

*A version of this paper appears in Umbach, Paul D. (Ed.) (2005). Survey research. Emerging issues. New directions for institutional research #127. (Chapter 3, pp. 33-50). San Francisco: Jossey-Bass.*

## Missing Data and Institutional Research

*Robert G. Croninger, Karen M. Douglas<sup>1</sup>*

Missing data are an unwanted reality in most forms of social science research, including institutional research. Like troublesome "guests" at a family gathering, missing data are a nuisance; they impose themselves on research designs and undermine the methodological assumptions of an analysis plan. The primary problems associated with missing data are the threats that they pose to a study's internal validity (primarily issues of statistical power) and external validity (being able to generalize results to a target population). Even when investigators employ appropriate strategies for coping with missing data, different approaches may lead to substantially different conclusions (Cohen, Cohen, West, and Aiken, 2003).

In this chapter we focus on one type of missing data - item non-response. Item non-response, as opposed to participant or unit non-response, occurs when only partial data are available for survey participants or subjects. Item non-response can occur for a multitude of reasons. Mistakes can be made in coding or data entry, respondents may fail or be unable to answer a set of survey items, or the study design may purposefully call for respondents to skip a section of items or answer different samples of items from a broader battery of survey questions. Regardless of how missing data occur, investigators must select one or more strategies for coping with missing data and consider how the use of these strategies may affect their results.

We discuss various types of missing data and describe a range of strategies that investigators can use to address them. Next we present a hypothetical institutional dataset with known forms of missing data, discuss ways of exploring and identifying types of missing data, and examine how different strategies for coping with missing data influence the estimation of parameters (such as means, coefficients, and error terms). Among the strategies that we consider are listwise deletion of cases, pairwise deletion, mean plugging, estimation of conditional means, imputation using the EM algorithm, and multiple imputations. We do not discuss three other approaches to missing data: reweighting of cases, hot deck imputation, and regression-based imputation. For the interested reader, Cohen, Cohen, West, and Aiken (2003) provide an overview of these techniques. We conclude with a series of recommendations for investigators about how to prevent missing data and cope with the occurrence of missing data in institutional studies.

### **Types of Missing Data**

All forms of missing data can be problematic, but some forms are more problematic than others. Three types of missing data are delineated in the literature on missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (NMAR). When missing data are MCAR, there is no discernable pattern to the missing values. The missing values for a variable, say  $X_1$ , are neither related to the variable itself nor related to any other variables in the analytic model ( $X_n$ ). Under this assumption, cases with valid values for  $X_1$  still have a reasonable probability of being a representative subset of values in the intended sample (and thus the original population). We say "reasonable" because even random samples can deviate from population parameters. Of course, even if missing data are MCAR, they may still pose a threat to the internal validity of a study. If the amount of missing data is substantial, the resulting loss of cases may considerably

weaken statistical power, increasing the likelihood of accepting the null hypothesis when it is actually false (committing a Type II error).

A somewhat weaker assumption is that data are missing at random (MAR). Under MAR, missingness on  $X_1$  is related to another variable ( $X_n$ ) in the analysis, but not to  $X_1$  itself. For example, suppose that we have a question about annual income and female respondents fail to answer this question at a rate higher than male respondents. So long as those male and female respondents who provided data are representative of the annual income earned by males and females in the intended sample, missing data can be described as missing at random. Under these conditions the inclusion of gender in a model provides accurate estimates of male and female income (though the overall population mean for income may be biased upward or downward depending on the difference between males and females in annual income). As with MCAR, if a substantial number of female respondents failed to provide income data, missing data may still weaken a study's statistical power, particularly regarding the effects of gender (or the relevant  $X_n$ ) on income.

A third type of missing data is missing not at random (NMAR). When data are NMAR, missing data are related to *both* the values of  $X_1$  and the values of one or more other variables ( $X_n$ ) in the analytic model. In the example of income and gender just discussed, we would characterize missing data as NMAR if female respondents failed to provide information about their annual income at a higher rate than male respondents *and* if those female respondents most likely to withhold information were females with the lowest annual incomes. Estimates of income for females will be biased upward, affecting not only estimates of income for females but estimates of income for the entire population. Moreover, as with MCAR and MAR, NMAR may also threaten a study's internal validity if the number of females who fail to provide information substantially reduces a study's statistical power.

MCAR and MAR are sometimes referred to as "ignorable" missing data, not because the investigator needs to do nothing about partial data but because there are known techniques for addressing these types of missing data. Schafer (1997) points out that the assumption of ignorability is more easily met when the dataset is rich and the analytic model is sufficiently complex because these conditions create opportunities to utilize a wider range of missing data strategies. NMAR, by contrast, is more problematic because it poses a clear threat to a study's external validity with no clear mechanism for addressing potential bias (such as including gender in an analytic model). Additional information about the sources or causes of missing data may facilitate an investigator's ability to address potential biases posed by NMAR (for example, knowing that low achieving women are less likely to report income than high achieving women). Unfortunately, investigators rarely know the sources or causes of missing data in a study.

