ELSEVIER

Journal of School Psychology 48 (2010) 5 – 37

Journal of School Psychology

An introduction to modern missing data analyses

Amanda N. Baraldi *, Craig K. Enders

Arizona State University, United States

Received 19 October 2009; accepted 20 October 2009

Abstract

A great deal of recent methodological research has focused on two modern missing data analysis methods: maximum likelihood and multiple imputation. These approaches are advantageous to traditional techniques (e.g. deletion and mean imputation techniques) because they require less stringent assumptions and mitigate the pitfalls of traditional techniques. This article explains the theoretical underpinnings of missing data analyses, gives an overview of traditional missing data techniques, and provides accessible descriptions of maximum likelihood and multiple imputation. In particular, this article focuses on maximum likelihood estimation and presents two analysis examples from the Longitudinal Study of American Youth data. One of these examples includes a description of the use of auxiliary variables. Finally, the paper illustrates ways that researchers can use intentional, or planned, missing data to enhance their research designs.

Keywords: Missing data; Multiple imputation; Maximum likelihood; Planned missingness

Introduction

Missing data are ubiquitous in quantitative research studies, and school psychology research is certainly not immune to the problem. Because of its pervasive nature, some methodologists have described missing data as "one of the most important statistical and design problems in research" (methodologist William Shadish, quoted in Azar, 2002, p. 70). Despite the important nature of the problem, substantive researchers routinely employ old

* Corresponding author.
E-mail address: Amanda.Baraldi@asu.edu (A.N. Baraldi).
ACTION EDITOR: Edward J. Daly III.

0022-4405/$ - see front matter © 2009 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsp.2009.10.001

standby techniques that have been admonished in the methodological literature. For example, excluding cases with missing data is a strategy that is firmly entrenched in statistical software packages and is exceedingly common in disciplines such as psychology and education (Peugh & Enders, 2004). This practice is at odds with a report by the American Psychological Association Task Force on Statistical Inference (Wilkinson & American Psychological Association Task Force on Statistical Inference, 1999, p. 598) that stated that deletion methods "are among the worst methods available for practical applications."

It is not a surprise that substantive researchers routinely employ missing data handling techniques that methodologists criticize. For one, software packages make these approaches very convenient to implement. The fact that software programs offer outdated procedures as default options is problematic because the presence of such routines implicitly sends the wrong message to applied researchers. In some sense, the technical nature of the missing data literature is a substantial barrier to the widespread adoption of sophisticated missing data handling options. While many of the flawed missing data techniques (e.g., excluding cases, replacing missing values with the mean) are easy to understand, newer approaches are considerably more difficult to grasp. The primary purpose of this article is to give a user-friendly introduction to these modern missing data methods.

A great deal of recent methodological research has focused on two "state of the art" missing data methods (Schafer & Graham, 2002): maximum likelihood and multiple imputation. Accordingly, the majority of this paper is devoted to these techniques. Quoted in the American Psychological Association's *Monitor on Psychology*, Stephen G. West, former Editor of *Psychological Methods*, stated that, "Routine implementation of these new methods of addressing missing data will be one of the major changes in research over the next decade" (Azar, 2002). Although applications of maximum likelihood and multiple imputation are appearing with greater frequency in published research articles, a substantial gap still exists between the procedures that the methodological literature recommends and those that are actually used in the applied research studies (Bodner, 2006; Peugh & Enders, 2004; Wood, White, & Thompson, 2004). Consequently, the overarching purpose of this manuscript is to provide an overview of maximum likelihood estimation and multiple imputation, with the hope that researchers in the field of school psychology will employ these methods in their own research. More specifically, this paper will explain the theoretical underpinnings of missing data analyses, give an overview of traditional missing data techniques, and provide accessible descriptions of maximum likelihood and multiple imputation. In particular, we focus on maximum likelihood estimation and present two analysis examples from the Longitudinal Study of American Youth (LSAY). Finally, the paper illustrates ways that researchers can use intentional missing data to enhance their research designs.

Theoretical background: Rubin's missing data mechanisms

Before we can begin discussing different missing data handling options, it is important to have a solid understanding of so-called "missing data mechanisms". Rubin (1976) and colleagues (Little & Rubin, 2002) came up with the classification system that is in use today: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). These mechanisms describe relationships between measured variables and the probability of missing data. While these terms have a precise probabilistic and

mathematical meaning, they are essentially three different explanations for why the data are missing. From a practical perspective, the mechanisms are assumptions that dictate the performance of different missing data techniques. We give a conceptual description of each mechanism in this section, and supplementary resources are available to readers who want additional details on the missing data mechanisms (Allison, 2002; Enders, 2010; Little & Rubin, 2002; Rubin, 1976; Schafer & Graham, 2002).

To begin, data are MCAR when the probability of missing data on a variable X is unrelated to other measured variables and to the values of X itself. In other words, missingness is completely unsystematic and the observed data can be thought of as a random subsample of the hypothetically complete data. As an example, consider a child in an educational study that moves to another district midway through the study. The missing values are MCAR if the reason for the move is unrelated to other variables in the data set (e.g., socioeconomic status, disciplinary problems, or other study-related variables). Other examples of MCAR occur when a participant misses a survey administration due to scheduling difficulties or other unrelated reasons (such as a doctor's appointment), a computer randomly misreads grid-in sheets, or an administrative blunder causes several test results to be misplaced prior to data entry. MCAR data may also be a purposeful byproduct of the research design. For example, suppose that a researcher collects self-report data from the entire sample but limits time-consuming behavioral observations to a random subset of participants. We describe a number of these so-called planned missing data designs at the end of the paper. Because MCAR requires missingness to be unrelated to study variables, methodologists often argue that it is a very strict assumption that is unlikely to be satisfied in practice (Raghunathan, 2004; Muthen, Kaplan, & Hollis, 1987).

The MAR mechanism requires a less stringent assumption about the reason for missing data. Data are MAR if missingness is related to other measured variables in the analysis model, but not to the underlying values of the incomplete variable (i.e., the hypothetical values that would have resulted had the data been complete). This terminology is often confusing and misleading because of the use of the word "random." In fact, an MAR mechanism is not random at all and describes systematic missingness where the propensity for missing data is correlated with other study-related variables in an analysis. As an example of an MAR mechanism, consider a study that is interested in assessing the relationship between substance use and self-esteem in high school students. Frequent substance abuse may be associated with chronic absenteeism, leading to a higher probability of missing data on the self-esteem measure (e.g., because students tend to be absent on the days that the researchers administered the self-esteem questionnaires). This example qualifies as MAR if the propensity for missing data on the self-esteem measure is completely determined by a student's substance use score (i.e., there is no residual relationship between the probability of missing data and self-esteem after controlling for substance use). As a second example, suppose that a school district administers a math aptitude exam, and students that score above a certain cut-off participate in an advanced math course. The math course grades are MAR because missingness is completely determined by scores on the aptitude test (e.g., students that score below the cut-off do not have a grade for the advanced math course).

Finally, data are MNAR if the probability of missing data is systematically related to the hypothetical values that are missing. In other words, the MNAR mechanism describes data

that are missing based on the would-be values of the missing scores. For example, consider a reading test where poor readers fail to respond to certain test items because they do not understand the accompanying vignette. Notice that the probability of a missing reading score is directly related to reading ability. As another example, consider a self-report alcohol assessment administered to high school students. MNAR data would result if heavy drinkers are more likely to skip questions out of fear of getting in trouble.

From a practical perspective, Rubin's missing data mechanisms operate as assumptions that dictate how a particular missing data technique will perform. For example, most of the missing data handling methods that researchers have relied on for many decades (e.g., deletion methods that exclude cases from the analysis) only produce accurate estimates when the data are MCAR. In the next section, we use a small artificial data set to illustrate the bias that can result when this assumption does not hold. In contrast, maximum likelihood and multiple imputation provide unbiased estimates when the data are MCAR or MAR, so these methods are more apt to produce accurate parameter estimates. Unfortunately, virtually every mainstream missing data technique performs poorly with MNAR data, although maximum likelihood and multiple imputation tend to fare better than most traditional approaches. Methodologists have developed analysis models for MNAR data (e.g., selection models and pattern mixture models), but these methods trade MAR for assumptions that are equally tenuous (if not more so). MNAR analysis methods are an important and ongoing area of methodological research, but these techniques are not yet well-suited for widespread use.

Of the three missing data mechanisms, it is only possible to empirically test the MCAR mechanism. Methodologists have proposed a number of MCAR tests, although these procedures tend to have low power and do a poor job of detecting deviations from a purely random mechanism (Thoemmes & Enders, 2007). In contrast, the MAR and MNAR mechanisms are impossible to verify because they depend on the unobserved data. That is, demonstrating a relationship (or lack thereof) between the probability of missingness and the would-be values of the incomplete variable requires knowledge of the missing values. Consequently, the MAR mechanism that underlies maximum likelihood and multiple imputation is ultimately an untestable assumption. Some authors argue that the MAR assumption may be tenable in school-based studies where student mobility is often the most common reason for attrition (Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997; Enders, Dietz, Montague, & Dixon, 2006), although there is ultimately no way to verify this contention. Although an MNAR mechanism is certainly possible in educational studies, this type of missingness is probably more common in clinical trials where attrition is a result of declining health or death.

Finally, it is important to point out that the missing data mechanisms are not characteristics of an entire data set, but they are assumptions that apply to specific analyses. Consequently, the same data set may produce analyses that are MCAR, MAR, or MNAR depending on which variables are included in the analysis. Returning to the previous math aptitude example, suppose that the parameter of interest is the mean course grade. Recall that the missing course grades are related to math aptitude, such that students with low-aptitude scores have missing data. Even though aptitude is not of substantive interest, the MAR assumption is only satisfied if this variable is part of the statistical analysis. Excluding the aptitude scores would likely bias the resulting mean estimate because the analysis is consistent with an MNAR mechanism (in effect, omitting the "cause" of missingness induces a spurious correlation between the probability of missing data and

course grades). Although the magnitude of the bias depends on the correlation between the omitted aptitude variable and the course grades (bias increases as the correlation increases), the analysis is nevertheless consistent with an MNAR mechanism. Later in the manuscript, we describe methods for incorporating so-called auxiliary variables that are related to missingness into a statistical analysis. Doing so can mitigate bias (i.e., by making the MAR mechanism more plausible) and can improve power (i.e., by recapturing some of the missing information).

