Chapter 24
Advances in Missing Data Models and Fidelity Issues of Implementing These Methods in Prevention Science

Shelley A. Blozis

Introduction

Data analysis with incomplete data is common in prevention research (Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997). Appropriate applications of methods to treat missing data are essential in ensuring unbiased parameter estimates, as well as in maintaining optimal efficiency in parameter estimates. The primary challenge for researchers if some data are missing concerns the particular analytic method that is deemed necessary for a data analysis, such as ANOVA, and the source or mechanism that gave rise to the missing data. That is, different analytic methods make different requirements of the data (e.g., analysis of variance requires complete data), and along with these requirements are specific assumptions about the source of the missing data. Careful planning of research studies when missing data are anticipated can greatly reduce the adverse effects of missing data and improve statistical inference. Knowledge of methods for handling missing data is therefore critical in the planning stages of a prevention study (Graham & Donaldson, 1993).

Missingness

In the missing data literature, whether *missingness* (i.e., whether a response is missing) is ignorable or nonignorable given a particular plan for data analysis is central to missing data problems. For a given data analysis, missingness is ignorable if the parameters of a data model (i.e., a model of the hypothesized relationships between the variables of interest) are independent of the mechanism that generated

S.A. Blozis (✉)
Department of Psychology, University of California, Davis, CA, USA
e-mail: sablozis@ucdavis.edu

Z. Sloboda and H. Petras (eds.), *Defining Prevention Science*, Advances in Prevention Science, DOI 10.1007/978-1-4899-7424-2_24,

the missing data. In other words, the missingness for a given problem is ignorable if the parameters of a data model are essentially the same whether or not the source of the missing data is taken into account. Thus, if the missing data mechanism is ignorable, the parameter estimates of a data model may be interpreted without concern for bias due to the missing data. Missingness is not ignorable if valid interpretation of a data model depends on the source of the missingness. Indeed, if the missingness is not ignorable and the missing data process is not addressed in the data analysis, parameter estimates may be biased. Several strategies have been developed to handle missingness that is not ignorable. To understand the role of missingness in a data analysis more generally, it is useful to first describe the types of data that arise when some data are missing. These types are described next.

Data Types

Little and Rubin (2002) provide a classification for data that are incomplete. For a given data set, the data are assumed to include both observed and missing values. The classification scheme depends on the relationships between the data, both the observed and the missing values, and whether data are missing or not. Thus, in addition to the data, one or more indicator variables are created to represent the missingness and are included in the data set. An indicator of missingness may, for instance, represent a particular pattern of missing data, such as whether an individual has completed participation in a prevention study, or a separate indicator variable may be created for each variable that has missing responses. A missingness indicator variable is usually set equal to 1 if an individual has a particular pattern of missing data and set equal to 0 otherwise. Different indicator variables may be created to represent the various missing data patterns that describe a particular data set (Schafer & Graham, 2002; also see Hedeker & Gibbons, 1997).

The first type of data is data that are missing completely at random (MCAR). Data are MCAR if the missingness is independent of both the observed and the missing data. In a prevention research study, data are MCAR if, for example, whether data are missing or not is independent of both the observed and the missing data. Examples of data that are MCAR are those for which missing data are planned as part of the data collection procedure, such as when participants are selected at random to receive one of multiple possible subsets of the study variables or to be measured according to one of several possible planned sets of measurement occasions of a longitudinal study. This type of missing data is referred to as *planned missingness* and may be used to reduce study costs or participant fatigue (Graham, Taylor, Olchowski, & Cumsille, 2006). If data are MCAR, then there are no systematic differences in the observed and missing values between those individuals with complete data and those with a pattern of missing data. Statistical inference from analytic methods that require complete data, such as ANOVA, is valid when data are MCAR. That is, if data are MCAR, complete-data methods yield unbiased estimates of parameters, although such approaches result in reduced sample sizes and statistical power.

The second type of data is data that are missing at random (MAR). Data are MAR if the missingness is independent of the missing data but is dependent on the observed data. More specifically, the missingness is not related to the missing data after accounting for relationships between the missingness and the observed data. In a prevention research study, for instance, in which measures of substance use are missing for some individuals, data are MAR if the likeliness that an individual will respond to an intervention survey concerning substance use is related to any variable other than the missing substance use outcome. Data are MAR if, for instance, older participants are less likely to respond to the substance use survey. In this example, the missingness depends on the age of the participant and not on the missing substance use data. More generally, the missingness in a given study may be related to one or more of the observed variables included in a data model.

