

## MAKING THE MOST OUT OF DATA ANALYSIS AND INTERPRETATION

# Multivariate modeling of missing data within and across assessment waves

AURELIO JOSÉ FIGUEREDO<sup>1</sup>, PATRICK E. McKNIGHT<sup>2</sup>,  
KATHERINE M. McKNIGHT<sup>2</sup> & SOURAYA SIDANI<sup>3</sup>

<sup>1</sup>*Ethology and Evolutionary Psychology and* <sup>2</sup>*Puget Sound Health Care System—Seattle, Washington, Evaluation Group for Analysis of Data, Department of Psychology, University of Arizona, USA* & <sup>3</sup>*Faculty of Nursing, University of Toronto, Toronto, Ontario, Canada*

### Abstract

*Missing data constitute a common but widely underappreciated problem in both cross-sectional and longitudinal research. Furthermore, both the gravity of the problems associated with missing data and the availability of the applicable solutions are greatly increased by the use of multivariate analysis. The most common approaches to dealing with missing data are reviewed, such as data deletion and data imputation, and their relative merits and limitations are discussed. One particular form of data imputation based on latent variable modeling, which we call Multivariate Imputation, is highlighted as holding great promise for dealing with missing data in the context of multivariate analysis. The recent theoretical extension of latent variable modeling to growth curve analysis also permitted us to extend the same kind of solution to the problem of missing data in longitudinal studies. Data simulations are used to compare the results of multivariate imputation to other common approaches to missing data.*

### Introduction

Missing data is a common problem in both cross-sectional and longitudinal research. Paradoxically, the problems encountered and the solutions implemented are hardly mentioned outside the statistical literature. Reading substantive journal articles, one would hardly suspect that missing data is a common problem. Unless one presumes that statisticians are particularly singled out by fate for affliction with missing data, one suspects that both the problems and solutions to problems of missing data are

glossed over in most published research reports. This may be due either to a widespread failure to appreciate the significance of the problems (e.g. to the representativeness of the sample) or a widespread lack of awareness of the solutions being implemented (often automatically and by default) by the statistical software.

For example, the output from commonly available software packages, such as SAS (SAS Institute, 1989) and SPSS (1997), make little mention of the fact that their default method of dealing with missing data is something called

---

Correspondence to: Dr Aurelio José Figueredo, Director, Ethology and Evolutionary Psychology, Evaluation Group for Analysis of Data, Department of Psychology, University of Arizona, USA.

Submitted 22nd February 2000; initial review completed 25th May 2000; final version accepted 20th July 2000.

listwise (or casewise) deletion. This factual observation is not meant as a criticism of that particular method. However, throughout many years of statistical consultation, we have found that a very large proportion of users do not even realize that many of their observations are being discarded, and have little notion of what the implications of that procedure might or might not be. The only message available in the output relating to this process is usually a statement of how many of the total cases were usable for each analysis. Precisely how and why the missing cases were lost and which ones in particular they might have been is something that is frequently nowhere in evidence. The typical concern of the average user is therefore limited to whether too many cases have been lost to preserve sufficient statistical power for the specified analysis. Although this may represent a valid concern in itself, it is usually nowhere near the entirety of the problem.

Furthermore, the gravity of the problems associated with missing data are greatly increased by the use of multivariate analysis, the use of which is clearly increasing. Multivariate statistics are "complete-data methods" (Rubin, 1987); they require complete data on all cases and on all variables and, therefore, automatically eliminate from the analysis cases with missing data. Although some multivariate statistical software packages, such as EQS for Windows (Bentler, 1989), have options for alternative methods of dealing with missing data, the invisible default is still listwise deletion. Paradoxically, as will be described in greater detail below, the increased number of variables needed in a multivariate analysis greatly multiplies the risk of losing cases due to missing data under procedures such as listwise deletion. Nevertheless, the case will be made here that multivariate analysis also holds greater promise for otherwise unavailable solutions to the problem of missing data.

This paper represents a summary of a larger multi-authored work, currently in preparation, which will provide an overview of the missingness problem and of the approaches and procedures for handling it in multivariate analyses. This larger work will build on the existing literature on missing data and present the strengths and limitations of each procedure, as indicated by previous research, and is intended to assist investigators in selecting the procedure that is most appropriate and useful for their particular

situation. Detailed empirical comparisons of the results of the different procedures will also be presented in support of these recommendations. Due to space constraints, however, the present paper will necessarily have to sacrifice comprehensiveness in favor of conciseness in its review of the different methods available and proceed as expeditiously as possible to the "bottom line" conclusions and recommendations derivable from our work. Furthermore, the present paper will emphasize the presentation of a general strategy that we have developed for the multivariate modeling of missing data that requires no more than standard statistical software and an intermediate level of statistical sophistication rather than specialized computer programs and advanced technical training.

### **The different patterns of missingness**

Three general patterns of missing data have been proposed as categorizing the different possible effects of subject non-response (Little & Rubin, 1981). These kinds of missing data have been referred to as data that are: (1) Missing Completely At Random (MCAR), (2) Missing At Random (MAR) and (3) Non-Ignorable Non-Responses (NINR).

Data are considered MCAR when the data with missing values are a random sample of all data values (Schafer, 1997a). In other words, MCAR means that missingness is not systematically related to anything else in the study; it occurs completely by chance. Two main conditions define MCAR: (1) the cases with missing data form a simple random subset of the total study sample, and the mean values on the variable(s) with missing data, as well as on any other measured variable, do not differ between the subgroup of cases with complete data and the subgroup of cases with incomplete data; and (2) the particular datum that is missing is not related to other study variables or to the experimental conditions (Brown, 1983; Rubin, 1987). For example, in survey research, the MCAR responses that are missing on one item are not associated with responses missing on other items or with some characteristics of the participants. In experimental studies, MCAR associated with attrition do not differ between the experimental and the control groups.

Data are considered MAR when the data are randomly missing from one variable, but are

none the less conditional on the values of some other observed variables (Rubin, 1987). That is, the probability that a value on a variable is missing may depend on the values observed/obtained on another variable or covariate (Schafer, 1997a). Thus, the reason for missingness is known and measured and therefore can be modeled.

Data are considered NINR when the above conditions for randomness are not met. Such missingness is inevitably biased and is indicated by systematic, significant differences in the mean values of measured/observed variables between the subgroup of cases with complete data and the subgroup of cases with incomplete data (Brown, 1983). This tends to occur when the factors presumably causing the missingness are related to the personal characteristics of participants.

### **The consequences of missingness**

The problem of missing data in multivariate analysis has serious implications that threaten the validity of conclusions. Where missing data are NINR, the problem of missingness presents an obvious threat to "internal" or statistical conclusion validity (cf. Cook & Campbell, 1979). What often goes unrecognized is that even where data are technically MAR, the problem of missing data might still present a threat to "external" validity or the generalizability of the findings. The results of multivariate tests are often based on the subset of cases with complete data. When the cases with complete data differ systematically from those with incomplete data, the results of multivariate tests are therefore biased. It is not always possible, or even accurate, to assume that cases with complete data do not differ from those with incomplete data on all measured and unmeasured variables. Systematic differences between the two subsets of cases renders the subset of cases with complete data, and thus included in the multivariate analysis, unrepresentative of the target population of interest. Consequently, although this condition does not threaten the statistical conclusion validity, the findings are, strictly speaking, only applicable to one portion of the population and their generalizability to the whole target population may be severely limited (Schafer, 1997a; Kaufman, 1988; Gibbons *et al.*, 1993). Finally, even where data are MCAR, the number of cases excluded from analysis using

Listwise Deletion can be far in excess of what a researcher may anticipate. By merely missing a value on a variable used in the analysis, the entire case is deleted, including the rest of the nonmissing data. A substantial portion of the sample is lost when missing values occur on several analyzed variables. When the number of cases available for multivariate analysis is decreased, the statistical power to detect significant effects is reduced, potentially leading to Type II errors (Ward & Clark, 1991). The chances for Type II error increases when the study sample is small to begin with, as may occur in experimental studies evaluating the efficacy of a treatment.

Addressing the problem of missing data before conducting the planned multivariate tests of statistical significance is essential to minimize the potential threats to the validity of the study conclusions. While several procedures have been proposed to address the problem, the selection of the most appropriate one should be based on an adequate understanding of the causes leading to the missingness in a particular study.

### **The many meanings of missingness**

A thorough discussion of the various meanings that may be attributed to missing data, presuming that they are not MCAR, must start with the question of whether the missing data have any substantive meaning at all. Cohen devised a statistical test where it could be determined whether the missing data were somehow biased with respect to the non-missing data (cf. Cohen & Cohen, 1983). The basic idea was that subject non-responsiveness on a particular item could itself constitute information with predictive validity for some criterion variable. A separate dummy variable could be used to code for the "missingness", which was set equal to 1 when the particular datum was missing and to 0 when it was not. The missing datum could afterwards be replaced with data imputation methods such as mean substitution. Mean substitution presumes that the non-respondents would have been similar to the respondents on the item in question, had they only but deigned to respond. Thus, the statistically expected and most stochastically probable response to impute for them would be the mean of that of all of the respondents. Cohen, however, suspected that nonrespondents might frequently be straying from the crowd. By marking them with the

dummy variable and including that variable as a predictor in a multiple regression equation along with the mean-substituted variable on which data were formerly missing, it would be possible to determine if they were indeed different on the criterion of ultimate interest. If the missing data dummy variable is a significant predictor, it means that the respondents and the non-respondents would very likely have differed systematically on that item had the complete data been available, as evidenced by the predicted deviation in the criterion.

