Out of Sight, Not Out of Mind: Strategies for Handling Missing Data

Eric R. Buhi, MPH, PhD; Patricia Goodson, PhD; Torsten B. Neilands, PhD

**Objective:** To describe and illustrate missing data mechanisms (MCAR, MAR, NMAR) and missing data techniques (MDTs) and offer recommended best practices for addressing missingness. **Method:** We simulated data sets and employed ad hoc MDTs (deletion techniques, mean substitution) and sophisticated MDTs (full information maximum likelihood, Bayesian estimation, multiple imputation) in linear regression analyses. **Results:** MCAR data yielded unbiased parameter estimates across all MDTs, but loss of power with deletion methods. NMAR results were biased towards larger values and greater significance. Under MAR the sophisticated MDTs returned estimates closer to their original values. **Conclusion:** State-of-the-art, readily available MDTs outperform ad hoc techniques.

**Key words:** missing data, MDTs, FIML, imputation, Bayesian analysis, health behavior research

No matter how carefully health behavior researchers plan their data collection when using survey methodologies, they will always be faced with missing data. Missingness—as the phenomenon is commonly referred to—may result from lost surveys, respondent refusal to answer survey questions (eg, questions may be too sensitive), skipped questions, illegible responses, procedural mistakes, computer malfunctions, or other reasons. Additionally, survey researchers interested in studying constructs over time via longitudinal methods are forced to deal with missing data resulting from attrition. For instance, participants in a youth survey study may withdraw from school and lose contact with researchers or intentionally drop out of the study due to issues directly related to their participation or to the survey's questions. When eligible participants do not take part in the study, the missing data represent survey nonresponse. Missingness can also occur, however, within returned surveys, in continuous variables such as age, income, and attitudes, or in categorical variables such as level of education, ethnicity, and behavior (eg, ever had sexual intercourse = yes/no). Survey researchers call missing values on items such as these item nonresponse. In this instance, partial data are still available from participants.

Switzer and Roth noted the “range of potential approaches to the problem of missing data is very broad, from ignoring the problem altogether to sophisticated mathematical techniques for predicting what data would have appeared in the missing cells” (p. 310). Several books,
as well as methods and statistics book chapters discuss the complex issues regarding managing missing data. General purpose statistical software, including SPSS (www.spss.com) and SAS (www.sas.com), and most structural equation modeling (SEM) packages including Amos (www.spss.com/amos), EQS (www.mvsoft.com), LISREL (www.ssicentral.com), and Mplus (www.statmodel.com), aid the analyst in addressing missing data. However, in order to maximize the utility of these statistical tools, the thoughtful health behavior researcher should become familiar both with potential missing data mechanisms in a data set (or types of missing data) and missing data techniques (MDTs—or the strategies for handling missingness) that provide the most accurate estimates across different scenarios. The purpose of this paper is, therefore, to (1) review 3 types of missingness (or missing data mechanisms) possible in data sets, (2) review techniques for managing missing data (MDTs), and (3) compare these MDTs, using simulated health behavior data, across the various mechanisms of missingness. This comparison will illustrate the application of the different MDTs to specific missing data mechanisms.

**Missing Data Mechanisms**

More important than mere familiarity with the various techniques for handling missing data is understanding the manner in which the needed information became “lost” or missing. This understanding is vital “because the different techniques for dealing with the problem make various assumptions about why they are missing” (p. 70). In other words, identification (or an informed guess) of the mechanisms by which incomplete data come to be missing can help data analysts select the optimal methods to address data missingness.

When data sets lack data points, 3 causes are usually at play: conditional randomness; complete randomness; and bias, or systematic reasons. These 3 causes lead to the classification of missing data into 3 types: data that are missing at random (MAR), missing completely at random (MCAR), and not missing at random (NMAR). An analyst may truly never know whether the data in hand are MAR, MCAR, or NMAR. In fact, all 3 types of missing data may be present in a single data set. However, analysts should carefully consider possible reasons that data are missing, prior to conducting any statistical procedures.

