Evaluating Interventions With Differential Attrition: The Importance of Nonresponse Mechanisms and Use of Follow-Up Data

John W. Graham and Stewart I. Donaldson

Evaluations of psychological interventions are often criticized because of differential attrition, which is cited as a severe threat to validity. The present study shows that differential attrition is not a problem unless the mechanism causing the attrition is inaccessible ( unavailable for analysis). With a simulation study, we show that conclusions about program effects (a) are unbiased when there is no differential attrition, even with usual complete cases analysis; (b) may be severely biased when based on usual complete cases analyses and there is differential attrition; (c) are unbiased when based on the expectation-maximization (EM) algorithm, even when there is differential attrition, as long as the attrition mechanism is accessible; and (d) are biased, even with the EM algorithm, when the attrition mechanism is inaccessible. Following Little and Rubin (1987), we advocate the collection of new data from a random sample of subjects with initially missing data. On the basis of these data, we propose a simple correction to the EM algorithm estimates. In our study, the correction produced unbiased estimates of program effects parameters, even with an inaccessible attrition mechanism and substantial differential attrition.

The number and variety of psychological interventions have risen dramatically over the past several decades. For example, psychologists concerned with improving working conditions have designed and implemented interventions that prevent poor mental health and promote high-quality reemployment for unemployed people (Vinokur, van Ryn, Gramlich, & Price, 1991), that improve various aspects of work performance (see Guzzo, Jette, & Katzell, 1985), and that enhance the quality of work life (see Mohrman, Ledford, Lawler, & Mohrman, 1986). Educational applications include programs that provide compensatory education for the underprivileged (see Hodges & Cooper, 1981; House, Glass, McLean, & Walker, 1978) and increase the social functioning of at-risk elementary school children (King & Kirschenbaum, 1990). Programs related to forensic psychology include domestic violence reduction (Sherman & Berk, 1984) and juvenile delinquency curtailment (see Lipsey, 1988). Some examples from the health area are interventions to prevent heart disease (Farquhar et al., 1990), encourage healthy life-styles (Nathan, 1984), and reduce the onset of adolescent substance use (for recent reviews see Flay, 1985; Hansen, 1990).

The prevalence and broad diversity of psychological interven-
ventions have created many challenges for those involved with their systematic evaluation. Although great strides have been made to increase the sophistication of the research designs used to evaluate interventions, methodological limitations are often cited as reasons to be skeptical of this type of evaluation research (e.g., Biglan & Ary, 1985; Biglan et al., 1987; Cook, 1985; Moskowitz, 1989). In this article we address the issue of participant attrition, which is sometimes referred to as the “weakest link” in intervention research (e.g., Hansen, Collins, Malotte, Johnson, & Fielding, 1985). We explore this issue within the context of a substance abuse prevention example. However, our arguments and procedures apply more generally to the broad domain of intervention research.

Participant attrition is a problem for virtually all research that requires measurement at two or more points in time. As Cook and Campbell (1979) have pointed out, attrition poses a threat to external validity (generalizability) as well as to the internal validity of a study. More recently, Biglan et al. (1987) as well as Moskowitz (1989) have argued that many prevention programs have shown apparently beneficial results that could be accounted for by participant attrition. Others have argued that the failure to find effects could be the result of attrition problems (e.g., Botvin, 1987).

In summarizing the earlier thinking regarding possible attrition artifacts in prevention research, Hansen et al. (1985) suggested four simple tests for determining whether attrition is or is not a threat to the validity of a particular study:

1. Are those who drop out different from those who stay on pretest values of the dependent variables and on demographic characteristics?

2. Do those who drop out show different patterns across time on the dependent variables than those who stay?

3. Are there differences in rates of attrition among conditions?

4. Are pretest scores for those who drop out different among conditions?
The first two questions address the issue of external validity, whereas Questions 3 and 4 address internal validity issues. In this article, we address primarily the issues relating to internal validity. However, many of our conclusions apply as well to issues of external validity.

For testing whether an observed beneficial program effect could be an attrition artifact, Hansen et al. (1985) suggested testing whether there were differential attrition rates in treatment and control groups. In substance use prevention research, it is almost always the case that those who drop out of a study have higher levels of use at the pretest than those who stay. Thus, if there are more "droppers" from the treatment group than from the control group, reported substance use among "stayers" is artificially lower in the treatment group.

The second internal validity issue (i.e., Question 4) raised by Hansen et al. (1985) is whether the pretest scores for droppers are different across conditions. Biglan et al. (1987) pointed out that this test is incomplete and suggested that it be modified to include differences among both the droppers and stayers. That is, Question 4 should be expanded to address the Program × Attrition Status interaction on the pretest measure of the main dependent variable. This tests specifically whether the differences between participants in the treatment and control groups depend on whether the participant is a stayet or a dropper.

If dropout rates are not different between treatment and control groups and if the Program × Attrition Status interaction is not significant (based on pretest data), then researchers have generally concluded that observed program effects are probably not due to attrition artifacts. In the present article we will show that this conclusion is correct in most instances.

However, if there is differential attrition based on pretest data (for the remainder of this article, unless otherwise noted, when we refer to differential attrition, we mean both different dropout rates in treatment and control groups and Program × Attrition Status interaction on pretest measures of the dependent variable), researchers have generally concluded there is attrition bias. We show that such conclusions may be incorrect for two reasons. First, tests such as those proposed by Hansen et al. (1985) and Biglan et al. (1987) were never intended as direct tests of attrition bias. Second, because these tests are based on pretest data, their conclusions may be very different from what one would hypothesetically obtain from a test involving the main dependent variable itself.