Although this diagnostic was originally proposed specifically for the data imputation method of mean substitution, we propose that it can be generalized to test empirically for systematic deviations from any imputed value, regardless of the method used. For example, the missing data dichotomy can be correlated both with other study variables and with each other to identify differences, if any, in the characteristics of cases with complete responses compared to those with incomplete responses. Besides serving as a diagnostic, this method can also be used to statistically control for missing data bias in multivariate analysis. The logic of its use in multiple regression introduced by Cohen can be generalized to multivariate models, such as path models and their associated measurement models. For example, within a multivariate measurement model or Confirmatory Factor Analysis (CFA), dummy variables can be created at the indicator level for each variable with missing data; the dummy variables can then be modeled to measure a latent variable reflecting missingness at the common factor or latent construct level (Muthen, Kaplan & Hollis, 1987). Within a manifest variable path analysis or Structural Equation Model (SEM), a missing data dichotomy may be incorporated into the structural component of the model to correct for any missing data bias. For example, univariate imputation of factor scores (e.g. mean substitution) may be combined with the dummy variable method in multivariate analysis by specifying structural pathways from the dummy variable coding for the missing factor to all other study variables in the model (cf. McCloskey, Figueredo & Koss, 1995). These dummy variable pathways to other study variables may be modeled as direct or may be modeled as indirect through the nonmissing common factors. Alternatively, the missing data dichotomy, whether at the indicator or the latent

variable level, can be used in multivariate analysis as a prior covariate, with its effects on the outcomes being residualized before testing the structural model. These practices are generally appropriate if the missingness is related to some characteristics of the respondents that may influence outcome achievement. Thus, the missing data dichotomy can be useful in statistical control of missing data bias when data are missing for reasons related to the characteristics of the participants.

### **The major approaches to handling missing data**

The most common approaches to dealing with missing data can be classified into two major categories: (1) Missing Data Deletion and (2) Missing Data Imputation.

#### *Deletion of missing data*

Common Missing Data Deletion methods often entail the indirect loss of non-missing data (such as by losing the entire case for which any one datum is missing), which functionally links the classic problem of missing data to that of small sample size, which we have at times referred to as the problem of *extremely* missing data. Two major alternative approaches to missing data deletion are Pairwise Deletion and Listwise (or Casewise) Deletion. In Pairwise Deletion, each bivariate correlation is estimated on all data available for each successive pair of study variables. Thus, Pairwise Deletion uses all available data in estimating each of the model parameters. This is an option usually available in standard software packages when computing correlation or regression coefficients. However, biased missingness and small samples may produce ill-conditioned sample correlation matrices.

Previous work (e.g. Raymond, 1986; Raymond & Roberts, 1987) has specified some of the conditions under which Pairwise Deletion can be effective, such as the following: (1) the missing data are MCAR, so that all pairwise complete data subsets are random samples of the same study target population; (2) the study sample is large, so that all pairwise complete data subsets are sufficiently large to adequately estimate the population parameters; (3) the percentage of missing data is low, usually less than 5%. Under the MCAR condition, the subgroups of

cases included in the estimation of different parameters are likely to represent the same population, because the central limit theorem holds with data that are MCAR. With Pairwise Deletion, however, any missing data biases that exist may produce subsamples for each different correlation coefficient based on slightly different subpopulations. Furthermore, if the multivariate analysis is to be conducted on a variance-covariance matrix (such as CFA or SEM), there is an increasing tendency with Pairwise Deletion, compared to with Listwise Deletion, that the determinant of the such a matrix will not be positive. What this means in practice is that the multivariate model will not be estimable with such data, and the analytical software simply will not run.

Listwise Deletion consists of eliminating cases with incomplete data on any of the variables included in the multivariate analysis: each bivariate correlation is estimated only on the subset of the data available for *all* possible pairs of study variables. As mentioned previously, it is the default option available with most multivariate statistical tests in almost all major statistical packages.

Previous work (e.g. Brown, 1983; Raymond, 1986; Raymond & Roberts, 1987) has identified some of the situations under which Listwise Deletion can be appropriate and efficient, such as the following: (1) the data are MCAR, since the parameters estimated are based on the subgroup of cases with complete data and this subgroup of cases is representative of the study target population; (2) the sample size is large enough so that discarding cases with incomplete data does not reduce statistical power dramatically; (3) the expected correlations among the variables are large, since data missing at either extremes of a score distribution would result in attenuated variance and covariance estimates; (4) error of measurement is minimal, since low scale reliability also attenuates the covariance estimates; and (5) the percentage of missing data is low on each of the variables included in the analysis, so that the total percentage (across variables and cases) is less than 5%. However, these conditions are quite similar to those specified for Pairwise Deletion, so that one is not provided with a pragmatic basis for selecting one data deletion method over the other, excepting the fact that Listwise Deletion generally eliminates more potentially useable data. On the other

hand, Listwise Deletion will rarely produce an "ill-conditioned" covariance matrix.

With Listwise Deletion, systematic biases in the data deleted may produce systematic biases in the data retained. This may result in a sample of usable cases that is unrepresentative of the study population, regardless of the amount of care initially taken to construct the original sampling frame. It is here that the structure of the missing data, or the lack thereof, becomes critically important and the difference between the effects of missing data on univariate and on multivariate analyses becomes most pronounced. If the pattern of missingness across different variables has a latent structure such that there are substantial correlations between the occurrence of missing data on one variable and the occurrence of missing data on another, then the missing data biases introduced by Listwise Deletion will be homogeneous and thus greatly amplified. If however, there are no substantial correlations between missingness on different variables, as in the case of truly "messy" missing data, then the biases introduced by Listwise Deletion may be so heterogeneous as to nearly cancel each other out. Nevertheless, we might also need to be concerned that listwise deletion may produce deviations from multivariate normality, which may constitute a problem for analytic methods such as ML, EM, and other model-based estimation procedures that rely on this condition (Little & Rubin, 1991).

#### *Detection of biases in multivariate missing data*

The recommended approach to detection of biases following Listwise Deletion therefore differs between univariate and multivariate analysis. In the univariate approach to missing data diagnosis, one may construct a different dummy variable to code for missingness of each separate variable with missing data. Using Pairwise Deletion diagnostically, one may then test the correlation of each dummy variable with every other study variable as well as that of each dummy variable with every other dummy variables coding for missing data. If there are multiple variables with missing data, as is often the case, the missing data dummy variables can also be correlated to each other. If the property of "missingness" is found to be correlated between different variables, factor analysis of their corresponding missing data dummy variables may

also be used to reveal the latent structure of the missing data (Richard Gorsuch, personal communication).

It should be noted that these procedures may grossly inflate the experiment-wide rate of Type I error ( $\alpha$ ), unless protective corrections (such as the Bonferroni procedure) are applied. However, such conservative tests may lack the statistical power to detect any biases if the number of dummy variables is high. Alternatively, one need only to construct a single dummy variable to code for two subsamples, the "listwise deleted" and the "listwise retained". Using Pairwise Deletion diagnostically, one may then test correlation of that single dummy variable with every other study variable and evaluate effect sizes of the single dummy variable on the other study variables as measures of the "bottom-line" Listwise Deletion bias. This alternative procedure may be useful because the different biases associated with different variables containing missing data may be of no practical interest. Whereas homogeneous missing data biases across study variables may tend to compound, the heterogeneous missing data biases of different variables may tend to cancel out, much as do "heterogeneous irrelevancies" in a meta-analysis. When missing data are not sufficiently "structured", a condition which we have referred to less politely as "messy" missing data, the biases on different variables with missing data may be quite heterogeneous. Thus, as mentioned above, we have found Listwise Deletion to be a great deal less hazardous than it sounds if the initial samples are sufficiently large and the missingness sufficiently "messy" or "unstructured", a condition which might produce a pattern of missing data that is virtually random. The major problem under Listwise Deletion then becomes whether the size of the remaining sample is sufficient to provide adequate statistical power, because "messy" missing data may sometimes cause the deletion of most (or even all!) of the original sample.

