The National Center for Post-Traumatic Stress Disorder

# PTSD Research Quarterly

VOLUME 12, NUMBER 2 ISSN 1050-1835 SPRING 2001

Published by:

The National Center for PTSD
VA Medical and Regional Office Center (116D)
215 North Main Street
White River Junction
Vermont 05009-0001 USA

(802) 296-5132
FAX (802) 296-5135

Email: ptsd@dartmouth.edu
http://www.ncptsd.org

Subscriptions are available from the Superintendent of Documents, P.O. Box 371954, Pittsburgh, PA 15250-7954.



In this issue:





## CONTEMPORARY APPROACHES TO MISSING DATA: THE GLASS IS REALLY HALF FULL

Daniel W. King, Lynda A. King,
& Peter S. Bachrach
National Center for PTSD
& Boston University School of Medicine
John J. McArdle
University of Virginia

A common dilemma in many types of research comes about from the need to deal with incomplete or missing data. Within a single data collection session, missing data often comes in the form of a participant's refusal to answer a specific interview question, skipped items or scales (either intentionally or unintentionally) on a paper-and-pencil survey, or perhaps an early termination of the session, with only partially completed tasks or instruments. In the case of longitudinal research with multiple data collection sessions, missing data may be more pervasive, with respondents providing data sporadically across time or dropping out of the study altogether. It might even be that some participants enter a longitudinal study late, such that their data are not available for the initial stages of the research.

Traditionally, the problem of missing data has been seen as a costly nuisance and sometimes even a "fatal flaw" to a research project. From a practical perspective, missing data often means more time and money to recruit study completers and achieve the target sample size. From a methodological perspective, the frequently-used strategies of case deletion (listwise or pairwise) or mean substitution have been shown to create more problems than they solve, posing threats to both statistical conclusion validity (low power and/or biased tests of significance) and to external validity (general ambiguity about the population to which inferences reasonably can be made).

Recent statistical work has allowed for an almost revolutionary shift in the manner in which missing data can be handled. More important, the execution of contemporary missing data techniques can obviate many practical and statistical concerns. Modern missing data strategies enhance efficiency, preserve resources, and guard against incorrect statistical inference. In fact, one could argue that the incorporation of corrective missing data analyses and even purposeful missing data designs shortly will be considered routine and expected by scientific editorial boards and funding review panels. In the sections that follow, we describe conditions under which data are missing and present some of the newer maximum-likelihood based methods. We then discuss the implementation of some of these methods in the context of stress and trauma research. We also provide resources for those wishing to learn more about the methods.

*Classes of incomplete or missing data.* According to Little and Rubin (1987), Rubin (1987), and Schafer (1997), among others, there are two fundamental conditions under which data are missing: *ignorable* and *nonignorable*. Under the ignorable condition, the presence or absence of data (a dichotomous "missingness" variable M) on a particular variable of substantive interest (Y) can be explained or predicted by one or more other variables in the data set ($X_1$, $X_2$, etc.), but not solely by the variable that is itself missing data (Y). Moreover, while the pattern of missing data (M) may be related to scores on the variable that is missing data (Y), scores on the other variables ($X_1$, $X_2$, etc.) are expected to mediate the relationship between the variable defining the pattern of missing data (M) and the variable with missing data (Y). Finally, we must assume there are no interactions between M and any of the X variables in the prediction of scores on Y. When these assumptions hold—as they do in many research situations—the missing information on Y is explained by the information contained in the X variables. We can be reasonably correct in estimating missing values of Y, since we are assured that the relationship between the Xs and Y is the same regardless of whether the Y data are present or absent. Therefore, we can use the known relationship between the Xs and the observed values of Y to derive the unobserved or missing values of Y. Rubin (1976) used the term *missing at random* to characterize this condition, and others have termed this *accessible* (Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997) or *recoverable* (McArdle, 1994).