An overview of traditional missing data techniques

Traditionally, researchers have employed a wide variety of techniques to deal with missing values. The most common of these techniques include deletion and single imputation approaches (Peugh & Enders, 2004). The goal of this section is to provide an overview of some of these common traditional missing data techniques and to illustrate the shortcomings of these procedures. To illustrate the bias that can result from the use of traditional missing data methods, we use the artificial math performance data set found in Table 1. We created this data set to mimic a situation where a math aptitude test determines participation in an advanced math course. The first column in Table 1 shows scores from this aptitude test. Students with high math aptitude scores participate in the advanced math

Table 1
Math performance data set.

Complete data		Observed data	Mean imputation	Regression imputation [a]	Stochastic regression imputation [a]	
Math aptitude	Course grade	Course grade	Course grade	Course grade	Random error	Course grade
4.0	71.00	–	81.80	65.26	7.16	72.42
4.6	87.00	–	81.80	68.22	0.73	68.95
4.6	74.00	–	81.80	68.22	12.01	80.23
4.7	67.00	–	81.80	68.71	−7.91	60.81
4.9	63.00	–	81.80	69.70	−4.07	65.63
5.3	63.00	–	81.80	71.68	27.41	99.09
5.4	71.00	–	81.80	72.17	25.76	97.93
5.6	71.00	–	81.80	73.16	2.76	75.92
5.6	79.00	–	81.80	73.16	−11.77	61.39
5.8	63.00	–	81.80	74.15	−0.56	73.59
6.1	63.00	63.00	63.00	63.00	–	63.00
6.7	75.00	75.00	75.00	75.00	–	75.00
6.7	79.00	79.00	79.00	79.00	–	79.00
6.8	95.00	95.00	95.00	95.00	–	95.00
7.0	75.00	75.00	75.00	75.00	–	75.00
7.4	75.00	75.00	75.00	75.00	–	75.00
7.5	83.00	83.00	83.00	83.00	–	83.00
7.7	91.00	91.00	91.00	91.00	–	91.00
8.0	99.00	99.00	99.00	99.00	–	99.00
9.6	83.00	83.00	83.00	83.00	–	83.00
Mean	76.35	81.80	81.80	76.12		78.70
Std. Dev.	10.73	10.84	7.46	9.67		12.36

[a] Imputation regression equation: $\hat{Y} = 45.506 + 4.938(\text{Aptitude})$.

class, whereas students below a cut-off do not. Consequently, the data are MAR because only those students with high math aptitude have advanced math course grades. The second column in Table 1 represents the course grades from the hypothetically complete data set (i.e., the data that would have resulted had all students enrolled in the advanced course), and Fig. 1 is a scatterplot of the complete data. The third column of Table 1 represents the observed course grade data (i.e., the course grades for students who scored high enough to enroll in advanced math). We use this small artificial data set to illustrate the impact of different missing data methods have on estimates of descriptive statistics.

Deletion methods

Deletion techniques are perhaps the most basic of the traditional missing data techniques. With listwise deletion (also called complete-case analysis or casewise deletion), cases with missing values are discarded, so the analyses are restricted to cases that have complete data. The major advantage of listwise deletion is that it produces a complete data set, which in turn allows for the use of standard analysis techniques. However, the disadvantages are numerous. Not surprisingly, deleting incomplete records can dramatically reduce the total sample size, particularly for data sets that include a large proportion of missing data or many variables. As a result, significance tests will lack power. More importantly, listwise deletion assumes that the data are MCAR (i.e., missingness is unrelated to all measured variables). When the MCAR assumption is violated – as it often is in real research settings – the analyses will produce biased estimates.

Pairwise deletion (also known as available-case analysis) is another commonly used deletion technique. With pairwise deletion, incomplete cases are removed on an analysis-by-analysis basis, such that any given case may contribute to some analyses but not to others (e.g., each element in a correlation matrix is based on a unique subsample; a set of ANOVA analyses use different subsamples). This approach is often an improvement over listwise deletion because it minimizes the number of cases discarded in any given analysis, but it still suffers from the same major limitation as listwise deletion, namely that the data are MCAR. Like listwise deletion, pairwise deletion can produce biased estimates when the data are inconsistent with an MCAR mechanism.

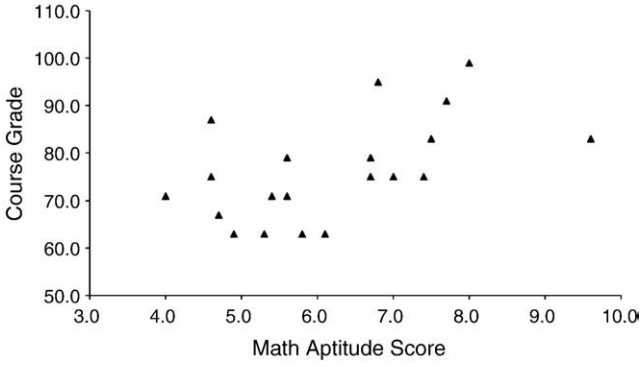

Fig. 1. Complete-data scatterplot of the math performance data in Table 1.

To illustrate the use of deletion methods, we return to the fictitious math performance data set in Table 1. Although the entire sample has data on the aptitude test, listwise deletion restricts all analyses to the cases with complete data (i.e., students that scored high on the aptitude test). Fig. 2 shows a scatterplot of the complete cases (because there are only two variables, the pairwise deletion scatterplot is identical to that of listwise deletion). The positive correlation (r=.55 in the hypothetically complete data set) suggests that the low-aptitude students would have obtained a low course grade, had they taken the class. Because listwise deletion uses the subsample of students with high aptitude scores, it systematically discards cases from the lower tails of both distributions. Consequently, the resulting means are too high and estimates of variability and association are too low. For example, the hypothetical complete data set has mean course grade mean of M=76.35, while the listwise deletion analysis yields an estimate of M=81.80. Similarly, the aptitude means are M=6.20 and M=7.35 for the complete-data and listwise analyses, respectively. Listwise deletion also attenuates estimates of variation and association. For example, the aptitude scores have a standard deviation of SD=1.41 in the complete data set, whereas the listwise estimate is SD=.97. Perhaps not surprisingly, restricting the variability of the data also attenuates the correlation between the two variables. Comparing the scatterplots in Figs. 1 and 2, the listwise deletion scatterplot is more circular than that of the complete data.

Single imputation methods

Single imputation refers to a collection of common traditional missing data techniques where the researcher imputes (i.e., "fills in") the missing data with seemingly suitable replacement values. Methodologists have outlined dozens of single imputation techniques, but we will focus on three of the more common approaches: mean imputation, regression imputation, and stochastic regression imputation. Again, we use the small artificial data set in Table 1 to illustrate the impact of single imputation.

Mean imputation replaces missing values with the arithmetic mean of the available data. Returning to the fictitious math performance data set, the mean course grade for the

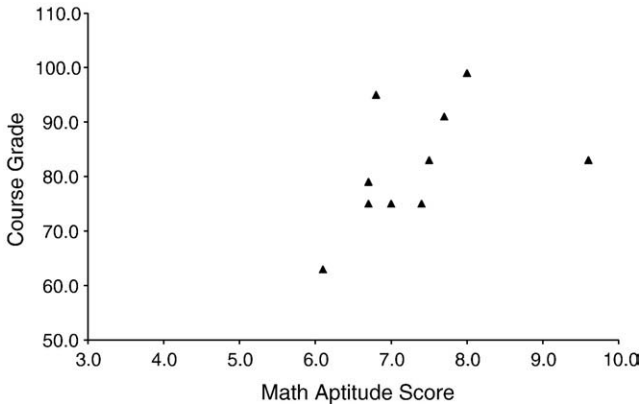

Fig. 2. Listwise deletion scatterplot of the math performance data in Table 1.

observed data is $M=81.80$. As shown in the fourth column of Table 1, this mean replaces the missing course grades, and the subsequent analyses treat these imputed values as if they are real data. Fig. 3 shows a scatterplot that illustrates mean imputation. In the figure, notice that the imputed values from mean imputation fall directly on a straight line that has a slope of zero. This implies that the subset of cases with imputed values has a correlation of zero between aptitude and course grades. Not surprisingly, filling in the data with a set of uncorrelated cases can dramatically attenuate the overall correlation estimate. For example, mean imputation gives a correlation of $r=.21$, whereas the complete-data correlation is $r=.55$. Filling in the missing scores with a constant value also attenuates the variability of the data. For example, the course grade standard deviation from mean imputation is SD=7.46 as opposed to SD=10.73 from the hypothetically complete data. Finally, with MAR data, mean imputation also produces biased mean estimates. In this example, the mean course grade is inflated because the missing values from the lower tail of the score distribution are replaced with an inappropriately high score; the hypothetical complete data set has mean course grade mean of $M=76.35$, while mean imputation yields an estimate of $M=81.80$.

As its name implies, regression imputation replaces missing values with predicted scores from a regression equation. In a bivariate analysis with missing data on a single variable, the complete cases are used to estimate a regression equation where the incomplete variable serves as the outcome and the complete variable is the predictor. The resulting regression equation generates predicted scores for the incomplete cases. To illustrate regression imputation, reconsider the math aptitude data in Table 1. Using the 10 complete cases to estimate the regression of course grades (the incomplete variable) on aptitude scores (the complete variable) yields the following equation: $\hat{Y}=45.506+4.938(\text{Aptitude})$. Substituting the aptitude scores for the incomplete cases into this equation yields predicted course grades, and these predicted values fill in the missing data. For example, Case 1 has an aptitude score of 4.0. Substituting this score into the previous regression equation yields an imputed course grade of 65.26. The fifth column of Table 1 shows the imputed values for

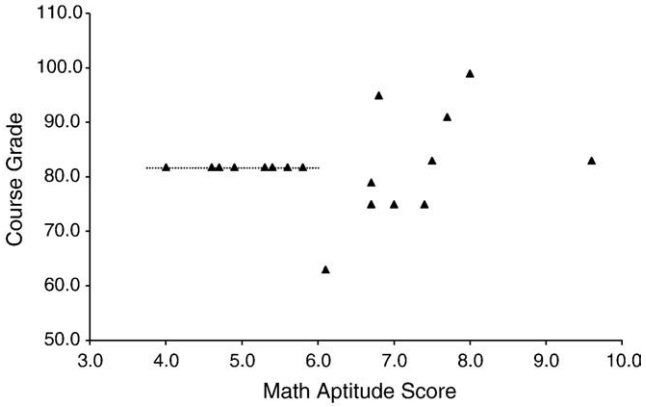

Fig. 3. Mean imputation scatterplot of the math performance data in Table 1.

the 10 incomplete cases. Like mean imputation, the replacement values function as real data in the subsequent statistical analyses.