We demonstrate that when there is differential attrition (based on pretest data), it is incorrect to conclude automatically that the internal validity of the study has been compromised; under these conditions there may or may not be a problem. When there is differential attrition, the key to the problem has to do with the mechanism causing the missing data (Dent, 1988; Heckman, 1979; Little & Rubin, 1987, 1989; Rubin, 1987). We show that when the data are not missing completely at random, it is important whether the missing data mechanism is accessible or not.

Defining Accessible Missing Data Mechanisms and Other Terms

Missing Data Mechanism

Knowing the missing data mechanism is equivalent to knowing what causes "missingness" in the dependent variable. Let the main dependent variable be denoted Y. Following Rubin (1987), let there be an indicator variable R, denoting whether Y is missing or not. If Y is present, R = 1; if Y is missing, R = 0. The cause of missingness on Y, denoted CR, is a variable that is causally related to R. CR may or may not be related to Y itself.

Accessible and Inaccessible Missing Data Mechanisms

By accessible, we mean that the cause of the missingness in Y (CR) is a variable that has been measured for all cases and is available for analysis. By our definition, the mechanism is accessible whether or not one chooses to make use of CR in the analysis. By inaccessible, we mean that CR has not been measured for everyone, or is otherwise unavailable for analysis. (These terms are similar to, but have important differences from, "ignorable" and "nonignorable" missing data mechanisms as described by Little and Rubin, 1987)

Data Missing Completely at Random

This term, as used by Little and Rubin (1987, 1989; Rubin, 1987), refers to the case for which CR is a random process (uncorrelated with other variables).

Complete Cases and All Cases

Complete cases refers to those subjects who have no missing data. All cases refers to all subjects involved in the study, including those with complete data for all variables and those with missing data for some variables. For many points we make below, we do not have complete data for all cases. Nonetheless, it is often important to compare a complete cases parameter estimate with the parameter estimate we would obtain, hypothetically, if we did have data for all cases.

Top-Down Versus Bottom-Up Perspectives

It is important to consider what perspective one has when making these definitions. Two important perspectives are (a) that of the person who created the data (the top-down perspective) and (b) that of the researcher, who has no way of knowing a priori what the true state of affairs is (the bottom-up perspective). The definitions we provide initially are from the top-down perspective. We realize that such definitions are of little help to the researcher who is trying to understand his or her data. Thus, in a later section, we consider issues relating to the bottom-up perspective.

We use a simple example to illustrate our points. We assume (a) a dummy-coded program variable (program = 1, control = 0), (b) a variable measured at the pretest (X), and (c) the measure of the main dependent variable (Y). In our example, we have complete data for the program and X variables, but missing data for the dependent variable, Y (attrition).

Little and Rubin (1987, 1989; Rubin, 1987) and Heckman (1979) have shown (a) that there is no bias associated with data that are missing completely at random, (b) that there is no bias for situations we describe as having accessible missing data mechanisms, as long as the accessible mechanism is taken into account, and (c) that there is bias for situations we describe as
having inaccessible mechanisms. Rather than reviewing their proofs here, we present below a conceptual rationale for these points and present a brief simulation to illustrate the points made, specifically as they apply to participant attrition.

In the sections that follow, we present brief conceptual explanations for (a) the lack of bias when \( C_{Y_{mis}} \) is a random process; (b) the lack of bias for the b-weight (\( X \) predicting \( Y \)) when \( C_{Y_{mis}} = X \) and \( Y \) is a variable that has been measured for all cases; (c) the bias for the means of \( X \) and \( Y \) when \( C_{Y_{mis}} = X \) and \( X \) has been measured for all cases; (d) the bias for the zero-order correlation, \( r_{Program \times Y} \), when \( C_{Y_{mis}} = X \) and \( X \) is measured for all cases but is not included in the model; (e) the lack of bias for the regression weight of program predicting \( Y \) when \( C_{Y_{mis}} = X \) and \( X \) is measured for all cases and is included in the regression model; and (f) the bias for the regression weight of program predicting \( Y \) when \( C_{Y_{mis}} = Y \) and \( Y \) is not measured for all cases, even when \( X \) is correlated with \( Y \) and \( X \) is included in the regression model.

\[ C_{Y_{mis}} \] is Random Process

Figure 1 presents hypothetical data showing distributions for the dependent variable, \( Y \), given values on another variable, \( X \), measured for all cases. First, look at the left side of Figure 1. If the \( C_{Y_{mis}} \) is a random process, then the \( Y \) values are deleted at random from each of the three distributions. It is clear that although values are omitted from these distributions, each distribution is a random subsample of the original. Because of this, the regression line (b-weight, \( X \) predicting \( Y \)) for the remaining (complete) cases is representative of the hypothetical b-weight involving all cases.

\[ C_{Y_{mis}} = X \] (X Measured for All Cases)

This is an example of an accessible missing data mechanism. The example illustrates why there is no attrition bias when the cause of the missingness is taken into account.

It is easiest to see the effect of attrition in this case if \( C_{Y_{mis}} = X \), where \( X \) is measured for all cases. For example, as shown on the left side of Figure 1, suppose the probability of missingness on \( Y \), \( p(Y_{mis}) = .1 \) if \( X = 1 \), \( p(Y_{mis}) = .4 \) if \( X = 2 \), and \( p(Y_{mis}) = .7 \) if

\[ X = 3 \]. Even though the missingness is not completely at random, it is easy to see that the \( Y \) scores are removed at random from within each level of \( X \). In other words, the \( X \) score partly determines from which column the data will be missing, but it does not determine which value within that column will be missing. Just as in the case of data that are missing completely at random, the distribution of \( Y \) scores within each level of \( X \) is a random subsample of the distribution for the hypothetical total sample. Thus, just as in the case of data missing completely at random, the regression line (b-weight, \( X \) predicting \( Y \)) is expected to be the same as the b-weight for the analysis with all cases, if we could calculate it. (The correlations are actually different in this situation, but the unstandardized b-weights are expected to be the same. Although the b-weights are unbiased, standard errors for the b-weights calculated in the usual ways may be biased.)