A more restrictive form of the ignorable condition is known as *missing completely at random*. Here, the above assumptions hold with the exception that M is not related to the X variables or to Y. Thus, the information from the X variables again can be used to estimate missing values of Y. Data are typically missing completely at random when the researcher intentionally plans to randomly exclude certain participants from selected assessments, either within

a single data collection occasion or across occasions. In this instance, there is deliberate use of modern missing data methods to reduce the burden on study participants and enhance the cost-effectiveness of the research: Each participant furnishes less information, but the information from the sample as a whole is preserved. In the real world of data collection, research having a planned missingness component is plagued with the usual difficulties of data omission or participant attrition. In the end, then, what might begin as a missing completely at random situation becomes a missing at random but still ignorable or recoverable one.

The nonignorable condition obtains when the missingness dichotomy (M) is related to scores on the variable having missing data (Y), but there is no X variable or combination of X variables that can accurately predict M or scores on Y. Very simply, the researcher has no available information that can assist in the estimation of the missing values of Y, and the kinds of missing data methods presented here may not be as useful.

A researcher never fully knows the conditions under which data are missing. Why, for example, some participants fail to respond to particular items or others fail to attend scheduled assessment sessions are likely attributable to a combination of ignorable and nonignorable factors. Nonetheless, current consensus (e.g., Graham et al., 1997; McArdle, 1994; McArdle & Hamagami, 1992; Schafer & Olsen, 1998) seems to favor the presumption of ignorability and advises the use of one or more of the maximum likelihood-based missing data techniques that are rapidly becoming available.

*Contemporary missing data methods*. We present two general approaches to maximum likelihood-based treatment of missing data that have appeared in the methodological literature. *The first approach is raw* or *direct maximum likelihood,* wherein statistics of interest (e.g., means, standard deviations, regression coefficients and their standard errors) are directly estimated for the full sample from the existing and limited data. That is, the desired numerical results are obtained, but there is no attempt to "fill in" values where data are missing.

An early but still viable method for direct maximum likelihood is the application of *multiple-group structural equation modeling (SEM),* first proposed by Joreskog and Sorbom (1979) and Horn and McArdle (1980, and later elaborated by Allison (1987) and Muthen, Kaplan, and Hollis (1987), among many others. The method requires that subsamples of the full sample be created on the basis of common patterns of missing data. For example, given four variables, participants having complete data on all four variables would form one subsample; participants missing data only on the first variable would form another subsample; those missing data only on the second variable would be another subsample; those missing data on both the first and second variable would be another subsample; and so on. Each subsample then becomes a separate group in the multiple-group SEM. Directions for parameterizing this type of model can be found in a number of sources, including Allison (1987), Muthen et al.(1987), Joreskog and Sorbom (1993), and Graham et al. (1997), and advances in the use of this method have been proposed by McArdle and his colleagues (e.g., McArdle, 1994; McArdle & Hamagami, 1992, 2001; McArdle & Woodcock, 1997). As with all SEM procedures, the results—values of parameter estimates—maximize the likelihood of the data for the full sample, which is a function of the likelihoods of the several subsamples. Thus, by partitioning the sample into different groups based on patterns of missing and complete data, this procedure uses information from all participants without any need to delete cases. Multiple-group SEM is available in most SEM software packages.

A disadvantage of multiple-group SEM is that large studies with many variables may produce numerous patterns of missingness and result in an overly complicated and implausible model in which there are more variables than cases for one or more groups. An alternative direct maximum likelihood method optimizes *likelihood at the level of the individual*, rather than the group (McArdle & Bell, 2000; Neale, Boker, Xie, & Maes, 1999). Again, the likelihood of the full sample is a function of the likelihoods of its components, in this case, the individual likelihoods, each of which is maximized using parameter estimates only for the data that individual brings to the study. Once more, the analyses take advantage of all of the data, no participants are discarded, and sample size remains protected. SEM software programs that employ individual-level direct maximum likelihood estimation include AMOS (Arbuckle, 1995), Mplus (Muthen & Muthen, 1998), and Mx (Neale et al., 1999).