Conversely, when missing data are sufficiently "structured", the sample can be subdivided into a manageable number of subsamples according to which data they have missing. In multivariate analysis, certain advanced procedures such as multisample analysis can then be used compare these subsamples for homogeneity with respect to shared model parameters. Multi-sample analysis is used in SEM (including CFA) for the

purpose of determining whether the parameter estimates hold for the subgroup of cases with different patterns of complete and incomplete data. The study sample is divided into groups corresponding to distinct patterns of missingness. The same model is tested, simultaneously, on the covariance matrices generated for each group. Equality constraints for various parameters across the groups could be imposed and empirically tested (Muthen *et al.*, 1987). Parameter estimates that are found equal across groups are considered stable, and these groups are considered comparable and, therefore, representative of the same population. A statistical rejection of such cross-sample equality constraints can be diagnostic of the missing data being biased. Evaluating cross-sample equality constraints in multi-sample analysis may be appropriate when: (1) the data are not necessarily MCAR, but instead may be MAR and therefore related to other relevant characteristics of the participants (cf. Erickson *et al.*, 1995); (2) the sample size is sufficiently large; (3) the proportion of missing data is large; and (4) the factor pattern and the structural relations are hypothesized to be the same for the groups of cases with different patterns of missing data.

#### *Univariate imputation of missing data*

If certain Missing Data Deletion methods run the risk of sacrificing much of the data you actually have, some Missing Data Imputation methods involve the creation of data which are otherwise nowhere in evidence. However, what have been called "Hot Deck" Data Imputation methods (cf. Little & Rubin, 1989) can themselves be classified into two major categories: (1) Between-Subjects Imputation, and (2) Within-Subjects Imputation. Between-Subjects Imputation involves assigning a score to a missing datum based on the scores of *other* subjects on the *same* variables. Within-Subjects Imputation involves assigning a score to a missing datum based on the scores of the *same* subjects on *other* variables.

Between-Subjects imputation is consistent with the nomothetic perspective for data analysis, which assumes that the best estimate of an individual's response is that which is most typical for the group. This imputation procedure consists of substituting the individual's missing value with that of some "average" value derived from

other members of the group. For example, the imputed value could be: (1) the mean for the study sample of the variable with missing data; (2) the mean of a randomly selected subgroup of the study sample; (3) the mean of a subgroup to which the case with incomplete data actually belongs (e.g. the mean of the experimental or the control group); (4) the value of the same variable observed in a randomly selected case; (5) the highest or lowest value of the same variable observed in any other case; and (6) normative population values for the variable with missing data, if available. In longitudinal designs, the mean group value obtained at one point in time is often used to impute missing data at that time for any given individual with missing data.

Although Between-Subjects Imputation is frequently used in analysis of variance, its potential for inflating the *F*-ratio should be taken into consideration. This imputation procedure may be of minimal effectiveness in multivariate correlation analyses, especially when the intercorrelations among variables are expected to be moderate-to-high, since it results in reduced variance and subsequently in biased (i.e. artificially low magnitude) estimates for the correlation coefficients (Raymond, 1986; Kaufman, 1988; Ward & Clark, 1991).

It is our position that Between-Subjects Imputation is fraught with peril because one must somehow justify the substitution by assuming that the case with the missing datum is similar on the variable of interest to other cases with nonmissing data. However, it is rarely substantiated that subjects missing and not missing data on the same variables are in any way similar. This assumption may have little justification if the missing data is systematically biased, but can be evaluated by the dummy variable method. We believe that Within-Subjects Imputation holds greater promise, especially for multivariate analysis, because one can more easily substantiate that the variable with the missing datum is similar to other variables for all cases with nonmissing data. For example, if the variables in question are parallel measures of the same latent construct, one can easily substantiate that variables missing and not missing data are somehow similar. This assumption can be directly supported by results of multivariate analysis.

Within-Subjects Imputation procedures have not yet been extensively tested for their effectiveness under different missing data conditions.

However, they are viewed as appropriate when the analysis is correlational since they do not reduce the variance-covariance estimates and when the proportion of missing data is relatively large. Within-Subjects Imputation procedures are also consistent with idiographic perspectives for data analysis. The particular value selected for the imputation of the missing datum may differ based on the design of the study. In a longitudinal design, for example, when score on an outcome variable is missing at one occasion within the series of measurements, all the following values have been proposed as likely candidates for missing data imputation: (1) the value on the occasion before or after the one with the missing value; (2) the average values on these two occasions (i.e. before and after the one missing); (3) the average of the case across all occasions with complete data. When the score on the outcome variables is missing at later occasions of measurement, related for instance to dropout or to early voluntary termination from the study, the following imputation procedures have been suggested: (1) Interpolation or Forecasting, where the scores for the missing values at later points in the series are predicted from earlier ones (Heyting *et al.*, 1992); (2) Zero-Implantation, where a value of zero is used to replace the missing scores at later points in the repeated measures series, which could be useful when cases withdraw from the study because of perceived lack of benefit and lack of effectiveness of the intervention (Weiss, 1991), for instance, when a person terminates treatment due to either complete treatment failure (including death, see: Nickel *et al.*, 1995; Raboud *et al.*, 1998) or complete treatment success (as when a former substance abuser has fully recovered and does not require further treatment or follow-up); and (3) Last-Observation Carried Forward (LOCF), where, for each missing score on the outcome variable, the case's last available score on the same variable is imputed for the missing one (Heyting *et al.*, 1992; also referred to as "last value carried forward" or LVCF in Lavori, 1992). All of these alternatives for longitudinal data ultimately rely on the similarities (or "intra-class correlations") within each individual subject of consecutive values on the same variable.

In cross-sectional designs, there are by definition no other values for that particular subject on the same variable. The regression method

circumvents this limitation by regressing the variable with missing data onto its most common covariates. These covariates are other variables that have been found in previous research to correlate highly with the variable presently containing missing data. Thus, the information available in the covariates is used to estimate the missing values for any given variable. Either of two statistical techniques can be used for this purpose: (1) simple regression, in which the missing data variable are regressed onto their covariate(s) using the cases with complete data, and the obtained solution for the regression equation is used to compute the missing values; and (2) iterative regression, in which the values on the variable with missing data are first imputed using another missing data procedure (mean substitution being the most frequently used) to create a complete dataset, the missing values are estimated and replaced using regression analysis, and the regression results are then used to repeatedly revise the missing value estimates until they show minimal change between such iterations.

*Multivariate imputation of cross-sectional missing data*

Although such univariate techniques are available for Within-Subjects Imputation, the advent of latent variable modeling makes a multivariate strategy particularly appealing. In multivariate analysis, it is important to distinguish between whether data are missing only on specific indicators or data are missing on all the indicators of a common factor. This latter case is an extreme example of the associations that may be found between the missing data dummy variables of different variables with missing data. In this case, the set of similarly affected variables may be the multiple indicators of a common factor. If this association is very high, there will be many cases in which it will be found that the entire common factor itself can be said to be missing, which is a condition that might be fraught with substantive meaning. Conversely, data that are missing on specific indicators but not others are much easier to impute based on available data and may not reflect much more than a problematical item or two. Thus, the distinction between missing factors and missing indicators constitutes a special case of the broader distinction between “structured” and “unstructured” missing data.

If only some, rather than all, of the items on a multiple-item scale or of the indicators on a common factor are missing, one can use the non-missing convergent items or indicators to estimate the missing ones. If the items or indicators are all expressed in a common metric, and if the scale or factor is known to be unidimensional, having acceptable internal consistency and convergent validity of indicators, a very easy way to do this is by imputing the missing item or indicator score as the mean of the non-missing item or indicator scores (McDermeit, Funk & Dennis, 1999). The composite scale value can then be computed using the sum of all the item or indicator scores, whether they were originally non-missing or subsequently imputed from the others. However, there are potential disadvantages to this procedure. If the scale is not definitely known a priori to be internally consistent and if confirmatory psychometric procedures are used following this kind of item-level imputation, using the mean of the non-missing items in the data for the imputed value of each missing item can inflate estimates of either interitem consistency (such as Cronbach’s alpha) or convergent validity (such as common factor loadings).

One way to circumvent this problem is to use the non-missing convergent items or indicators to directly estimate an imputed score for the underlying scale or latent variable itself. Thus, if only the reconstructed scale or latent variable score is of ultimate interest, one need not even impute a value for the missing item or measured variable score. An additional advantage of this method is that, for any observations on which complete data are available, these procedures will also automatically compute the scale or latent variable scores in the normal way, rendering this procedure seamless with that of scale construction for the sample as a whole. Furthermore, a minimum number of non-missing items or indicators (e.g. three or more) can also be specified to avoid imputing scale or latent variable scores with what might be deemed an insufficient number of nonmissing items or indicators. This minimum acceptable number may depend on several psychometric criteria, such as item or indicator reliability and convergent validity. There is also an implicit tradeoff between the relative wisdom of using scale or latent variable scores based on reduced numbers of items or indicators, entailing some loss of

scale reliability and perhaps validity, and that of using samples with reduced numbers of cases, entailing some loss of statistical power and perhaps generalizability, due to loss of usable values for scale or latent variable scores on which an insufficient number of items or indicators are deemed to be available. The relative merits of these alternatives must be considered separately for each situation because there is no general rule which can be readily formulated to cover all the possible scenarios.

We have named this general approach Multivariate Imputation (MVI), and tentatively classify it as a Missing Data Imputation procedure. We are tempted to put it into a different category altogether, along with Multisample Analysis, and call it something like "Missing Data Avoidance", because it works around missing data rather than literally trying to replace it. However, we decided to leave it in with other missing data imputation procedures because the underlying latent variable score (whether psychometric or chronometric) is being imputed based on incomplete information. Nevertheless, no indicator score is ever actually replaced as in a true data replacement procedure. MVI is a technique for working within the confines of the data you actually have. Sample SPSS and SAS commands for these procedures are provided in the appendix.