**Missing at random (MAR).** Switzer and Roth noted, “Most studies [involving] missing data assume that the data are [missing at random, or MAR],” which is “probably the least troublesome pattern” (p. 318). When data are MAR, incomplete data arise not from the missing values themselves; rather, the missingness is a function of some other observed variables in the data set (ie, those variables for which the study has data). For example, suppose members of one sex (V1) are less likely to disclose their weight (V2) on a questionnaire. The probability, then, that V2 is missing depends on the value of V1, for which data are available. In this case, the missing data for V2 are MAR because it is being a man or woman (and not how much they weigh) that explains the missing data. Similarly, in cohort studies, baseline covariates and outcomes can often be used to meet the MAR assumption for participants who are lost to follow-up. For instance, gender and participant weight measured at time 1 may be correlated with subsequent attrition or failure to report weight at post-baseline measurements in a longitudinal study of gender and weight change. Because the factors leading to attrition are complex, a full treatment of the topic of missing data methods for longitudinal studies is beyond the scope of this paper. Interested readers are referred to Little, Wothke, Verbeke and Molenberghs, and Molenberghs and Verbeke. Data that are MAR are termed ignorable, because when this pattern occurs, the analyst can ignore the reason(s) data are missing and employ a missing data technique to manage the problem.

**Missing completely at random (MCAR).** A special case of MAR, when the data are missing completely at random (MCAR), occurs when the probability of missingness is unrelated to both the observed variables (ie, those for which the study has data) and the variables with missing values (ie, those for which the study has no or incomplete data). An example of MCAR data occurs when a study participant fails to return for follow-up questioning for reasons unrelated to the study, such as bad weather, an illness, or a death in the family. Another example
occurs when a computer malfunction precludes retrieval of some participants’ data values but not other participants’ values. These circumstances occur completely at random—there is no relationship between them and the study’s topic, design, procedures, or measures. Similar to data that are MAR, MCAR data are ignorable; thus, the analyst can ignore the reason(s) the data are missing and employ a MDT. The interested reader can refer to Heitjan and Basu\textsuperscript{19} for more in-depth treatment of MAR and MCAR mechanisms.

**Not missing at random (NMAR).** Data that are not missing at random (NMAR)—in other words, data made missing by systematic influences—may present complex issues for analysts who decide to use certain missing data techniques, as NMAR is the most problematic pattern of missingness. NMAR as a missing data mechanism means that “the probability of missingness is related to values that are themselves missing”\textsuperscript{14} (p. 70). For instance, people who weigh more (V2) may be less likely to disclose their weight on a survey. Thus, if nondisclosure of weight is related to the participant’s weight and unrelated to all other observed variables in the analysis, the probability that information on a person’s weight is missing depends solely on the value of V2, which is unavailable, because it is missing. In another example, a high school student may not answer a series of survey questions due to the sensitive nature of the items (eg, sexual or delinquent behavior questions). The probability of a data set containing missing data on sexual behavior items may be directly related to the degree of sexual behavior in the survey sample or the perceived consequences that may result if a person were to answer these items, or to both factors simultaneously.

Data that are NMAR are nonignorable because standard data analysis models cannot properly take into account incomplete data that arise from an NMAR missing data mechanism. In situations where analysts suspect missing data are due to one or more NMAR mechanisms, analysts would need to directly model the missing data mechanism as part of data analyses of interest. Such modeling relies upon strong, untestable assumptions about how the incomplete data came to be missing, however.\textsuperscript{4} Rather than attempting to model the missing data process, a more fruitful approach health behavior researchers can take to deal with NMAR issues is to include additional measures during the study design phase to enhance the likelihood of incomplete data meeting the MAR assumption (see Schafer & Graham\textsuperscript{15}).

**Missing Data Techniques**

The methods for handling missing data generally fall under 3 categories: deletion, direct estimation, and imputation techniques. Listwise and pairwise deletion techniques are encompassed in the first category. These techniques discard cases during an analysis if they contain missing data. As the default MDT in many general purpose statistical software packages, deletion methods are easy to employ and do not require a wealth of statistical expertise. Thus, they are frequently used. By contrast, direct estimation approaches such as full information maximum likelihood (FIML) and fully Bayesian analysis use all available information in the data, including the observed values from cases with data on some, but not all, variables, to construct parameter estimates and standard errors. In the case of Bayesian analysis, the researcher also has the option to introduce prior information into the analysis. Several methods for managing missing data fall under the category of imputation: mean substitution (or single imputation, which includes hot deck imputation) and multiple imputation (MI). According to Allison,\textsuperscript{4} “the basic idea is to substitute some reasonable guess (imputation) for each missing value and then proceed to do the analysis as if there were no missing data” (p. 11). For instance, mean substitution replaces missing values for a variable with the variable mean score. Hot deck imputation replaces the missing value with values from similar respondents. MI, a more sophisticated technique, replaces missing values with a complex combination of single and multiple imputation results. Below, we describe these techniques in further detail.

