To Impute or not Impute: That's the Question

Paper Methodological Advice

Lodder, P. (Paul)
University of Amsterdam, Amsterdam, The Netherlands

Researchers often encounter missing values in their datasets, yet they frequently use suboptimal methods to tackle this missing data problem. This might be explained by either ignorance of more sophisticated missing data techniques, or anxiety about the complexity of those methods. The main goal of this paper is to present a number of more sophisticated missing data techniques in non-statistical terms, so that it can be a guide for researchers or methodological consultant who want to tackle a missing data problem. We first give an overview of older missing data treatments and illustrate their limitations and possible usefulness. After that we wll focus on four modern missing data treatments: mean imputation, regression imputation, multiple imputation and a maximum likelihood technique. We illustrate the usefulness of each technique with a practical example.

Imagine a researcher who studies the relationship between having a mental disorder and being abused as child. The researcher creates a survey in which participants are among others asked to indicate whether they have been abused as a child and whether they have a mental disorder. As soon as the data has been collected, the researcher wants to analyze it but discovers a lot of missing values in the dataset on the child abuse question. The situation described above is not uncommon in scientific research, especially in sciences that have humans as their primary research unit.

When analyzing data, missing values in your dataset can be problematic because most statistical analyses are not designed to handle these missing values. For this reason, researchers often ask statistical consultants how to solve this problem. One easy solution is to ignore the participants with missing data altogether and perform your analyses on a limited part of the dataset. However, this solution – called *listwise deletion* or *complete case analysis*– may lead to biased, underpowered, or unreliable parameter estimates (Schafer & Graham, 2002; Wothke, 2000). In fact, one study showed that listwise deletion leads to a decrease in statistical power between 35% (if 10% of the data is missing) and 98% (if 30% of the data is missing)(Raaijmakers, 1999).

To overcome these limitations, statisticians have started to invent more sophisticated techniques. In this article we give an overview of these new techniques in non-statistical terms and illustrate the advantages and limitations of each technique. We use examples whenever possible and aim to make this article a consultant's companion in the treatment of missing data problems.

Types of missing data

When thinking about missing data, it is useful to make distinctions between different types of missing data. First, we can distinguish between unit nonresponse and item nonresponse. In case of unit nonresponse, the data of an entire unit is missing. For example, a programming error that the prevented the computer from saving the data. Item nonresponse, on the other hand, occurs when only a part of the unit data is missing.

Figure 1. Missing data prevents researchers to see the entire story that their data has to tell.

I would like to thank Herman Ader, Don Mellenbergh and the students of the Methodological Advice course for providing me with useful feedback on earlier drafts of this paper.

▶ E-mail: **p.lodder@uva.nl**

Within item nonresponse, we can make another useful distinction between a missing variable and a missing item. The difference between the two is that a missing item is *a part* of a set of items that are meant to measure a concept (e.g. an item in a 20 item depression questionnaire), whereas a missing variables is already a concept in itself (e.g. age, or gender) and does therefore consist of only one measure.

The last distinction we will make concerns the randomness of the missing data (Rubin, 1976). If the missingness does not depend on any information in the dataset, then we can call the data *missing completely at random* (MCAR). With MCAR, the probability of missingness is unrelated to any information in the dataset and therefore considered a random process, much like a coin flip. If, on the other hand, the missingness *does* depend on information in the dataset, then the missing data belongs to either one of two types. The data is *missing not at random* (MNAR) if the missingness depends on the missing variable or item itself, for instance if a questionnaire item is too sensitive to answer. In contrast, the data is *missing at random* (MAR) if the missingness does not depend on the item or variable itself, but on other observed information in the dataset. An example of MAR is that the missingness of a questionnaire item depends on the gender or the age of the respondents (Little & Rubin, 1989).

These different types of missing data are important to keep in mind, because they determine which statistical treatments of the missing data can effectively be used. MNAR is often considered to be the worst missing data type, because it most often leads to biased parameter estimates in your statistical analyses. With MCAR and MAR, this is less of a problem, but those missing data mechanisms might lead to a loss of statistical power (Graham, 2009).

The amount of missing data is another consideration that determines the most optimal treatment of missing data (Ader, Mellenbergh, & Hand, 2008). If more than 25% of the data is missing and researchers apply modern treatments to impute the missing data, then they should always compare the results of their subsequent analyses with the results they would have obtained if they had used complete case analysis. If the results differ, then we can conclude this to be the result of the imputation strategy and this makes it less likely that the results will be publishable.