Although the strategy of borrowing information from the complete variables is a good one, regression imputation also produces biased parameter estimates. Fig. 4 shows the regression imputation scatterplot from the artificial math aptitude data. Notice that the imputed values fall directly on a straight line with a non-zero slope. This implies that the subset of cases with imputed values has a correlation of one between aptitude and course grades. Not surprisingly, filling in the data with a set of perfectly correlated cases can overestimate the overall correlation estimate. For example, regression imputation gives a correlation of $r = .72$, whereas the complete-data correlation is $r = .55$. Because the imputed values fall directly on a straight line, the filled-in data also lack variability that would have been present in the hypothetically complete data set. Consequently, the imputation procedure attenuates estimates of variability. For example, the course grade standard deviation from regression imputation is SD = 9.67 as opposed to SD = 10.72 from the hypothetically complete data. Although regression imputation biases measures of association and variability, it does produce unbiased estimates of the mean when the data are MCAR or MAR.

Both mean imputation and regression imputation lead to bias because they fail to account for the variability that is present in the hypothetical data values. In an attempt to remedy this issue, some researchers use a modified version of regression imputation called stochastic regression imputation. Like standard regression imputation, a regression equation generates predicted values for the incomplete variables. However, after computing the predicted values, a random error term is added to each predicted score and the resulting sum is used in place of missing values. The error term is a random number generated from a normal distribution with a mean of zero and a variance equal to the residual variance from the preceding regression analysis. Returning to the small artificial data set in Table 1, the regression of course grades on aptitude scores has a residual variance of 106.40. The sixth column of the table shows the residual terms for the 10 incomplete cases. Again, these residual terms are randomly drawn from a normal distribution with a mean of zero and a

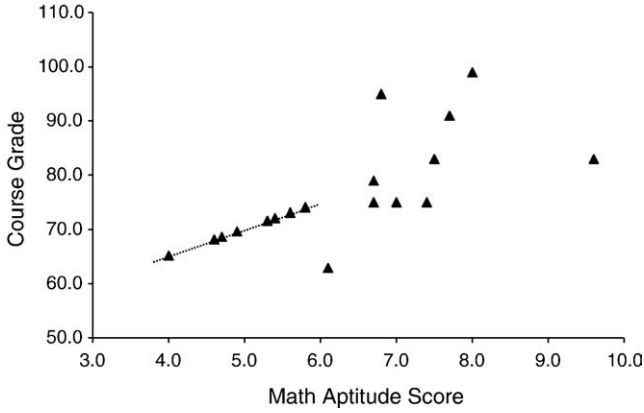

Fig. 4. Regression imputation scatterplot of the math performance data in Table 1.

variance of 106.40. Adding the residuals to the predicted scores from regression imputation generates the imputed values. The right-most column of Table 1 shows the filled-in data for the incomplete cases.

Fig. 5 shows the stochastic regression imputation scatterplot from the artificial math aptitude data. As evident in the plot, the imputed values no longer fall directly on a regression line, so the addition of the random error terms effectively restores lost variability to the data (the distances between the data points and the dashed regression line are the random residual terms). In fact, the stochastic regression scatterplot closely resembles the complete-data scatterplot in Fig. 1. It ends up that this slight modification to regression imputation produces parameter estimates that are unbiased under both the MCAR and MAR assumptions (the same is true for the two modern missing handling approaches in the next section). Despite yielding comparable estimates to maximum likelihood and multiple imputation, stochastic regression imputation provides no mechanism for adjusting the standard errors to compensate for the fact that the imputed values are just guesses about the true score values. Consequently, the standard errors are inappropriately small, and significance tests will have excessive Type I error rates.

Modern missing data techniques

Maximum likelihood estimation and multiple imputation are considered "state of the art" missing data techniques (Schafer & Graham, 2002) and are widely recommended in the methodological literature (e.g. Schafer & Olsen, 1998; Allison, 2002; Enders, 2006). These approaches are superior to traditional missing data techniques because they produce unbiased estimates with both MCAR and MAR data. Furthermore, maximum likelihood and multiple imputation tend to be more powerful than traditional data techniques because no data are "thrown out." This article will focus on introducing maximum likelihood and multiple imputation at a broad and conceptual level. Readers interested in an in-depth treatment are encouraged to seek more detailed sources (e.g., Enders, 2006, 2010; Schafer, 1997; Schafer & Olsen, 1998; Little & Rubin, 2002). Despite their advantages, it is

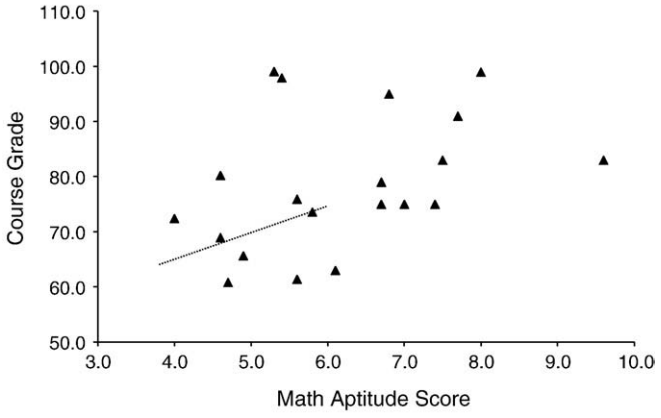

Fig. 5. Stochastic regression imputation scatterplot of the math performance data in Table 1.

important to reiterate that maximum likelihood and multiple imputation are not a perfect fix. They too will yield biased parameter estimates when the data are MNAR, although the magnitude of this bias tends to be far less than the bias that results from traditional techniques.

Multiple imputation

Multiple imputation creates several copies of the data set, each containing different imputed values. Analyses are then carried out on each data set using the same procedures that would have been used had the data been complete. Analyzing each data set separately yields multiple sets of parameter estimates and standard errors, and these multiple sets of results are subsequently combined into a single set of results. These three steps (imputing the data, analyzing the data, and pooling the results) are sometimes referred to as the imputation phase, the analysis phase, and the pooling phase, respectively. This section gives a brief description of each of these steps. A variety of sources give additional details on multiple imputation (Allison, 2002; Enders, 2010; Rubin, 1987, 1996; Schafer & Olsen, 1998; Schafer, 1997; Sinharay, Stern, & Russell, 2001).

The first part of a multiple imputation analysis is the imputation phase. The imputation phase generates a specified number of data sets, each of which contains different estimates of the missing values (20 data sets is a good rule of thumb; Graham, Olchowski, & Gilreath, 2007). Various algorithms have been proposed for the imputation phase, but the data-augmentation procedure is arguably the most widely-used approach for normally distributed data. Data augmentation imputes data sets using a two-step iterative algorithm. The first step, the imputation step (I-step), is procedurally identical to stochastic regression imputation. Specifically, estimates of the means and the covariances are used to construct a set of regression equations that predict the incomplete variables from the complete variables. These regression equations produce predicted scores for the missing values, and a normally distributed residual term is added to each predicted value to restore variability to the data. The filled-in data are carried forward to the posterior step (P-step), where Bayesian estimation principles are used to generate new estimates of the means and the covariances (the parameters that are the building blocks of the I-step regression equations). Conceptually, the P-step computes the means and the covariances from the filled-in data set, then adds a random residual term to each of the resulting estimates. This procedure creates a new set of parameter values that randomly differ from those that were used to create the imputed values in the preceding I-step. Using these updated parameter values to construct a new set of regression equations in the next I-step produces a new set of imputations, the values of which also differ from those at the previous I-step. Repeating this two-step process many times yields multiple copies of the data set, each of which contains unique estimates of the missing values.

To illustrate the imputation phase, we will again use the small math performance data set from Table 1. Table 2 shows four imputed data sets created with the data-augmentation algorithm. To be clear, using only four data sets is usually not sufficient, but we arbitrarily chose this number for illustration purposes only. To begin, notice that the cases with missing course grades have different imputed values in each data set (e.g., Case 1 has imputed values of 51.48, 67.91, 69.38, and 72.45), whereas the cases with complete data

Table 2
Imputed course grades from multiple imputation procedure.

Observed data		Imputed course grades			
Math aptitude	Course grade	Data Set 1 [a]	Data Set 2 [b]	Data Set 3 [c]	Data Set 4 [d]
4.00	–	51.48	67.91	69.38	72.45
4.60	–	59.53	62.59	74.19	57.38
4.60	–	62.34	59.77	67.43	46.47
4.70	–	68.45	53.56	71.39	56.99
4.90	–	75.47	63.79	72.54	85.96
5.30	–	81.81	57.16	70.99	68.71
5.40	–	61.05	90.47	56.25	74.11
5.60	–	77.72	46.92	69.14	52.91
5.60	–	71.49	70.79	73.89	72.44
5.80	–	68.36	59.98	67.04	77.53
6.10	63.00	63.00	63.00	63.00	63.00
6.70	75.00	75.00	75.00	75.00	75.00
6.70	79.00	79.00	79.00	79.00	79.00
6.80	95.00	95.00	95.00	95.00	95.00
7.00	75.00	75.00	75.00	75.00	75.00
7.40	75.00	75.00	75.00	75.00	75.00
7.50	83.00	83.00	83.00	83.00	83.00
7.70	91.00	91.00	91.00	91.00	91.00
8.00	99.00	99.00	99.00	99.00	99.00
9.60	83.00	83.00	83.00	83.00	83.00
Mean	81.80	74.79	72.55	75.51	74.15
SE	10.84	12.18	14.53	10.49	13.81

[a] Imputation regression equation: $\hat{Y} = 6.03(\text{Aptitude}) + 33.92$.
[b] Imputation regression equation: $\hat{Y} = 5.11(\text{Aptitude}) + 38.49$.
[c] Imputation regression equation: $\hat{Y} = 5.62(\text{Aptitude}) + 41.15$.
[d] Imputation regression equation: $\hat{Y} = 6.40(\text{Aptitude}) + 31.54$.

have constant values. Conceptually, the imputed values represent random samples from a distribution of plausible replacement values for the missing data. The fact that the imputed values differ across data sets illustrates the crucial differences between multiple imputation and stochastic regression imputation (or any single imputation technique, for that matter): rather than placing stock in any one set of imputed values, multiple imputation uses a number of filled-in data sets to account for the uncertainty in the missing data. As we described previously, a slightly different regression equation (or set of equations) fills in each data file, and the footnote of the table gives the equations from the example analysis. You can think of these equations as a random sample from a population of plausible regressions that describe the relationship between aptitude and course grades. Mathematically, the differences among the four equations result from the random perturbations that occur at each P-step.