With this example, we have shown conceptually that if a single variable \( X \) is the cause of missingness on \( Y \) then the b-weight (\( X \) predicting \( Y \)) is not biased by attrition. An important generalization is that this lack of bias holds if \( C_{Y_{mis}} = Program \), that is, under these conditions, the b-weight (program predicting \( Y \)) is unbiased (again, the standard error for this b-weight, calculated in the usual ways, may be biased).

Although the b-weights are not biased under these conditions, the mean for \( X \) is biased if we use only complete cases. In our example, on the left side of Figure 1, there is a much higher probability of missingness on \( Y \) if \( X = 3 \) than if \( X = 1 \). That is, too many 3s will be missing from any analysis involving complete cases. In this case, then, the mean of \( X \) based on complete cases will be lower than the mean of \( X \) based on all cases. Furthermore, because \( X \) and \( Y \) are correlated, the mean for \( Y \) based on complete cases will also be lower than it would be (hypothetically) based on all cases.

\[ C_{Y_{mis}} = X \] With Differential Attrition

This is also an example of an accessible missing data mechanism. However, in this case we illustrate how attrition bias is introduced if \( C_{Y_{mis}} \) is not taken into account.

As before, suppose that \( C_{Y_{mis}} = X \) and that \( X \) is measured for all cases. Suppose we have two groups, program and control. Suppose further, as shown in Figure 1, that in the program group, attrition is more likely for high values of \( X \), whereas in the control group, attrition is more likely for low values on the pretest. This is the classic case of differential attrition (e.g., Cook & Campbell, 1979).

As we noted previously, because more high scores are missing from the program group, the complete cases mean of \( X \) for that group will be too low. Also, because more low scores are missing from the control group, the complete cases mean of \( X \) for that group will be too high. These mean differences translate into a spurious correlation between the group membership dummy variable and \( X (r_{program \times X}) \) and for \( r_{program \times Y} \) to the extent that \( X \) and \( Y \) are correlated. In this case, the spurious correlation makes an ineffective program look effective. The opposite form of differential attrition could also occur, leading an ineffective program to appear ineffective.
\[ C_{Y_{\text{meas}}} = X, \ r_{XY} = 0 \]

Suppose that \( C_{Y_{\text{meas}}} = X \), where \( X \) is measured for all cases, that \( r_{XY} = 0 \), and that there is differential attrition. Under these conditions, we have a situation similar to that depicted in Figure 2. Even though there is substantial differential attrition (as in the previous example), it is easy to see that the correlation \( r_{\text{Program} \cdot Y} \) is not biased for the complete cases analysis. It is true that program cases are much more likely to drop out if they have high values on \( X \). Thus, as before, the mean of \( X \) within the program group will be too low. However, because the \( r_{XY} = 0 \), the expected mean for \( Y \) is the same for every level of \( X \). Thus, the expected mean for \( Y \) is the same for complete cases and the hypothetical all cases analysis. To generalize, a variable, \( C_{Y_{\text{meas}}} \), that causes missingness in the dependent variable, \( Y \), produces attrition bias only to the extent that \( C_{Y_{\text{meas}}} \) is correlated with \( Y \).

This point has important implications. Again, suppose \( C_{Y_{\text{meas}}} = X \), where \( X \) is measured for all cases. Regression analyses partial out the effects of the predictor variable, \( X \), on \( Y \). The residual from the regression analysis is that part of \( Y \) that is uncorrelated with \( X \). If we plot the distributions of \( Y \) (residual) for each level of \( X \), it looks like something like that shown in Figure 2, even when the raw \( X \) and \( Y \) plots look like the distributions shown in Figure 1. Because the correlation \( r_{XY_{\text{model}}} = 0 \), there is no bias for the correlation \( r_{\text{Program} \cdot Y_{\text{residual}}} \). More generally, the regression weight for the program effect with complete cases is an unbiased estimate of the population regression weight, when \( C_{Y_{\text{meas}}} = X \) and \( X \) is included in the regression model, even when there is substantial differential attrition.

We have illustrated effects of accessible missing data patterns with a simple example involving a program variable and a single covariate, \( X = C_{Y_{\text{meas}}} \). It should be obvious, however, that the basic pattern would hold for any number of covariates. If \( C_{Y_{\text{meas}}} \) is a function of several variables and all of these variables are included in the model as covariates, then the total cause of missingness is unrelated to the residualized \( Y \), and there is no bias in the regression weight for the program variable predicting \( Y \). If \( C_{Y_{\text{meas}}} \) is a function of several variables but not all are measured, then the bias due to the measured causes will be controlled by the regression analysis involving complete cases. However, as we point out below, to the extent that the unmeasured causes are correlated with \( Y \), the mechanism is inaccessible and will cause bias.

### Inaccessible Missing Data Mechanisms

If \( C_{Y_{\text{meas}}} = Z \) and \( Z \) has not been measured for all cases or is otherwise unavailable for analysis, we say that the missing data mechanism is inaccessible. However, even in this case, there is no bias problem if there is no differential attrition between treatment and control groups. This is easy to show conceptually. Suppose \( C_{Y_{\text{meas}}} = Y \) itself (e.g., high \( Y \) values are missing with higher probability) but that the missing data pattern is the same for program and control. In this situation, the means for the two groups will be similarly affected by the attrition. Thus, there is no bias for the correlation between the program variable and the posttest.