**Listwise deletion.** Listwise deletion, frequently referred to as complete case analysis, involves excluding from the analysis entire cases with missing values for any variable. For example, if each respondent fails to answer just one questionnaire item, the software package, using listwise deletion, would discard all
Handling Missing Data

There are 2 potential problems with this method. First, listwise deletion results in a loss of statistical power, or an inflation of type II error. Because this method includes only complete cases in the analysis, it can reduce sample size substantially and diminish analysts’ ability to find statistically significant effects. Second, because survey takers may choose not to answer certain questions, completely eliminating these people from the analysis may introduce bias.

Pairwise deletion. Pairwise deletion, also referred to as available case analysis, uses all available data for each variable to compute means and variances. When conducting correlational analyses (i.e., bivariate correlations, multiple regression), however, all available pairs of values are used. Thus, due to varying response rates for different survey items, the resulting values of analyses are products of different subsets of the same sample.

There are 2 major weaknesses associated with pairwise deletion. First, due to the differing numbers of observations used to estimate components of the covariance matrix... pairwise deletion can—according to Pigott—produce estimated covariance matrices that are implausible, such as estimating correlations outside the range of -1.0 to 1.0. Even if the correlation values appear plausible and fall within the range of -1.0 to 1.0, pairwise deletion may yield nonpositive definite correlation or covariance matrices, which can be problematic when using LISREL or similar SEM software. The second weakness with this technique is similar to that of listwise deletion: pairwise deletion diminishes sample size and, thus, may inflate type II error. Schafer and Graham noted, “If a missing-data problem can be resolved by discarding only a small part of the sample, then [pairwise deletion] can be quite effective” (p. 156). However, if many of the data are discarded, analysts will face a substantially reduced sample size resulting in wasted data and a loss of statistical power.

Mean substitution. Mean substitution—perhaps “the most widely used estimation technique”—involves replacing missing values for cases with the variable mean score. A strength of this method (sometimes referred to as single imputation) is that, by replacing the missing value with an actual value, the analyst is able to increase the sample to its original size, solving the “wasted data” issue created by using deletion methods. A second strength is its ease of use; no specialized computing software or statistical expertise is required because basic functions in general purpose statistical packages can easily execute the mean substitution procedure. However, for Pigott, “while this strategy allows the inclusion of all cases in a standard analysis procedure, replacing the missing values with a single value changes the distribution of that variable by decreasing the variance that is likely present” (p. 365). Tanguma observed that the variances and covariances resulting from imputing a single value will be downwardly biased, yielding attenuated correlations among variables.

Hot deck imputation. Hot deck imputation is a variant of the mean substitution technique described above. This technique involves replacing a respondent’s missing value with a value from a similar respondent. A similar respondent is someone who shares the same patterns of response for a group of matching variables (determined by the analyst). The hot deck method is a common practice in survey research, and is used by the US Census Bureau in its Current Population Survey and the Substance Abuse and Mental Health Services Administration’s National Survey on Drug Use and Health (formerly called the National Household Survey on Drug Abuse). A weakness of hot deck imputation is its assumption of no differences between nonrespondents and respondents. Another weakness is that specialized software or extensive programming in standard statistical software is needed to conduct hot deck imputation, and thus, this technique is not widely available. Interested readers can refer to Westsat for an example of such specialized software.