As an important aside, the imputed values in Table 2 were not drawn from successive I-steps. In a multiple imputation analysis, it is important that the imputed values in one data

set be independent of the imputed values from other data sets. In practice, this is accomplished by allowing a large number of iterations to lapse between each saved data set. For example, we allowed 200 data-augmentation cycles (i.e., 200 I- and P-steps) to lapse between each of the data sets in Table 2. That is, we allowed the algorithm to run for 200 iterations, saved the first data set, allowed the algorithm to run for another 200 iterations, saved the second data set, and so on. The exact spacing of the saved data sets is an important practical issue that arises in a multiple imputation analysis. Additional details on this nuance of the procedure are available from a variety of sources (Enders, 2010; Schafer, 1997; Schafer & Olsen, 1998).

Following the imputation phase is the analysis phase. In the analysis phase, a statistical analysis is performed on each data set using the same techniques that would have been used had the data been complete. The research question dictates this analysis and no changes need to be made as a result of the missing data. Returning to the math performance data set, suppose that it is of interest to estimate the mean course grade. An estimate of the mean and its corresponding standard error would be calculated from each imputed data set. As can be seen from the bottom of Table 2, each filled-in data set yields a slightly different mean estimate.

The analysis phase yields several estimates of each parameter and standard error. Finally, in the pooling phase, the estimates and their standard errors are averaged into a single set of values. Rubin (1987) outlined formulas for pooling the parameter estimates and standard errors. Pooled parameter estimates are calculated by taking the arithmetic mean of the estimates from each data set. Returning to the math performance example, the multiple imputation mean estimate is simply the average of the four estimates in Table 2, which is $\overline{M}=74.25$. Notice that this estimate is quite similar to that of the hypothetically complete data set ($M=76.35$) and is a dramatic improvement over the estimates from the traditional approaches.

Pooling standard errors is slightly more complicated because it involves the standard errors from the imputed data sets (i.e., within-imputation variance) as well as a component that quantifies the extent to which the estimates vary across data sets (i.e., between-imputation variance). More specifically, the within-imputation variance is the arithmetic average of the squared standard errors, as follows,

$$W = \frac{\sum SE_t^2}{m}, \tag{1}$$

where t denotes a particular imputed data set and m is the total number of imputed data sets. The between-imputation variance quantifies the variability of the estimates across data sets,

$$B = \frac{\sum(\hat{\theta}_t - \overline{\theta})^2}{m-1}, \tag{2}$$

where $\hat{\theta}_t$ is the parameter estimate (e.g., the mean course grade) from filled-in data set t, and $\overline{\theta}$ is the average parameter estimate (e.g., $\overline{M}=74.25$). Notice that Eq. (2) is simply the sample variance formula, where the parameter estimates serve as the data points. Finally,

the pooled standard error combines the within- and between-imputation variance, as follows,

$$SE = \sqrt{W + B + B/m}. \tag{3}$$

One of the problems with single imputation techniques is that they underestimate standard errors because they treat the imputed values as real data. Multiple imputation solves this problem by incorporating the between-imputation variance in the standard error (this term represents the additional noise that results from filling in the data set with different estimates of the missing values). In this way, multiple imputation standard errors explicitly account for the fact that the imputed values are fallible guesses about the true data values.

To illustrate the standard error computations, reconsider the math aptitude example. The analysis phase yields four standard error estimates (see the bottom row of Table 2). The within-imputation variance is the average of the squared standard errors, as follows:

$$W = \frac{\sum SE_t^2}{m} = \frac{148.41 + 211.08 + 110.00 + 190.64}{4} = 165.03. \tag{4}$$

Next, the between-imputation variance quantifies the extent to which the mean estimates varied across the four data sets. Substituting the four mean estimates into Eq. (2) gives

$$B = \frac{\sum(\hat{\theta}_t - \overline{\theta})^2}{m-1}$$

$$= \frac{(74.79-74.25)^2 + (72.55-74.25)^2 + (75.51-74.25)^2 (74.15-74.25)^2}{4-1} = 1.59. \tag{5}$$

Finally, the pooled standard error is

$$SE = \sqrt{W + B + B/m} = \sqrt{165.04 + 1.60 + 1.60/4} = 12.92. \tag{6}$$

Although the analysis and pooling phases appear to be very tedious, multiple imputation software packages tend to automate these steps, so it is rarely necessary to compute the pooling equations by hand.

Maximum likelihood estimation

Maximum likelihood estimation is the second modern missing data technique that methodologists currently recommend. Like multiple imputation, this approach assumes multivariate normality and MAR data. However, the mechanics of maximum likelihood are quite different. Rather than filling in the missing values, maximum likelihood uses all of the available data – complete and incomplete – to identify the parameter values that have the highest probability of producing the sample data. This section provides a non-technical overview of the mathematics behind estimation. As a disclaimer, it is important not to let the seemingly complex technical details deter you from using this missing data handling

approach. Software packages that implement maximum likelihood tend to be user-friendly and do not require knowledge of the mathematical intricacies in this section. Nevertheless, it is useful to have some idea about the inner workings of the "black box." Maximum likelihood estimation routines are readily available in structural equation modeling software packages, and we illustrate an analysis later in the manuscript.

At a broad level, maximum likelihood estimation identifies the population parameter values that have the highest probability of producing the sample data. This estimation process uses a mathematical function called a log likelihood to quantify the standardized distance between the observed data points and the parameters of interest (e.g., the mean), and the goal is to identify parameter estimates that minimize these distances. This is conceptually similar to ordinary least squares estimation, where the goal is to identify the regression coefficients that minimize the collective distances between the data points and the predicted scores.

Understanding the basic mechanics of maximum likelihood estimation is easiest in a univariate context, and the basic ideas readily extend to multivariate data. To begin, the log likelihood for a sample scores is

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right]. \quad (7)$$

Although the equation for the log likelihood function is rather complex, the most important components are familiar. The term in brackets is known as the probability density function and is the mathematical equation that describes the shape of the normal curve. Collectively, the bracketed terms quantify the relative probability of obtaining a single score from a normally distributed population with a particular (unknown) mean and standard deviation. The summation symbol adds the relative probabilities into a summary measure (the sample log likelihood) that quantifies the probability of drawing the entire sample from a normally distributed population.

In terms of understanding estimation, the crucial component of Eq. (7) is the squared *z*-score that appears in the exponent of the function. Substituting parameter values (i.e., the mean and the standard deviation) and a score into this component of the function gives a value that quantifies the standardized distance between that data point and the mean. Score values that are close to the mean produce a small *z*-score and a large log likelihood (i.e., relative probability) value, whereas scores that are far from the mean produce larger *z*-scores and smaller log likelihoods. The parameter estimates that maximize the sum of the individual log likelihood values are the so-called maximum likelihood estimates. Of course, the difficulty is determining exactly which values of the population mean and standard deviation produce the highest log likelihood. In most situations, an iterative algorithm repeatedly substitutes different parameter values into the equation until it identifies the estimates that minimize the distance to the data (i.e., produce the highest log likelihood value, or probability).

It is easiest to demonstrate how the log likelihood function works if we first pretend that the population parameter values are known. Continuing with the math performance data set, assume that the course grade variable has a population mean and variance of $\mu = 75$ and $\sigma^2 = 114$, respectively. If the population parameters are known, then there is only one

variable in the equation — the y_i score values in the squared z-score portion of the function. Substituting a score of 71 into Eq. (7) yields a log likelihood value of −3.36, and substituting a score of 87 yields a log likelihood value of −3.92. Note that a score of 71 is closest to the population mean (and hence the center of the distribution), so the resulting log likelihood (i.e., relative probability) value is higher than the log likelihood for a score of 87. Said differently, a score of 71 has a higher probability of originating from a normal distribution with a mean of 75. The log likelihood scale expresses individual probabilities on a negative metric, and higher values (i.e., less negative values) are associated with scores that provide better fit to the parameters (i.e., a score that is closer to the mean). Repeating this calculation for the entire sample and summing the individual log likelihood values yields the sample log likelihood. Essentially, the sample log likelihood is a measure that quantifies the fit of the entire data set to particular set of parameter values.

In reality, the population parameters are unknown and need to be estimated from the data. During estimation, maximum likelihood "auditions" different parameter values by substituting them into the log likelihood function and computing the sample log likelihood until it finds the estimates that best fit the data. That is, it tries out different estimates with the goal of identifying the values that produce the largest sample log likelihood, or the largest probability. Similar to ordinary least squares estimates, the resulting maximum likelihood estimates minimize the standardized distances to the data points.

Eq. (7) is the log likelihood formula for univariate estimation. In multivariate estimation, the basic logic is the same but matrices replace the scalar values. For example, the matrix expression of the squared z-score component becomes

$$(\mathbf{Y}_i-\boldsymbol{\mu})^T\sum^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}). \tag{8}$$

In the matrix formula, $\boldsymbol{\mu}$ represents the mean population vector, $\boldsymbol{\Sigma}$ is the population covariance matrix, $\mathbf{Y}_i$ is a vector that contains the scores for a single individual, the T symbol represents the matrix transpose (i.e., rows become columns, columns become rows), and -1 denotes the inverse (i.e., the matrix analog to division). Although the nuisances of matrix algebra require the formula to look slightly different than its univariate counterpart, Eq. (8) still yields a squared z-score. In this case, the squared z-score quantifies the standardized distance between a *set* of scores for a particular individual and the population means. Despite this additional complication, the underlying mechanics of estimation remain the same: an iterative algorithm auditions different sets of parameter values until it identifies those that produce the largest log likelihood value (i.e., the best fit to the sample data).