Keeping these considerations in mind, we will now give a short overview of the historical solutions to the missing data problem. We do not provide an exhaustive overview, but aim to describe the most commonly used techniques and illustrate their limitations and possible usefulness.

Historical solutions

The best treatment of missing data is its prevention. When designing a study, researchers should always try to think about situations that might cause their data to be missing. They can then either try to avoid those situations, or collect additional data on the reasons for the potential missingness. However, if researcher are not able to prevent the missing data, then they can use a number of techniques to treat this problem.

Most statistical software has a built in option to treat missing data with two different methods: *listwise deletion (or complete case analysis)* and *pairwise deletion (or available case analysis)*. Both methods involve deleting missing values from the dataset, but they differ in the extent to which they delete. Listwise deletion deletes each unit that contains a missing value. So even if the dataset contains an infinite number of columns with missing data in only one column, then listwise deletion still excludes the entire unit from further analyses (hence the name *complete* case analysis. Pairwise deletion, on the other hand, does not exclude the entire unit, but uses from every unit as much data as possible. This means that if a unit has missing values, then we can still use its non-missing values in statistical analyses.

A disadvantage of listwise deletion is that it can lead to biased parameter estimates, because the group with complete data will most likely differ from the group with missing data. (Schafer & Graham, 2002; Wothke, 2000). Another disadvantage is that deleting entire cases might lead to a decrease in statistical power, especially if most respondents have at least one missing value. This second disadvantage is solved by the pairwise deletion treatment, because it lets us still use the available data of each unit, even if they all have at least one missing value. However, a drawback of this pairwise deletion technique is that it can still lead to biased parameter estimates, because we compute them based on different sets of units (Schafer & Graham, 2002).

The drawbacks mentioned above do not necessarily imply that we should never use listwise– and pairwise deletion. There are still situations in which both can be useful and in which their disadvantages are minimized.

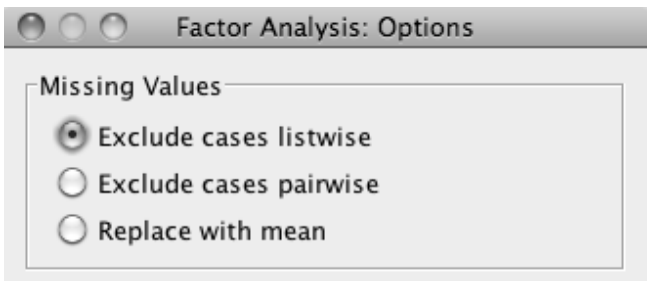

Figure 2. Most statistical software (in this case: SPSS) include listwise and pairwise techniques to treat missing data.

For instance, if departures from MCAR are not extreme, then the bias might be trivial. However, it is difficult to judge the size of the bias in practice. Independent of the missing data mechanism, a rule of thumb is that if less than 5-10% of the data is missing, then pairwise deletion does no longer pose a major threat to statistical power. Moreover, if the source of the missingness is known and it is not in the missing variable itself (e.g. if the data is MAR), then we can even include the variable that causes the missingness as a covariate in our model, which reduces the bias of the parameter estimates (Graham & Donaldson, 1993). The kind of study researchers want to conduct also influences whether a deletion treatment is appropriate. If researcher conduct an exploratory factor analyses on a large number of variables and do not plan to publish the data, then they can apply pairwise deletion without alarming consequences (Graham, 2009).

The situations described above show that we can still use pairwise or listwise deletion methods, although under strict rules. The two case deletion techniques are certainly the most time efficient solution, but most researchers advise to use more sophisticated *imputation based* methods to treat missing data, because they often result in more power and less bias (Rubin, 1976; Little & Rubin, 1989; Schafer & Graham, 2002; Graham, 2009). The next section gives an overview of the most used imputation methods and illustrates their advantages and disadvantages.

Modern solutions

An alternative to the deletion methods are techniques that try to replace the missing data with plausible values based on characteristics of the non-missing data. The practice of replacing missing data with new values is called *data imputation* (Rubin, 1976). The reason for imputing data is not that we want to estimate what a respondent would have said if the data would have been present. Rather, we impute the data to get a better estimate of the distribution underlying the data. The focus of data imputation is therefore on population level and not on unit level. The various imputation techniques aim to provide accurate estimates of population parameters such as means, variances and correlations (Weiner, Schinka, & Velicer, 2012). The main advantages of imputation techniques are that power is not decreased, because the data imputation retains the entire sample size. Another advantage is that after imputation, researchers are most often allowed to use standard analysis methods and software without having to worry about the previously missing data (Schafer & Graham, 2002). Below we elaborate on a number of imputation methods and their limitations.

