward to rule out the infants' unseen scores as a potential source of MNAR missingness. In other cases, however, the distinction may not be so clear. Take, for example, the missing psychosocial disability data described earlier, which could plausibly have been caused (a) by the patients' scores on external variable(s) like pain or stress, as in Figure 2.1B, (b) by patients' unseen disability scores, as in Figure 2.1C, or (c) by a combination of both, as in Figure 2.1D. In such cases, it may be necessary to conduct a sensitivity analysis, as described earlier, estimating a variety of MAR and MNAR models and comparing their results. If the pattern of results remains similar across these analyses, then the consequences of assuming a particular mechanism may make little difference. But if the results of these competing analyses differ, it may be useful to present all model results side-by-side so that readers can compare the effects of making different missing data assumptions. Although this side-by-side approach sacrifices the parsimony of choosing and displaying only one final model, it increases transparency and preserves a more detailed account of the models' possible results in the published record.

Next, we recommend providing the details of the missing data handling methods applied to each substantive analysis of interest, including all software packages used (along with their version numbers) and any specific settings invoked in the process. Because different analysis models may contain unique features (e.g., categorical predictors, product terms, or multilevel random effects) that require special consideration and because the variables included in different models may be acted upon by different missing data mechanisms (requiring different auxiliary variables), it follows that *missing data handling strategies should generally be tailored to, and reported with respect to, each specific analysis model one intends to run*. When using FIML estimation, researchers should report how they incorporated auxiliary variables into the model (as saturated correlates? extra dependent variables?), how they incorporated incomplete predictors into the likelihood function,[5] as well as any special estimation procedures required to address specific features of the substantive model (e.g., using the factored FIML approach described earlier to address interactions).

When using multiple imputation, researchers should report exactly what variables were included in each imputation model. At a minimum, this should include all variables from the substantive model, regardless of whether the variables are complete or missing (leaving a variable out of the imputation model results in imputations that assume its correlation with all other model variables is 0) along with any auxiliary variables identified earlier. It is also critical to report any special imputation procedures used to accommodate particular features of a substantive model, for example, using model-based imputation methods to generate imputations appropriate for a moderation analysis, as described above. Following this, one should report how many imputed data sets were analyzed and pooled (as stated earlier, 20 imputed data sets is a good general rule, but more are better, if computational time allows).

Additionally, one should report convergence information for any imputation models run. For example, one might report that, "An initial diagnostic run indicated that the worst (highest) potential scale reduction factor across all model parameters dropped below 1.05 after approximately 2,000 iterations, suggesting converge of

[5]Although we have not emphasized specific software commands in this chapter, we note that one can freely estimate the distributions of predictor variables and incorporate them into the multivariate normal likelihood function of Equation 3 by declaring all exogenous variable means and variances in the MODEL statement in Mplus, or by setting the argument fixed.x = FALSE in the lavaan() function in lavaan and explicitly referencing all exogenous means, variances, and covariances in the lavaan model syntax.

the MCMC chains." Although this could seem like nuts-and-bolts technical information not worth including in a published paper, we believe that providing this information is important because it serves to subtly combat a widely held misconception that one can multiply impute one's data without assessing convergence, simply trusting one's default software settings. This misconception is especially troubling because default settings in software packages are often dramatically insufficient (e.g., the SPSS multiple imputation routine uses only five iterations, when hundreds or thousands may be necessary).

Finally, when using multiple imputation to address missing data, it is sometimes necessary to address readers' (or reviewers'[6]) potential unease with this approach by clearly and succinctly explaining how the method works and rebutting possible misconceptions. Toward this end, one might assure readers that although the notion of filling in missing values may seem to conjure images of fraudsters purposefully editing values in a data file in order to make an analysis appear significant, the synthetic values generated in a multiple imputation analysis are driven by the data, not by the researcher, with no guarantee that they will lead to flattering model results (e.g., if the correlation between a pair of variables in the data is 0, the correlation between imputed values generated for these variables is also 0). It can also be helpful to remind readers that resorting to listwise deletion methods to avoid imputing missing values will necessarily result in discarding potentially large amounts of real data provided by participants who responded to some—but not all—measures. Viewed in this way, multiple imputation, like FIML, is as much—if not more—about preserving and using all observed values in a data set as it is about filling in missing values. In fact, despite their surface-level differences (e.g., explicitly filling in missing values with imputations generated from a multivariate normal distribution vs. implicitly assuming a multivariate normal distribution is likely to have produced the unseen values), both multiple imputation and FIML estimation are based on the same underlying assumptions and are well-known to produce near-identical results under the same input (as shown by Collins et al., 2001, and as reflected in the results of our analysis examples presented earlier).

In sum, although far from exhaustive, we hope that this brief list of reporting recommendations helps researchers think through how best to describe the results of their missing data analyses to readers who may vary in their levels of familiarity and comfort with these analyses. Because expertise in missing data analysis among one's readers cannot be taken for granted, we recommend that researchers err on the side of reporting too much information rather than too little; all else being equal, it seems better to describe each and every step in a missing data analysis thoroughly than to risk incorrectly assuming that some details can be treated as "common knowledge" and left unsaid. Providing such clear and detailed descriptions of each missing data analysis not only helps readers understand and, hopefully, accept the need for these analyses in a particular study but also provides them with a template for reporting such analyses in their own future work.