Up until now, we have sidestepped the issue of missing data handling, but the basic ideas behind maximum likelihood estimation change very little when the data are incomplete. Importantly, the log likelihood function does not require complete data. With missing data, an individual's squared z-score is computed using whatever data are available for that person, and the fit for the entire sample (i.e., the sample log likelihood) is simply a weighted sum of the individual fit values. In this way, maximum likelihood estimates the parameters without discarding data and without filling in the data.

To illustrate how the log likelihood function works with missing data, reconsider the small math aptitude data set in Table 1. The subset of students with complete data

contributes two scores (math aptitude and course grade) to the calculation of the log likelihood, whereas the students with missing course grades only contribute math aptitude scores. For example, the squared *z*-score formula for the first subject is as follows:

$$(\mathbf{Y}_i-\boldsymbol{\mu})^T\sum^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}) = (4.0-\mu_{MA})^T\frac{1}{\sigma^2_{MA}}(4.0-\mu_{MA}), \tag{9}$$

where μ_{MA} and σ^2_{MA} are the unknown population mean and variance, respectively, of the math aptitude scores. Note that this student's course grade is missing, so it is not included in the calculation. As a second example, consider the last person in the data set. This student has data on both variables, so the resulting squared *z*-score formula is computed as follows:

$$(\mathbf{Y}_i-\boldsymbol{\mu})^T\sum^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}) = \left(\begin{bmatrix}9.6\\83.0\end{bmatrix}-\begin{bmatrix}\mu_{MA}\\\mu_{CG}\end{bmatrix}^T\right)\begin{bmatrix}\sigma^2_{MA} & \sigma_{MA,CG}\\\sigma_{CG,MA} & \sigma^2_{CG}\end{bmatrix}^{-1}\left(\begin{bmatrix}9.6\\83.0\end{bmatrix}-\begin{bmatrix}\mu_{MA}\\\mu_{CG}\end{bmatrix}\right). \tag{10}$$

Although it is not at all obvious, including the partially complete data records in the estimation process produce very different estimates than you would get by excluding the incomplete cases. Conceptually, estimation borrows information from the complete variables en route to identifying the most likely values of the population parameters. For example, consider the bivariate data set in Table 1. Because the two variables are positively correlated, the presence of a low-aptitude score in the log likelihood computations (e.g., the squared *z*-score in Eq. (9)) implies that the missing course grade would have also been low, had it been observed. Consequently, including the partial data records in the estimation process results in a downward adjustment to the course grade mean (recall that listwise deletion produced an estimate that was too high because it ignored the data in the lower tail of the score distributions). Specifically, we used a structural equation modeling program to estimate the descriptive statistics from the data in the table and obtained a mean estimate of $M=76.12$. Notice that this estimate is quite similar to that of the hypothetically complete data set ($M=76.35$) and is a dramatic improvement over the estimates from the traditional approaches. Again, it is important to reiterate that maximum likelihood neither discards data nor does it impute the data. It simply estimates the parameters of interest using all of the available data.

Maximum likelihood and multiple imputation both yield unbiased parameter estimates with analyses that satisfy the MAR assumption. From a practical standpoint, this means these approaches will produce accurate estimates in situations where traditional approaches fail (e.g., the analyses from the data in Table 1). Even when the data are MCAR, maximum likelihood and multiple imputation significance tests will tend to be more powerful than those from other techniques because the incomplete variables essentially borrow strength from the complete variables (e.g., when two variables are correlated, the observed data carry information about the missing data). The major downfall of the two modern approaches is that they rely on the untestable MAR assumption. Even with this potential short-coming, maximum likelihood and multiple imputation are superior methods for handling missing data because traditional techniques also yield biased estimates with MNAR data. Furthermore, some methodologists have argued that routine departures from

MAR may not be large enough to cause serious bias in the resulting estimates (Schafer & Graham, 2002).

Analysis example 1

The previous bivariate analysis example used a fictitious and somewhat unrealistic data set. This section describes analyses of a real-world data set that uses maximum likelihood to estimate correlations, means, standard deviations, and a multiple regression model. Here, we have chosen to limit our analytic example to maximum likelihood estimation because this procedure is often easier to implement than multiple imputation. Every commercially available structural equation modeling software package offers maximum likelihood missing data handling, so this framework will be familiar to many researchers. As an aside, maximum likelihood and multiple imputation tend to produce similar estimates, so choosing between the two methods is largely a matter of personal preference. Although there are some situations where multiple imputation provides flexibility that maximum likelihood does not, many analyses that are common to school psychology research (e.g., correlation, regression, ANOVA, structural equation models, longitudinal growth curve analyses) are quite easy to estimate in the structural equation modeling framework. Analysis examples that involve multiple imputation can be found elsewhere in the literature (Allison, 2002; Enders, 2006, 2010; Peugh & Enders, 2004; Schafer & Olsen, 1998).

This illustrative example uses a sample of 1094 cases extracted from the LSAY data set. This data set includes complete information on ethnicity (we represent this variable as two dummy codes: Hispanic = 1 and Caucasian/African American = 0, African-American = 1 and Caucasian/Hispanic = 0). The data set is also comprised of the following incomplete variables: parental math encouragement in 9th grade (16.2% missing), parental college encouragement in 9th grade (14.8% missing), parental academic encouragement in 9th grade (16.2% missing), and grade 12 math scores (33.4% missing). Note that the encouragement variables are ordinal scales, but we treat these variables as though they are continuous for the purposes of illustration. Readers who are interested in recreating this analysis example can contact the authors of this manuscript for a copy of the data set and the computer code.

The ultimate goal of this analysis is to determine whether the relationship between the parental encouragement variables and math scores is different across ethnic groups. That is, we wanted to determine whether ethnicity moderates the association between parental encouragement during middle school and later academic achievement. To do so, we performed a series of regression analyses that added sets of variables in sequential blocks. Block 1 was comprised of the two ethnicity dummy variables. In Block 2, we added the three continuous parental encouragement variables. Finally, in Block 3, we added the six interaction terms that were computed by multiplying the ethnicity dummy codes by the parental encouragement variables. Consistent with recommendations from the multiple regression literature (e.g., Aiken & West, 1991), the continuous predictor variables were centered at their respective grand mean prior to computing the interaction terms (see the Fairchild and McQuillin paper in this issue for more details on moderated regression).

The regression model from the final block is as follows:

$$\begin{aligned} Math\,12 = {} & B_0 + B_1(Hispanic) + B_2(AfriAmer) + B_3(Collenc9) \\ & + B_4(MathEnc9) + B_5(AcadEnc9) + B_6(Hispanic)(CollEnc9) \\ & + B_7(Hispanic)(MathEnc9) + B_8(Hispanic)(AcadEnc9) \\ & + B_9(AfriAmer)(CollEnc9) + B_{10}(AfriAmer)(MathEnc9) \\ & + B_{11}(AfriAmer)(AcadEnc9) + e. \end{aligned} \tag{11}$$

The model in Eq. (11) is depicted by the path diagram in Fig. 6. In path diagrams, single-headed straight arrows denote regression coefficients, double-headed curved arrows represent correlations, rectangles denote manifest (observed) variables, and ellipses represent latent (unobserved) variables (Bollen, 1989; Kline, 2005). Because a residual (error) term is essentially a collection of unobserved influences, the residual term is represented as an ellipse. We used the Mplus software program (Muthén & Muthén, 2004) to estimate the multiple regression model in Fig. 6. The syntax for two of the analyses is given in Appendices A and B.

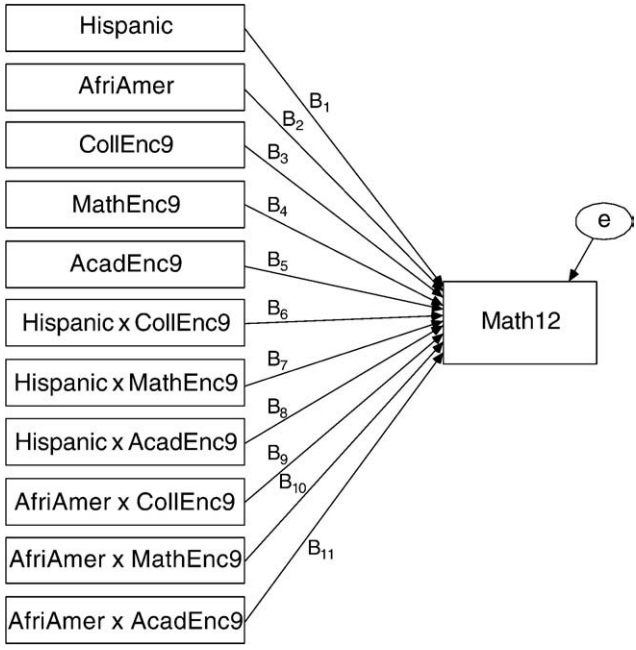

Fig. 6. Path diagram of Block 3 regression model. Single-headed straight arrows denote regression coefficients. Note that the correlations among the predictor variables (i.e., curved arrows connecting each pair of variables on the left side of the model) are omitted from the graph in order to reduce clutter.

In the first block of the regression model, we included the two ethnicity dummy variables. The coefficients and their standard errors are provided in the top portion of Table 3. Despite the shift from ordinary least squares to maximum likelihood estimation, the interpretation of the regression coefficients does not change. Specifically, the coefficients in the table represent mean differences between Hispanics and African Americans relative to Caucasians (the group with zero codes on both dummy variables). By virtue of our coding scheme, the regression intercept (i.e., $B_0 = 71.06$) is the math achievement mean for Caucasians. The coefficient for the Hispanic dummy code ($B_1 = -8.89$, $SE = 1.74$, $p < .001$) indicates that the Hispanic mean was nearly nine points lower than the Caucasian mean. Similarly, the coefficient for the African American dummy code ($B_2 = -10.14$, $SE = 1.54$, $p < .001$) indicates a math grade difference of nearly ten points between African Americans and Caucasians. In total, the ethnicity dummy codes explain 6.6% of the variability in 12th grade math scores.

In Block 2, we added the three parental encouragement variables. The regression coefficients and their standard errors are provided in the middle section of Table 3. The B_1 and B_2 coefficients are mean differences, controlling for the parental encouragement variables. The B_3 coefficient represents the expected change in 12th grade math performance for a one-unit increase in 9th grade college encouragement, controlling for

Table 3
Regression model estimates from data analysis 2.