On the other hand, when there is differential attrition and the mechanism is inaccessible, the parameter estimates are biased. We present below a brief conceptual explanation for this.

When \( C_{Y_{\text{meas}}} = Y \) itself, the missing data mechanism is, by definition, inaccessible. The mechanism is also inaccessible if \( C_{Y_{\text{meas}}} = Z \), where \( Z \) is an unmeasured variable for some or all cases. We have already shown that there is no attrition bias (for \( r_{\text{Program} \cdot Y} \)) if \( Z \) is uncorrelated with \( Y \). Thus, attrition bias for \( r_{\text{Program} \cdot Y} \) occurs only when \( r_{Z \cdot Y} \) is different from zero. Because of this, with inaccessible mechanisms it is convenient to think that \( C_{Y_{\text{meas}}} = Y \) to the extent that \( Z \) and \( Y \) are correlated.

Suppose that the missing data on \( Y \) is the same basic pattern shown in Figure 1, except that the cause of the missingness is \( Y \) rather than \( X \). Two important things happen in this situation. First, the distributions of \( Y \) within each level of \( X \) are no longer random representations of the hypothetical all cases distribution. A simple example will illustrate this point. Suppose that all the \( Y \) data on the left side of Figure 1 are missing if \( Y > 3 \). It is clear that this kind of missing data pattern would seriously alter (i.e., truncate) the distribution of \( Y \) scores, especially for \( X = 3 \). The mean \( Y \) score for \( X = 3 \) would be lower and the regression line for \( X \) predicting \( Y \) would be flatter, indicating a slightly lower \( r_{XY} \) correlation.

The second important problem with inaccessible mechanisms is also illustrated by Figure 1. The mean of \( Y \) for the program group will be lower for complete cases than for the hypothetical all cases sample, and the mean for \( Y \) for the control group will be higher for complete cases (both as before). However, because \( C_{Y_{\text{meas}}} = Z \) and not \( X \), partialing \( X \) out of \( Y \) does not partial the cause of missingness out of \( Y \). Thus, the cause of missingness is still related to the \( Y_{\text{residual}} \), and the mean for the residualized \( Y \) is still spuriously lower in the program group than in the control group.

### Traditional Data Analysis Solutions With Accessible Missing Data Mechanisms

As we have shown, even when the missing data mechanism is accessible, standard correlation analyses with complete cases produce biased estimates of \( r_{\text{Program} \cdot Y} \) when there is differential attrition and \( C_{Y_{\text{meas}}} \) is some variable other than program. However, we have also shown above that when \( C_{Y_{\text{meas}}} = X \), where \( X \) is
measured for all cases, regression analyses involving \( X \) produce unbiased estimates of program effects. In general, regression-based analyses that include possible causes of attrition are recommended over simple correlation analyses between program and posttest variables.

The EM Algorithm

The EM algorithm (expectation–maximization; Dempster, Laird, & Rubin, 1977; Little & Rubin, 1987) is an iterative estimation procedure that has proven to be an excellent method for conducting data analysis with missing data. Our implementation of Little and Rubin’s (1987) procedure begins with the E-step, which involves collecting sums and sums of squares and cross-products as the data are read into the program (EMCOVEXE; Graham, Hofer, & MacKinnon, 1991).

For sums, if the data value is present, the value is added to the sum for that variable. If the data value is missing, the best estimate for the value is added (in this case the best estimate is based on a regression equation with all other variables as predictors).

For sums of squares and cross-products, if either data value is present, the square or cross-product is based on the actual values or the actual value for one and the best estimate of the other value (as above). If both data values are missing, the square or cross-product is based on the best estimates of the values (as above) plus a penalty term. Little and Rubin (1987, p. 143) defined this term as follows:

\[
C_{\text{pen}} = \text{Cov} (y_{i0}, y_{i1} | y_{i0}, \Theta (0))
\]

That is, the penalty term for iteration (i) is equal to the covariance of the two variables \( y_{i0}, y_{i1} \) given the nonmissing data \( y_{i0} \) and the estimated covariance matrix (\( \Theta \)) at iteration (i). For the simple attrition situation in which data are missing only for the main dependent variable, this penalty term is the bottom, right-hand element of the matrix resulting from the sweep operation.

For this version of the EM algorithm, the M-step involves simply calculating the covariance matrix elements based on the sums and sums of squares and cross-products obtained from the E-step. Little and Rubin (1987) then suggested using the sweep operator (e.g., Goodnight, 1978) as a computationally convenient method for obtaining regression weights from a covariance matrix. The regression weights from this part of the procedure are used in the next E-step to obtain the best estimates for missing data values. A by-product of the sweep operation is the penalty term to be used in the next iteration when calculating sums of squares and cross-products when both data points are missing. The E- and M-steps are repeated until the change in the estimates of the covariance matrix are sufficiently small and judged to be unimportant (i.e., when the convergence criterion has been met).

EM Algorithm Analysis With Accessible Mechanisms

When the missing data mechanism is accessible, analyses based on the EM algorithm (Graham et al., 1991; Little & Rubin, 1987) produce unbiased estimates of all regression coefficients (including the effect of program on \( Y \)), as long as all causes of missingness are included in the analysis. This is true even when there is substantial differential attrition. These facts are especially easy to see for the simple but common case in which one has complete data for the program variable and \( C_{YM} = X \) (pretest) and missing data only for \( Y \) (the main dependent variable). In this case, the EM algorithm solution reduces to the traditional regression solution as presented earlier.

