Comments on the Analysis of Data with Missing Values

T. Mark Beasley, Guest Editor
St. John's University

In the **Orsak, Mendro, and Weerasinghe** article (*pp*. 3-12), the authors search for an acceptable methodology for estimating missing student post-test scores within a school effectiveness analysis. It appears that the current methodology involves Listwise Deletion of data which has notable problems, especially when data are missing on a systematic basis. Thus, the authors attempt to answer, "How could we effectively rank the school of interest without complete data for its constituents?" (*p*. 3). More succinctly, this question could be posed as, "Can we find a method better than Listwise Deletion for calculating School Effectiveness Indices (SEIs)?" Thus, it would seem that Listwise Deletion should have been included as a method for handling missing data. Listwise Deletion is noted for simply reducing statistical power when data are missing randomly (Hartley & Hocking, 1971). In this case, however, the statistical power of a test statistic is not of interest. Rather, the accuracy of predicted values used to replace missing values is the central issue. Thus, although the properties of Listwise Deletion could be examined in terms of SEI accuracy, these properties could not be investigated at the *data* level. With multiple variables, Listwise Deletion may lead to a severe loss of complete-case data. Thus, one may assume that any unbiased estimate would be better than nothing at all (Frane, 1976). When data are missing systematically, however, serious biases may occur (Little & Rubin, 1987). Therefore, one response to this article is another question: "What if missing data is correlated with the index of SEI?" For example, "Do low performing schools have more missing data?" Or, "Is missing data correlated to other factors such as SES?"

Another important issue involves whether more complex imputation models provide better estimates when higher percentages of data are missing. The authors conclude that the more complex models (especially HLM) provide more accurate estimation of the original data for greater percentages of missing data (see *p*. 11). Intuitively this seems reasonable; however, despite these claims, this increased accuracy does not manifest itself to an overwhelming extent in the results. Perhaps the similarities among these regression-based approaches can potentially be attributed to the replaced data being initially missing on a random basis. Furthermore, concerning the SEI correlations, it must be considered that the replaced (missing) values are entered into a second linear composite to compute SEIs. In general, quantitative estimates based on sums should be unbiased if data are missing randomly. Therefore, based on the Central Limit Theorem, SEIs should be normally distributed and unbiased asymptotically when data are missing randomly. This may also help explain the similarity of OLS and HLM when the correlation of their respective SEIs is examined.

Another point of contention is that a clear distinction between *statistical models* and *estimation procedures* is necessary. Although not frequently elaborated at the MLR: GLM SIG (except by Randy Schumacker), HLM can be performed using GLM interaction terms and Ordinary Least Squares (OLS) solutions. Dayton's (1970) excellent chapter on nested designs elaborates this approach. Therefore, the distinction between a HLM and OLS regression solution is, in many cases, the difference in what algorithm is used to estimate parameters. The confusion arises because the most noted HLM software uses Empirical Bayes (EB) estimation, whereas most linear regression modules in other statistical softwares provide an OLS estimation of parameters. The authors should consider this issue when claiming that the "three models indicate that HLM is more suitable for estimating missing data than OLS or the average school score. This advantage must be gained by HLM's adjustments for school trends in comparison to overall trends for student score" (*p*. 11).

First of all, regression procedures that include interaction terms can make adjustments for school (Level 2 or Outer Level) trends. Furthermore, the results for Models 1 (HLM) and 1A (OLS) are only slightly different which can be attributed to the HLM and OLS models being identical random effects models. Although fixed effects linear regression models make no assumptions about the form of the predictor variables, when predictor variables are treated as random effects, as in this case, normality is assumed and the distributional shape of the predictor is critical in terms of the accuracy and efficiency of the regression model. Importantly, the predictor variable in Models 1 and 1A (MATH95) is probably close to being normally distributed. Therefore, the distinction between the EB and OLS estimators would not be expected to be great.

By contrast, the authors report that the "which is best" decision leaned more clearly to HLM for the Model 2 analysis. However, this may not be attributable to the HLM approach. Rather it may be due to EB estimation procedure. That is, Models 2 and 2A do not have "nested" or hierarchical structures. Thus, the difference in the results for these random

effects models may be due to the superiority of the EB estimators over OLS for the added predictor variables (i.e., Percent block poverty (POV) and Percent block college (COL)), both of which are likely to be skewed. Thus, only the Model 3 results are convincing in demonstrating a definite advantage of the HLM approach with EB estimation. The possibility remains, however, that this advantage could potentially dissipate if interaction terms are created so that the OLS regression could model the hierarchical structure of the data. Therefore, OLS regression should still be considered as a viable method for estimating missing data. Fortunately, both Mundrom and Whitcomb (*pp*. 13-19) as well as Brockmeier et al. (*pp*. 20-39) also investigate the properties of regression-based imputation procedures.