Methods that are valid when data are MCAR or MAR are then suitable choices for a data analysis. Such methods include structural equation models and mixed-effects models that rely on maximum likelihood (ML) estimation procedures and make use of all of the observed data. In a longitudinal study, for instance, data are MAR if the missingness is related to the observed response measured prior to dropout but, conditional on this relationship, the missingness cannot be related to the missing data at the time of dropout or thereafter. Some types of planned missingness may also generate situations in which data are MAR (as opposed to MCAR), such as if study participants are selected for follow-up assessments based on their scores at a previous occasion, such as a pretest score. Here, the missingness is related to the observed pretest score, and conditioning on that relationship, may not be related to the missing score.

The third type of data is data that are not missing at random (MNAR). Data are MNAR when the missingness is related to the missing data even after taking into account the relationships between the missingness and the observed data. In a prevention study, data are MNAR if individuals with high levels of substance use are less likely to report their levels of use. In this case, the missingness depends on the missing substance use data. If data are MNAR, an analysis that is based on the assumption that data are MCAR or MAR may result in biased parameter estimates that may then lead to wrongful conclusions about the relationships among the variables of a data model. Thus, data that are MNAR present a missing data problem that is not ignorable. Indeed, the consequences of ignoring data that are MNAR are a source of great concern for researchers (Molenberghs & Kenward, 2007).

Data Analysis with Missing Data

The need to address missing data in a data analysis has resulted in a variety of approaches to missing data, including analysis that is based on only cases with complete data, in addition to a variety of imputation techniques that aim to generate complete data so that complete-case procedures may be applied. Imputation methods include both single and multiple imputation techniques. Understanding the requirements and possible consequences of the different methods is essential to making informed decisions about a missing data method. Schafer and Graham (2002) provide a comprehensive review of many ad hoc missing data methods and consequences of their applications. Here, imputation methods and likelihood-based methods are reviewed.

Single imputation methods are those that aim to replace missing data with a single value. Common approaches include mean substitution in which missing values are replaced by the mean of the observed scores and regression substitution in which missing scores are replaced by values predicted by the observed data. Once more commonplace in the treatment of missing data, single imputation methods are generally avoided due to the serious problems associated with them, including biased parameter estimates and errors in statistical inference. Importantly, these ad hoc methods do not address the uncertainty involved in replacing the missing data. That is, imputed values represent estimates of the missing values, but ad hoc methods were not designed to address this uncertainty. A consequence of ignoring this uncertainty is that the standard errors of related parameter estimates tend to be underestimated, consequently leading to an increase in the type I error rate.

Multiple imputation (MI) is a technique in which missing data are replaced by multiple imputed values (Rubin, 1978, 1987). A common implementation of MI is by using a computer simulation process in which a set of observed data provide information about the relationships among a set of variables that form the basis of an imputation model. Missing values are then replaced by multiple random draws from the assumed distributions of the variables of the imputation model. The result is a set of multiple imputed data sets that are then analyzed individually using a complete-case analysis procedure. Parameter estimates obtained from each analysis are averaged across the set of results to obtain a single set of estimates. Standard errors of the estimates are generated to take into account variation in the estimates both between and within the imputed data sets. Thus, the standard errors take into account the added variation in the parameter estimates that is due to the imputation procedure. It should be noted that under MI the estimated standard errors of the parameter estimates will generally be greater than those that would have been produced had the data set been complete.

Not all of the variables that are included in an imputation model need to be included in a data analysis, but all of the variables to be included in a data analysis should be included in the imputation model. That said, MI requires that the data used for the imputation process are MCAR or MAR. Finally, it is also important to note that MI is model based. That is, MI is done under assumptions made about the predictive distributions of the missing data. Consequently, a different imputation model may result in a different set of imputed data.

Likelihood-based methods offer an alternative to complete-data methods and, similar to MI procedures, relax the assumption of MCAR. Full information maximum likelihood (FIML), a type of ML estimation that relies on raw data rather than sufficient statistics for the data analysis, may be used to estimate structural equation models and mixed-effects models, for instance, and provide unbiased estimates when data are MCAR or MAR. By making use of all observed data, these methods do not suffer from a reduction in the precision of parameter estimates and statistical power, problems that have been well documented for analytic approaches that rely on only complete cases.