Multiple imputation. Single imputa-
tion techniques such as mean substitution and hot deck imputation cannot account for the uncertainty introduced by imputing data for values that are missing. Representing a more sophisticated MDT, multiple imputation (MI) involves replacing the missing value with 2 or more imputed values. MI generates m multiple data sets where the observed values are identical across the data sets, but imputed values vary in value. It is this variability from one imputed data set to another that enables an MI-based analysis to properly factor in the uncertainty involved in imputing missing values. MI is used by the Centers for Disease Control and Prevention in its AIDS surveillance systems and in the National Health and Nutrition Examination Survey conducted by the US National Center for Health Statistics.

MI is characterized by 2 major strengths: First, it corrects the lowered variance problem created by the mean substitution technique. Second, it allows analysts to examine the variance due to the imputation process and even compute statistical significance correctly.

Graham et al offered a fairly nontechnical explanation of the 3-step MI process:

First, one creates m imputed data sets, such that each data set contains a different imputed value for every missing value. The value of m can be anything greater than 1, but typically ranges from 5 to 20. Second, one analyzes the m data sets, saving the parameter estimates and standard errors from each. Third, one combines the parameter estimates and standard errors to arrive at a single set of parameter estimates and corresponding standard errors.

A key advantage of MI is that once the imputed data sets are created, the researcher can use them in almost any type of analysis ranging from simple descriptive statistics through regression methods and complex multivariate statistical analyses.

Although Little and Rubin point out that “the only disadvantage of MI over single imputation is that it takes more work to create the imputations and analyze the results” (p. 86), MI is readily available to SAS 9 users through PROC MI and PROC MIANALYZE and to SPSS users through the optional Amos structural equation modeling module. Researchers using these packages can thereby capitalize on using MI as a more state-of-the-art MDT. Some stand-alone MI packages, such as the NORM program, are available free of charge on the World Wide Web. A documentation of the NORM program and others is provided by Schafer in his text on incomplete multivariate data methods. A readable how-to guide may be found in Schafer and Olsen.

It is important to note that multiple imputation—conducted through SAS PROC MI and PROC MIANALYZE—is based on the assumption that the data arise from a multivariate normal distribution. Chen et al and Canchola et al warned that ordinal variables with highly skewed distributions and categorical variables, such as sexual orientation, do not adhere to this distributional assumption. Findings arising from analyses of multiply imputed data sets when the joint multivariate normality assumption is violated are robust to this assumption violation, however, as long the statistical models used to subsequently analyze the imputed data properly account for the data’s nonnormality. Schafer cites simulation studies and provides his own simulation evidence illustrating the robustness to nonnormality of imputation generating models that assume joint multivariate normality when the number of missing data is moderate (eg, < 50%) and the amount of nonnormality in variables not severe. Moreover, recent versions of NORM and PROC MI include variable transformation and categorical variable rounding utilities that may further improve the performance of multiple imputation conducted under the assumption of joint multivariate normality. Nonetheless, researchers must take caution and examine distributional assumptions before choosing MI methods that assume joint multivariate normality of the data. Other MDTs, which we discuss below, including maximum likelihood estimation of categorical variable analysis models and special MI methods for categorical and mixed categorical and continuous variable data sets, may prove to be more fruitful when variables exhibit extreme departures from normality. For a comprehensive discussion of MI, see Rubin and Schafer.

**Maximum likelihood methods.** Maxi-
Handling Missing Data

Maximum likelihood (ML) is a general statistical estimation approach that is used often in logistic and linear regression models,\(^4\) as well as in factor analysis and latent class analysis.\(^{15}\) ML parameter estimates for cases with complete and incomplete data can be obtained by using direct ML estimation (ie, full information ML [FIML]; see Arbuckle\(^9\) and Wothke\(^{13}\)). FIML is a direct method in the sense that model parameters and standard errors are estimated directly from available data. The FIML algorithm directly estimates model parameters; it does not impute missing values.\(^{31}\) Cases with incomplete data are included in computations, however, and all available data are employed by the ML algorithm to obtain optimal parameter estimates. The major SEM packages including Amos 7, LISREL 8, and Mplus 4 have the capability to run FIML for linear models with continuous outcomes;\(^ {32}\) Mplus 4 features a similar suite of estimators for models involving ordered categorical, nominal, and Poisson distributed outcomes.\(^ {33}\) Although it is quite complex in terms of computation, FIML is extremely easy to use and transparent to the researcher. For instance, in Amos, researchers can invoke FIML estimation with a single mouse click.