Some readers may be familiar with a landmark article on the analysis of longitudinal psychiatric data by Gibbons and Hedeker and colleagues (Gibbons et al., 1993). In that article, the authors overviewed special characteristics of longitudinal data. They also cautioned against the use of the traditional statistical methods of end-point analysis (in which the dependent variable is simply the difference between a participant's baseline and last available data point, regardless of the latter's timing) and mixed effects and multivariate repeated-measures analysis of variance (where listwise deletion is frequently applied to accommodate missing data). In addition, they objected to the restrictive assumptions required to use repeated-measures analysis of variance with longitudinal data. In this and a series of follow-up works (e.g., Hedeker & Gibbons, 1994, 1997), these methodologists advocated the use of random-effects regression, which uses the direct maximum likelihood approach for data that are missing in a dependent variable measured across time. Random-effects regression is also employed by Bryk and Raudenbush (1992) within their hierarchical linear modeling framework. Though random-effects regression does not accommodate missingness within predictors or covariates, it has one distinct advantage over other direct maximum likelihood methods. Not only can participants be missing data for one or more assessments in the time series, but there are no restrictions regarding when the assessments are made. Thus, the researcher gains a lot of flexibility in planning and executing

a study: All participants need not be assessed on a rigid schedule.

*The second general approach to the treatment of missing data is imputation*, for which actual values for the missing information are derived and then employed in the calculation of parameter estimates and standard errors. The proper use of imputation mandates a recognition of two sources of uncertainty in the imputed value for a missing datum: uncertainty in the form of sampling variability and uncertainty regarding the correctness of the imputed value itself. Accordingly, Rubin (1987) and Little and Rubin (1987) recommended that two or more imputations be performed, resulting in the creation of two or more complete data sets. Using this *multiple imputation* method, whatever analyses are of interest to the researcher (analysis of variance, multiple regression, etc.) are conducted on each data set. Next, parameter estimates and standard errors are saved and combined in a rather straightforward manner, using formulas proposed by Rubin (1987). Parameter estimates are merely averaged, and their standard errors are a weighted composite of between and within variability. (Recent work has merged the multiple imputed sets using multiple-group SEM [McArdle & Hamagami, 2001]). Test statistics then can be calculated and represent the findings of the study. No cases are deleted, and findings take advantage of information from all participants, even those who supplied only partial data. The number of required imputations is not very large, usually between three and five and rarely more than ten.

One technique for arriving at multiple imputations of missing data relies on what is known as *propensity scores*, as represented in the SOLAS software program (Statistical Solutions, 1999). To offer a simplified explanation: A missingness variable M is created for a particular variable Y that has missing data. A number of predictors ($X_1$, $X_2$, etc.) are identified by the researcher, and M is regressed on these X variables. The program then assigns each participant a predicted probability of missingness index. Participants are arranged into percentile groups (quintiles, by default in SOLAS) according to their predicted probability of missingness. Participants with missing Y values within specific percentile groups are each assigned the value of another member of their group, selected at random with replacement. This procedure is repeated for as many independent imputations as the analyst desires, thus yielding multiple complete data sets. These data sets are, in turn, submitted to separate data analytic procedures, followed by a synthesis of parameter estimates and standard errors.

Another technique for obtaining multiple imputations is *data augmentation*, as implemented by Schafer and his associates in a number of specialized programs distributed through The Pennsylvania State University. Very briefly and perhaps too simplistically, data augmentation uses an iterative or multistage process in which information about the relationship between other variables (the Xs) and data that are available on the Y variable is used to predict scores for the missing Y data. The predicted value is used to create a distribution of possible scores for each person who is missing data on Y. From this distribution, a random value is selected and is used in the recalculation of the prediction equation, which produces predicted values from which another distribution of missing data values is developed for each person. This process continues until a convergence criterion is achieved. The predicted values at each step are saved, and from these the imputed value is randomly selected, and the first imputation is complete. The process begins again and is repeated for as many times (imputations) as the analyst requests.

Though the direct maximum likelihood-based and multiple imputation approaches to missing data appear quite different, in fact, they produce parameter estimates and standard errors that are fully comparable when certain assumptions (i.e., multivariate normality and ignorable condition) are correct. There may be some reasons to select one over the other (see, e.g., Enders, 2001, and Schafer & Olsen, 1998). A third approach, the EM algorithm followed by bootstrap estimation of standard errors, is overviewed by Graham et al. (1997). Studies have demonstrated the effectiveness of missing data methods in reproducing findings highly comparable to those that would have been obtained had the full data set been available. Moreover, any of these more contemporary maximum likelihood-based procedures is superior to the traditional choices of case deletion and mean substitution.