Mean imputation

One of the easiest ways to impute missing data is to replace each missing value with the mean of non-missing values of the variable or item. The simplicity of this method is also its disadvantage: the distribution of the imputed variables can get highly distorted and the variance underestimated, because each missing value is assigned the same imputation value. Simulation studies show that mean imputation indeed yields highly biased parameter estimates (Graham, Hofer, & Piccinin, 1994; Graham, Hofer, & MacKinnon, 1996; Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997). However, some studies point out that the limitations of mean imputation are almost absent if less than 10% of the data is missing and when the correlations between the variables are low (Raymond, 1986; Tsikriktsis, 2005). Therefore, we advise not only use mean imputation under these strict conditions and if other more sophisticated imputation techniques are not possible. Two techniques similar to mean imputation are median and modus imputation. Those methods were invented to account for imputation of not normally distributed data, but they suffer from the same limitations as mean imputation and are therefore generally not recommended.

Regression imputation

Another way in which we can impute data is by predicting the missing value from one or more non-missing values. Suppose for instance that a variable Y contains missing values, and a set of other variables X and Z do not contain missing values. We can then predict the missing Y values by using the non-missing X and Z values as predictors in a regression analysis of Y on X and Z. We then replace all missing Y's with the predicted Y's. We can also extend this method by using several non-missing variables to estimate a propensity score. We can then use this propensity score to predict the missing Y values.

An advantage of regression imputation over mean imputation is that the former is better able to preserve the distribution shape. Contrary to mean imputation, regression imputation can also be used when more than 10% of the data is missing and when the data contains highly correlated variables (Little & Rubin, 1989). A disadvantage of this method, however, is that it might still lead to biased parameter estimates, especially with MNAR and MAR mechanisms (Schafer & Graham, 2002). Although regression imputation forms the basis of the more complicated methods described below, we only recommend to use the technique if the missing data mechanism is MCAR (Tsikriktsis, 2005).

Maximum Likelihood

A missing data treatment that overcomes the above mentioned problem of a biased covariance structure is the Expectation Maximization (EM) algorithm (Little & Rubin, 1989). An EM algorithm produces maximum likelihood (ML) estimates of the missing values. The algorithm alternates between two steps: the expecta-

tion (E)-step and the maximization (M)-step. In the E-step, the algorithm uses the observed data and current parameter estimates to compute the expected value of the incomplete data. In the M-step, the algorithm uses the expected values obtained in the E-step, to calculate a mean vector and covariance matrix of the data. Subsequently, a maximum likelihood function will be maximized to provide new parameter estimates. These new parameter estimates are again used in the E-step and this process repeats itself until it converges to stable parameter estimates. Note that different solutions are possible, because the algorithm can converge to local maxima and might be therefore be unable to find the real maximum.

A drawback of this maximum likelihood (ML) based technique was that it assumes the missing data to be missing at random (MAR). However, more recent methods can use ML to even estimate parameters under MNAR mechanisms (Liou, 1998). Another limitation of the EM algorithm is that it does not automatically provide standard errors of the parameter estimates. This makes it difficult to use EM for studies in which hypothesis are tested. However, a workaround to this problem is to obtain an estimate of these standard errors by using bootstrap procedures (Efron & Efron, 1982; Graham et al., 1997). Despite this workaround, Graham (2009) still advises to withhold from the EM technique if the missing data concerns a hypothesis testing study. Instead, the multiple imputation technique (described below) can more effectively be used. Studies that do not test hypotheses and do not so much need a standard error, such as coefficient alpha analyses, can effectively use the EM algorithm to impute missing data (Enders & Peugh, 2004; Graham, Cumsille, & Elek-Fisk, 2003).

Multiple imputation

Missing data treatments like mean imputation and regression imputation treat missing data by imputing a single value for each missing data point. A disadvantage of these method is that the error of these imputations is not incorporated. For instance, the regression imputations all lie directly on the regression line, while in reality there most often is some error variance that causes the data points to deviate from the regression line. Another disadvantage of single imputation techniques is that a single imputation does not represent the uncertainty associated with the missing value.(Graham, 2009).