Parameter	Estimates	*SE*	*p*
Block 1 ($R^2 = .066$)			
B_0 (Intercept)	71.062	0.530	<.001
B_1 (Hispanic)	−8.893	1.741	<.001
B_2 (AfriAmer)	−10.144	1.535	<.001
Block 2 ($R^2 = .211$)			
B_0 (Intercept)	71.101	0.494	<.001
B_1 (Hispanic)	−8.110	1.642	<.001
B_2 (AfriAmer)	−11.878	1.440	<.001
B_3 (CollEnc)	3.597	0.464	<.001
B_4 (MathEnc)	3.912	0.613	<.001
B_5 (AcadEnc)	0.137	0.219	0.53
Block 3 ($R^2 = .215$)			
B_0 (Intercept)	71.104	0.493	<.001
B_1 (Hispanic)	−15.504	4.823	0.00
B_2 (AfriAmer)	−6.995	5.872	0.23
B_3 (CollEnc)	3.488	0.514	<.001
B_4 (MathEnc)	3.973	0.679	<.001
B_5 (AcadEnc)	0.082	0.246	0.74
B_6 (Hispanic × CollEnc)	1.548	1.901	0.42
B_7 (Hispanic × MathEnc)	−1.211	2.297	0.60
B_8 (Hispanic × AcadEnc)	1.012	0.694	0.14
B_9 (AfriAmer × CollEnc)	−0.113	1.442	0.94
B_{10} (AfriAmer × MathEnc)	0.244	2.006	0.90
B_{11} (AfriAmer × AcadEnc)	−0.669	0.738	0.36

ethnicity and the two other encouragement variables. Likewise, the B_4 and B_5 coefficients represent the effects of math encouragement and academic encouragement, respectively, after controlling for all other variables. As seen in the table, college and math encouragement both had significant partial regression coefficients ($p < .001$ for both variables), but the effect of academic encouragement was non-significant ($p = .53$). Including the encouragement variables increased the R^2 statistic from .066 to .211.

In the final block, we added the six interaction terms to the model. Estimates for this model are presented in the bottom portion of Table 3. The B_1 and B_2 coefficients are still mean differences, but they now partial out the encouragement variables and the interaction terms. In the presence of the interaction terms, the B_3, B_4, and B_5 coefficients change meaning. Specifically, these terms reflect the partial regression slopes for Caucasians (i.e., the group with zero codes on the two dummy variables). The B_6, B_7, and B_8 coefficients quantify the difference between the regression slopes for Hispanic and Caucasian students. For example, B_6 1.55 can be interpreted to mean that the partial regression coefficient for the college encouragement variable is 1.55 points higher in the Hispanic subsample. In a similar vein, the B_9, B_{10}, and B_{11} coefficients represent slope differences between African Americans and Caucasians. As evident in Table 3, this model contains many non-significant coefficients, including all of the interaction terms. In addition, adding the interaction terms to the model increased the R^2 statistic from .211 to .215.

In a typical regression analysis, researchers routinely begin by examining the omnibus F-test. For example, using the F-statistic to test the incremental contribution of each block in a hierarchal regression analysis is a common procedure. In an ML analysis, the likelihood ratio statistic is an analogous test. Like the F-statistic, the likelihood ratio is a test that compares the relative fit of two nested models. Nested models can take on a variety of different forms, but a common example occurs when the parameters from the nested model are a subset of the parameters from the full model. In our example, the Block 2 regression model is nested within the Block 3 regression model, and the Block 1 regression model is nested within the Block 2 regression model (Block 1 is also nested within Block 3). The difference between the log likelihood values from two nested models provides the basis for a significance test, as follows:

$$LR = -2(logL_{\text{Nested}} - logL_{\text{Full}}), \tag{12}$$

where $\log L_{\text{Nested}}$ and $\log L_{\text{Full}}$ are the sample log likelihood values from the nested (e.g., Block 2) and full (e.g., Block 3) models, respectively. The probability value for the likelihood ratio statistic is determined using a chi-square distribution, where degrees of freedom are calculated as the difference between the number of estimated parameters from each model. For example, the Block 2 regression model has six fewer parameter estimates than the Block 3 regression model (i.e., B_6, B_7, B_8, B_9, B_{10}, and B_{11}), so the likelihood ratio statistic that compares the incremental contribution of the interaction terms would follow a chi-square distribution with six degrees of freedom. When using maximum likelihood estimation, the likelihood ratio test can be used in place of the familiar F-tests. We illustrate the use of the likelihood ratio test in the context of the previous regression analysis. Table 4 details the results of the likelihood tests.

Table 4
Log likelihood and likelihood ratio tests from analysis example 1.

Model	Estimates	log*L*	LR	*df*	*p*
Intercept only	2	−11,042.442			
Block 1	4	−11,011.886	61.112	2	<.001
Block 2	7	−10,944.763	134.246	3	<.001
Block 3	13	−10,942.436	4.654	6	0.59

First, we compare the regression model with the two ethnicity variables (Block 1) to a nested model that contains only the regression intercept (i.e., a model that constrains all regression coefficients to zero during estimation). Using the likelihood ratio statistic to compare these models is analogous to using an *F*-test to determine whether the R^2 value from the first block ($R^2 = .066$) is significantly different from zero. To begin, the first row of Table 4 gives the sample log likelihood value from a model where the regression coefficients are constrained to zero (i.e., the intercept-only model). This "empty" model includes only two parameter estimates: the regression intercept (i.e., grand mean of 12th grade math scores) and the variance of math scores. Estimating the Block 1 model adds two parameter estimates (the regression coefficients for the two ethnicity dummy codes). As discussed earlier, log likelihood values are scaled such that higher (i.e. less negative) values reflect better fit. Consequently, the increase from −11,042.44 to −11,011.89 indicates that the ethnicity variables improve model fit relative to the empty model. Substituting the log likelihood values into Eq. (12) yields a likelihood ratio statistic of LR = 61.11. If the null hypothesis is true (i.e., the nested model is equivalent to the full model), then the likelihood ratio test is distributed as a chi-square statistic with two degrees of freedom because the Block 1 model has two additional parameter estimates (i.e., B_1 and B_2). Referencing LR to a chi-square distribution with two degrees of freedom returns a *p*-value that is less than .001, indicating that the fit of the model with the ethnicity variables is substantially better than the fit of the intercept-only model. Said differently, the set of predictor variables in the Block 1 model improves fit relative to a model with no predictors. Notice that this is the same interpretation that can be drawn from an omnibus *F*-test that assesses the increment in R^2 due to the block of ethnicity codes.

When performing hierarchal regression analyses, researchers often use *F* change tests to examine the incremental predictive power of subsequent blocks. The likelihood ratio test can also be used for this purpose. To illustrate, we used the likelihood ratio test to compare the Block 2 model (i.e., the model that adds the three parental encouragement variables) to the Block 1 model (i.e., the model that includes only the two dummy codes). As found in the third row of Table 4, the sample log likelihood for the Block 2 model is −10,944.76, which is larger (less negative) than the corresponding value for the Block 1 model (i.e., Block 2 provides a better fit to the data). Substituting the log likelihood values into Eq. (12) yields LR = 134.25, and referencing this statistic to a chi-square distribution with three degrees of freedom (Block 2 has three additional regression coefficients) yields $p < .001$. From this, we can conclude that the addition of the encouragement variables improves model fit above and beyond the ethnicity variables alone. Again, this likelihood ratio test is akin to an *F*-statistic that evaluates the increment in R^2 from Block 1 to Block 2. Lastly, we used the sample log likelihood values to determine if Block 3 improves model fit over

Block 2 (i.e., to assess the incremental contribution of the interaction terms). As seen in Table 4, this model comparison yields a likelihood ratio statistic of LR = 4.65. Referencing LR to a chi-square distribution with six degrees of freedom (Block 3 adds six interaction terms) results in $p = .59$. From this, we can conclude that adding the interaction terms does not significantly improve model fit. Conceptually, this is identical to testing the incremental improvement in R^2 that results from including the interaction terms.

Before proceeding, it is important to note that structural equation modeling software packages are not uniform in their treatment of missing predictor variables. Specifically, some software programs exclude cases that have incomplete data on explanatory variables, while others do not. The Mplus program that we used does allow for missing predictor variables in many models, but taking advantage of this capability requires special care, particularly when likelihood ratio tests are involved. In the previous description of the analyses, we somewhat incorrectly stated that we "added" variables in each block. When the predictor variables have missing data, two models that differ in the number of predictors (e.g., Block 1 versus Block 2) are no longer nested, making the likelihood ratio test invalid. To get around this problem, it is necessary to start with the final model (i.e., the Block 3 model) and work backwards. Specifically, we started by estimating the Block 3 model in Fig. 6. Next, we estimated the Block 2 model by constraining (i.e., fixing) the interaction regression coefficients to zero during estimation. Finally, we estimated the Block 1 model by constraining the interaction and the encouragement coefficients to zero during estimation. This process is conceptually equivalent to adding variables in a step-by-step fashion, but keeping the same set of variables in each of the models ensures that the models are nested and the likelihood ratio tests are valid. Enders (2010) gives additional details on handling missing data on explanatory variables.

Using auxiliary variables to fine-tune a maximum likelihood analysis

Recall from the earlier discussion of missing data mechanisms that MAR is not a characteristic of the entire data set, but is a situation that depends on the variables included in the analysis. Since maximum likelihood and multiple imputation require the MAR assumption, adding so-called auxiliary variables to an analysis can help fine-tune the missing data handling procedure, either by reducing bias or by increasing power. Auxiliary variables are additional variables not required to answer the research question, but are chosen based on their potential correlations with missingness or their correlations with incomplete analysis variables. Adding auxiliary variables may improve the chances of satisfying the MAR mechanism and, as such, these variables have the potential to improve the quality of the resulting estimates. In addition, auxiliary variables that are highly correlated with the incomplete analysis model variables can restore some of the power loss that occurs due to missing data. For this reason methodologists recommend an inclusive analysis strategy that incorporates a number of auxiliary variables into the analysis model (Collins, Schafer, & Kam, 2001; Graham, 2003; Rubin, 1996; Schafer & Graham, 2002).