MVI is not to be confused with Multiple Imputation (MI), which is another strategy which does not fall neatly into our "Between-Subjects" versus "Within-Subjects" scheme. MI is a simulation-based approach to the handling of missing data, which relies on replacing each missing datum with more than one simulated value (Schafer, 1997a). This produces multiple reconstructed versions of the complete data. The empirical distributions of these different versions of the complete data can be used as a basis for inferential statistics, such as the estimation of confidence intervals or hypotheses testing. MVI, on the other hand, does not support this additional function. Nevertheless, there are still various controversies surrounding the use of these algorithms, such as the correct estimation of the applicable sample sizes and standard errors, and certain problems in convergence where the percent of missing information is large (Little & Rubin, 1989). Furthermore, not one of these techniques for handling missing data has yet been widely adopted by practicing data analysts (Schafer, 1997a). A brief discussion of the vari-

ous software packages, modules and codes currently available for MI is also provided in Appendix I.

In contrast, MVI can be easily implemented with commonly available software and is not computationally intensive. MVI procedures ultimately rely for their validity on the theory and practice surrounding the factor score estimation procedure known as "unit weighting" (Gorsuch, 1983). Because of this equal weighting of convergent indicators, all provided that all these components are expressed in the same metric, no multiple regression procedures are usually deemed necessary for the estimation of the missing values. By dispensing with sophisticated estimation algorithms which require adequate statistical power for a sufficient level of precision, these simpler algorithms are also amenable for use with smaller sample sizes, a related problem which is taken up again below.

To fully justify the use of MVI in cross-sectional data, we therefore digress briefly into a discussion of the estimation of factor scores (cf. Wiggins, 1973; Gorsuch, 1983; Figueredo, Cox & Rhine, 1995). The scores assigned to latent variables in traditional factor analysis are often called "differentially weighted" factor scores to emphasize the fact that the indicator variables are not weighted equally in the composite. These differential weights are ultimately based on the different factor loadings ( $\lambda$  coefficients) assigned to the indicators by the factor model, indicating their different degrees of convergent validity with respect to the latent construct. The problem is that, with moderate sample sizes, the standard errors for different factor loadings are typically quite large. These estimates are therefore known to be quite "unstable" or sample-specific, making it seldom possible to discriminate between similar factor loadings. This situation only worsens when the sample sizes are smaller. An alternative method is referred to as the estimation of "unit-weighted" factor scores, in which all significant indicators weighted equally (i.e. 1.0). These have nevertheless been found by numerous Monte Carlo studies to be correlated on the order of 0.95 with differentially-weighted factor scores, and therefore to be largely intersubstitutable with them. Furthermore, they have been found to be more generalizable than differentially-weighted factor scores across independent samples. They are also much easier to calculate, the most common

method being the simple summation of the standardized scores of the salient indicators.

The procedure for estimating unit-weighted factor scores is quite simple: (1) the factor model must be theoretically specified a priori (i.e. confirmatory); (2) all the indicators must be expressed in a common metric (e.g.  $Z$ -scores); (3) the factor scores can be estimated as arithmetic means of the indicator scores, because means represent linear transformations of summed "unit weighted" factor scores but remain invariant with respect to the number of indicators actually used; and (4) standardizing the new factor scores eliminates any superficial differences between means and sums anyway. The convergent validity of these scores can be readily examined using the following steps: (1) the bivariate correlations of the factor scores with the indicator scores represent the "factor structure" (lambda coefficients); (2) the bivariate correlations of the factor scores with each other represent "factor intercorrelation matrix" (the phi matrix); and (3) the bivariate correlations of the factor scores with their own arithmetic mean may be used to represent any "higher-order factor structures" hypothesized. As in confirmatory factor modeling, the bivariate correlations representing factor loadings can be tested for statistical significance. However, these bivariate methods will not provide any overall measures of goodness-of-fit of the data to the common factor model.

Nevertheless, unit-weighted factor scores are ideal for MVI and for the detection of associated biases, which unfortunately might remain with us following imputation. For example, they can be used to test the generalization of factor structures across what would otherwise become the listwise-retained and the listwise-deleted subsamples as follows: (1) using listwise deletion, construct and validate factor models (e.g. by CFA) for listwise-retained data; (2) estimate unit-weighted factor scores as the means of standardized ( $Z$ ) scores of significant indicators; (3) compute unit-weighted factor scores for the listwise-retained data using complete data; (4) impute unit-weighted factor scores for the listwise-deleted data using the values of all non-missing indicators; (5) using pairwise deletion diagnostically, construct factor structures for both subsamples by correlating unit-weighted factor scores to indicator variables and factor intercorrelation matrices by correlating unit-

weighted factor scores to each other; (6) compare the factor structures and phi matrices of the listwise-deleted subsample to those of the listwise-retained subsample; (7) where the factor structures and phi matrices cross-validate across subsamples, one may recompute the unit-weighted factor scores using MVI for the recombined sample, because MVI will automatically "compute" factor scores for complete data and "impute" factor scores for incomplete data; and (8) where the factor structures and phi matrices do not cross-validate well across subsamples, alternative missing data procedures should be implemented.

Several clarifications and qualifications of this otherwise straightforward procedure should be noted. For step 1, the initial listwise deletion should also be used to delete any cases for which the number of indicators was deemed to be below the minimum acceptable for MVI, if such a minimum is to be subsequently applied. Also in step 1, CFA may be used if the available sample size after listwise deletion is sufficiently large and there exists sufficient theory to support prespecifying a common factor model. If there is insufficient or inadequately substantiated theory, Exploratory Factor Analysis (EFA) may be used instead (Gorsuch, 1983). Alternatively, one may choose to apply any number of more traditional psychometric methods of scale construction and validation (e.g. Wiggins, 1973). In step 5, specifying a minimum acceptable number of non-missing indicators for MVI will avoid the possible artifact of occasionally correlating an indicator with itself where all others are missing. For step 6, one may choose to apply any number of more traditional psychometric methods of factor comparison: specifically, those that are used to compare factor structures based on the same set of indicator variables across independent samples (e.g. Gorsuch, 1983).

#### *Multivariate imputation of longitudinal missing data*

The recent theoretical extension of latent variable modeling to growth curve analysis allows us to apply a variant of MVI to missing data in longitudinal studies. The latent growth curve parameters, conceived of as "chronometric" factors (as opposed to traditional "psychometric" factors), can be estimated for each subject by using the non-missing data from the remaining time points in each series. Although both these

applications entail the loss of some degree of reliability and validity for the imputed latent variable scores, this can be accounted for in the final analysis by weighting the estimated factor scores by the number of measurements (whether they be convergent indicators or successive time points) actually used. By analogy, the application of this logic to the problem of missing data in longitudinal studies has also been called the "meta-analytic" approach to growth curve analysis (Figueredo *et al.*, 2000) because it is essentially equivalent to a weighted meta-analysis of multiple longitudinal case studies (cf. Hedges & Olkin, 1985).

This way of handling missing data in longitudinal studies is also essentially equivalent to the procedures automatically implemented for handling missing data in Random Regression Models (see Hedeker & Mermelstein, 2000), such as those implemented in SAS PROC MIXED. As in Random Regression Models (RRMs), these procedures can also be used with non-randomly missing data, as by including terms in the model that can potentially adjust for missing data. For example, one could define subgroups of individuals based on the number of weeks they were measured, estimate a treatment effect for each subgroup, and examine whether these effects are consistent. One can also use covariates to predict non-response (Gibbons *et al.*, 1993). These additional strategies may become important if the missing data results from non-random causes, for example, where dropping out from the study may be due to either lack of improvement or very rapid improvement. The most fundamental differences between the meta-analytical approach, on one hand, and both RRM and Hierarchical Linear Models (HLMs), on the other hand, are as follows: (1) the meta-analytical approach to growth curve analysis estimates the Level 1 and Level 2 growth curve models in two sequential steps, whereas most RRM and HLMs perform both operations within a single step; (2) the meta-analytical approach uses the true intercepts of each individual growth curve as an estimate of each subject's pre-treatment baseline, whereas RRM and HLMs use their means over time, which may be affected by differential rates of change and, hence, by post-treatment effects; and (3) RRM and HLMs are both complex univariate models, whereas the meta-analytical approach can also accommodate multivariate modeling, such as

SEM, if one uses the weighted covariance matrix of the Level 1 growth curve parameters as the data input for the Level 2 growth curve analysis.