Although MCAR, MAR, and NMAR provide useful distinctions for talking about missing data, it is very difficult, and perhaps even impossible, to determine the types of missing data that exist in a dataset or an analytical model. Because most, if not all, of the population parameters for variables are unknown in a study (we do studies to estimate these parameters), it is impossible to determine with certainty if missing data are MCAR, MAR, or NMAR. Furthermore, delineation of types of missing data is difficult with a large number of variables. Large-scale studies probably include examples of all three types of missing data, the consequences of which depend on the amount of missing data, the size of the sample, the specification of an analytical model, and the patterns of missing data associated with specific variables. When the amount of missing data is relatively small and the sample size is relatively large, determining which type of missing data exists is less important - most strategies for coping with missing data will provide the same results, regardless of whether missing data are MCAR, MAR,

or NMAR. However, as the amount of missing data increases and the sample size decreases, the choice of how to address missing data is potentially more consequential.

### **Strategies for Coping with Missing Data**

Most strategies for coping with missing data assume either MCAR or MAR. A few strategies may work fairly well, even when data are NMAR if other variables provide sufficient information about missingness to estimate underlying relationships. Although both MCAR and MAR require strong assumptions about missing data, these assumptions may be reasonable for many studies. For example, research designs may purposefully collect specific information from a random subset of survey respondents and then use the information to generalize to other respondents in a study. Or the design may call for each respondent to answer a random subset of items, which may then be used to estimate a set of plausible values for respondents across all items. Even when missing data are not the result of research design, it may be reasonable to assume that missing data are at least MAR. Studies of item non-response on surveys suggest that respondents fail to provide information for a wide range of reasons (Krosnick, 2000), and these reasons, when taken together, may create "noise" but not substantial bias in a study.

We divide the strategies for coping with missing data into two broad categories: (1) traditional strategies appropriate with MCAR or MAR and (2) emergent strategies that are appropriate for MCAR, MAR, and sometimes even NMAR. The traditional strategies that we consider include listwise deletion, pairwise deletion, single imputations (such as mean plugging and estimation of conditional means), while emergent strategies include imputation using the EM algorithm and multiple imputations. These newer strategies, though computationally more complex, have the advantage of handling multiple types of missing data reasonably well. Even when missing data are NMAR, these

newer strategies for coping with missing data may still provide a satisfactory solution to the potential bias that NMAR introduces into a study.

**Traditional Strategies.** Careful inspection of missing data patterns for a set of variables may reveal that an exclusion strategy is an appropriate approach to addressing the absence of information for some cases. If the majority of missing data is exclusive to a variable of debatable importance or a small set of cases in the sample, investigators may decide to drop the variable or drop the cases from the study. Although loss of information is never the optimal solution to missing data, if the number of variables and cases involved is relatively small, say less than three percent, and the sample size exceeds 200 (Cohen, Cohen, West, and Aiken, 2003), excluding variables and cases from a study is often a prudent response to the problems posed by missing data. Even if missing data remain after exclusion, the remaining missingness may be addressed by other missing data strategies.

*Listwise Deletion.* If missing values can be characterized as MCAR or MAR, then listwise deletion of values is an appropriate strategy (Allison, 2002). Listwise deletion (a more extreme case of exclusion) involves deleting any case from the sample that has missing values for variables that an investigator intends to use in a study. The resulting data are complete in that every case provides full information for the analysis. When missing values are MCAR, the resulting dataset has a high probability of being a representative subsample of the larger or full dataset. Analyses done with the reduced dataset should provide unbiased estimates of any parameters of interest, and the standard errors, though larger due to the reduction in sample size, are, nonetheless, appropriate for the reduced dataset.

Even when missing data deviate from the assumption of MCAR, listwise deletion may yield reasonable estimates of population parameters. In the

case of MAR, model specification can often be used to successfully address any potential bias in the restricted sample. Providing that the variables associated with missing data can be identified, and "missingness" is not related to the dependent variable, MAR can be thought of as a special case of stratified sampling - that is, specific groups are either under- or over-represented in the dataset. If these groups are included in the analytic model, it is possible to derive accurate regression coefficients and preserve the validity of a study's findings.

Although deletion is (a) easy to do, and (b) appropriate for any type of statistical analysis, its primary drawback is the possibility of depleting the statistical power of a study. Moreover, even under the best of conditions, listwise deletion may yield unsatisfactory results as the amount of missing data increases, increasing the likelihood that the reduced sample will deviate from the full sample. Alternative approaches to coping with missing data are responses to the fundamental recognition that loss of data always threatens the internal and external validity of a study. Moreover, as we argued above, it is impossible to know for certain if missing data are MCAR, MAR, or NMAR, and most studies probably include a mix of all three types of missing data. To address possible loss of statistical power and information, investigators have developed alternative techniques of coping with missing data that utilize information about the relationship between variables so as to preserve a greater proportion of cases.