For this simple case, the EM algorithm also produces unbiased estimates of all covariance elements, even with differential attrition, as long as \( C_{YM} \) is included in the analysis. This is true because the EM algorithm makes use of \( C_{YM} \) in estimating the zero-order correlation \( r_{XY} \), whereas the traditional zero-order correlation, \( r_{XY} \), is biased because it does not take \( C_{YM} \) into account.

For the case in which one has complete data for the program and \( C_{YM} = X \) variables and missing data only for \( Y \), traditional analyses are just as good as the EM algorithm. Unfortunately, the researcher cannot know in advance if \( C_{YM} = X \). Also, with more complicated missing data patterns (e.g., for patterns involving some data missing from some pretest covariates), the EM algorithm offers clear advantages. If we assume that the cause of missing data is either random or is included in the analysis, then the EM algorithm simply calculates the unbiased estimates of the entire covariance matrix in a single analysis. Although the same results could conceivably be found with more traditional analyses, it would be a tedious process. Finally, as we discuss later, the EM algorithm offers clear advantages for data with inaccessible missing data mechanisms.

Data Analysis Solutions for Inaccessible Missing Data Mechanisms

As we showed earlier, if the missing data mechanism is inaccessible and there is differential attrition, parameter estimates relating to program effects can be seriously biased when based on complete cases regression analyses. Because the EM algorithm is equivalent to the regression analysis in the simple attrition situation, even the EM algorithm estimates are biased when the mechanism is inaccessible.

The only way we know to analyze data with inaccessible missing data mechanisms is to model the mechanism (Little & Rubin, 1987; Rubin, 1987). One approach is to make reasonable guesses about the mechanism based on what one believes the true mechanism to be. Rubin (1987) discussed one solution that involves imputing the missing values and then modifying the imputed score to account for the supposed missing data mechanism. For example, one suggestion is to add a percentage (say 20%) of the imputed value to the imputed value for those students with missing data.

Depending on the research situation, this approach may or may not be useful. If one makes conservative and liberal guesses about how much to add to the imputed values and if the research conclusions based on these guesses remain about the same, then the procedure could be extremely useful. However, if the research conclusions are widely different based on liberal and conservative guesses, then the procedure is of little value. Still, there may be information that allows the researcher to make more precise guesses about the missing data mechanism even if that mechanism is not completely known.
Little and Rubin (1987) and Rubin (1987) proposed another general solution to the problem of inaccessible missing data mechanisms. In the present article, we follow their lead with a relatively simple procedure that should be of help in many attrition analyses involving multiple regression (or analysis of covariance) involving continuous variables.

The solution is to make the inaccessible mechanism accessible by collecting additional data, ideally at random, from the subjects originally having missing data. Even a small additional sample will provide enormous information about the missing data mechanism. We continue with our simple (but common) example (two predictors, program and X; missing data only for the main dependent variable, Y). Also, recall that for this example, the EM algorithm solution is the same as a simple regression solution.

Assume that we have three samples: Sample 1 is the sample of subjects with originally nonmissing data; Sample 2 is the small random sample of subjects with originally missing data, but for whom we have collected posttest data; and Sample 3 is the sample of subjects whose posttest data are still missing. In Sample 1, the regression equation is

\[ \hat{Y} = b_0 + b_1 \text{Program} + b_2 X. \]  

We could also estimate the same model in Sample 2 rather than Sample 1. Then we would have

\[ \hat{Y}^* = b'_0 + b'_1 \text{Program} + b'_2 X. \]  

Conceptually, the solution to inaccessible missing data amounts to using the regression from Sample 2 (Equation 2) as the best estimate of missing data in Sample 3. If we were actually imputing (replacing) missing data in Sample 3, this procedure would be easy to implement. However, single imputations are known to produce highly biased parameter estimates (Graham et al., 1991; Little & Rubin, 1987; Rubin, 1987), so use of the EM algorithm is desirable. However, because the EM algorithm does not actually impute values, but instead calculates the covariance matrix directly, this conceptually straightforward solution is computationally difficult to implement.

However, it can easily be shown that the conceptually straightforward solution is the same as making a simple correction to the Sample 1 equation (Equation 1) as follows. First, we apply Equation 1 from Sample 1 to the data in Sample 2, giving us

\[ \hat{Y}^* = b_0 + b_1 \text{Program} + b_2 X. \]  

Substituting, we have

\[ \hat{Y}^* = b'_0 + b'_1 \hat{Y}^*, \]  

where \( \hat{Y}^* \) is the predicted score when Equation 1 is applied in Sample 2, and \( b'_0 \) and \( b'_1 \) are the intercept and regression weight when \( \hat{Y}^* \) is used to predict the actual Y value in Sample 2.

The correction for inaccessible mechanisms, shown in Equation 4, is easy to implement in the EM program. In the previously described E-step, when values were missing, the sums (and sums of squares and cross-products) were based on the best estimate of the missing data. For the inaccessible mechanism correction, we simply add at each E-step the new information from Equation 4. The correction involves substituting the best estimate of the value (\( \hat{Y}^* \)) into Equation 4 and using \( \hat{Y}^* \) as the best estimate at each E-step rather than the uncorrected value (\( \hat{Y}^* \)).

**Simulation Study**

We performed a brief simulation to illustrate the points we have made. The basic model used was the same as for the simple attrition example. We have a dummy-coded program variable (Program), a pretest variable (X), and the main dependent variable (Y). Program was a dichotomous variable with values 0 = control and 1 = program. X and Y were ordinal variables each with four levels and approximately uniform distribution.