**Bayesian methods.** Over the past 10 years, researchers in the social sciences and medicine have been increasingly using Bayesian statistical methods.\(^ {34,35}\) Classical statistical methods rely on comparisons of statistics generated from sample data to hypothesized population parameters that are each assumed to be an unknown but fixed value. In contrast to classical statistical theory, the Bayesian framework relies on the notion of subjective probability.\(^ {36}\) In short, Bayesian analysts incorporate preexisting evidence and beliefs about a parameter into what is called a prior distribution of probabilities. In other words, in the Bayesian analysis framework, a population’s parameters will vary depending upon its subjective probability. Once data are collected, a mathematical routine derived from Bayes’ theorem is used to combine the prior distribution with the new data to generate an updated posterior distribution, on which conclusions are then drawn.\(^ {37}\) Conveniently, if the analyst assumes a uniform (also referred to as a diffuse or noninformative) prior distribution in which each value of the distribution is postulated to have an equally likely chance of occurring, the posterior distribution is asymptotically equivalent to the likelihood of the observed data. In Bayesian analysis, incomplete data values are viewed as additional parameters to be estimated, subject to the constraints of the analyst’s model and the selected prior distributions of model parameters. According to Burton et al,\(^ {37}\) “Bayesian-based inferences are more intuitive and can lead to more appropriate decisions” (p. 320) than traditional null hypothesis significance testing. One reason for the increased use of Bayesian methods has been major improvements in computing power. Bayesian estimation for continuous, normally distributed and ordered categorical outcome variables is available in the Amos structural equation modeling program.

**Which Missing Data Technique is Optimal? An Illustration**

No one missing data technique is optimal for every missing data situation. As stated earlier, understanding the conditions in which data are MAR, MCAR, or NMAR is pivotal “because the different techniques for dealing with the problem make various assumptions about why they are missing”\(^ {14}\) (p. 70). For instance, listwise deletion, pairwise deletion, and mean substitution techniques assume that all incomplete data arise from an MCAR process. If the missing data do, in fact, fall under the MCAR category, then using these ad hoc methods may result in nonbiased parameter estimates.\(^ {9,13,38}\) If any data are missing due to a MAR or NMAR mechanism, however, results from subsequent analyses employing these ad hoc methods will be biased.

FIML, Bayesian estimation, and MI assume MCAR or MAR data. If data missingness is due to one or both of these mechanisms, results from analyses employing FIML, Bayesian methods, or MI will generally be unbiased. If NMAR missingness is present, no standard analysis method—which we described previously—can fully remove bias. Muthén et al\(^ {38}\) argued that FIML will result in less bias under NMAR than will listwise deletion, pairwise deletion, and related ad hoc methods, however. For most conditions, mean substitution is not recommended as an accurate missing data technique.\(^ {3,10}\)
A very strict assumption that typically does not hold fully in most studies occurs when researchers assume missing data arise from a MCAR situation. MAR scenarios are much more likely in social and behavioral science research, especially if researchers measure constructs such as social desirability and need for cognition or reading comprehension in their surveys. These types of constructs, for which it is often possible to obtain complete data, can be useful in predicting who does or does not complete survey instruments. According to Schafer and Olsen,28 "in the vast majority of studies, principled methods [such as MI and FIML] that assume MAR will tend to perform better than ad hoc procedures such as listwise deletion or imputation of means" (p. 553).

To illustrate these missing data mechanisms and how they may impact results when employing various MDTs, we generated simulated data from a population with known parameter values. From the SAS Technical Support website, we first downloaded mvn.sas, a program that simulates bivariate or multivariate normal data. We supplied the program with a sample correlation matrix and mean vector consisting of zeros with a sample size of n = 300 (a typical sample size found in health behavior survey research), in order to create 3 continuous variables with a joint multivariate normal distribution. The generated complete data set consisted of 3 columns (variables): COL1, COL2, and COL3. In this illustration, COL1 represents alcohol use behavior, COL2 represents depression, and COL3 represents the perceived norm toward alcohol use. We next created 3 different data sets containing missing values for COL1. To illustrate a MCAR scenario, we simply deleted the last 25% of COL1 values—a completely random process because the computer-generated data were themselves generated via a completely random process. To represent a MAR scenario, we dropped 25% of COL1 values, defined by each case's COL2 and COL3 scores, with higher values of COL2 and lower values of COL3 being more likely to have missing values. Finally, to illustrate a NMAR scenario, we dropped 25% of COL1 values, defined only by their original COL1 values, with lower values of COL1 being more likely to be missing.