Under certain conditions, analyses based on MI and FIML have been shown to yield similar results (Schafer, 1997). If the two methods are applied to the same set of variables and a sufficiently large number of imputed data sets are generated when using MI, the two methods yield consistent results. Both methods have been shown to yield superior results relative to single imputation methods because these methods provide a better representation of the data (see Little & Rubin, 2002). Multivariate analysis using FIML or MI procedures is possible using major software programs, such as SAS. Several more specialized software programs have also made these procedures available, including those designed for the estimation of structural equation models, such as LISREL (Jöreskog & Sörbom, 2006) and Mplus (Muthén & Muthén, 1998–2010).

Ignorable and Nonignorable Missing Data

As discussed earlier, a central issue in performing an analysis with missing data concerns whether or not the missingness is ignorable. In summary, for complete-data methods, missingness is ignorable if data are MCAR but not MAR or MNAR. For MI and likelihood-based methods, missingness is ignorable if data are MCAR or MAR but not if data are MNAR. Two key issues arise from this information: The first concerns how to handle a data analysis when the missing data process is not ignorable. The second issue concerns how to evaluate assumptions of missingness because all three of the data types (MCAR, MAR, and MNAR) involve an assumption about the relationship between the missingness and the missing data, a relationship that cannot be empirically tested. That is, given that the missing data are not available for study, it is not possible to evaluate the relationships between the missingness and the missing data.

Data Analytic Approaches to Nonignorable Missingness

Several major frameworks have been developed for the analysis of data that are MNAR. Generally, these may be grouped into those that treat the missing data prior to the data analysis and those that handle the missing data and the data model simultaneously. MI procedures fall under the first type and may be used to treat data that are MNAR by incorporating information from correlates of the data, referred to as auxiliary variables, into the imputation process. This strategy may then lead to an ignorable missing data problem for the data analysis. Methods that fall under the

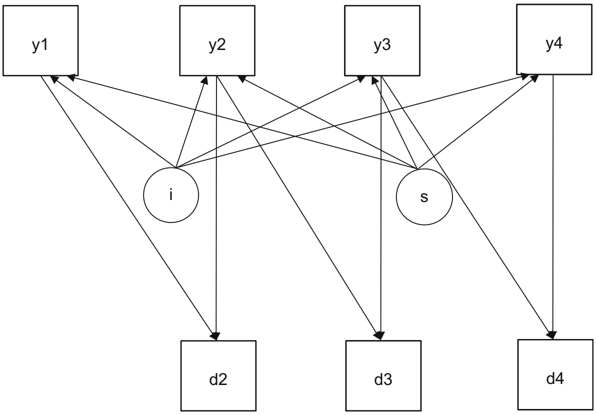

Fig. 24.1 Nonignorable missingness in a selection model

second type involve either an explicit model for the missing data process or an extension of a data model to directly include correlates of the data.

MI with Auxiliary Variables. MI may be used to address data that are MNAR by including correlates of the data in the imputation process. If auxiliary variables provide the information needed about the data, then the data are what is referred to as auxiliary variable MAR (or A-MAR) and the missing data process becomes ignorable (Daniels & Hogan, 2008). Data analysis may then be performed on the imputed data using only the variables of the data model. Using MI with auxiliary variables can provide superior results to MI methods that do not take advantage of these variables (Rubin, 1997).

MNAR Models. Nonignorable missingness may be addressed directly by extending a data model to include a model for the missing data process. Such models include selection models, shared-parameter models, and pattern-mixture models. In a selection model, the missingness depends on the data. In a longitudinal study, for instance, a selection model may be specified in which the missingness depends on the missing response at the time of dropout and the observed responses prior to dropout, as shown in Fig. 24.1 (Diggle & Kenward, 1994). In Fig. 24.1, indicators of missingness (e.g., dropout status) at occasions 2, 3, and 4, denoted by the variables *d2*, *d3*, and *d4*, depend on the observed response at the previous occasion (assuming an individual has not dropped from the study by that point) and the current measure that is missing if the individual has dropped or observed if the individual remains in the study by that occasion.