We created a master data set \( (N = 500) \). The correlation matrix and regression equation for the master data set are presented in Table 1. Next, several additional data sets were created with four missing data patterns. Two missing data factors (type of mechanism and attrition status) were fully crossed, producing four conditions: (a) accessible mechanism, differential attrition; (b) inaccessible mechanism, differential attrition; (c) accessible mechanism, no differential attrition; and (d) inaccessible mechanism, no differential attrition.

For each of these combinations, we generated 20 data sets such that there were no missing data for Program or X and missing data for Y according to the particular combination of missing data factors. The missing data were generated as follows.

The accessible mechanism was produced by forcing missing data on Y as a function of scores on X (plus a random component). All missing data were created such that higher levels of X were more likely to be missing than low levels of X. For conditions with no differential attrition, the probability of being missing on Y was the same for program and control conditions (when \( p = .10, .30, .50, .70 \) for \( X = 1, 2, 3, \) and 4, respectively). For conditions with differential attrition, the probability of having missing values on Y was different for program and control conditions (when \( p = .10, .30, .50, .70 \) and \( p = .10, .15, .20, .25 \) respectively for \( X = 1, 2, 3, \) and 4, respectively). For conditions involving the inaccessible mechanism, missing data on Y were created exactly as described above, except that the cause of missing data was the score on Y rather than the score on X.

For each of these 80 data sets, we obtained correlations based on complete cases (listwise deletion), regression parameters based on complete cases, correlations based on the EM algorithm (using EMCOX; Graham et al., 1991), and regression parameters based on the EM algorithm. In addition, for the inaccessible mechanism, differential attrition pattern, we ob-

<table>
<thead>
<tr>
<th>Variable</th>
<th>Program</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>-0.25</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>-0.96</td>
<td>0.598</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: \( N = 500 \). Regression parameters for the dependent variable (Y) were as follows: intercept = 1.327; program = -.180; and X = .596.
tained a 20% random sample of the cases originally missing and analyzed those data using our correction for inaccessible missing data mechanisms (a modified version of the EMCOVEXE program was used). The results for this correction are also presented.

Finally, to illustrate the effects of obtaining different amounts of data originally missing, we conducted a separate simulation study in which 50 master data sets (each with \( N = 500 \)) were created with approximately the same population parameter values previously used \( (r_{12} = -.004, r_{13} = -.075, r_{23} = .522) \). For each sample, approximately 30% (an average of \( N = 138 \)) of the cases were set to be missing. We then obtained 5%, 10%, 15%, 20%, 30%, and 40% random samples of the initially missing cases and computed the adjusted EM estimates for the correlation parameters.

Simulation Results

**Main Simulation**

The results for the correlation parameter estimates appear in Table 2. There are several findings of interest in Table 2. First, the correlation parameter estimates based on the EM algorithm and complete cases were unbiased when there was truly no differential attrition. Second, when there was differential attrition, correlation parameter estimates were always biased when they were based on analysis of complete cases. Third, even with substantial differential attrition, EM algorithm estimates of the correlation parameters were unbiased as long as the missing data mechanism was accessible. EM algorithm estimates of the correlation parameters were biased only when there was both differential attrition and an inaccessible missing data mechanism. Finally, even with substantial differential attrition and an inaccessible missing data mechanism, the simple correction to the EM algorithm (denoted EM, in Tables 2 and 3) produced unbiased estimates of the correlation parameters.

The results for the regression parameter estimates appear in Table 3. The results for regressions mirror those for the correla-

<table>
<thead>
<tr>
<th>Differential attrition?</th>
<th>Accessible</th>
<th>Inaccessible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>EM</td>
<td>Complete</td>
</tr>
<tr>
<td>Yes</td>
<td>.000</td>
<td>-.246*</td>
</tr>
<tr>
<td>Estimate</td>
<td>.011</td>
<td>.009</td>
</tr>
<tr>
<td>SE</td>
<td>-.008</td>
<td>.004</td>
</tr>
<tr>
<td>No</td>
<td>.014</td>
<td>.016</td>
</tr>
</tbody>
</table>

Note. \( N = 500 \). Estimates greater than twice the standard error are considered statistically significant. Complete refers to analysis based on complete cases (listwise deletion); EM refers to estimates based on the expectation-maximization algorithm; EM<sub>C</sub> is the correction to EM estimates based on the sample of previously missing cases.

**Results for Different Size Samples**

In Tables 2 and 3, we included the results based on a correction of the EM algorithm calculated from a 20% random sample of cases initially missing. In Table 4, we expand somewhat the results for the correction, showing the results of a separate simulation study in which different random samples were drawn from those initially missing. As shown in Table 4, 5%, 10%, 15%, 20%, 30%, and 40% random samples all yield unbiased estimates of the key parameter estimates. However, as one

<table>
<thead>
<tr>
<th>% Sampled</th>
<th>n</th>
<th>( r_{13} )</th>
<th>SE</th>
<th>( r_{23} )</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>138</td>
<td>-.075</td>
<td>.041</td>
<td>.522</td>
<td>.032</td>
</tr>
<tr>
<td>40</td>
<td>55</td>
<td>.003</td>
<td>.046</td>
<td>-.006</td>
<td>.035</td>
</tr>
<tr>
<td>30</td>
<td>41</td>
<td>-.002</td>
<td>.043</td>
<td>-.018</td>
<td>.047</td>
</tr>
<tr>
<td>20</td>
<td>28</td>
<td>.004</td>
<td>.045</td>
<td>-.026</td>
<td>.052</td>
</tr>
<tr>
<td>15</td>
<td>21</td>
<td>.002</td>
<td>.051</td>
<td>-.022</td>
<td>.057</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>-.012</td>
<td>.062</td>
<td>-.017</td>
<td>.069</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>.015</td>
<td>.088</td>
<td>-.021</td>
<td>.084</td>
</tr>
</tbody>
</table>