The first analysis involved the complete data set, where we regressed COL1 (alcohol use behavior) onto COL2 (depression) and COL3 (perceived norm toward alcohol use). This initial investigation served as the benchmark against which we could compare our analyses of data sets containing missing values. The first row in Table 1 (labeled "Regression") contains these complete data analysis results. Next, we subjected the MAR, MCAR, and NMAR data sets to multiple imputation with 10 imputations created, followed by SAS PROC REG analyses of each of the imputed data sets and the use of PROC MIANALYZE to combine the results from the 10 PROC REG analyses. We then exported the MAR, MCAR, and NMAR data sets into SPSS 14 and employed listwise deletion, pairwise deletion, and mean substitution techniques. Finally, we exported the 3 data sets into Amos 7, which offers FIML estimation of incomplete multivariate normal data as well as Bayesian estimation. Here, we set up the regression analysis using Amos Graphics and fit the regression model to the MAR, MCAR, and NMAR data. For each analysis we report the regression estimates and respective 95% confidence intervals for the COL2 and COL3 regression coefficients; where available, we also report traditional p-values associated with the null hypothesis that the relevant parameter is zero in the population from which we drew our sample.

Results from the analyses described above can be found in Table 1. Moving from left to right in the table, readers can see that the MCAR situation yielded unbiased estimates across the 6 MDTs, but somewhat of a loss of statistical power for the COL3 predictor with the deletion and mean substitution techniques. The MAR situation resulted in slightly downwardly biased parameter estimates using MI, FIML, and Bayesian estimation. Interestingly, when we employed listwise deletion (in the MAR situation), the COL3 predictor (perceived norm toward alcohol use) actually resulted in a reversed sign, from negative to positive, losing significant explanatory power in the theorized direction. Pairwise deletion, under the MAR situation, resulted in slightly biased (larger) values and inflated standard errors for the COL3 predictor. In the NMAR situation, the results were substantially biased towards bigger values, more significance, and inflated standard errors for COL2, and a general loss of statistical
power for the COL3 predictor with pairwise deletion and mean substitution techniques. Consistent with the statistical simulation literature, our analysis illustrated an instance in which the listwise deletion technique did not work very well under the MAR situation, yet FIML, MI, and Bayesian estimation worked quite well. Given the ready availability of the state-of-the-art techniques in easy-to-use software programs, there seems little reason to continue using ad hoc methods to handle missing data, unless the number of missing data is so small that the resulting bias is inconsequential (eg, 5%; see Roth). However, results from this illustration also demonstrate the risks of analyzing NMAR data using methods that assume incomplete data arise from exclusively MAR or MCAR missingness mechanisms. Generally, across all 3 patterns of missingness, the MI, FIML, and Bayesian estimation procedures restored parameter estimates, confidence intervals, and statistical significance closer to the original (complete data) values.

Table 1
Illustration of Multiple Linear Regression Analysis Using Missing Data Techniques

<table>
<thead>
<tr>
<th></th>
<th>Complete Data (n = 300)</th>
<th>MCAR (25% missingness on COL1, arising from a random process)</th>
<th>MAR (25% missingness on COL1, defined by COL2 &amp; COL3 scores)</th>
<th>NMAR (25% missingness on COL1, defined only their original COL1 values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COL2</td>
<td>COL3</td>
<td>COL2</td>
<td>COL3</td>
</tr>
<tr>
<td>Regression</td>
<td>.233 (.059)</td>
<td>-.132 (.052)</td>
<td>.193 (.069)</td>
<td>-.127 (.053)</td>
</tr>
<tr>
<td>Multiple</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>imputation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIML</td>
<td>-</td>
<td>-</td>
<td>.215 (.064)</td>
<td>-.127 (.053)</td>
</tr>
<tr>
<td>Bayesian estimation</td>
<td>-</td>
<td>-</td>
<td>.213 (.064)</td>
<td>-.127 (.053)</td>
</tr>
<tr>
<td>Listwise deletion</td>
<td>-</td>
<td>-</td>
<td>.213 (.065)</td>
<td>-.108 (.058)</td>
</tr>
<tr>
<td>Pairwise deletion</td>
<td>-</td>
<td>-</td>
<td>.219 (.065)</td>
<td>-.129 (.059)</td>
</tr>
<tr>
<td>Mean substitution</td>
<td>-</td>
<td>-</td>
<td>.216 (.065)</td>
<td>-.121 (.053)</td>
</tr>
</tbody>
</table>