SUMMARY

The goal of this chapter was to provide an overview of maximum likelihood estimation and multiple imputation, two major missing data handling strategies with strong support from the methodological literature. Both approaches have developed since the first edition of this handbook, and the types of analyses that researchers can perform is broader than ever. Given the same data and assumptions, maximum likelihood and multiple imputation usually produce indistinguishable results, so the choice of method often boils down to practical considerations and personal

[6]In some cases, when concerns are raised during the review process, this clarifying information may be most appropriate to include in a response to peer reviewers. In other cases, however, as when one believes that readers in their subfield are broadly unfamiliar with these methods, it may be appropriate to include this information in the main document.

preference—the analysis examples illustrated this conclusion. In truth, the most important consideration isn't which method to use, but rather the composition of the analysis model. In general, any analysis that features an incomplete interaction term, curvilinear effect, random slope, or other type of nonlinearity requires newer factored regression methods, whereas "classic" versions of maximum likelihood and multiple imputation are well suited for analyses that do not have these special features. Models with mixtures of categorical and numeric variables are a second example where factored regression specifications are useful, and multiple imputation is generally more flexible for these types of problems.

In closing, we note that this tailored, analysis-by-analysis approach to thinking about missing data handling also implies an important underlying principle: *The fundamental goal of the missing data handling approaches discussed throughout this chapter is to accurately and appropriately adjust the results of a target analysis for the likely influences of missing data.* That is, although the problems caused by missing data may originate in the form of missing scores in one's data file, these problems ultimately manifest themselves in the form of potentially distorted, untrustworthy estimates in one's statistical models, and it is these estimates—not the missing scores, themselves—whose accuracy is at stake in a missing data analysis. This suggests, for example, that an "impute first, decide the analysis later" approach is rarely viable[7] and is never wise, as such an approach is fundamentally backwards: the synthetic values generated in a multiple imputation analysis are not intended to function individually as perfect proxies for participants' missing raw scores[8] but, rather, to function together to adjust the estimates of a specific statistical analysis model that one intends to fit to the filled-in data. Thus, to paraphrase a well-known idiom (Covey, 1989), researchers are well-advised to "begin with the analysis in mind," treating the specific features of each substantive model as the "true north" that guides all subsequent missing data handling decisions. By approaching the task in this way, we believe that researchers will be able to confidently identify the most appropriate methods for addressing missing data in every analysis of interest.

References

Arbuckle, J. N. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Lawrence Erlbaum Associates. Inc.

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

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

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

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

Covey, S. R. (1989). *The seven habits of highly effective people*. Simon & Schuster.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. SAGE.

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models.

[7]Except in the luckiest cases, for example, when a single set of agnostic multiple imputations might serve the goals of multiple subsequent analyses. We note that it is this scenario that is assumed by the current default imputation settings in SPSS, potentially leading users to infer that a single set of imputations should serve the goals of any conceivable analysis.

[8]Indeed, perfectly estimating participants' individual true scores is well-known to be a statistically intractable problem even with complete data (e.g., Steiger and Schönemann, 1978).

Structural Equation Modeling, *15*(3), 434–448. https://doi.org/10.1080/10705510802154307

Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). Guilford Press.

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, *25*(1), 88–112. https://doi.org/10.1037/met0000228

Enders, C. K., & Keller, B. T. (2021). *Blimp user's manual* (Version 3) [Computer software]. Applied Missing Data. www.appliedmissingdata.com/multilevel-imputation.html

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gomer, B., & Yuan, K.-H. (2021). Subtypes of the missing not at random missing data mechanism. *Psychological Methods*, *26*(5), 559–598. https://doi.org/10.1037/met0000377

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*(1), 80–100. https://doi.org/10.1207/S15328007SEM1001_4

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

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the R package mdmb: A flexible sequential modeling approach. *Behavior Research Methods*, *53*(6), 2631–2649. https://doi.org/10.3758/s13428-020-01530-0

Hayes, T. (2019). Flexible, free software for multilevel multiple imputation: A review of blimp and jomo. *Journal of Educational and Behavioral Statistics*, *44*(5), 625–641. https://doi.org/10.3102/1076998619858624

Hoel, P. G. (1984). *An introduction to mathematical statistics* (5th ed.). Wiley.

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

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

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*(4), 1195–1199. https://doi.org/10.1037/a0015665

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley.

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods*, *25*(2), 157–181. https://doi.org/10.1037/met0000233

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

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

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

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, *41*(1), 143–153. https://doi.org/10.1177/0165025415618275

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://r-project.org/

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

Raykov, T., & West, B. T. (2016). On enhancing plausibility of the missing at random assumption in incomplete data analyses via evaluation of response-auxiliary variable correlations. *Structural Equation Modeling*, *23*(1), 45–53. https://doi.org/10.1080/10705511.2014.937848

Robitzsch, A., & Luedtke, O. (2021). *Package 'mdmb'* (R package version 1.5–8) [Computer software]. https://cran.r-project.org/package=mdmb

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. https://doi.org/10.1002/9780470316696

Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling*, *16*(3), 477–497. https://doi.org/10.1080/10705510903008238

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling*, *19*(3), 477–494. https://doi.org/10.1080/10705511.2012.687669

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall. https://doi.org/10.1201/9781439821862

Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the behavioural sciences* (pp. 136–178). Jossey-Bass Publishers.

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 Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680

von Hippel, P. T. (2018). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, *49*(3), 699–718. https://doi.org/10.1177/0049124117747303

Widaman, K. F. (2006). III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, *71*(3), 42–64. 10.1111/j.1540-5834.2006.00404.x

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

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