*Pairwise Deletion.* As with listwise deletion, if data are missing MCAR, pairwise deletion yields unbiased estimates of population parameters and covariance structures for variables in an analytic model. Unlike listwise deletion, though, pairwise deletion utilizes all of the available data in an analysis, preserving information and the statistical power associated with a study. When data are present for any two variables, pairwise deletion utilizes this information to construct means, standard

deviations, and a correlation matrix for all of the variables of interest in a study. Each statistic is an unbiased estimate of the population parameter for the full dataset, even though the number of cases for each statistic may vary substantially. Once these summary measures are computed, they can then be used as input for appropriate statistical procedures, such as ordinary least squares regression.

While pairwise deletion provides some advantages over listwise deletion, it also brings with it some disadvantages. A major disadvantage is that there is no clear way to unambiguously compute the standard errors, since errors vary with the sample size on which the summary statistics are based. The actual sample size ( $n$ ) is somewhere between the minimum and the maximum number of cases for variables so estimates of error are ambiguous. Another disadvantage is that pairwise deletion does not always result in a "positive definite" covariance structure (Allison, 2002). Because the maximum range of covariance varies across pairs of variables, it is possible for the coefficient of determination ( $R^2$ ) to exceed one (or 100 percent). The possibility of nonsensical covariance structures increases as patterns of "missingness" vary across variables in an analytic model. The more the sample sizes vary across pairs of variables, the greater the likelihood that pairwise deletion will fail to produce a "positive definite" covariance structure (an occurrence that is not always clearly identified by statistical software). For this reason, investigators advise caution in the use of pairwise deletion of missing data in analyses.

*Single Imputation.* An alternative to listwise and pairwise deletion is to use available data to estimate a value for cases with missing data. This strategy preserves a larger number of cases and creates a common sample size for all variables in the analysis. The simplest strategy is to impute the value of a missing case as being the mean for that variable. If all of the missing values are handled in this way, the treatment of missing values can

be thought of as orthogonal across the variables in a study. An alternative is to estimate conditional values for missing data based on relationships with other variables in a study. The investigator might create a matrix of variables that includes race, gender, and high and low values on some continuous measures of interest. The mean values in a cell (for example, the average earnings of white females with 16 or more years of education) can then be substituted for relevant cases with missing data. By using more information in the dataset, the investigator increases the possibility of accurately estimating the value for missing data. A shortcoming of the conditional mean approach, however, is that estimations with partial information may not be a substantial improvement over mean plugging.

One problem with either approach to imputation is that the standard error associated with a variable is likely to be too small. This is obviously true when the grand mean is used to estimate values, but it is also true when more complex strategies are used to estimate a value. For each strategy, the amount of information that we have for an analytic model is not equivalent to the number of cases that we have in a dataset. To address this, investigators have recommended using a dummy-coded variable for imputed data in analytic models (1 = imputed, 0 = not). By doing so, investigators partition the effect associated with a variable between observed and imputed values, as well as provide a direct test of whether observed cases differ from imputed cases in their effects. This strategy can be easily extended to multiple variables with imputed values for missing data.

The primary advantages of single imputation strategies are that they preserve more of the data and it is possible to test directly whether imputed and observed values for variables differ in their relationship to a dependent variable. If missing data are MCAR, these procedures typically yield unbiased estimates of population parameters and preserve a study's validity; if they are not, estimates may be biased, even under conditions of MAR. If

the investigator does not know how missingness is related to other variables in the study, the imputed values may also be biased. Moreover, the standard errors associated with imputed values are always underestimated and the corresponding test statistics are always over estimated (Allison, 2002). Without incorporating uncertainty about the actual value of missing data, imputation strategies always result in standard errors that are smaller than should be expected by chance.

**Emergent Strategies.** Recent developments in statistical methodology, coupled with the rapidly expanding capacity of computers to perform complex mathematical tasks, provide two new techniques for addressing missing data: the EM algorithm and multiple imputation strategies. The EM algorithm and multiple imputation are currently available in a number of statistical analysis packages and specialized programs; some of these programs are available free of charge from their authors.

*Imputation with EM Algorithm.* "EM" stands for "Expectation-Maximization," and although it is a mathematically complex approach, the basic logic is as follows. The EM approach asks the question, "What parameter estimates are most likely given the data that were observed?" In answering this question, the parameters of interest are estimated from the available data. Then these parameters are used to estimate possible values for the data that are missing. Next, the parameters are re-estimated using the available data and estimated values. Then the missing values are re-estimated based on the new parameters. This iterative process continues until the difference between estimates of parameters from one cycle to the next is deemed inconsequential.

The EM approach takes advantage of the relationships among variables in the dataset to form "best guesses" for missing values. It is typically designed to maximize estimation of summary statistics, however, and not the values for individual cases. Its strength is in estimating summary

statistics and the covariance structure in the data. Although estimations with the EM algorithm often provide substantial advantages over traditional approaches, the EM algorithm still does not take into consideration that the estimations are of unknown precision - that is, that estimations represent one possible solution in an uncertain range of possibilities. Multiple imputation strategies address this shortcoming.