Individual growth curve analysis uses chronometric latent constructs to examine the individual's pattern of change in a variable over the occasions of measurement (cf. Collins & Horn, 1991). The pattern of change is described by a trajectory that reflects the way in which the individual's value on the variable changes over time. The intercept is the best estimate of the initial status of each individual at baseline and the slope is the parameter indicating the average rate of change over time (Rogosa, Brandt & Zimowski, 1982; Rogosa & Willett, 1985). The rate of change can be estimated with a linear regression equation in which the values on the variable is regressed on the time variable; the slope and intercept parameters may be estimated separately for each individual in the sample (Petrinovich & Widaman, 1984). Non-linear trends can also be examined. Missing data at any occasion of measurement does not present a problem, since growth curve analysis does not require the same data collection design for each individual. The number and spacing of measurement occasions may vary across cases, since the focus is on the true process of change (Bryk & Raudenbush, 1992) rather than on a linear transformation of the values observed over time. Thus, MVI makes use of all the longitudinal observations available for each subject, which may greatly improve the reliability of the estimated growth curve parameters and thus the statistical power of the analysis (cf. Sutcliffe, 1980).

#### *Limitations of multivariate imputation*

Nevertheless, it is important to recognize the limitations of MVI. The first of these is that MVI uses the information contained within multiple convergent items or indicators, implying that it can only be applied in multivariate analyses involving either scales with multiple convergent items or latent variables with multiple convergent indicators, including repeated measures of the same chronometric constructs over time. Secondly, MVI works with missing items or indicators, and does not work with entirely missing scales or common factors. All indicators of a scale or common factor might none the less be missing, suggesting a systematic pattern

of non-response. Alternatively, the number of non-missing indicators might be less than the specified minimum, producing what is essentially the same condition. We therefore put forward the modest proposition that MVI is a very useful technique in multivariate analysis with missing data, but that it is no panacea. Sometimes, in spite of our best efforts, we are going to lose some of the data, and then we may have to deal with the problems of a reduced sample size.

### **Small samples: the problem of extremely missing data**

In many cases, the problem of small sample sizes can be construed as a problem of extremely missing data. For example, the dummy variable method of handling missing data, whether applied in the univariate or the multivariate case, is not generally recommended when the proportion of missing data exceeds 20% of the total sample size (Cohen & Cohen, 1983). As mentioned above, a large proportion of missing data is generally deemed to be a threat to the validity of most methods of data imputation. Common Missing Data Deletion methods, such as Listwise and Pairwise Deletion, may result in a small dataset even though we started with what appeared to be a big one. From this it follows that a large proportion of data that is missing completely at random is functionally equivalent to a "small" (or at least "smaller") sample size. Where a large proportion of the data is missing completely at random, the only problem with data deletion methods is that the remaining sample may be of inadequate size. This is because data that are missing completely at random do not compromise the representativeness of the original sample.

We may now reasonably ask just what defines a "small" sample size? Several operational definitions of small sample sizes have been proposed. Most of these principles have been put forward in relation to how many subjects are required to perform valid statistical analyses. The majority of these rules represent the number of subjects required as relative to number of predictor variables in the model to be tested (e.g. 20:1; cf. Pedhazur, 1982). Others make this ratio of subjects to predictors contingent on the exploratory versus the confirmatory nature of model (e.g. 20:1 versus 10:1; cf. Cohen &

Cohen, 1983). In structural equations modeling, where a clear distinction is necessary between the number of included variables and the number of model parameters, this ratio is made relative to number of free parameter estimates (e.g. 5:1, cf. Bentler, 1989). A dissenting view is put forward by those who argue for an absolute rather than a relative definition of adequate sample size, making the minimum number of subjects independent of the number of predictor variables and equal to the absolute number needed to stabilize the correlation coefficient (e.g. 75–100; cf. Gorsuch, 1983).

All but the latter principle imply that the definition of a small sample size is determined by how many measures are to be used in the analytical model. Working forwards from the minimal number of measures we would need to adequately answer the practical questions posed by the study, we may deduce that the minimal number of cases that need to be sampled is some predetermined multiple (as specified above) of the number of measures that would be required. However, application of this algorithm might reveal that some program evaluations simply cannot be accomplished with the resources available. Working backwards from the number of study participants actually available, we may deduce that the number of measures to be analyzed should be "no more than are warranted" by our available sample size, as estimated by the rules given above. However, the application of this inverse algorithm might seriously limit the scope of our inquiry, to the extent that the results of the final version of the study would be of extremely limited utility. What we are going to defend is the proposition that the correct answer to the question of how many measures to obtain is not "no more than are warranted" by the available sample size, but instead "as many as you can get", subject to certain caveats detailed below, and that this contrary principle applies especially with smaller sample sizes!

### *Sources of variance*

To understand why we would propound such an apparently heretical notion, it is useful to revisit the relevant implications of one of the fundamental principles in statistical theory related to the notion of statistical power, the central limit

theorem. The Central Limit Theorem defines the sampling variance of the mean as follows:

$$S_{\bar{x}}^2 = S_i^2/N$$

where  $S_{\bar{x}}^2$  represents the sampling variance of the mean,  $S_i^2$  represents the variance of the individual observations, and  $N$  represents the sample size or number of individual observations. The sampling variance of the mean is the square of the standard error of the mean, which is used in both the computation of confidence intervals in parameter estimation and in the calculation of critical values for hypothesis tests of statistical significance. As a matter of practical importance, the way that this relation is typically perceived by social scientists is as follows:

$$S_{\bar{x}}^2 = S_i^2/N$$

with  $N$ , the sample size, being the term of most critical importance and most amenable to our control to enhance our statistical power. Unfortunately, in health services research it is often not as easy to control our sample size as it might be in, for instance, experimental psychology. We therefore propose that this implication of the Central Limit Theorem is better understood as follows:

$$S_{\bar{x}}^2 = S_i^2/N$$

The astute reader will no doubt observe that how bold or italic we make the letters in our notation of the formula has no bearing on the outcome of the computations. Nevertheless, what we are proposing is not a change in the way we do the math, but a change in perspective from what has become the commonplace conception of the problem.

The change in perspective that we advocate entails revisiting the essential relationship between variance and data aggregation. There are at least two major sources of variance in our data: (1) sampling error, which is indeed, as explained above, largely a function of  $N$ , and (2) measurement error, which often constitutes over 50% of the variance between scores. A typical reliability that is deemed acceptable by most in the social sciences is 0.70, meaning that fully half of the observed variance in the data is due to pure measurement error. This is to be distinguished from sampling error, which is usually due to real individual differences, depending on the design of the study. The reason that this is

called "error" at all is that inadequate sampling of individuals, given those real differences, may cause us to make erroneous estimates of the population parameters. Given a random sampling of individuals from the population of interest, this sampling error is not expected to be either a systematic underestimate or overestimate of the parameters, but a truly stochastic effect of our "luck of the draw" on any given sampling occasion. Nevertheless, the lion's share of the excess observed variance in our data is often due to measurement error and not to sampling error. That measurement error, in addition to true individual differences, shows up in the  $S_i^2$  portion of the equation given above.

#### *The psychometric principle of aggregation*

One possible way to act to reduce this excess measurement error is intentionally selecting measures with higher reliability. However, such measures are often either not available in program evaluation or their reliabilities are unknown. Another way to approach the same kind of solution is to use the psychometric principle of aggregation (cf. Nunally, 1978). Much has been both theorized and empirically demonstrated with respect to this principle in the field of personality assessment (e.g. Mischel, 1973, 1983, 1984; Epstein, 1979, 1980, 1983; Mischel & Peake, 1982), but a detailed discussion of those particular issues is beyond the scope of this paper. This principle is also implicit in such well-established concepts as the Spearman-Brown Prophecy Formula, the doctrine of Multiple Operationism, and the use of multiple convergent indicators for construct validation (Campbell & Fiske, 1959). However, these principles are usually applied in the field of multivariate analysis which, it is commonly said, requires even higher sample sizes than usual for inferential statistics. Can we possibly be compounding our heresy by suggesting that these be applied to small samples?

The fundamental reason that multivariate methods, such as factor analysis, require so many cases is that they require a large number of parameter estimates. In exploratory factor analysis, for example, the number of parameter estimates is at least equal to the number of indicators times the number of common factors, because every factor is allowed a loading on every indicator. In confirmatory factor analysis,

even assuming perfect factorial simplicity, the number of parameter estimates is at least equal to the number of indicators, corresponding to one factor loading for each indicator. Both models are thought to require a minimum number of cases equal to some multiple of the number of parameter estimates required. On the other hand, it is reasonable to question the practical need for so many parameter estimates. Once the factor model has been correctly specified (i.e. once the significant indicators have been identified), the principal function of differentially estimated factor loadings appears to be the estimation of differentially weighted factor scores. Thus, if one is willing to use unit-weighted factor scores, one may question why so many parameter estimates are needed at all.

As mentioned above, even with what are normally considered “adequate” sample sizes, the standard errors for the different factor loadings are typically so large that it is seldom possible to discriminate, with any satisfactory degree of confidence, any more than between the “large” ones and the “small” ones, also called the “salient” and the “hyperplane” loadings (Gorsuch, 1983). As sample sizes become smaller, and standard errors progressively larger, discriminating between differentially estimated factor loadings becomes increasingly hopeless. It therefore stands to reason that one should, perhaps, not even attempt to estimate any differential loadings with smaller samples. If this logic is valid, it implies that common factor modeling with small samples (and, perhaps, with many samples formerly considered sufficiently large) need only support a single parameter estimate, assumed to be equal across all prespecified indicators, otherwise known as estimating unit-weighted factors. This model simplification, occasioned by “inadequate” sample sizes, reduces the need for data dramatically.