*Examples of modern missing data strategies in stress and trauma research.* In this section, we summarize a variety of applications in which direct maximum likelihood or multiple imputation methods were employed. The first are from two studies of PTSD and its correlates that used the same sample of male and female Gulf War veterans. In both studies, participants were assessed on two occasions, immediately upon return to the United States from the Gulf region and again 18 to 24 months later. The first study (King et al., 2000) was interested in how PTSD symptomatology might influence self-reported accounts of trauma exposure. The second study (Erickson, Wolfe, King, King, & Sharkansky, 2001) concerned the association between PTSD and depression. Both used a cross-lagged panel design, with each of two variables measured on each of two occasions, and structural equation modeling. Those who conduct such longitudinal research on veteran populations are well aware of the issue of attrition and inability to locate participants over repeated assessments. Such was the case with this cohort. At the initial assessment, there were 2,942 participants; at the second assessment, the number had dropped to 2,295. Obviously, the lower value would be sufficient for any analyses, but with listwise deletion there is the question of external validity. Do we wish to generalize findings only to those veterans who made themselves available on both occasions? The missing data were judged to satisfy the ignorable condition, and the Mplus software program (Muthen & Muthen, 1998) was used for the structural equation modeling. The effective sample size was maintained at 2,942.

A thoroughly detailed demonstration of the propensity score approach to multiple imputation is supplied by

Lavori, Dawson, and Shera (1995). They described a randomized, double-blind clinical trial to test the efficacy of two drugs for treating panic disorder, along with a placebo. Accordingly, there were three groups, who were treated and assessed over an 8-week period. The full sample size at the outset of the clinical trial was 1,168. Attrition was a problem in all three conditions, but especially in the placebo condition, where the number of participants decreased from 391 to 220, a 44% loss. To restore the data to the full complement, Lavori et al. proceded from the earliest assessment where missing data appeared (week 3), and then sequentially performed waves of 10 multiple imputations for each of the subsequent assessments in the study, thus filling in incomplete data through week 8. They concluded that they were able to fully exploit a complete data set and pointed out that they had produced a veridical multiply imputated database available to future researchers.

Another application of multiple imputation involved a study examining the long-term health and adjustment of a group of repatriated Vietnam-era aviators who were held as prisoners of war (POWs), relative to a group of matched control aviators who were not captured (Keane et al., 2001). One set of analyses concerned captivity status as a moderator of the relationship between neurocognitive functioning assessed in the 1970s and intelligence test scores two decades later. At the initial assessment, neurocognitive functioning data were obtained for 119 POWs and 98 controls. By the time intelligence was measured in the 1990s, the sample size dropped to 42 POWs and 38 controls. Using captivity status, 1970s intelligence scores, and age at time of the second assessment as covariates, the SOLAS propensity score method was used to multiply impute 10 datasets, resulting in complete contemporary intelligence data for all participants. Thus, sample size was restored. Hierarchical moderated multiple regression analyses were performed on each data set, and the SOLAS rollup function combined the results across the 10 datasets.

Schnurr, Spiro, Aldwin, and Stukel (1998) were interested in the association between trauma exposure and physical health outcomes. They drew 1,079 cases from a large-scale longitudinal study of World War II and Korean Conflict veterans and examined the course of physical symptoms over some 30 years. As might be expected, there were missing data in the assessments of physical symptoms over time, with the number of assessments ranging from 2 to 10; 95% of the sample completed at least 5 assessments. Despite the discrepancy in the number of assessments, these researchers were able to extrapolate trends across the full interval of ages 30 to 75 years to model the course of physical symptoms. Sample size remained at 1,079, and inferences were generalizable to the full population referenced by this sample. The analytic tool used by Schnurr et al. was the generalized estimating equation technique, a SAS macro created by Zeger and Liang (1986) that employs the direct maximum likelihood approach to missing data. The class of programs (MIXREG, MIXOR, etc.) developed by Hedecker and Gibbons (1994, 1996) and the Hierarchical Linear Modeling (HLM; Bryk & Raudenbush, 1996) software packages also are appropriate for the study of individual change or trajectory over time where data are missing in the time-dependent variable.