*Multiple Imputation.* In multiple imputation approaches, a number of values are estimated for each missing value. The end result is a number of datasets in which the available data are all the same, but the missing values are replaced with different plausible values in each dataset. The technique used to generate these estimates for missing values is a simulation-based approach called Markov Chain Monte Carlo (MCMC) estimation. Stated very simplistically, in the multiple imputation approach, a distribution of plausible values is estimated for each missing data point, and then in each imputation one value is randomly selected from this distribution. The researcher creates a number of imputed datasets, and then performs analyses of interest on each dataset. Parameter estimates can then be combined across each of these analyses to provide better estimates and a picture of the variability of these estimates among the various imputed datasets. Although the number of imputed datasets is determined by the researcher, as few as five imputed datasets will frequently provide satisfactory estimates (Schafer, 1997). The primary advantage of multiple imputations is that it incorporates uncertainty found in observed data when imputing values. More technical explanations of these newer strategies, along with a comparison with traditional approaches, is found in Little and Rubin (2002), Schafer (1997), and Schafer and Graham (2002).

### **An Example**

To examine the possible consequences of using different strategies for coping with missing data, we constructed a dataset of 1,000 cases using variables from an alumni survey conducted by a community college. We generated a hypothetical dataset using similar variables; therefore, the substantive findings reported here are purely fictional and presented merely to illustrate different missing data techniques. The primary analytic question is whether there are differences in alumni satisfaction with their educational experiences based on an alumnus' reason for enrollment (Academic = 1 if enrolled in pursuit of an academic degree, other reasons = 0), as well as an alumnus' GPA, age, and gender (Male = 1 if male, 0 otherwise). We standardized the three continuous measures around their means and standard deviations ( $M = 0$ ;  $SD = 1$ ). The proportional representation for respondents who said that their reason for attending had been to pursue an academic degree and who identified themselves as male was 52 percent and 56 percent respectively. The correlations among variables are shown in Table 1.

\*\*\*\*\* Insert Table 1 Here \*\*\*\*\*

Our original 1,000 cases included complete data for all variables. We next set about creating a second dataset that contained various types and patterns of missing data. We preserved all cases for our intended dependent variable, Satisfaction. We could have created missing data for the dependent variable, as well, but decided against it since it is more common in institutional research to simply delete any case that does not have a value for the dependent variable. Nonetheless, investigators should recognize that missing data strategies other than listwise deletion can be used successfully with dependent variables as well. We randomly deleted information about gender for 50 cases (5 percent); these missing data can be characterized as MCAR. We next deleted information about the reason for enrollment for 300 of the cases for alumni who said that their intent was to pursue an academic degree (30 percent). Cases with missing values on this

variable were higher on GPA and Satisfaction, and more likely to be female. Therefore, the Academic variable in our example is NMAR. Next we deleted the value for Age for the 200 oldest alumni (20 percent), another example of NMAR. Finally, we deleted 250 values of GPA for cases with the lowest values on Satisfaction (25 percent). This is an example of MAR because missingness is dependent on another variable (Satisfaction), but not on GPA itself.

The resulting dataset includes a substantial amount of missing data and provides a somewhat rigorous test for traditional and emergent strategies for coping with missing data. Table 2 shows the number of cases exhibiting each pattern of missing values for the variables in the example. The majority of cases have values for at least three variables, and missing values form what Schafer and Graham (2002) refer to as an "arbitrary" pattern of missing values (in contrast to variables missing for a block of cases, as would occur when there are skip patterns on a survey). Although no single variable is missing more than 25 percent of its cases, the reduced dataset has complete information for only about one third (371 cases) of the full dataset. While we do not think that this is an uncommon occurrence in institutional research involving multiple variables with varying amounts of missing data, we do believe that this scope of missing data represents a considerable challenge for an investigator. Moreover, missing values for the variables Academic and Age are missing systematically -- all missing values on Academic are those who enrolled for academic reasons, and all missing values on Age are the oldest alumni. Missingness on GPA is strongly related to Satisfaction, given that the values that we deleted included all of the least satisfied alumni. These patterns of missingness are fairly extreme. A more realistic assumption for NMAR would be that missing data for these variables are biased but less extremely so than represented here. We focus on these more extreme conditions because we believe that they have a greater likelihood of

revealing how well different approaches to missing data estimate a known set of parameters.

\*\*\*\*\* Insert Table 2 Here \*\*\*\*\*

**Baseline Estimates.** We used the full dataset to construct a baseline set of estimates, or "truth," about whether there were differences in alumni satisfaction based on an alumnus' reason for enrolling at the college, GPA, age, and gender. Table 3 presents the means and standard deviations for variables, the regression coefficients, and the corresponding 95 percent confidence intervals around the regression coefficients. The means for the variables are the same as those reported above. The confidence intervals for these univariate statistics are .12 SD for the three standardized, continuous variables and .06 SD for the two dummy-coded variables. The reported confidence intervals represent a range of possible parameter estimates approximately two standard deviations (plus or minus) from the baseline parameters.