In a shared-parameter model, a special case of a selection model, the missingness depends on the random coefficients of a data model. The missingness may, for example, depend on the random intercept and slope of a mixed-effects model that is used to characterize longitudinal data, as depicted in Fig. 24.2 (Wu & Carroll, 1988). In Fig. 24.2, indicators of missingness at occasions 2, 3, and 4 are denoted by the variables *d2*, *d3*, and *d4*. Each indicator of missingness is specified to depend on the random intercept and random slope, represented by the variables

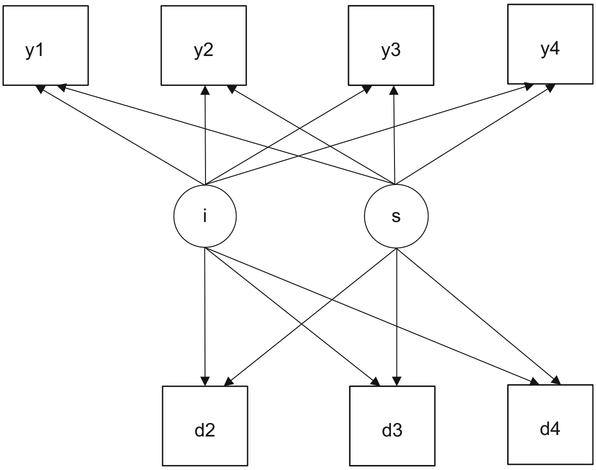

Fig. 24.2 Nonignorable missingness in a shared-parameter model

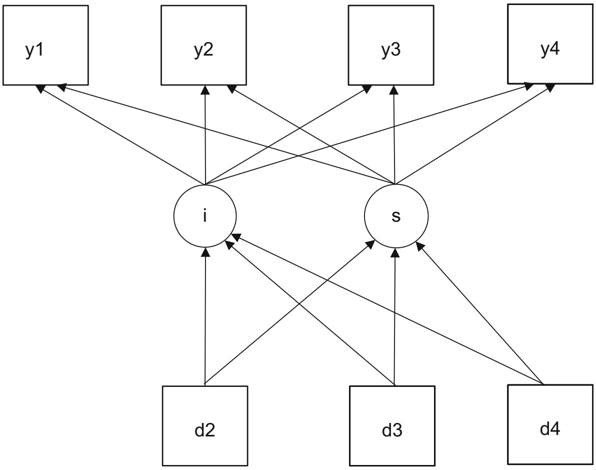

Fig. 24.3 Nonignorable missingness in a pattern-mixture model

i and *s*, respectively. In this example, the missingness depends on the expected value of an individual's response at a given measurement occasion, such as at the start of a prevention study, as well as on an individual's expected rate of change in the response over the study period.

In a pattern-mixture model, the response is specified to depend on the missingness. In a pattern-mixture random-effects model, for instance, the random intercept and slope of a model for a longitudinal response may depend on an indicator of missingness, such as an indicator of dropout, as shown in Fig. 24.3 (Hedeker & Gibbons, 1997). In Fig. 24.3, indicators of missingness at occasions

2, 3, and 4, denoted by the variables *d2*, *d3*, and *d4*, moderate the random intercept and random slope, represented by the variables *i* and *s*, respectively. In this model, an individual's expected response status at a given measurement occasion and expected rate of change in the response over time may be moderated by the dropout status.

Unlike MI methods, the MNAR models described require an explicit model for the missing data process. That is, the relationship between the missingness and the observed and missing data must be explicitly defined. This can present a challenge to researchers because the specific mechanism that gave rise to the missing data may not be known. Thus, it is often recommended that multiple missing data models, theoretically driven, be specified and fitted to the data so that the analysis does not rely on any single missing data model. In this way, competing ideas about a missing data process may be evaluated.

Models that Include Auxiliary Variables. Data that are MNAR may also be handled by fitting a data model that is extended to include correlates of the data. That is, the data model is extended to allow one or more auxiliary variables to correlate with variables of the data model. In this way, these models do not require an explicit model for the missing data process as required by selection, shared-parameter, and pattern-mixture models, and the data model is not altered with regard to the hypothesized relationships that define the data model. Estimation of models that include auxiliary variables may be carried out using FIML. Similar to MI implemented with auxiliary variables, information about the missing data is drawn from the auxiliary variables. Unlike MI, however, adding auxiliary variables as correlates into a data model does not require the generation of multiple data sets that must then be analyzed and summarized through additional steps.

Graham (2003) describes a saturated correlates model in which auxiliary variables are allowed to correlate with the exogenous variables, the residuals of the manifest dependent variables, the residuals of any indicators of latent variables, and each other. An example of a saturated correlates model applied to a longitudinal measure is shown in Fig. 24.4. In Fig. 24.4, a set of auxiliary variables are shown to correlate with the residuals that result from the regressions of the four measured responses ($Y_1, \ldots, Y_4$) on the random intercept (I) and slope (S), as well as with the random intercept and slope that represent the exogenous variables of the longitudinal data model.