#### *Integration of multivariate imputation with multivariate aggregation*

We therefore see that unit-weighting of latent variable scores proverbially “kills two birds with one stone”: first, it permits the use of MVI under the most commonly occurring conditions of missing data. Secondly, it permits the use of what we have dubbed MVA, the small-sample strategy described in this section, for those conditions of missing data where the reduction of

our original data to a smaller sample might become practically unavoidable. Unit-weighting automatically provides a convenient fall-back option (MVA) for circumstances when MVI alone might not otherwise preserve a sufficiently large proportion of our original sample. Furthermore, the MVA procedure is implicit in MVI, and therefore does not require us to apply any additional procedures. The same sample SPSS (1997) and SAS (1989) language provided in Appendix I for MVI will automatically perform both functions (MVA and MVI) as occasioned by the presence of complete or of incomplete data.

Finally, the operating characteristics of both MVI and MVA have transparent implications for the strategic design of studies where the collection of either incomplete or small samples can be plausibly anticipated. Both strategies rely on the use of a multiplicity of indicators for each latent construct of interest. Paradoxically, both strategies are designed for situations under which the collection of data on that many variables would be otherwise expected to be fraught with peril: (1) “messy” (“unstructured”) missing data, and (2) small sample size. Nevertheless, the optimal conditions for the application of both MVI and MVA suggest the counterintuitive recommendation that we try to collect data on as many manifest indicators or parallel measures of the same constructs as possible, *especially* when we can plausibly anticipate incomplete samples, smaller samples, or some combination of both.

#### *Limitations of multivariate aggregation*

There are certain limitations of this combined MVI/MVA strategy which should also be noted. First, it is possible that including more measures might actually lead to disproportionately more missing data in questionnaires due to increasing the respondent burden. Adding measures should therefore be carried out with some moderation, keeping respondent burden in mind. Secondly, mindlessly increasing the number of indicators soon reaches a point of diminishing returns. Thirdly, any added indicators should be sufficiently distinct from existing ones to avoid the problem of creating “bloated specifics” instead of true common factors. In spite of these various caveats, however, we have found in verbal presentations of this proposition that many of our colleagues find some of our brazen assertions

somewhat hard to take. To substantiate our seemingly extravagant claims, we have performed a variety of data simulations to demonstrate the effectiveness of our recommendations under a broad spectrum of empirical conditions.

### The relative performance of MVI and MVA in missing data and small sample simulations

This finally brings us to the empirical question of which missing data procedure works best with real data, and under what conditions of missingness some procedures might work better than others. We have conducted a variety of data simulations starting with an initially complete set of real empirical data (so that we could know the "right" answer) but generating missing data systematically to test the performance of the various alternative missing data procedures under different multivariate patterns of missingness. In one such study, the relative performance of these procedures was assessed under varying conditions of usable sample size, degree of missing data bias, number of indicators with missing data, and amount of simulated measurement error.  $M$  was defined as the proportion of the data experimentally deleted (in eight increments of 5% each, from  $M = 5\%$  to  $M = 40\%$ );  $B$  was defined as a dichotomy for whether the missing data was randomly deleted (from anywhere in the entire distribution) or non-randomly deleted (from only the upper quartile of the distribution);  $K$  was defined as a dichotomy for whether only one or all of the indicators of a latent common factor were missing; and  $E$  was defined as the proportion of random error added to each score, taken from a random variable with the same mean and standard deviation as each original variable (in 10 increments of 10% each, from  $E = 0\%$  to  $E = 90\%$ ).  $M$ ,  $B$ ,  $K$  and  $E$  were varied independently in 100 independent samplings for each  $M*B*K*E$  condition, yielding a total of 32 000 simulations for this study. Unfortunately, the orthogonalized design of these parametric data manipulations produced certain artifacts which affected our final conclusions.

Nevertheless, a comparison of missing data deletion and missing data imputation techniques yielded the following results. First, Pairwise Deletion produced highly variable and often pathological results under different missing data conditions (e.g. by producing non-positive-

definite covariance matrices). Secondly, Listwise Deletion performed surprisingly well under a wide array of missing data conditions. In fact, all else being equal (*ceteris paribus*), Listwise Deletion even appeared to outperform MVI in most missing data conditions. In reality, however, all else is seldom equal because Listwise Deletion typically produces much more missing data than MVI (e.g. with "messy" missing data). The fact that this was not reflected in the results our data simulations was an artifact of the orthogonalized design. Nevertheless, MVI produced conservative but robust estimates under most missing data conditions and held up well under varying degrees of missing data bias, but broke down under conditions of very high measurement error (e.g. 70%). However, when both missing data bias and measurement error were very high, none of these techniques worked very well.

To demonstrate the value of MVA with smaller samples, we again conducted a series of simulations starting with a complete set of real empirical data. These data simulations involved the random selection of progressively smaller number of subjects from the initial sample of complete data, and the random selection of progressively decreasing numbers of indicator variables from a psychometrically homogeneous multiple-item scale. In one such study, for example, the scale used was a 20-item composite with a Cronbach's alpha of 0.92, which was not substantially modifiable, either upwards or downwards, by dropping any of the items chosen. This new scale alone, which we shorten here to  $X1$ , predicted 55% of the variance in our criterion variable, which we here call  $Y$ . However, to monitor the effects of data aggregation on the influence of other collinear variables in the model, another statistically significant predictor, here called  $X2$ , was also used in the prediction equation for the purposes of simulation. In these data simulations,  $N$  represented the number of subjects randomly selected in each sample (in 15-subject increments, from  $N = 15$  to  $N = 120$ ),  $K$  represented the number of items randomly selected to be used in the composite (in four-item increments, from  $K = 4$  to  $K = 20$ ).  $N$  and  $K$  were varied independently in 30 independent samplings for each  $N*K$  condition, yielding a total of 1200 simulations for this study.

Hierarchical multiple regressions were then used to estimate the relative impact of independently perturbing  $N$  and  $K$  on various parame-

ters of the prediction equation (relating  $X1$  and  $X2$  to  $Y$ ). In spite of regularly assigning causal priority to  $N$ , the degree of data aggregation ( $K$ ) used in the 20-item scale ( $X1$ ) proved to be an exceedingly important factor, compared with sample size ( $N$ ), in the statistical power of the prediction equations, as reflected in such indices as the magnitude of the parameter estimates, the standard errors and the coefficients of determination of the prediction equation. For example, the effect (unstandardized regression weight) of  $K$  on the effect size ( $B1$ ) of  $X1$  on  $Y$  was over 200 times the absolute magnitude of that of  $N$  on  $B1$ . As expected from the principle of disattenuation for the unreliability of a fallible predictor (Cohen & Cohen, 1983), increasing the measurement reliability of  $X1$  (by increasing  $K$ ) substantially increased the estimated magnitude of the effect of  $X1$  on  $Y$ . Interestingly, the effect (unstandardized regression weight) of  $K$  on the effect size ( $B2$ ) of  $X2$  on  $Y$  was over 60 times the absolute magnitude of that of  $N$  on  $B2$ . As expected, increasing the measurement reliability of  $X1$  (by increasing  $K$ ) substantially reduced the estimated magnitude of the correlated and thus competing predictor,  $X2$ . Similarly, the unstandardized regression weight of  $K$  on the root-mean-squared error ( $RMSE$ ) of the prediction equation was over forty times the absolute magnitude of that of  $N$  on the  $RMSE$ . As anticipated, increasing the measurement reliability of  $X1$  (by increasing  $K$ ) greatly reduced the errors of prediction, as reflected in the average absolute size of the regression residuals,  $RMSE$ . These results generally support our proposition that MVA can be an extremely important tool in compensating for small sample size in statistical models.

The following points summarize the principal results of our explorations, both theoretical and empirical, of the uses of MVA for smaller samples: (1) unreliability and invalidity of measurement are usually larger sources of error variance than sampling error; (2) use of unit-weighted factor scores derived from validated factor structures can substantially enhance both reliability and validity of measurement; (3) Monte Carlo simulations show that values of validated unit-weighted factor scores are highly stable across random samples of decreasing size drawn from the same population (equivalent to the results of Listwise Deletion with increasingly large amounts of data missing completely random); (4) use of unit-weighted sample scores also per-

mits MVI to prevent a small sample from getting even smaller due to missing data; (5) intelligent use of MVA can make small samples yield statistically significant results; and (6) all dimensions of the Cattell "Data Matrix" (cf. Gorsuch, 1983), and not merely the number of "cases" available, co-determine the total number of observations and thus the aggregate statistical power of a multi-dimensional sampling design. Furthermore, hierarchical multiple regressions on simulated data indicated that: (1) the number of indicators used in the measurement model for the latent variable was overwhelmingly more influential than sample size in affecting the structural model parameters; (2) excessive concern over the ratio of cases to variables in latent variable structural equation models is probably not warranted; (3) more indicator variables for multivariate constructs are manifestly more desirable than larger sample sizes; (4) where one variable is collinear with another, as in a mediational model, too few indicators for a latent variable can skew the results towards favoring the confounding variable in the structural model by inflating its effect size; and (5) MVA works well as a solution for the problem of small sample sizes.