\*\*\*\*\* Insert Table 3 Here \*\*\*\*\*

The regression results are presented in the bottom half of the table, and represent the effect sizes of each variable. Because the dependent variable is standardized on its mean and standard deviation, the regression coefficients are in standard deviation (SD) units. The effect sizes represent a proportion of a SD change in satisfaction for every unit change in the independent variables. Effect sizes provide a useful way of comparing and characterizing the magnitude of changes in dependent variables associated with independent variables specified in one or more models. For an explanation and descriptions of standard deviation units and the use of effect sizes in social science research, see Rosenthal, Rosnow, and Rubin (1999). Bolded coefficients in italic are statistically significant at  $p < .05$  or lower. Our analytic model, based on our hypothetical data, indicates

that there are differences between alumni in their satisfaction with their educational experiences. The least satisfied alumni are males (- .53 SD) and alumni who said that they enrolled in the college to pursue an academic degree (-.41 SD). Alumni with GPAs one standard deviation above the mean report higher levels of satisfaction (.40 SD higher), whereas Age is unrelated to the level of satisfaction reported by alumni. The range for the 95 percent confidence intervals are similar for the intercept (.17 SD), as well as the coefficients for GPA (.20 SD), Academic (.20 SD), and Male (.21 SD). The range for the 95 percent confidence interval for Age is smaller at .10 SD.

We next ran the same analysis on the dataset with partial information using different strategies for addressing missing data. Listwise, pairwise, mean plugging and conditional means were computed using SPSS; the estimates for the EM approach and multiple imputations were computed using NORM.<sup>2</sup> We used two criteria for judging the validity of our parameter estimates under different missing data techniques - *decision accuracy* and *parameter precision*. We defined decision accuracy, specifically with regards to the regression coefficients, as the extent to which we would reach the same substantive conclusions that we reached for the analysis we conducted with complete data. Used in this way, judgments of accuracy reflect whether findings of statistically significant relationships among variables would change<sup>3</sup> with the use of different missing data strategies. We examined the sign for the regression coefficients and the results of the hypothesis tests to evaluate each missing data strategy for this criterion. We defined parameter precision to be the extent to which the missing data strategies provide parameter estimates that fall within the 95 percent confidence intervals<sup>4</sup> for the same estimates using complete data. Parameter precision, then, is concerned with how well various missing data strategies estimate the value of the mean and regression coefficients.

**Comparison of Strategies.** Table 4 compares univariate statistics and regression coefficients for the six different strategies for coping with missing data (listwise deletion, piecewise deletion, mean plugging, conditional means, imputations with the EM algorithm, and multiple imputations). The first column of figures in Table 4 shows the baseline parameter estimates that would be obtained using the complete dataset (see Table 3). As in Table 3, bolded coefficients in italic are statistically significant at  $p < .05$  or lower. Shaded cells represent means or regression coefficients that fall outside of the 95 percent confidence interval established by the parameter estimates for the complete dataset (see Table 3).

\*\*\*\*\* Insert Table 4 Here \*\*\*\*\*

*Decision Accuracy.* All six missing data strategies yield the same substantive conclusions as those we made for the regression analysis based on complete data. In each analysis, we would conclude that female alumni, alumni with higher GPAs, and alumni who enrolled for non-academic reasons are more satisfied with their educational experiences than male alumni, alumni with lower GPAs, and alumni who enrolled for academic reasons. In none of the analyses would we conclude that alumni satisfaction varies with alumni age - older alumni in all six analyses are no more (or less) likely to be satisfied than younger alumni. Even when relatively large amounts of data are missing, and some missing data are NMAR (as in this example), available strategies for coping with missing data can yield consistent regression results. The missing data strategies traditionally used by institutional investigators are relatively robust if the resulting sample size is modestly large and the standard errors are relatively small compared to the regression coefficients (as is the case in this example). The smallest statistically significant coefficient in our example is .40 SD and none of the standard errors is greater than .05. If some of the relationships in our analytic

model had been borderline significant, there might have been greater divergence in accuracy between strategies.

*Parameter Precision.* If we consider the univariate statistics first, the only parameter consistently estimated with precision by all strategies is the mean for Male, the variable for which missing data are MCAR. All of the strategies estimate the mean for Satisfaction with precision except listwise deletion. However, since Satisfaction has no missing data, this is not much of an accomplishment. The EM algorithm and multiple imputations succeed in estimating with precision the mean for GPA (created to be MAR), but not for Academic and Age (both NMAR). The largest deviations from the "true" mean hover around  $-.33$  SD for Age, regardless of the missing data strategy employed, while the smallest deviations is  $.13$  SD for GPA when we used listwise deletion. Clearly, when missing data introduce bias either directly or indirectly into a dataset, it is difficult to estimate univariate statistics with precision without knowing why and how data are missing.