Note.

a CIs provide information for point null hypothesis testing—if the interval includes zero, the coefficient enclosed by the interval is not significant at P < .05—but they also document the precision of the parameter estimate—narrow intervals signify greater precision, whereas wider intervals signify less precision.

b Bayesian estimation in Amos does not produce p-values, but does produce 95% CIs that may be interpreted in the same way as FIML-based 95% CIs when the data analyst uses a uniform or non-informative prior distribution for model parameters. For instance, in our example the Bayesian 95% CIs are virtually identical to those produced by the Amos FIML algorithm, so essentially the same substantive conclusions may be drawn using the Bayesian and FIML methods.

c When employing listwise and pairwise deletion methods, the sample size was reduced to n = 240.

d COL1 = alcohol use behavior, COL2 = depression, COL3 = perceived norm toward alcohol use

*** P ≤ .001
** P ≤ .01
* P ≤ .05
CONCLUSION

As missing data pervade health behavior research, scholars are guaranteed to face the problem at some time in their data collection efforts. Unfortunately, the predicament is one for which there are no ideal solutions, and the researcher must choose—among several complex techniques—the one that optimizes accuracy of estimated parameters, while avoiding error inflation. The purpose of this paper was to make available to health behavior researchers—especially to those using survey methodologies—a primer on techniques for managing missingness. Thus, this paper described 3 mechanisms of missing data (MAR, MCAR, and NMAR), and presented ad hoc methods—listwise deletion, pairwise deletion, and single substitution—as well as more state-of-the-art techniques—MI, FIML, and Bayesian estimation—for managing missing data under MCAR and MAR conditions. Further, we illustrated, using simulated health behavior data sets, how these missing data techniques compare across the various mechanisms of missingness in a simplified, but prototypical health behavior research data analysis. Our illustration reflected the findings from many data simulation studies published in the statistical literature: sophisticated MDTs (FIML, Bayesian estimation, and MI) generally outperform the older ad hoc approaches (listwise deletion, pairwise deletion, and mean substitution). Based on the results from those simulation studies, we suggest that health behavior researchers employ sophisticated MDTs whenever possible in situations where the amount of missing data is not trivially small (ie, < 5% of the sample). FIML and Bayesian methods are the easiest to use for model types and outcomes supported by software programs that feature these estimation methods; for all other analysis problems, MI is a useful approach.

Finally, at the risk of stating the obvious, it is important to remember that the best way of handling missing data in health behavior research is by means of prevention. Taking necessary precautions to ensure optimal response rates and item-response with little or no missing information is not merely ideal, but a goal for which every researcher should strive with his or her research design and data collection methods. An important component of this prevention effort may also include adequate preparation for data that, despite researchers’ best preventive efforts, turn out to be incomplete: measuring variables that, when included in the data help them meet the MAR assumption in the analysis, can be a helpful strategy. For instance, we suggest researchers collect as much data as possible and collect auxiliary information. In the example of measuring weight data, researchers should collect information related to correlates of weight and use proxy measures of weight in order to minimize MNAR scenarios. Researchers may also consider social desirability issues regarding variables of interest and whether respondents can adequately understand questions being posed.

Regardless of the efforts employed to prevent or prepare for missingness, as stated previously, health behavior researchers will likely face the problem and the need to choose among alternative techniques. Becoming familiar with available options and understanding their applicability are important first steps. Our wish is to facilitate accomplishing these goals.

REFERENCES

10. Graham JW, Cumsille PE, Elek-Fisk E. Meth-
Handling Missing Data

...