Mundrom and Whitcomb (*pp*. 13-19) note that physicians often use empirically derived classification functions to make important decision concerning the treatment or transport of the patient. Unfortunately patient data is often missing. In situations where a classification function or prediction equation is being estimated, missing data may lead to less statistical power or biased estimates. By contrast, in the medical decision scenario, the classification function has already been derived. Therefore, missing data preempts the decision. To use the classification function for making a decision about a patient's status, the missing data MUST be replaced. This differs from the Orsak et al. article in that SEIs could be estimated if missing data were deleted. Thus, Mundrom and Whitcomb examine efficient ways to estimate a replacement value for a classification function when a patient has missing data.

Because of the practical nature of this problem, parsimony is an issue. That is, a physician who uses the classification function wants the best prediction with the least effort or complexity. To examine the problem, Mundfrom and Whitcomb systematically deleted each value of an existent data set ($N = 99$) then replaced each deleted value with one of three values (Mean Substitution, Hot-Deck imputation, Multiple Regression imputation). Next, the data were submitted to two different classification functions in order to examine which missing data approach was better in terms of making the "correct" decision. This procedure was completed for each variable in the classification function. (see *p*. 15) This problem in missing value analysis, the methodology, and the results lead to many speculations and comments.

In general, Mean Substitution is considered one of the worst things to do when data are missing. This distrust is based on the use of Mean Substitution in developing statistical models not its use as a decision making tool. Typically, Mean Substitution is criticized because it gives no leverage to the replaced values (Frane, 1976). When there is a substantial number of missing values, mean substitution reduces the average leverage (i.e., Pearson correlation). Mean Substitution also reduces the average squared deviation (i.e., variance) which may create a restriction of range issue. The Mean Substitution method in this application, however, is a *ceteris paribus* approach. That is, all things being equal, what is the decision? This is because each coefficient is partialled and the predicted value of any score at the mean of a variable does not raise or lower the predicted value (i.e., regression surfaces always intersect the centroid). Thus, the approach implies, "If we do not know the information, let's substitute the mean because it will not influence the decision or predicted value."

In practice, the Hot-Deck imputation procedure involves randomly selecting a data value from the existent distribution of the variable for replacement. Therefore as the authors note, the results vary from one selection to another. This is the *danger* of using the Hot-Deck procedure especially with variables with large dispersion. In terms of this study, one would never know whether in practice a physician would select the same value (in one replication) as did the simulation researcher. To address this issue, the authors aggregated the results of the Hot-Deck imputation over 1,000 replications. However, the average of 1,000 replication makes the results of the Hot-Deck procedure identical to Mean Substitution asymptotically. That is, with 99 values and 1,000 replications, the average Hot-Deck imputed value should be the *expected value* of the variable which *IS* the Mean Substitution procedure. Therefore, investigating the properties of Hot-Deck imputation is problematic given the authors' simulation methodology. This issue could be addressed by randomly generating multiple (e.g., 1,000) samples of 99, rather than using one sample of 99 repeatedly. Furthermore, because Hot-Deck imputation involves replacing the missing datum with a randomly selected value from the existent data set, it tends to rely on the shape of the distribution. If the variable is normally distributed the randomly selected value is likely to be near the mean and Hot-Deck imputation should perform similarly to Mean Substitution. When the Hot-Deck results were aggregated over 1,000 replications the results tended to be similar to Mean Substitution regardless of distributional shape because of the Central Limit Theorem. The Multiple Regression approach performed surprisingly poorly relative to the other two procedures. This truly makes it unattractive given that it is the most complicated of the three procedures.

From a realistic perspective, the relative costs of making a Type I (sending the patient to a city hospital) or Type II (keeping the patient in the rural hospital) should also be considered. It could be

beneficial to replace the missing value with an extremely low, but plausible value (i.e., best case scenario) and then with an extremely high plausible value (i.e., worst case scenario). Then the physician could evaluate whether the decision changes based on these extremities. Similarly, one might investigate that given all the existent data, at what point does the decision change and how plausible is that replacement value? Of course this approach would be dependent on the variability and predictive importance of the variable. However, all three of the imputation procedures are dependent on these two factors. For example, the authors note that the Syncope variable was least affected by any imputation method. Perhaps this was because it was the strongest partial predictor or because it had the least variance. Certainly, it would seem that Mean Substitution and Hot-Deck imputation may not work well with variables with a great deal of dispersion. However, it would seem that some data may be so crucial (strong partial correlation) that a valid or accurate decision can *NOT* be made without it. In such a case, classification accuracy would be a function of the "importance" of the predictor. The performance of the imputation procedures for missing data on these important or crucial predictors should be investigated. Also, the variability of predictor variables should be examined because Mean Substitution may not perform as well with highly disperse predictors. Thus in general one must ask, "Would the imputation procedure perform differently if the variables were of different importance (had differing partial relationships)?" As was also the case with the Orsak et al. article, one must wonder whether in reality the data would be missing on a random basis. For example, is the fact that the patient has a missing Heart Sound Reading indicative of some other factor (e.g., the type of insurance coverage)? From a practical perspective, it would be important to examine whether there are "proxy" variables that are not in the final regression solution (or classification function) but could be used to impute missing values. Such variables do not necessarily have to be related to the outcome (else they would be in the regression solution), but they should be related to the predictors so that they can take their place and be used to impute missing values. As was also the case with Orsak et al., the Mundfrom and Whitcomb should certainly consider examining how Mean Substitution and other imputation procedures perform when data is missing systematically.