The bottom half of the table provides estimates of precision for the regression results. Although each of the missing data strategies provides accurate parameter estimates for the regression analysis, the estimates frequently fall outside of the 95 percent confidence intervals for the full data. The parameter estimated with the greatest precision is the coefficient for Age, and this is largely because Age is not related to the dependent variable. Nonetheless, there are some noticeable differences in the precision of parameter estimates yielded by the different approaches. Both the estimates based on the EM algorithm and multiple imputations are more precise than the estimates based on traditional approaches to missing data. All of the estimates using multiple imputations fall within the 95 percent confidence interval for the full data, and only the estimate for the intercept diverged from the 95 percent confidence interval for the analysis that used imputations based on the EM algorithm. Although the assumptions

about missing data for the EM algorithm and multiple imputations are the same as traditional approaches, these strategies yielded more precise parameter estimates, even for variables we specified as NMAR.

The traditional strategies performed noticeably less well on this criterion. Of the four approaches that we examined, only pairwise deletion provided precise estimates for a majority of the parameters (3 out of 5). Mean plugging provided precise estimates for two out of five parameters, conditional means provided precise estimates for one out of five parameters, and listwise deletion provided precise estimates for none of the parameters. Although not included in Table 4, we introduced a series of dummy-coded variables indicating whether missing data had been imputed through mean plugging or conditional means into the model to see if this would improve the precision of estimates. The dummy-coded variables for missing data were statistically significant for GPA, Academic, and Age, but not statistically significant for Male. While the dummy-coded variables improved the precision of the intercept, it did not improve the precision for any of the other coefficients. The precision for Academic actually became worse, exceeding the 95 percent confidence interval, when mean plugging with imputed data indicators was the strategy used to address missing data.

These findings are based on a single, hypothetical dataset. A more thorough investigation of the consequences of employing different missing data strategies requires replication of findings across multiple datasets with different specifications for missing data, sample size, analytic models, and baseline relationships. Changes in any of these specifications could alter, perhaps substantially, these findings. To our knowledge, there is very little literature that would allow us to speculate about the robustness of these findings, particularly how they might vary across different datasets and alternative specifications. Nonetheless, our findings are consistent with those reported by Cohen, Cohen, West, and Aiken (2003), who found only

small differences between traditional techniques for addressing missing data in terms of substantive conclusions (accuracy), and Schafer and Graham (2002), who used simulations to demonstrate the advantages (primarily precision) of the EM algorithm and multiple imputations compared to more traditional missing data techniques. With this in mind, we offer a few recommendations to institutional researchers faced with the problem of missing data.

### **Recommendations for Coping with Missing Data**

What can institutional researchers do to cope with missing data? First, take the methodological problems associated with missing data seriously and plan for its occurrence in research designs. While the strategies that we have examined focus on coping with missing data *after* they occur, some of the most successful strategies may be those employed by investigators *before* missing data occur. A substantial amount of research has been done on methods for improving the quality of survey data (see Groves, Couper, Lepkowski, Singer, and Tourangeau, 2004). These studies have sought to understand how such things as sample design, item construction, item placement, and survey follow-up procedures influence the quality of data available to investigators, including biases introduced by missing data. Careful attention to the details of research design can limit not only the amount of item non-response that investigators must address when analyzing data but also the amount of unit non-response in a dataset (an even more serious problem for some forms of institutional research).

Second, examine closely missing data when it occurs and seek to understand its patterns and possible causes. Although we remain skeptical about the ability of investigators to determine precisely the types and mix of missing data that occur in an analytic model or institutional dataset, examinations of missing data patterns can provide useful insights about the

challenges posed by missing data, the consequences of specific strategies (such as listwise deletion), and sometimes even the sources of missing data. Randomness of missing data can be studied by examining the shape of univariate distributions, correlating missing value indicators, and comparing the means of variables for cases with and without missing data (see Hair, Tatham, Anderson, and Black (1998) for a discussion of exploratory techniques with missing data). Minimally, an investigator should determine the extent to which available data for a study diverges from the larger institutional dataset so as to determine and report the likely parameters for external validity.

Investigators may also find it helpful to examine carefully the patterns of missing data in an entire dataset (not just those variables that are candidates for a particular study). If investigators discover persistent patterns in missing data (for example, particular types of survey items that have large amounts of non-response or sources of institutional data that consistently under report valuable information), steps can be taken to address these problems. Surveys can be altered or redesigned to limit missing data; data monitoring procedures can be implemented to enhance the quality of data and address lapses in data collection before they create fatal flaws in an institutional dataset. Investigators may also find it beneficial to design small studies that target specific areas of desirable information plagued by missing data issues. Focus groups with potential respondents and representatives of organizations that report data can provide useful insights into how to improve data collection, redesign surveys, and reduce missing data.

Finally, investigators should familiarize themselves and become proficient with the full range of missing data strategies that can be utilized in institutional research. With regards to these strategies, our examination of approaches to handling missing data provides both good and bad

news for investigators. Although we caution that our findings are based on a single simulated dataset, the good news is that when missing data disrupt your research plans, there are relatively good strategies for dealing with them. If the goal is to provide accurate regression results, available techniques for coping with missing data often provide valid coefficients, providing the sample size is relatively large and the ratio of the "true" standard error to the "true" parameter estimate is relatively small. Even when the variables of interest include a mix of missing data types, traditional techniques often provide substantive conclusions that are not dependent on the type of strategy used to address incomplete data. Nonetheless, we do not claim that these strategies will always and under any circumstances yield the same substantive results or that the treatment of missing data is inconsequential.