Collins, Schafer, and Kam (2001) conducted a data simulation study to better understand the role of auxiliary variables in a data analysis when data are incomplete. In their study, auxiliary variables contributed most to the analysis when they had strong correlations with the missingness and the amount of missing data was more than 25 %. In some cases, auxiliary variables were also helpful when the percentage of missing data was below 25 %. Overall, the study provided evidence that suggested that using more auxiliary variables, as opposed to restricting their number, was most helpful in reducing biases in the parameter estimates of the data model and improving efficiency in the parameter estimates. Taking a more liberal approach by including more auxiliary variables may help to ensure that particularly helpful auxiliary variables are included. They also noted that auxiliary variables need not correlate with the missingness but rather may be correlated with the data.

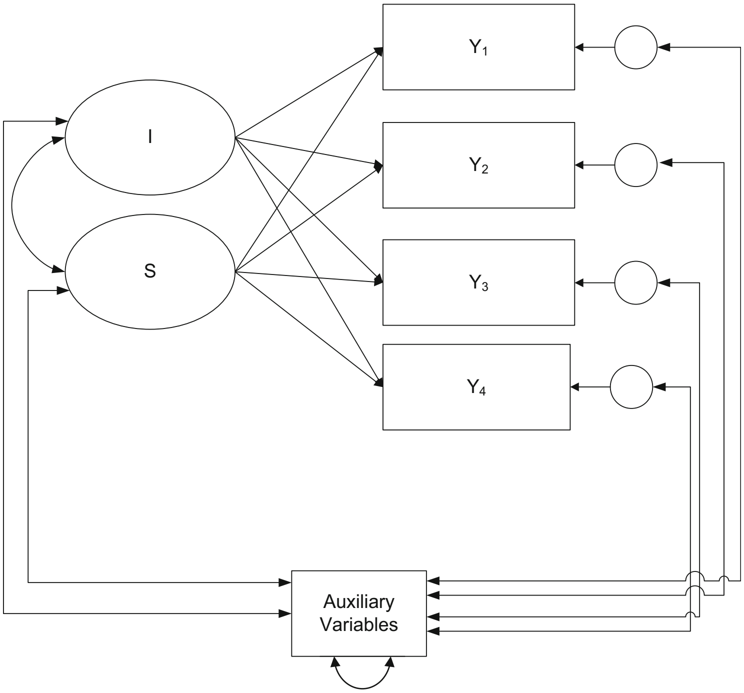

Fig. 24.4 Auxiliary variables in a saturated correlates model

Sensitivity of Parameter Estimates to Assumptions About Missingness

Unless data are generated with known characteristics such as by a simulation process, it is not possible to test for a given data set any assumption about the relationship between the missingness and the missing data. Indeed, any effort to evaluate any of the three data types (MCAR, MAR, and MNAR) can be carried out using only observed data. Thus, conclusions about the type of data involved in any case are only speculative. Given this, a recommended strategy in a data analysis when the missingness may not be ignorable is to perform what is called a sensitivity analysis.

Sensitivity analysis is often used to study whether changes in parameters estimates or statistical inference result after a model or the data have been modified, such as by making different distributional assumptions about data or making changes in the data (such as removing a case from a data set to study the influence of an individual's data on parameter estimates). In a data analysis involving missing data, a sensitivity analysis may be performed to evaluate whether parameter estimates depend on the specific assumptions made about the data (Little & Rubin, 2002). As noted earlier, a missing data process is not likely to be known in practice. Thus, a recommendation is to consider multiple missing data treatments (Molenberghs & Kenward, 2007). This may involve applying different missing data frameworks (e.g., applying both a selection model and a pattern-mixture model), as well as formulating different missing data models (e.g., one model may assume that the missingness depends on the measured outcome at the occasion immediately prior to dropout, whereas another model might assume that the missingness depends on the measured outcome only at the start of the study).

Planning for Missing Data in Prevention Research

In prevention research, data analysis is very often done using methods that require complete data and so also require that data are MCAR. In other cases, methods that may make use of all of the observed data, such as structural equation models, require that data are MCAR or MAR. If whether or not data are missing is related to the missing values, however, inference from methods that require that data are MCAR or MAR can be problematic. If the missing data process is not ignorable, possible consequences include drawing false conclusions about the magnitude or strength of the relationship between measures or drawing wrongful conclusions about the effectiveness of a treatment program. Such problems may be avoided with careful planning in the beginning stages of data collection to address issues surrounding either potential sources of missing data or by making plans to measure variables, near or at the start of a study, that are likely to be correlated with the missingness or the missing data.