Returning to the previous LSAY example, 9th grade math performance is an ideal auxiliary variable because it is highly correlated with the 12th grade math achievement. Nearly one-third of the 12th grade scores are missing, so including the 9th grade scores in the analysis could improve power because the incomplete variable can essentially borrow

strength from a variable with far fewer missing values (approximately 6% of the 9th grade scores are missing). One option is to add the 9th grade scores as an additional predictor variable, but this is a bad solution because it accommodates the auxiliary variable by altering the substantive interpretation of the model parameters (i.e., partialling out a variable that would not have appeared in the analysis had the data been complete). Graham (2003) outlined two structural equation modeling strategies for incorporating auxiliary variables into a maximum likelihood analysis — the extra dependent variable model and the saturated correlates model. The basic goal of both approaches is to use a series of correlations to work the auxiliary variables into the analysis without altering the substantive interpretation of the parameters. We describe the saturated correlates model here, and interested readers can consult Graham (2003) for details on the extra dependent variable model. As an aside, no special procedures are required to incorporate auxiliary variables into a multiple imputation analysis. The auxiliary variables simply appear as predictors in the imputation phase, and these additional variables are ignored during the subsequent analysis phase.

In an analysis that involves a set of manifest variables (i.e., a statistical model with no latent variables), Graham's (2003) rules for specifying a saturated correlates model are as follows: correlate an auxiliary variable with explanatory variables, other auxiliary variables, and the residual terms of the outcome variables. Returning to the LSAY regression analysis, we chose three auxiliary variables: 9th grade math scores, a variable that quantifies the amount of math and science resources in the home, and mother's level of education (we discuss the rationale for these choices later). Fig. 7 shows a path model diagram of the saturated correlates model. Note that we excluded the interaction terms from the model because the previous analysis indicated that they were unnecessary.

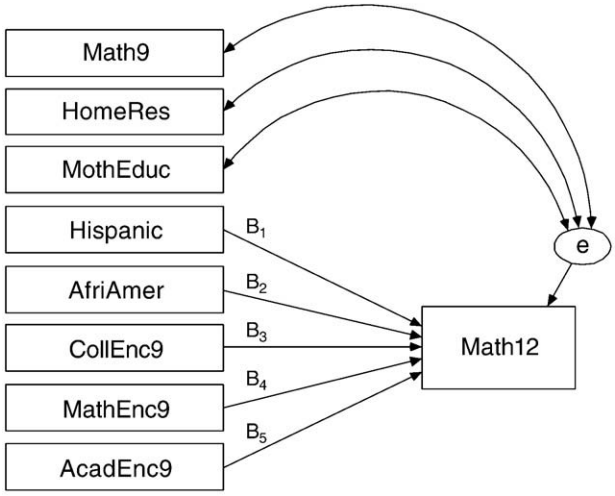

Fig. 7. Path diagram of the auxiliary variable regression model. Note that the correlations among the variables have been omitted from the diagram in order to reduce clutter. The diagram should include curved arrows (i.e., correlations) that connect every pair of variables on the left side of the model (i.e., curved arrows connecting each pair of predictor variables, and curved arrows connecting the auxiliary variables to the predictor variables).

It is important to note that the saturated correlates model transmits the information from the auxiliary variables to the analysis model variables without affecting the interpretation of the parameter estimates. Consequently, the interpretation of the regression coefficients is the same as it was in the previous analysis (e.g., the B_3 coefficient represents the expected change in 12th grade math performance for a one-unit increase in 9th grade college encouragement, controlling for ethnicity and the two other encouragement variables). Adding the auxiliary variables to the model can change the estimated value of the coefficients (e.g., by removing bias or reducing random error) and their standard errors, but the substantive interpretation of the regression slopes is unaffected. As a final note, the rules for incorporating auxiliary variables change slightly when the analysis involves latent variables. See Graham (2003) or Enders (2010, chapter 5) for additional details.

Analysis example 2

To demonstrate the use of auxiliary variables in a maximum likelihood analysis, we will continue with the previous LSAY example. In this particular example, we chose three auxiliary variables: 9th grade math scores, a variable that quantifies the amount of math and science resources in the home, and mother's level of education. In general, a useful auxiliary variable is a potential cause or correlate of missingness or a correlate of the incomplete variables in the analysis model (Collins et al., 2001; Schafer, 1997). In most situations, identifying variables that predict the propensity for missing data involves some educated guesswork. Student mobility is a common cause of attrition in school-based research studies (Enders et al., 2006; Graham et al., 1997), as is socioeconomic status. For this reason, we chose to include mother's education because it a reasonable indicator of socioeconomic status. We chose the other two auxiliary variables primarily because they were correlated with the incomplete analysis variables. In particular, the 9th grade achievement test is an ideal auxiliary variable because it is very highly correlated with 12th grade test ($r = .86$). The home resources variable is somewhat less effective as an auxiliary variable because it has modest correlations with the analysis model variables (r's between .19 and .42). The primary purpose of this analysis is to illustrate the use of auxiliary variables, so we did not thoroughly consider the wide range of possible options for auxiliary variables. Space limitations preclude a thorough discussion of the variable selection process, but for additional details consult Enders (2010, chapter 5).

Table 5 compares the parameter estimates obtained for the Block 2 regression model with and without the inclusion of the auxiliary variables (because the auxiliary variable parameters are not of substantive interest, we omit them from the table). We used the Mplus software program (Muthén & Muthén, 2004) to estimate the auxiliary variable model in Fig. 7. The syntax for the analysis is given in Appendix C. Mplus is particularly useful in this regard because it has built-in facilities for including auxiliary variables. Notice that the auxiliary variables altered some of the estimates, while others changed very little. For example, the intercept (i.e., the Caucasian achievement mean) changed by approximately 9/10 of a standard error unit, whereas the math encouragement slope changed by only 1/10 of a standard error. It is difficult to draw definitive conclusions about the differences in the coefficients because the changes could result from a reduction in bias or a reduction in random error.

Table 5
Comparison of estimates with and without auxiliary variables.

Parameter	With auxiliary		Without auxiliary	
	Estimates	*SE*	Estimates	*SE*
B_0 (Intercept)	70.671	0.457	71.101	0.494
B_1 (Hispanic)	−8.127	1.540	−8.110	1.642
B_2 (AfriAmer)	−11.599	1.358	−11.878	1.440
B_3 (CollEnc)	3.140	0.434	3.597	0.464
B_4 (MathEnc)	3.972	0.565	3.912	0.613
B_5 (AcadEnc)	0.031	0.204	0.137	0.219

Notice also that including the auxiliary variables decreased the standard errors for every parameter estimate in the model. At first glance, the reduction in the standard errors may seem relatively small. For example, the standard error of the math encouragement slope decreased from .613 to .565. Because the magnitude of the standard errors is directly related to sample size, it is possible to express this reduction as a function of *N*. All things being equal, the reduction in the math encouragement standard error is equivalent to the reduction that would have resulted from increasing the total sample size by nearly 18%! A similar conclusion holds for the other model parameters, such that the standard error reductions were commensurate with sample size increases of 12% to 18% (the magnitude of the reduction depends largely on each variable's correlations with the auxiliary variables). When given a choice between including a small set of auxiliary variables and collecting more data, most researchers would undoubtedly choose the former option. It is important to point out that the large reduction in standard errors was due to the fact that one of the auxiliary variables (9th grade achievement) was strongly correlated with the analysis model variables, particularly the outcome variable. Had we not included this variable in the analysis, the standard error reductions would have been much smaller (e.g., after rerunning the analysis with two auxiliary variables, the standard error reductions were commensurate with increasing the total sample size by 1% to 2%). This underscores the important point that the most useful auxiliary variables are those that are highly correlated with the incomplete analysis model variables.

Using missing data to your advantage: planned missingness designs

Thus far, we have described methods for dealing with unintentional missing data. The development of modern missing data techniques has made planned missing data research designs a possibility. The idea of intentionally creating missing data may feel counterintuitive because missing data is generally thought of as a nuisance and something to avoid. However, the reader may already be familiar with planned missing data designs. For instance, in a classic experimental design, subjects are randomly divided into a treatment or a control condition. A participant's unobserved response to the unassigned condition (e.g. a control participant's response to the treatment or a treatment subject's response to the control condition) is actually an example of MCAR missing data. The possibility of analyzing data using maximum likelihood or multiple imputation affords new opportunities for implementing research designs that use intentional missing data to

minimize data collection burden. This section describes a few such possibilities, and readers are encouraged to consult Graham et al. (2006) for a detailed description of planned missing data designs.

Certain constructs may be prohibitively expensive to administer to the entire sample. For example, consider a study that examines behavioral problems in a sample of elementary school children. A questionnaire is inexpensive to administer to the teachers and the parents, but it may not be as reliable as a more comprehensive evaluation based on observations. Using a planned missing data design, the researcher could collect questionnaire data from the entire sample but restrict the time-consuming behavioral observations to a random subsample of participants. As a second example, consider a year-long study of reading comprehension, where researchers are interested in measuring standardized reading outcomes at the end of the spring term. If time restrictions impede the end-of-year assessment, the researchers could instead administer the test to a random subsample of students. In both of the previous examples, the instinctual course of action may be to analyze the data using only the subsample of students that have scores on the expensive measure. However, with maximum likelihood and multiple imputation, the entire sample can be used for the analyses, including the cases that have missing data on the expensive measure (in effect, the inexpensive measure serves as an auxiliary variable in the analysis). Doing so can dramatically increase power relative to analysis based on the subsample of students with complete data.

In many research contexts that employ self-report questionnaires, the number of survey items can quickly become excessive. The constraint on questionnaire length may be due to time restrictions (e.g., a survey must be administered during the first class period of the school day) or participant engagement (e.g., short attention span or low motivation). Respondent burden is particularly relevant for school-age kids that might lack the attention span required to fill out a long survey. When faced with the dilemma of having too many survey items, researchers often go through the difficult task of determining which questions are expendable. Rather than shortening the survey, researchers should consider the 3-form design proposed by Graham and colleagues (Graham, Hofer, & Mackinnon, 1996; Graham et al., 2006) as a way to overcome this dilemma.

The 3-form design minimizes respondent burden through planned missing data. This design creates three questionnaire forms, each of which is missing a different subset of items. The design divides the item pool into four sets (X, A, B, and C) and allocates these sets across three questionnaire forms, such that each form includes X and is missing A, B, or C. A layout of the basic design can be found in Table 6. As an illustration of the 3-form

Table 6
Missing data pattern for a 3-form design.