On the contrary, if the goal is to provide precise estimates of parameters, the news is not so good. Traditional approaches to addressing missing data provide imprecise means, intercepts and regression coefficients. In our opinion, this is not a trivial matter. Despite the relative accuracy of the different approaches, the absence of precision may lead investigators to different conclusions about population parameters and even the relative importance of variables in an analytical model. In the example that we provided, the "true" estimate for the influence of GPA on satisfaction is .40 SD (a moderately strong influence); however, when using traditional approaches for coping with missing data, the parameter estimate is as low as .21 SD (mean plugging) and never exceeds .31 SD (pairwise deletion). Similarly, the parameter estimates for Academic ranges from a low of  $-.29$  SD (listwise deletion) to a high of  $-.57$  SD (conditional means). Such differences are not trivial, even if investigators arrive at the same conclusions regarding hypothesis testing and the direction or relationships.

Fortunately, newer strategies for coping with missing data yield not only accurate but more precise parameter estimates than traditional strategies. Even though these newer approaches are computationally complex, they are relatively robust to multiple types of missing data, even some patterns of missing data that violate assumptions of MCAR and MAR. Although we do not think that traditional strategies for addressing missing data should be abandoned by investigators, especially since these strategies often provide an efficient and relatively straightforward approach to addressing partial information, we strongly encourage investigators to expand their missing data repertoire to include these newer techniques. Imputations with the EM algorithm and multiple imputations provide robust estimates of parameters for multiple types of missing data. By using these techniques appropriately, institutional researchers can provide policymakers with not only accurate but more precise estimates of many parameters of interest.

Robert G. Croninger is Associate Professor in the Department of Educational Policy and Leadership at the University of Maryland.

Karen M. Douglas is a Doctoral Candidate in the Department of Measurement, Statistics and Evaluation at the University of Maryland.

## References

- Allison, P. *Missing Data*. Thousand Oaks, CA: Sage, 2002.
- Cohen, J., Cohen, P., West, S., and Aiken, L. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3<sup>rd</sup> edition). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 2003.
- Groves, R., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. *Survey Methodology*. Hoboken, NJ: John Wiley & Sons, 2004.
- Hair, J., Anderson, R., Tatham, R., and Black, W. *Multivariate Data Analysis* (5<sup>th</sup> edition). Upper Saddle River, NJ: Prentice Hall, 1998.
- Krosnick, J. "Survey Research." *Annual Review of Psychology*, 2000, 50, 537-567.
- Little, R., and Rubin, D. *Statistical Analysis with Missing Data* (2<sup>nd</sup> edition). Hoboken, NJ: John Wiley & Sons, Inc, 2002.
- Rosenthal, R., Rosnow, R., and Rubin, D. *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. New York: Cambridge University Press, 1999.
- Schafer, J. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC, 1997.
- Schafer, J., and Graham, J. "Missing Data: Our View of the State of the Art." *Psychological Methods*, 2002, 7(2), 147-177.
- von Hippel, P. "Biases in SPSS 12.0 Missing Value Analysis." *The American Statistician*, 2004, 58(2), 160-164.

Table 1: Bivariate correlations for dataset with no missing values

	<i>Satisfaction</i>	<i>GPA</i>	<i>Academic</i>	<i>Age</i>	<i>Male</i>
<i>Satisfaction</i>	1.00				
<i>GPA</i>	.50	1.00			
<i>Academic</i>	-.32	-.16	1.00		
<i>Age</i>	.18	.27	-.06	1.00	
<i>Male</i>	-.39	-.21	.20	-.11	1.00

Table 2: Missing data patterns<sup>a</sup>

<i>Number of Cases</i>	<i>Satisfaction</i>	<i>GPA</i>	<i>Academic</i>	<i>Age</i>	<i>Male</i>	<i>Variables Observed</i>
371	1	1	1	1	1	5
183	1	1	0	1	1	4
182	1	0	1	1	1	4
97	1	1	1	0	1	4
57	1	1	0	0	1	3
29	1	0	0	1	1	3
25	1	0	1	0	1	3
17	1	1	0	1	0	3
12	1	1	1	1	0	4
7	1	1	1	0	0	3
6	1	0	0	0	1	2
6	1	1	0	0	0	2
5	1	0	1	1	0	3
1	1	0	0	1	0	2
1	1	0	1	0	0	2
1	1	0	0	0	0	1

<sup>a</sup>A value of "0" indicates that data values are missing; "1" indicates that data are observed .