Multiple imputation techniques solve this problem by randomly drawing multiple imputations from a distribution of imputations and also by introducing additional error variance to each imputation. The variability among the m imputations represents the uncertainty of the imputation. Researchers can indicate how many sets of imputations they want to randomly draw from the population. The multiple imputation technique results in an imputed dataset for each drewed set of imputations. Researchers can then perform analyses on each of those datasets and aggregate the results. (Rubin, 1976). Similar to the maximum likelihood technique, a disadvantage of multiple imputation is that it assumes the data to be missing at random (MAR). Nevertheless, multiple imputation has been studied extensively and most people agree that it is a very powerful technique (Rubin, 1976; Tsikriktsis, 2005; Donders, van der Heijden, Stijnen, & Moons, 2006; Van Ginkel, Van der Ark, & Sijtsma, 2007; Graham, 2009). However, some studies also show that in some instances the maximum likelihood technique performs better, for instance in factor-analysis studies (Fay, 1992; Bernaards & Sijtsma, 1999).

Practical example

We performed a simulation study to illustrate the effectiveness of different imputation methods. We used a dataset of 508 participants who all filled out the Beck's Depression Inventory (BDI). This questionnaire measures the extent to which someone can be diagnosed as suffering from a depression. We assessed the reliability of the BDI and proved to be a reliable measure because Cronbach's alpha was equal to 0.8.

To test the effectiveness of data imputation, we need a dataset with missing values. To achieve this, we used the software *R* to simulate the missing values. We simulated the missing values to be either missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). In the MAR simulations, we let the missingness depend on data characteristics other than the items themselves. We simulated men to have a higher chance on missing data than women, because men in general are thought to have difficulty in expressing their emotional states and therefore possibly a higher reluctance in answering the BDI questions. In the MNAR simulations, we simulated a different chance of missingness for each BDI item. For each item, we decide whether or not we thought it could be a question that is too sensitive to answer for some people. For each of the three missing data mechanisms, we simulated four different percentages of missing data: 10%, 20%, 30% and 40%. In total, this lead to twelve different datasets with missing values.

We treated the missing values of each of those twelve datasets with four different missing data techniques (i.e. mean imputation, regression imputation, multiple imputation, maximum likelihood imputation). We wrote our own R-function to do the mean imputations, we used the R-package *Mice* to perform the regression- and multiple imputations (m=5 imputed datasets), and we used the R-package *Lavaan* to apply the maximum likelihood technique based on the expectation maximization (EM) algorithm. Appendix A shows both the R-codes used to simulate the datasets with missing values, as well as the R-codes used to impute the missing data with each of the four missing data techniques.

After we used to different techniques to impute each dataset, we calculated the Cronbach's alpha of each