Form	Item sets			
	X	A	B	C
1	✓	–	✓	✓
2	✓	✓	–	✓
3	✓	✓	✓	–

Note. A check mark denotes complete data.

design, suppose a researcher is interested in administering four questionnaires to a sample of sixth graders and each questionnaire has 30 items. The resulting questionnaire battery would include 120 items. Past experience might suggest that the attention span required to respond to the entire questionnaire set would be too great for a sixth grader, but the students could realistically respond to a shorter questionnaire with 90 items. Using the 3-form design, the respondent burden can be overcome by assigning one questionnaire to each of item groups X, A, B and C. In this design, each sixth grader will only be required to respond to 90 items, but the researcher will be able to analyze the data based on the entire set of 120 items. The 3-form design is flexible and can be adapted to research needs. However, the 3-form design requires careful planning as there are a number of important nuances in implementation (e.g., optimizing power by properly allocating questionnaires to the forms, constructing the forms in a way that allows for the estimation of interactive effect). Readers interested in additional details on the 3-form design can consult Graham et al. (2006) and Enders (2010).

Respondent burden is also a major obstacle in many longitudinal studies. Graham, Taylor, and Cumsille (2001) describe variations of the 3-form design that can be applied to longitudinal studies. The basic idea is to split the sample into a number of random subgroups and impose planned missing data patterns, such that each subgroup misses a particular wave (or waves) of data. The idea of purposefully introducing missing data is often met with skepticism, but Graham et al. (2001) show that planned missing data designs can be more powerful than complete-data design that use the same number of data points. This has important implications for designing a longitudinal study. For example, suppose that each assessment (i.e., data point) costs $50 to administer and your grant budget allows you to collect a total of 1000 assessments. Graham et al.'s results suggest that collecting complete data from *N* participants will actually yield less power than collecting incomplete data from a larger number of respondents.

As with all research designs, researchers must weigh both the costs and benefits of introducing planned missing data to determine if the strategy is appropriate for their research goals. If cost and time considerations pose severe limitations on the ability to collect data, planned missingness (in conjunction with maximum likelihood or multiple imputation) has the potential to greatly enhance the research process. Planned missing data designs involve a number of nuances and require careful planning to get right (e.g., allocating the missingness in a way that does not decrease power), but these designs will undoubtedly become increasingly popular in the coming years.

Discussion

Historically, researchers have relied on a variety of ad hoc techniques to deal with missing data. The most common of these ad hoc techniques include deletion methods or techniques that attempt to fill in each missing value with a single substitute. Some of the traditional missing data techniques require strict assumptions regarding the reason why data are missing (i.e., the MCAR mechanism) and only work in a limited set of circumstances. Others (e.g., mean substitution) never work well. As a result, the use of these ad hoc techniques may result in biased estimates, incorrect standard errors, or both (Little & Rubin, 2002). Despite recommendations from APA and an extensive methodological literature on the subject,

these rather archaic techniques are still quite common in published research studies (Peugh & Enders, 2004; Bodner, 2006; Wood, Whitec & Thompson, 2004). One of the goals of this manuscript is to illustrate the potential problems with traditional missing data techniques and to reinforce the recommendation that researchers should discontinue their use. Traditional methods that assume an MCAR mechanism are virtually never better than maximum likelihood and multiple imputation, even when the MCAR mechanism is plausible (e.g., because they lack power). For this reason, we argue that these techniques should be abandoned.

The methodological literature currently recommends two so-called modern missing data handling techniques, maximum likelihood and multiple imputation. These approaches are advantageous because they require less strict assumptions and provide researchers with sophisticated ways to address missing data that mitigate the pitfalls of traditional techniques. Multiple imputation creates several copies of the data set, each containing different imputed values. Analyses are subsequently carried out on each data set using the same procedures that would have been used had the data been complete. Analyzing each data set separately yields multiple sets of parameter estimates and standard errors, and these multiple sets of results are ultimately combined into a single set of results. The mechanics of maximum likelihood are quite different. Rather than filling in the missing values, maximum likelihood uses all of the available data – complete and incomplete – to identify the parameter values that have the highest probability of producing the sample data. Maximum likelihood and multiple imputation tend to produce similar estimates, so choosing between the two methods is largely a matter of personal preference. Although there are some situations where multiple imputation provides flexibility that maximum likelihood does not, many analyses that are common to school psychology research (e.g., correlation, regression, ANOVA, structural equation models, factor analysis models, longitudinal growth curve analyses) are quite easy to estimate with maximum likelihood. Consequently, our analysis examples focused on maximum likelihood estimation.

In recent years, software programs have undergone dramatic improvements in the number of and type of missing data analyses that they are capable of performing. Every commercially available structural equation modeling software package now implements maximum likelihood estimation, and this is the approach that we used for our analysis examples. Many school psychology researchers are already familiar with structural equation modeling packages or have ready access to these programs, so making the switch to more principled missing data handling approaches typically involves very little effort. In fact, implementing maximum likelihood estimation is usually as simple as specifying a single keyword, clicking a radio button, or adding a single line of code to a structural equation program (some packages now perform this routine by default). Multiple imputation arguably has a steeper learning curve than maximum likelihood, but it too is readily available in familiar software packages. For example, several freeware programs (e.g., Schafer's NORM package) are available for download on the Internet, and both SPSS and SAS now offer multiple imputation routines.

Given the widespread availability of missing data software and the ease with which maximum likelihood and multiple imputation can be implemented, we strongly recommend that school psychology researchers abandon old standby procedures in favor of these modern approaches. This recommendation has strong theoretical and empirical support from the methodological literature and is also consistent with recent recommendations from

the American Psychological Association (Wilkinson & American Psychological Association Task Force on Statistical Inference, 1999).

Appendix A

Mplus program for analysis 1 Block 3 regression model.

```
title:
Analysis 1 example: Block 3 regression model;
data:
! specify data file;
file is jsp_lsay_data.dat;
variable:
! specify order of variables in data file;
names are
id hispanic afriamer collenc9 mathenc9 acadenc9 math12
behprob math9 homres momses momeduc male;
! select variables for analysis model;
usevariables are
hispanic — math12
hcencint hmencint haencint
acencint amencint aaencint;
! center continuous predictor variables;
centering = grandmean(collenc9 mathenc9 acadenc9);
! specify missing value code;
missing are all (−99);
define:
! compute interaction terms;
hcencint = hispanic * collenc9;
hmencint = hispanic * mathenc9;
haencint = hispanic * acadenc9;
acencint = afriamer * collenc9;
amencint = afriamer * mathenc9;
aaencint = afriamer * acadenc9;
analysis:
! invoke maximum likelihood missing data handling;
type = missing h1;
model:
! specify regression equation;
math12 on hispanic afriamer
collenc9 mathenc9 acadenc9
hcencint hmencint haencint
acencint amencint aaencint;
output:
! request descriptive statistics in output file;
sampstat;
```

Appendix B

```
Mplus program for analysis 1 Block 2 regression model.
title:
Analysis 1 example: Block 2 regression model;
data:
! specify data file;
file is jsp_lsay_data.dat;
variable:
! specify order of variables in data file;
names are
id hispanic afriamer collenc9 mathenc9 acadenc9 math12
behprob math9 homres momses momeduc male;
! select variables for analysis model;
usevariables are
hispanic — math12
hcencint hmencint haencint
acencint amencint aaencint;
! center continuous predictor variables;
centering = grandmean(collenc9 mathenc9 acadenc9);
! specify missing value code;
missing are all (−99);
define:
! compute interaction terms;
hcencint = hispanic * collenc9;
hmencint = hispanic * mathenc9;
haencint = hispanic * acadenc9;
acencint = afriamer * collenc9;
amencint = afriamer * mathenc9;
aaencint = afriamer * acadenc9;
analysis:
! invoke maximum likelihood missing data handling;
type = missing h1;
model:
! specify regression equation;
math12 on hispanic afriamer
collenc9 mathenc9 acadenc9;
! constrain interaction terms to zero;
math12 on
hcencint@0 hmencint@0 haencint@0
acencint@0 amencint@0 aaencint@0;
output:
! request descriptive statistics in output file;
sampstat;
```

Appendix C

```
Mplus program for analysis 2 regression model.
title:
Analysis 2 example: regression with auxiliary variables;
data:
! specify data file;
file is jsp_lsay_data.dat;
variable:
! specify order of variables in data file;
names are
id hispanic afriamer collenc9 mathenc9 acadenc9 math12
behprob math9 homres momses momeduc male;
! select variables for analysis model;
usevariables are
hispanic — math12;
! center continuous predictor variables;
centering = grandmean(collenc9 mathenc9 acadenc9);
! specify missing value code;
missing are all (−99);
! select auxiliary variables;
auxiliary = (m) math9 homres momeduc;
analysis:
! invoke maximum likelihood missing data handling;
type = missing h1;
model:
! specify regression equation;
math12 on hispanic afriamer
collenc9 mathenc9 acadenc9;
output:
! request descriptive statistics in output file;
sampstat;
```

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage Publications.

Azar, B. (2002). Finding a solution for missing data. *Monitor on Psychology*, *33*, 70.

Bodner, T. E. (2006). Missing data: Prevalence and reporting practices. *Psychological Reports*, *99*, 675−680.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330−351.

Enders, C. K. (2006). A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic Medicine*, *68*, 427−436.

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

Enders, C., Dietz, S., Montague, M., & Dixon, J. (2006). Modern alternatives for dealing with missing data in special education research. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Advances in learning and behavioral disorders*, *Vol. 19* (pp. 101–130). New York: Elsevier.

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

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325−366). Washington, DC: American Psychological Association.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*, 197−218.

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*, 206−213.

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

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*, 323−343.

Kline, R. B. (2005). *Principals and practice of structural equation modeling*, 2nd Ed. New York: Guilford.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*, 2nd Ed. Hoboken, NJ: Wiley.

Muthen, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *51*, 431−462.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus: The comprehensive modeling program for applied researchers — Users guide.* Los Angeles, CA: Muthén & Muthén.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525−556.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health*, *25*, 88−117.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581−592.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473−489.

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147−177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545−571.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*, 317−329.

Thoemmes, F., & Enders, C. K. (2007, April). *A structural equation model for testing whether data are missing completely at random.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wilkinson, L. American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594−604.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials Review*, *1*, 368−376.