Note. Data were based on 50 master data sets, each with \( N = 500 \). For each data set, an average 138 were set to missing. Correlations for no missing data (100% sampled) are actual correlations, not deviations. Standard errors are shown for each correlation deviation.
might expect, smaller samples were associated with more variability around the estimates. In this case, because there were 50 different master data sets, the measures of variability approximate the variability around the population values. That is, the standard errors shown in parenthesis in Table 4 are approximations of the missing data standard errors around the parameter values. For the smaller correlation corresponding to program effects in our simulation ($r_{13} = -.075$), the 20% ($N = 28$), 30% ($N = 41$), and 40% ($N = 55$) samples, each yielded standard errors that were rather similar to those based on no missing data. For the larger correlation ($r_{23} = .522$), however, only the standard error for the 40% ($N = 55$) sample approached the standard error calculated with no missing data.

Discussion

In this article, we have demonstrated very clearly that early thinking about attrition was incomplete. In the past, researchers have believed steadfastly that differential attrition meant that results (for example, beneficial program effects) were necessarily contaminated by attrition bias. In fact, the results for the correlation estimates would appear to bear out this early thinking. On the other hand, when one makes use of the EM algorithm or even when one uses regression analyses rather than correlation analyses, we have shown that there are situations in which there are clearly no problems resulting from differential attrition. Furthermore, it was clear from these results that in the true absence of differential attrition, even inaccessible mechanisms are not a problem.

The problem of differential attrition in the context of inaccessible attrition mechanisms, however, may or may not be a problem. Our results showed clearly that correlation and regression analysis, either by complete cases or EM algorithm, were seriously biased. However, our correction for the EM algorithm, based on collection of additional data from a random sample of previously missing cases, proved to be very useful in this situation.

The difficult part of the correction to the EM algorithm for inaccessible missing data mechanisms is not the statistical part. The major difficulties in using this procedure lie in the collection of the new data. We are not insensitive to the problems associated with collecting such data. Nonetheless, we suggest that researchers build such additional data collection into their projects whenever possible. The results we have shown here justify the expense of such additional data collection.

Although our simulations were based on a relatively simple analysis of a covariance model involving only three variables, it should be noted that this model is very common in evaluation research (e.g., for evaluating alcohol and drug prevention interventions). Although we did not present such results here, the conclusions drawn in the present study are easily generalizable to analysis of covariance models involving any number of covariates.

The Bottom-Up View of Attrition: What Is a Researcher to Do?

Everything we have discussed so far has been from the perspective of someone who knows whether there is differential attrition (on the main dependent variable) and whether the missing data mechanism is accessible or not. Unfortunately, the researcher never has this information. So what should a researcher do?

The problem facing the researcher with attrition is what the key parameter estimates would be if there had been no attrition. Because one does not have the data that are missing, one must make inferences based on the available information. The inferences one makes have to do with the missing data mechanism. The general point is to make inferences about the mechanism with the best precision possible.

Several strategies are available. Some strategies allow the researcher to be extremely confident about the inference made, whereas other strategies allow inferences about which the researcher has less confidence. The main difference between strategies is the degree of confidence the researcher may have regarding the missing data mechanism and ultimately regarding the important research conclusions. We now consider some of these strategies.

Collect additional data. Our recommendation is to collect new data whenever possible. The only way one can be sure about the missing data mechanism is to sample the missing data. Of the various strategies for collecting new data, the best is to obtain a random sample of the previously missing data. If one is able to obtain a random sample, one is able to have very high confidence about the missing data mechanism, as well as about study conclusions.

In our main simulation, we used a 20% random sample of those previously missing. In general, sampling theorists point out that the absolute number of cases sampled is more important than the percentage of cases sampled. Samples must be large enough (e.g., greater than about N = 30) to guarantee that the sample represents the population with sufficient precision (Backstrom & Hursh-César, 1981). Larger samples allow more precise parameter estimates (smaller standard errors), whereas smaller samples would allow less precise estimates. The results of our brief follow-up simulation illustrate these ideas. For the most important correlation (that corresponding to program effects), obtaining a sample of $N = 28$ (20% in this case) or greater yielded standard errors around the population parameter values that were roughly the same as the standard error obtained with no missing data.

In determining the size of the sample to obtain, it may reduce to an issue of statistical power. If power is not a concern, then even relatively small random samples would be sufficient to provide unbiased estimates of key program effect parameters. However, if statistical power is important, then it will be important to draw larger samples. In the current simulation, a little more than a quarter of the cases were missing. In this case, randomly sampling 30 cases from those initially missing appears to be sufficient for most purposes. Whether this will apply uniformly for other amounts of missing data remains to be shown.

In addition, one must balance the need for precision against the costs of collecting additional data. One way to determine the appropriate size of the sample would be to estimate what it would cost to obtain measures from the selected sample of previously missing cases (the difficult sample). For example, assume it costs $5 per subject for normal measurement proce-
ANALYZING DATA WITH DIFFERENTIAL ATTRITION

dures, and the researcher estimates that procedures required to
collect data from this difficult sample would cost five times as
much ($25 per subject). With these assumptions, measuring 20% of
the difficult sample would cost about the same as using
standard measurement procedures for the entire sample. The
advantage would be that for the same costs, the data obtained
are much more likely to generalize to the population of those
initially with missing data.

Use theory, logic, or prior data. In some situations, the re-
searcher may be rather certain, on the basis of strong theory or
on the basis of past research with similar populations, that the
major cause of missingness on the dependent variable is some
variable that has been measured or is some variable that is
unrelated to the dependent variable. In such situations, the re-
searcher may be justified in simply arguing that the missing
data mechanism is accessible and using standard missing data
procedures for accessible missing data mechanisms.