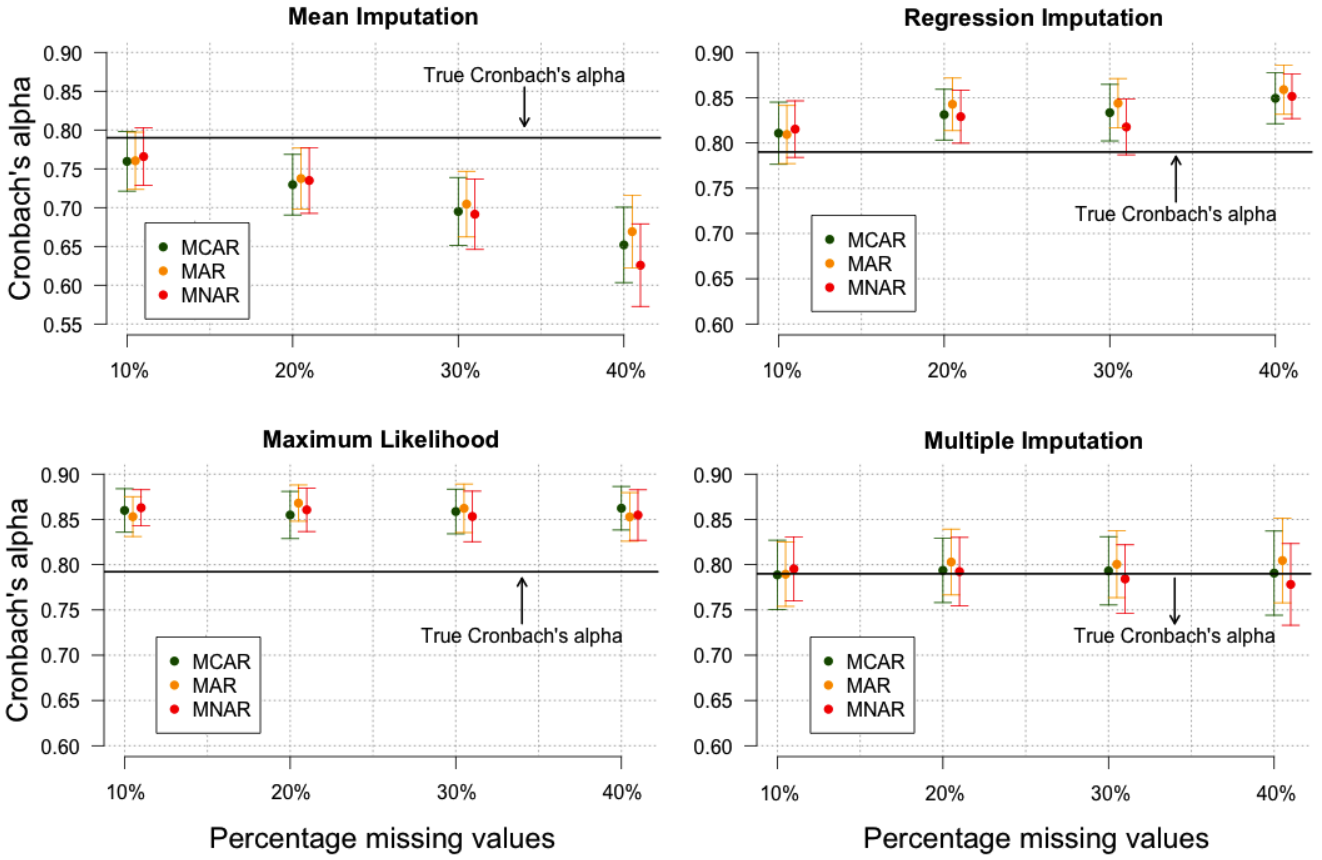

Figure 3. For each of the four modern missing data treatments (Mean imputation, Regression imputation, Maximum Likelihood, and Multiple imputation): the Estimated Cronbach Alpha's (with 95% CI) for four different percentages of missing data (10%, 20%, 30%, 40%) under three different missing data mechanisms (MCAR, MAR, MNAR)

dataset, together with a confidence interval based on the standard error of Cronbach alpha. Figure 3 shows four plots: for each of the four imputation techniques the estimated values of Cronbach's alpha, together with their 95% confidence interval. The x-axis displays the percentages of missing values in the datasets, the y-axis shows the value of Cronbach's alpha and the different line colors indicate under which mechanism the missing data was simulated.

First, the mean imputation plot shows that this imputation technique leads to biased estimates of Cronbach's alpha with increased proportions of imputed missing values. At a low proportion of missing values, the Cronbach's alpha estimate is still unbiased, but as the proportion of missing data increases, the Cronbach's alpha gets underestimated. There seem to be no difference in performance with respect to the different missing data mechanisms. Secondly, the regression imputation plot shows that with a low proportion of missing data, the Cronbach's alpha is estimated without bias. However, as the proportion of missing data increases, the technique leads to an overestimation of Cronbach's alpha. Again, the missing data mechanisms do not seem to influence the performance of this data imputation technique. Third, the maximum likelihood plot shows that this technique always overestimates Cronbach's alpha, independent of the proportion missing values. This implies that, given the present dataset, the maximum likelihood technique yields consistent, yet biased results. The missingness mechanism does not influence the performance of this imputation technique. Fourth, the multiple imputation plot shows that for each proportion of missing data and for each missing data mechanism, the multiple imputation technique is able to provide unbiased estimates of Cronbach's alpha.

We conclude from our simulations that the multiple imputation technique is most effective in treating the missing data. The proportion and mechanism of missingness does not seem to influence its effectiveness. The other three missing data techniques all provide biased estimates of Cronbach's alpha, especially at higher proportions of missing data.

Advice to researchers

The most important advise we can give to researchers is to prevent the occurrence of missing values in their dataset. As already said, we encourage researchers to think about situations that potentially cause missingness and then try to avoid those situations. If it cannot be avoided then researchers should measure information about the situations that most likely cause the missingness to occur. In this way, researchers can prepare themselves for missing data analysis, because they will have more information to treat the missing data with.