### Summary and conclusions

We have reviewed the problem of missing data in multivariate analysis of both cross-sectional and longitudinal data from a variety of perspectives. We have discussed the lack of protracted attention that this problem is generally given in the substantive research literature and engaged in informed speculation on the probable reasons for that state of neglect. In response to this alarming state of affairs, we have tried to clarify the serious implications that missing data can have for both the internal and external validity of research results and thus justify the amount of effort truly required to adequately address this problem. To put the situation in perspective, we have tried to explicate the various interpretations that can be made regarding the occurrence of missing data and what it might and might not be able to tell us about the substantive problem at issue. To be able to test hypotheses regarding the possible substantive significance of missing data, we have reviewed and extended the applicability of some available but underutilized diagnostic procedures for determining the possible impact of missing

data upon the results of a study. We then reviewed the various patterns that have been observed in the occurrence of missing data and discussed the relative merits and limitations of the most common procedures for dealing with missing data in terms of these general patterns. Based on these comparisons, we developed a general strategy, which we called Multivariate Imputation (MVI), for maximally exploiting the unique characteristics of multivariate analysis in dealing with missing data in both cross-sectional and longitudinal studies. We also explored the value of a related technique which we called Multivariate Aggregation (MVA), implicit in the logic of MVI, in compensating for the problems of small sample size, which often result from the use of missing data deletion methods under conditions of "messy" missing data. Finally, we tested the validity of these recommendations on systematically altered incomplete subsets of an originally complete set of empirical data. These experimentally generated incomplete subsets of real research data were designed to simulate a wide array of parametric conditions under which missing data may naturally occur and enable us to compare the performance of various alternative missing data procedures, including MVI, in recovering the known parameters of the complete data. The results of these data simulations generally supported our theoretical expectations regarding the advantages of MVI under a wide variety of missing data conditions. However, we were surprised to also find that Listwise Deletion performed at least as well as MVI under a broad range of conditions provided no substantially greater loss of data was associated with that procedure. Nevertheless, we deemed it unlikely that Listwise Deletion could in actual practice preserve usable sample sizes comparable to those typically salvaged by MVI, and attributed the marginal superiority of Listwise Deletion in our data simulation experiments to an artifact of our orthogonalized design. Finally, we observed that under the more extreme conditions of error and bias, none of the missing data procedures evaluated could adequately recover the known parameters of the complete data. This last observation should serve as a check against overconfidence in the ability of any of our current methods to compensate for excessive loss or unreliability of data. Applying these data simulation techniques to the performance of MVA with smaller samples, we found that the number

of indicators measured for a latent variable was overwhelmingly more important in correctly estimating a variety of key model parameters than the sample size itself. MVA can therefore serve as a built-in fail-safe mechanism for circumstances under which MVI would otherwise be unable to preserve a sufficient proportion of the original sample. In spite of these admitted limitations, there is some justification for a certain degree of guarded optimism regarding the advantages of MVI under the more commonly experienced parametric conditions for the occurrence of missing data in both cross-sectional and longitudinal research.

## References

- BENTLER, P. M. (1989) *EQS: structural equations program manual* (Los Angeles, CA, BMDP Statistical Software).
- BROWN, C. H. (1983) Asymptotic comparison of missing data procedures for estimating factor loadings, *Psychometrika*, 48, 269–291.
- BRYK, A. S. & RAUDENBUSH, S. W. (1992) *Hierarchical Linear Models. Applications and data analysis methods* (Newbury Park, CA, Sage).
- CAMPBELL, D. T. & FISKE, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin*, 56, 81–105.
- COHEN, J. & COHEN, P. (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Hillsdale, NJ, Lawrence Erlbaum).
- COLLINS, L. M. & HORN, J. L. (1991) *Best Methods for the Analysis of Change: recent advances, unanswered questions, future directions* (Washington, DC, American Psychological Association).
- COOK, T. D. & CAMPBELL, D. T. (1979) *Quasi-experimentation: design and analysis issues in field settings* (Boston, MA, Houghton Mifflin).
- EPSTEIN, S. (1979) The stability of behavior, I. On predicting most of the people much of the time, *Journal of Personality and Social Psychology*, 37, 1097–1126.
- EPSTEIN, S. (1980) The stability of behavior, II. Implications for psychological research, *American Psychologist*, 35, 790–806.
- EPSTEIN, S. (1983) Aggregation and beyond. Some basic issues on the prediction of behavior, *Journal of Personality*, 51, 360–391.
- ERICKSON, J. R., STEVENS, S., MCKNIGHT, P. & FIGUEROA, A. J. (1995) Willingness for treatment as a predictor of retention and outcomes, *Journal of Addictive Diseases*, 14, 135–160.
- FIGUEROA, A. J., COX, R. L. & RHINE, R. (1995) A generalizability analysis of subjective personality assessments in the Stumptail macaque and the Zebra finch, *Multivariate Behavioral Research*, 30, 67–197.
- FIGUEROA, A. J., BROOKES, A. J., LEFF, S. & SECHREST, L. (2000) A meta-analytic approach to growth curve analysis, *Psychological Reports*, 87, 441–465.

- GIBBONS, R. D., HEDEKER, D., ELKIN, I., WATERNAUS, C., KRAEMER, H. C., GREENHOUSE, J. B., SHEA, T., IMBER, S. D., SOTSKY, S. M. & WATKINS, J. T. (1993) Some conceptual and statistical issues in analysis of longitudinal psychiatric data, *Archives of General Psychiatry*, 50, 739-750.
- GORSUCH, R. L. (1983) *Factor Analysis* (Hillsdale, NJ, Lawrence Erlbaum).
- GREGORICH, S. (1999) EM\_COVAR.SAS. Written in SAS Macro language. <http://sites.netscape.net/gregorich/missing.html>
- HEDEKER, D. & MERMELSTEIN, R. J. (2000) Analysis of longitudinal substance use outcomes using ordinal random-effects regression models, *Addiction*, 95(suppl. 3), S381-S394.
- HEDGES, L. V. & OLKIN, I. (1985) *Statistical Methods for Meta-analysis* (New York, NY, Academic Press).
- HEYTING, A., TOLBOOM, J. T. & ESSERS, J. G. (1992) Statistical handling of drop-outs in longitudinal clinical trials, *Statistics in Medicine*, 11, 2043-2061.
- KAUFMAN, C. J. (1988) The application of logical imputation to household measurement, *Journal of the Market Research Society*, 30, 453-466.
- LAVORI, P. W. (1992) Clinical trials in psychiatry: should protocol deviation censor patient data? *Neuropharmacology*, 6, 39-48.
- LITTLE, R. & RUBIN, D. B. (1989) The analysis of social science data with missing values, *Sociological Methods and Research*, 18, 292-326.
- McCLOSKEY, L. A., FIGUEREDO, A. J. & KOSS, M. P. (1995) The effects of systemic family violence on children's mental health, *Child Development*, 66, 1239-1261.
- McDERMEIT, M., FUNK, R. & DENNIS, M. (1999) *Data cleaning and replacement of missing values*, LI SPSS Training Series (Bloomington, IL, Chestnut Health Systems), [www.chestnut.org/li/tools](http://www.chestnut.org/li/tools).
- MISCHEL, W. (1968) *Personality and Assessment* (New York, NY, Wiley).
- MISCHEL, W. (1973) Toward a cognitive social learning conceptualization of personality, *Psychological Review*, 80, 252-238.
- MISCHEL, W. (1983) Alternatives in the pursuit of the predictability and consistency of persons: stable data that yield unstable interpretations, *Journal of Personality*, 51, 578-604.
- MISCHEL, W. (1984) Convergences and challenges in the search for consistency, *American Psychologist*, 39, 351-364.
- MISCHEL, W. & PEAKE, P. K. (1982) Beyond déjà vu in the search for cross-situational consistency, *Psychological Review*, 89, 730-755.
- MUTHEN, B., KAPLAN, D. & HOLLIS, M. (1987) On structural equation modeling with data that are not missing completely at random, *Psychometrika*, 52, 431-462.
- NICKEL, J. T., SALSBERY, P. J., CASWELL, R. J., KELLER, M. D., LONG, T. & O'CONNELL, M. (1995) Quality of life in nurse case management of persons with AIDS receiving home care, *Research in Nursing and Health*, 19, 91-99.
- NUNNALLY, J. C. (1978) *Psychometric Theory* (New York, McGraw-Hill).
- PEDHAZUR, E. J. (1982) *Multiple Regression in Behavioral Research: explanation and prediction* (New York, NY, Holt, Rinehart & Winston).
- PETRINOVICH, L. & WIDAMAN, K. F. (1984) An evaluation of statistical strategies to analyze repeated-measures data, in: PEEKE, H. V. S. & PETRINOVICH, L. (Eds) *Habituation, Sensitization and Behavior*, pp. 156-201 (New York, NY, Academic Press).
- RABOUD, J. M., SINGER, J., THORNE, A., SCHECHTER, M. T. & SHAFRAN, S. D. (1998) Estimating the effect of treatment on quality of life in the presence of missing data due to drop out and death, *Quality of Life Research*, 7, 487-494.
- RAYMOND, M. R. (1986) Missing data in evaluation research, *Evaluation and the Health Professions*, 9, 395-420.
- RAYMOND, M. R. & ROBERTS, D. M. (1987) A comparison of methods for treating incomplete data in selection research, *Educational and Psychological Measurement*, 47, 13-26.
- ROGOSA, D. R., BRANDT, D. & ZIMOWSKI, M. (1982) A growth-curve approach to the measurement of change, *Psychological Bulletin*, 92, 726-748.
- ROGOSA, D. R. & WILLET, J. B. (1985) Understanding correlates of change by modeling individual differences in growth, *Psychometrika*, 50, 203-228.
- RUBIN, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys* (New York, NY, John Wiley).
- SAS INSTITUTE, INC. (1989) *SAS Language and Procedures: usage*, version 6, 1st edn (Cary, NC, SAS Institute).
- SCHAFFER, J. L. (1996a) CAT: multiple imputation for multivariate categorical data, software library for SPLUS. Written in SPLUS and Fortran-77. <http://www.stat.psu.edu/~jls/>
- SCHAFFER, J. L. (1996b) MIX: multiple imputation for mixed continuous and categorical data, software library for SPLUS. Written in SPLUS and Fortran-77. <http://www.stat.psu.edu/~jls/>
- SCHAFFER, J. L. (1996c) NORM: multiple imputation of incomplete multivariate data under a normal model, software library for SPLUS. Written in SPLUS and Fortran-77. <http://www.stat.psu.edu/~jls/>
- SCHAFFER, J. L. (1997a) *Analysis of Incomplete Multivariate Data* (London, UK, Chapman & Hall).
- SCHAFFER, J. L. (1997b) PAN: multiple imputation for multivariate panel data, software library for SPLUS. Written in SPLUS and Fortran-77. <http://www.stat.psu.edu/~jls/>
- SCHAFFER, J. L. (1998) LMM: some improved procedures for linear mixed models, software library for SPLUS. Written in SPLUS and Fortran-77. <http://www.stat.psu.edu/~jls/>
- SCHAFFER, J. L. (1999) NORM: multiple imputation of incomplete multivariate data under a normal model, software for Windows 95/98/NT, Version 2. Written in Fortran-95 with Windows extensions. <http://www.stat.psu.edu/~jls/>
- SPSS SCIENCE (2000a) *SPSS for Windows*, release 10 (Chicago, IL, SPSS Science).
- SPSS SCIENCE (2000b) *SYSTAT for Windows*, release 9 (Chicago, IL, SPSS Science).
- STATISTICAL PROGRAM FOR THE SOCIAL SCIENCES (SPSS) (1997) *SPSS Base 7.5 Syntax Reference Guide* (Chicago, IL, SPSS), <http://www.spss.com/>