For example, suppose we have an adolescent drug prevention
or learning enhancement study and the major cause of attrition
is that families have moved away from the area. Suppose we
know, on the basis of school records, that the student was in no
trouble and that the attrition was very likely due simply to nor-
tmal transiency in the population. Suppose further that we
could include a measure of transiency (e.g., How many times
has your family moved in the past five years?) in our main
analyses. In this situation we would have greater confidence
that the major cause of missingness was accessible and that the
effects of attrition were minimal.

The only problem with this general strategy, even when the
argument is plausible, is that one cannot know that the mecha-
nism is accessible. Thus, one's level of confidence is relatively
lower than if one could obtain a new random sample of previ-
ously missing cases. The extent to which confidence is lower
depends, of course, on the strength of the theory or other argu-
ment that the missing data mechanism is accessible.

Other strategies for modeling the mechanism. Although it is
best if the new data are collected from a random sample of
those initially missing, there may be conditions under which
nonrandom samples are of value in making inferences about
those still missing. For example, suppose one collects data from
one additional sample with limited effort and then collects data
from a second sample with greater effort and then collects data
from a third additional sample with still greater effort. When
data are collected with different levels of effort, it may be possi-
ble to make reasonable projections about those who remain
missing. Strategies such as these should be explored in future
research.

Even when a researcher is unable to collect additional data,
there may be additional relevant data available. For example,
another study with similar subject characteristics may have col-
lected data from those initially missing from the posttest. On
the basis of such information, the researcher could make the
same kind of adjustment as suggested above for data collected
from the random sample.

Alternatively, one could make use of simulation data in con-
junction with data from previous studies to make reasonable
guesses regarding the population under study. Although ap-
proaches such as this may involve somewhat less confidence,
they may be better than making inferences based on the most
conservative and liberal assumptions possible.

Finally, one could simply make the most conservative possible
inferences. If beneficial program effects are still observed un-
der these conditions, one can be confident about these con-
clusions.

Make inferences based on data available for all cases. If it is
absolutely impossible to collect additional data of any sort, then
one can make inferences based on data that are available (e.g.,
pretest data). The strategies outlined by Hansen et al. (1985)
and Biglan et al. (1987) fall into this category. The disadvantage
of using these strategies is that the inferences are based on
correlates of the dependent variable rather than on the depen-
dent variable itself. If the available measures are highly corre-
lated with the dependent variable, then these analyses could
provide reasonable estimates of important research conclu-
sions under some conditions. However, if the available mea-
sures are not highly correlated with the posttest, then this type
of strategy is of less value.

As an example of how the Hansen et al. (1985) and Biglan et
al. (1987) strategies may be limited, we examined how well they
performed in our simulation. Of the 80 data sets in our simula-
tion, the Program \times Attrition Status interaction term was signif-
ificant (for the pretest variable) for 28 data sets. However, there
was attrition bias for only 8 of these 28 data sets. That is, when
the interaction was significant, the test was correct only 29% of
the time. This problem occurred despite the relatively large
correlation ($r = .60$) between $X$ and $Y$.

On the other hand, the interaction was nonsignificant for 52 of
the 80 data sets in our simulation. There was attrition bias for
12 of these 52 datasets. That is, when the interaction was non-
significant, the test was correct 77% of the time.

In summary, Hansen et al.'s (1985) and Biglan et al.'s (1987)
strategy may be helpful if its result is no differential attrition,
but it is not helpful if its result is that there is differential attri-
tion. Although our simulation was by no means a systematic
evaluation of these procedures, it is clear that they are limited as
a general solution for attrition problems.

Dent's (1988) procedure. One final procedure that makes
use of data that are available (e.g., pretest data) was described by
Dent (1988). Dent suggested using available data to predict
missingness on the $Y$ variable ($Y_{mis}$):

$$
\hat{Y}_{mis} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k
$$

The single predicted score for $Y_{mis}$ is then included in all
program effects analyses, for example:

$$
\hat{Y} = b_0 + b_1\text{Program} + b_2\hat{Y}_{mis}
$$

This procedure can correct for any accessible causes of attri-
tion, but does not correct for the inaccessible causes. Unfortu-
ately, without collecting additional data, one can never know
the extent to which controlling for the accessible causes con-
trols for all causes of missingness on $Y$. Still, one may be able
to make use of data collected from a similar sample, strong theory,
or common sense to argue that the inaccessible causes of mis-
ningness are minimal in a particular case. The effect of such
alternatives to collecting additional data would be to reduce the
researcher's confidence in the obtained results. The extent of
reduction in confidence would, of course, vary from study to study.

Applications

We have discussed participant attrition in relation to a substance use prevention example. However, the implications of this article are far reaching. When substantial participant attrition is discovered in any study, the procedures described above give researchers a clear strategy for approaching the problem. In general, the EM algorithm, and even standard regression procedures, can often provide unbiased conclusions, even in the face of substantial differential attrition. In addition, the strategy of collecting a random sample of those with initially missing data and correcting the EM algorithm estimates provides unbiased conclusions in all attrition situations.

We agree completely with those who argue that minimizing attrition in the first place is important (e.g., Capaldi & Patterson, 1987). However, we also believe that eliminating attrition entirely in intervention research may ultimately be an impossibility. Therefore, it is our hope that when elimination of attrition is not possible, procedures such as the ones outlined in this article will enhance the usefulness of evaluation data from psychological interventions regardless of the attrition problems involved.

References


Received April 16, 1991
Revision received June 19, 1992
Accepted June 19, 1992.