Table 3: Baseline estimates using complete data ("Truth")

<i>Variables</i>	<i>Parameter<sup>a</sup> Estimate</i>	<i>SD</i>	<i>SE</i>	<i>95 % Confidence Intervals</i>	
				<i>Low</i>	<i>High</i>
<b>Univariate Statistics</b>					
Satisfaction	0.000	1.000	0.032	-0.063	0.063
GPA	0.000	1.000	0.032	-0.063	0.063
Academic	0.520	0.500	0.016	0.488	0.552
Age	0.000	1.000	0.032	-0.063	0.063
Male	0.555	0.497	0.016	0.523	0.587
<b>Regression Results<sup>b</sup></b>					
Intercept	<b><i>0.507</i></b>		0.044	0.421	0.593
GPA	<b><i>0.404</i></b>		0.027	0.352	0.457
Academic	<b><i>-0.410</i></b>		0.051	-0.511	-0.309
Age	0.028		0.026	-0.023	0.079
Male	<b><i>-0.529</i></b>		0.052	-0.631	-0.426

<sup>a</sup> The first five rows provide univariate statistics for the variables while the last five rows provide the intercept and regression coefficients.

<sup>b</sup> Bolded coefficients in italic are statistically significant at the  $p < .05$  or lower.

Table 4: Effects of using different missing data strategies on means and regression coefficients<sup>ab</sup>

<i>Variables</i>	<i>Baseline Estimates<sup>c</sup></i>	<i>Listwise Deletion<sup>d</sup></i>	<i>Pairwise Deletion<sup>e</sup></i>	<i>Mean Plugging<sup>f</sup></i>	<i>Conditional Means</i>	<i>EM Estimates</i>	<i>Multiple Imputations</i>
<b>Univariate Statistics (Means Only)</b>							
Satisfaction	0.000	0.287	0.000	0.000	0.000	0.000	0.000
GPA	0.000	0.125	0.210	0.210	0.176	0.000	0.004
Academic	0.520	0.677	0.743	0.743	0.733	0.724	0.726
Age	0.000	-0.312	-0.348	-0.348	-0.343	-0.332	-0.332
Male	0.555	0.547	0.555	0.555	0.553	0.552	0.554
<b>Regression Results (Coefficients Only)</b>							
Intercept	<i>0.507</i>	<i>0.665</i>	<i>0.639</i>	<i>0.718</i>	<i>0.769</i>	<i>0.621</i>	<i>0.578</i>
GPA	<i>0.404</i>	<i>0.235</i>	<i>0.314</i>	<i>0.213</i>	<i>0.251</i>	<i>0.392</i>	<i>0.398</i>
Academic	<i>-0.410</i>	<i>-0.294</i>	<i>-0.463</i>	<i>-0.481</i>	<i>-0.569</i>	<i>-0.433</i>	<i>-0.376</i>
Age	0.028	-0.056	0.020	0.060	0.054	-0.007	-0.001
Male	<i>-0.529</i>	<i>-0.413</i>	<i>-0.638</i>	<i>-0.694</i>	<i>-0.684</i>	<i>-0.566</i>	<i>-0.560</i>

<sup>a</sup> Shaded parameter estimates are outside of the 95 percent confidence intervals for complete data (see Table 3). <sup>b</sup> Bolded coefficients in italic are statistically significant at the  $p < .05$  or lower. <sup>c</sup> Baseline estimates are based on complete data as shown in Table 3. <sup>d</sup> Listwise deletion has a sample size of 371. <sup>e</sup> Pairwise deletion has a sample size that ranges from 1000 for Satisfaction to 700 for Academic. <sup>f</sup> Mean plugging, conditional means, EM estimates, and multiple imputations have sample sizes of 1,000.

## Endnotes

---

<sup>1</sup> The authors for this chapter contributed equally to its preparation. Special thanks to Michelle Appel, Senior Research Analyst in the Office of Institutional Research and Planning at University of Maryland for information on issues of interest to institutional researchers, and to Gayle Fink, Director-Planning, Research and Evaluation at The Community College of Baltimore County, for sharing the dataset on which our simulated data are based.

<sup>2</sup> NORM is a program to do multiple imputation developed by J. L. Schafer. This software can be downloaded from <http://www.stat.psu.edu/~jls/misoftwa.html>. Attempts to use SPSS for EM estimation provided inconsistent results, a finding further explicated in von Hippel (2004). NORM is well documented but does require a number of additional steps in data preparation and implementation. NORM uses a normal distribution, and although best suited to continuous variables can accommodate dummy-coded variables. Several similar programs are in development for use with categorical variables, mixtures of categorical and continuous variables, datasets with interaction terms and multi-level designs. As with all statistical procedures, the user bears the responsibility for obtaining a basic understanding of the MCMC approach and associated diagnostics.

<sup>3</sup> Obviously, decision accuracy depends on the criterion for rejecting the null hypothesis. The results that we report in this chapter are based on the conventional criterion of  $p < .05$ .

<sup>4</sup> The means and coefficients that we report for each missing data technique are estimates based on a different sample. The 95 percent confidence interval represents the range of means and coefficients that we would expect in 95 percent of the samples drawn from the same hypothetical population that we used to calculate our baseline parameters. If a coefficient deviates from the range, it is statistically different from the coefficient that we report for the baseline model.