Summary

Prevention science research often results in data that are not complete. Many commonly used statistical procedures including regression analysis and analysis of variance generally require complete data. Consequently, missing data on at least one variable for an individual leads to an exclusion of the individual from analysis, resulting in a reduced overall sample size and statistical power, as well as increases in type I and type II errors. In some cases, analysis of only complete data can result in biased parameter estimates and wrongful statistical inference.

Given the need for multiple assessments over time in prevention science research, researchers are increasingly in need of appropriate statistical methods to handle missing data. Naturally, as the number of assessments over time increases, there is often a greater likeliness that some data will be missing, usually due to participant attrition. Structural equation models and mixed-effects models, methods that may rely on likelihood-based procedures for estimation, allow for missing data. These methods in their usual application require that data are MCAR or MAR. In a mixed-effects model, for instance, whether or not data are missing can depend on the measured outcome when it is observed, such as observations made prior to the time when a participant drops out of a study. Conditional on this relationship, the missingness under MAR is assumed to be independent of the missing data (Laird, 1988). This is very different from a complete-data procedure, such as ANOVA, that under MCAR assumes that the missingness is independent of both the missing data and the observed data.

Structural equation models and mixed-effects models that address a nonignorable missing data process, as described here, have been well developed, and many of the procedures may be carried out using several commercially available software packages (e.g., Blozis et al., 2013; Xu & Blozis, 2011). A pattern-mixture random-effects model, for instance, may be used to examine differences in responses according to different patterns of missing data, such as testing whether the outcome measure for those who dropped from a treatment program versus those who completed the treatment changed on average according to different rates.

Multivariate analysis in general carried out using likelihood-based methods allows for the analysis of incomplete data so that all individuals may be retained for analysis. Likelihood-based methods, now a standard in the field, are valid when data are MCAR or MAR. In prevention science research, missing data can create an added challenge if data are MNAR. In these cases, analytic methods that assume data are MCAR or MAR may yield biased results. This paper reviewed state-of-the-art methods that may be used when data are MNAR. These methods include multiple imputation with auxiliary variables, selection models, shared-parameter models, pattern-mixture models, and models that include auxiliary variables that are correlates of the variables of a data model.

A major challenge in dealing with missing data is that assumptions about missingness cannot be completely tested. This is due to the fact that the missing data are not available for study. A preliminary analysis that suggests no differences between individuals with complete versus incomplete data with regard to background characteristics (e.g., gender, age) or the observed values of the measured outcomes, for instance, does not guarantee that data are MCAR because MCAR also specifies that the missing data are independent of the missingness. A common strategy for dealing with this problem of uncertainty concerning the status of the data is to perform a sensitivity analysis of the parameter estimates of a data model under different assumptions about the missing data.

A sensitivity analysis may involve the use of different missing data strategies, such as fitting different models that make explicit assumptions about the missing data process, as would be done when fitting a selection model, shared-parameter model, or a pattern-mixture model, or fitting a data model that has been extended to include correlates of the missingness or missing data. Whatever methods are selected for study, it is important to keep in mind that the study of missing data is carried out using only observed data. Thus, conclusions about whether the missing data process is ignorable or not are not certain. Furthermore, in cases in which the missingness is not ignorable but the true missing data process is not captured in the data analysis, it may not be possible to detect that a nonignorable missing data process is operating.

References

Blozis, S. A., Ge, X., Xu, S., Natsuaki, M. N., Shaw, D. S., Neiderhiser, J., et al. (2013). Sensitivity analysis of multiple informant models when data are not missing at random. *Structural Equation Modeling, 20*, 283–298.

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.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Diggle, P. J., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics, 43*, 4–73.

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

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of followup data. *Journal of Applied Psychology, 78*, 119–128.

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. Bryant, M. Windle, & S. 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., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323–343.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*, 64–78.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80 for Windows (computer software)*. Lincolnwood, IL: Scientific Software International, Inc.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7*(1–2), 305–315.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

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

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.

Rubin, D. B. (1978). Multiple imputations in sample surveys. In *Proceedings of the survey research methods section* (pp. 20–34). Alexandria, VA: American Statistical Association.

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

Rubin, D. B. (1997). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473–489.

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

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

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

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