- STATISTICAL SOLUTIONS, INC. (1998) *SOLAS for Missing Data Analysis*, version 1 (Cork, Ireland, Statistical Solutions).
- SUTCLIFFE, J. P. (1980) On the relationship of reliability to statistical power, *Psychological Bulletin*, 88, 509–515.
- WARD, T. J. & CLARK, H. T. (1991) A re-examination of public-versus private school achievement: the case for missing data, *Journal of Educational Research*, 84, 153–163.
- WEISS, D. J. (1991) A behavioral assumption for the analysis of missing data: the use of imputed zeros, *Journal of Social Behavior and Personality*, 6, 955–964.
- WIGGINS, J. S. (1973) *Personality and Predictors: principles of personality assessment* (Reading, MA, Addison-Wesley Publishing Company).

## Appendix I

This appendix contains some sample codes for the implementation of Multivariate Imputation (MVI) in both SPSS and SAS, as well as a brief listing of currently available software packages, modules and codes for Multiple Imputation (MI). This listing includes bibliographical citations for publications and addresses for websites where more information regarding these software packages, modules or codes may be found.

A simple algorithm for the replacement of missing item or indicator scores with the mean of the non-missing item or indicator scores has been written in SPSS (1997), where “var4” represents the scale item with the occasional missing values, “var1” to “var3” are other scale items with complete data, and “scale” is the reconstructed four-item scale (McDermeit *et al.*, 1999):

```
compute replace = var4.
do if (missing(replace)).
  compute replace = rnd(mean(var1 to var4)).
end if.
compute scale = sum(var1,var2,var3,replace).
```

Note that this algorithm is exclusively for imputing values on “var4” and presumes that no data are missing for “var1” to “var3”.

An alternative algorithm for directly imputing the scale or latent variable score from the non-missing item or indicator scores without directly imputing a score for the missing item or indicator is also available within SPSS (1997). Where “var1” to “var4” are the scale items, on any of which data might be missing, and “scale” is the reconstructed four-item scale, this procedure is extremely simple (McDermeit *et al.*, 1999):

```
compute scale = mean(var1 to var4).
```

Note that this algorithm works equally well for any missing values in “var1” to “var4”. The minimum number of nonmissing items or indicators (e.g. three or more) deemed sufficient to impute the scale or latent variable score can be specified using the following variant of this command:

```
compute scale = mean.3(var1 to var4).
```

A similar procedure exists in SAS (1989), which can be used to define a scale value within the DATA step:

```
scale = mean(of var1 var2 var3 var4);
```

Note that this algorithm also works equally well for any missing values in “var1” to “var4”. Although there appears to be no simple variant of this function to specify a minimum acceptable number of non-missing items, the following additional commands can be used to accomplish the same objective, where “nscale” is the number of non-missing items on this scale for each case:

```
nscale = n(of var1 var2 var3 var4);
if nscale < 3 then nscale = .;
```

While we are at it, we can also add a line to construct Cohen’s missing data dichotomy for any scale values left missing by this procedure, where “dscale” is the dummy variable representing missingness on this scale for each case:

```
If scale = . then dscale = 1; else dscale = 0;
```

Alternatively, variants of this command can be written to diagnose the effects of any number of nonmissing items from which the scale value has been imputed:

```
If nscale = 3 then dscale3 = 1; else dscale3 = 0;
If nscale = 2 then dscale2 = 1; else dscale2 = 0;
If nscale = 1 then dscale1 = 1; else dscale1 = 0;
```

These dummy variables can be used to diagnose empirically whether imputing scale values with differing numbers of constituent items might have had any effect on the results. These diagnostics might be useful because if no minimum number is specified, either of these SPSS or SAS algorithms will automatically compute the scale value based on whatever number of non-missing items is available. For any observations on which complete data are available, these procedures will also automatically compute the scale values in the normal way. Note that the following alternative SPSS (1997) and SAS (1989) commands, respectively, which appear logically equivalent on the surface, will instead produce Listwise Deletion of the entire scale for any missing item:

```
*SPSS:
  compute scale = (var1 + var2 + var3 + var4)/4.
*SAS:
  scale = (var1 + var2 + var3 + var4)/4;
```

Simple correlation procedures in SAS can be used to generate both the unit-weighted factor structure and the unit-weighted factor intercorrelation matrix (roughly corresponding to the lambda and phi matrices in confirmatory factor modeling), implementing Pairwise Deletion of any missing values by default. These two procedures can be implemented in SAS with the following command language, where “scale1” and “scale2” are two different four-item scales, composed, respectively, of “var1” to “var4” and “var5” to “var8”:

```
*Unit-Weighted Factor Structure:
proc corr;
  var scale1 scale2;
  with var1 var2 var3 var4 var5 var6 var7 var8;
```

```
*Unit-Weighted Factor Intercorrelation Matrix:  
proc corr;  
var scale1 scale2;
```

These SAS procedures can be used with the Listwise Retained as well as with the Listwise Deleted subsamples, permitting a direct comparison of their results, because the same commands will work with either complete or incomplete data. These procedures will automatically test all of these correlations for statistical significance but do not provide any overall measures of goodness-of-fit of the data to the common factor model.

There are various algorithms currently available for ML, usually based on some form of ML (Maximum Likelihood) estimation, including EM (Expectation Maximization) and FIML (Full Information Maximum Likelihood). The EM algorithm is implemented in both the SPSS and HLM software packages, the

FIML algorithm is implemented in both the AMOS and MX multivariate software packages. In addition, there are now numerous stand-alone software programs, comprehensive statistical package modules and software codes available for handling missing data. Stand-alone software programs include NORM (Schafer, 1999) and SOLAS (Statistical Solutions, Inc., 1998). Comprehensive statistical package modules include those for SPSS (SPSS Science, 2000a) and SYSTAT (SPSS Science, 2000b). Specialized software codes for existing statistical software packages include CAT (Schafer, 1996a), MIX (Schaefer, 1996b), NORM (Schafer, 1996c), PAN (Schafer, 1997b) and LMM (Schafer, 1998) for S-Plus and EM for SAS (Gregorich, 1999). The technical features and relative benefits of each package is beyond the scope of the current paper.