Our final example from the stress and trauma domain is an application of planned missingness for the purpose of reducing data collection time in a national telephone risk and resilience survey of Gulf War veterans (King, King, & Vogt, 2001). The design is cross-sectional. For this study, we were constrained by budget and practicality to a 45-minute interview, judged as a maximum of about 250 items or questions that could be administered by telephone. Our ideal item set, however, numbered approximately 400. We adapted Graham, Hofer, and MacKinnon's (1996) multiple-form design by constructing six separate but systematically overlapping interview forms, with random assignment of items to forms. Each veteran was administered 5/8 of the full item set, thus achieving our goal of no more than 250 items to each participant. Multiple imputation will be applied to this partial data set to fill in or simulate complete data for subsequent analyses.

*Missing data resources*. The Pennsylvania State University Methodology Center website (http://methcenter.psu.edu/homepage.shtml) provides a wide-ranging selection of materials pertaining to missing data techniques, including an extensive reference list, downloadable publications and technical reports, announcements of conferences and workshops, and free software for multiple imputation. Regarding the latter, at this website are housed Graham and Hofer's EMCOV program and Schafer's data augmentation software series of NORM, PAN, MIX, and CAT. The website also provides links to other sites that feature software that can handle missing data.

Another website related to missing data is Multiple Imputation Online: http://www.multiple-imputation.com. This site offers content about the principles of multiple imputation and contact information for experts. The site also lists several other macros for missing data analyses and programs that are not discussed here, such as MICE by van Buuren and Oudshoorn and AMELIA by Honaker, Joseph, King, Scheve, and Singh, both freeware.

SOLAS is a commercially available missing data program. It provides two different multiple imputation procedures (one of which is the propensity score method described previously), data visualization tools (that graphically present the types and patterns of missing data), and a set of univariate and multivariate statistical techniques (descriptives, *t*-tests, analysis of variance, and multiple regression). Once the imputation process is completed and selected analyses conducted, SOLAS automatically combines results using the Rubin (1987) algorithms. SOLAS affords easy data importation from and exportation to most commonly used software packages. It is worth noting that SAS is about to release two procedures, PROC MI and PROC MIANALYZE, thereby making the full range of the SAS system amenable to multiple imputation methods.

Two of the direct maximum likelihood SEM programs that facilitate missing data analyses at the individual level, AMOS (Arbuckle, 1995) and Mx (Neale et al., 1999), have graphical-user interfaces. Thus, the analyst need only draw the appropriate path diagrams and not necessarily be versed in the syntax for the programs. The Mx program, after drawing the model, will produce and save a script file with the matching syntax for future use. Mx is freeware, obtainable from Neale's Virginia Commonwealth University website: http://views.vcu.edu/mx/. The AMOS program and Mplus (Muthen & Muthen, 1998) provide likelihood ratio-based tests of overall model-data fit. It is important to note that these programs, and all SEM programs, can be easily adapted to accommodate any of the analytical procedures subsumed by the general linear model, as examples, randomized group analysis of variance or covariance and simple bivariate and multiple regression.

Regarding random-effects regression resources, Hedeker's website (http://tigger.uic.edu/~hedeker/mixreg.html) contains the original MIXREG program and its variations (MIXOR, MIXNO, and MIXPREG), which are free and downloadable. This website also provides references and downloadable publications by Hedeker and Gibbons and their collaborators. The HLM website is http://www.ssicentral.com/hlm/hlm.htm. It contains examples of HLM analyses, reference lists, and an online index to the Bryk and Raudenbush text (1992). A feature in the latest version of HLM (5.2) is the automated analysis of multiply-imputed data, in which correct parameter estimates and standard errors are calculated on data sets multiply imputed from other software.

The three primary books on missing data, those by Little and Rubin (1987), Rubin (1987), and Schafer (1997) are fairly technical. The Graham et al. (1997) book chapter provides a more accessible survey of missing data methods, and Schafer and Olsen's (1998) article gives a nice overview of the data augmentation approach to multiple imputation. Two additional recommended readings for the novice are those by Enders (2001) and Schafer (in press).

## PILOTS UPDATE

National Center for PTSD (116D)
VA Medical and Regional Office Center
215 North Main Street
White River Junction, Vermont 05009-0001