If the final dataset contains missing values, then researchers should first try to find out what the type of missing data is they are dealing with. How much missing data is missing? Is the data missing on item or on unit level? Is the missing value an item within a questionnaire, or is it a variable or construct in itself? Is the mechanism that is causing the data to be missing completely at random? Or can we explain the missingness with other variables in our dataset?

Little and Rubin (1989) showed how we can find out whether the mechanism behind the missing data is MCAR. To do this, we can split the data in two groups. One group contains respondents without missing values on a particular variable or item, while the other group contains respondents who do have missing values. Subsequently, we compare both groups on different characteristics (e.g. gender, age, education, nationality) and look whether those characteristics significantly differ between groups. If this is the case, then we conclude that the data is not missing completely at random. A second technique to find out if the missing data is completely random is to create a new variable that indicates for each respondent whether data is missing or not. We can then perform a logistic regression with the missingness variable as dependent variable and individual characteristics such as age as predictors. If one or more predictors has a significant effect then we can conclude the data to be not missing completely at random.

An important question in the treatment of missing values is which technique to use to achieve the most optimal results. The effectiveness of different techniques depends on multiple factors: the amount of missing values in the data, the distribution of the data, the correlation structure of the data, the randomness of the missing values and the possible mechanisms underlying the missingness. It is not always possible to get insight into all these different factors. Therefore, researchers should try to gather as much information as possible about them and use the present article to make an informed choice. Researchers are also referred to other excellent reviews on missing data treatments (Schafer & Graham, 2002; Tsikriktsis, 2005; Donders et al., 2006). Finally, if they are not able to choose a method themselves, then researchers are encouraged to bring their data to a methodological consultant who can fill the missing data they need to make an informed choice.

```
### Missing Data Simulation ###

# Missing data percentages
m=c(1/10,2/10,3/10,4/10)

# MAR probability for each gender
gp[males,]<-rep(0.4,(ni-1))
gp[females,]<-rep(0.6,(ni-1))

# MNAR probability for each item
pmiss3=c(0.7,1,1.3,0.7,1.1,0.9,1.5,1,1.7,1,1,
         1,0.3,1,0.3,0.7,0.5,0.5,1.6,0.8,1.5)


### Missing Data Treatments ###

# Mean Imputation
impute.mean <- function(x) replace(x, is.na(x), mean(x, na.rm = TRUE))
for(i in 1:(dim(dataset)[2])){
  dataset[,i]<-impute.mean(dataset[,i])}

# Regression Imputation
library(mice)
mice(dataset, m=1, method="norm.predict")

# Maximum Likelihood
library(lavaan)
cfa(model, dataset, missing = "ML")

# Multiple Imputation
library(mice)
mice(dataset, m=5, method="norm")
```

Figure 4. Appendix A: Most important R-codes used in simulation study

References

Ader, H. J., Mellenbergh, G. J., & Hand, D. J. (2008). *Advising on research methods: a consultant æĎĂ companion*. Johannes van Kessel Publ.

Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, *34*(3), 277–313.

Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087–1091.

Efron, B., & Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans* (Vol. 38). SIAM.

Enders, C. K., & Peugh, J. L. (2004). Using an em covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling*, *11*(1), 1–19.

Fay, R. E. (1992). When are inferences from multiple imputation valid. In *Proceedings of the survey research methods section of the american statistical association* (Vol. 81, pp. 227–32).

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549–576.

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. *Handbook of psychology*.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, *78*(1), 119.

Graham, J. W., Hofer, S. M., Donaldson, S. I., MacKinnon, D. P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. *The science of prevention: Methodological advances from alcohol and substance abuse research*, 325–366.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*(2), 197–218.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. *NIDA research monograph*, *142*, 13–13.

Liou, M. (1998). Establishing score comparability in heterogeneous populations. *Statistica Sinica*, *8*(3), 669–690.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*(2-3), 292–326.

Raaijmakers, Q. A. (1999). Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement*, *59*(5), 725–748.

Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation & the health professions*, *9*(4), 395–420.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in om survey research. *Journal of Operations Management*, *24*(1), 53–62.

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research*, *42*(2), 387–414.

Weiner, I. B., Schinka, W. A., & Velicer, W. F. (2012). *Handbook of psychology, research methods in psychology*. Wiley.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data.