Thus, before regression procedures can be applied in practical decision-making situations, there is a need for studies like the one conducted by **Brockmeier, Kromrey, and Hines** (*pp*. 20-39) that address the issue of systematically missing data. However, the issue of predictor variability and/or importance becomes a concern when interpreting their findings. As is often noted, if data are missing at random then the reduced number of cases is simply a power issue and most methods yield similar results (Little & Rubin, 1987). This again leads to the questions that have been asked about the two previously reviewed articles: "How do researchers know when data is missing?" and "How can they be sure that the pattern of missing data is random?"

Most substantive researchers agree that data is rarely missing on a random basis. Despite this consensus, however, the authors accurately eschew the all too common avoidance of investigating the extent and nature of missing data. Rather, many researchers choose to simply delete missing data either purposely because it is convenient or inadvertently because it is the default of most statistical software. As researchers and statistics educators, we should reinforce that data screening is not simply ritualistic behavior that we learned in graduate school. Rather, carefully examining the data for outliers and missing data patterns is paramount in terms of researchers becoming familiar with their data and investigating whether any missing data may create a bias in the interpretation of their results. Specifically, one can determine whether the data is missing systematically by examining whether a dummy-coded variable (e.g., 1 = nonmissing, 0 = missing) is related to other collected variables. If it is related to variables that will be potentially included in the regression model then systematically missing data may result in biased parameter estimates and ultimately to a specification error. If the dummy-code is related to variables not in the model (e.g., SES), external validity may be limited. In either case, the interpretation of the results is compromised.

After concluding that the missing data pattern is systematic, one of many missing data approaches may be selected. Thus, the authors examine the properties of several of these procedures. As is the case with most newer advances in statistical methodology, however, multiple imputation and maximum likelihood approach are not utilized frequently due to lack of accessible software. Likewise, stochastic imputation is not frequently used either which may also be due to a lack of software accessibility. Thus, it is important that the authors included their algorithms in the Appendix (*pp*. 38-39). Possibly, the trend to ignore missing data will reverse with new statistical modules such as SPSS 8.0 Missing Value Analysis. Based on my experience, however, such a convenient module is alarming because of the potential for misuse.

In terms of their methodology, I must sympathize with these researchers because there are so many variables that can be manipulated when simulating a regression model. For preliminary work, I agree with the author's decision to investigate

the standardized regression model. If raw score models were investigated then other variables such as the variance of the predictors could be manipulated thus increasing the number of simulation conditions and in general making investigation and interpretation more complicated. Also, in educational research, standardized models are more common; however, the authors should consider that many researchers would be interested in how these approaches to handling missing data would affect the *Y*-intercept. In any case, the accuracy of estimating the standardized regression parameters of β_1, β_2 and Population R^2 (i.e., ρ^2) in a two-predictor model was investigated. In terms of Monte Carlo studies, statistical hypothesis testing, and therefore investigating whether Type I error rates remain near an expected nominal alpha level, has been the bread-and-butter of simulation researchers. Furthermore, given that statistical hypothesis testing is not going away any time soon (see Robinson & Levin, 1997), I would suggest that the authors consider simulating complete and partial null structures and then investigating Type I error rates for each parameter. However, given the task at hand (i.e., estimation accuracy) perhaps coverage probabilities for confidence intervals constructed for each parameter could suffice. This would allow an investigation of whether systematically missing data biases the accuracy of parameter estimates and the coverage probabilities of their confidence intervals. To elaborate, if a 95% confidence interval is constructed in multiple replications, the confidence interval should cover the population parameter 95% of the time regardless of its value (i.e., whether it is a null or non-null structure). By taking this approach, one could examine the potential bias in: (a) coverage probabilities (i.e., Does the confidence interval cover the population parameter?); (b) power (i.e., Does the confidence interval cover 0 with a non-null structure?); and (c) Type I error rate (i.e., Does the confidence interval cover 0 with a null structure?).

Despite the absence of a null structure, the authors do present two interesting regression structures. For the 6th grade data, there is a "dominant" predictor (see Table 1, *p*. 21). By contrast, both predictors are equally related to Y in both a zero-order and partial sense for the 9th grade data (see Table 2, *p*. 21). Thus, the issue of predictor variability and/or importance becomes a concern in the interpretation of the results. I have taken the liberty of constructing a very simple summary table of the results for estimating population R^2. Interestingly, Listwise Deletion tended to underestimate ρ^2 when the predictors were equally related to *Y*. Having a large percentage of data that are missing above the mean for the predictor variables created a more serious underestimation, possibly because these missing values have the most influence or leverage. Furthermore, this situation creates a restriction of range problem.

When one predictor was "dominant," one of the predictors tended to "take over" in terms of estimating ρ^2 as a summary measure. That is, there seems to have been some compensatory process. Similarly, it would seem that Pairwise Deletion would lead to a compensation because the remaining *X*-*Y* coordinates that are not affected by the missing data are still used. The results, however, showed that Pairwise Deletion typically underestimated ρ^2.

Similar to many other studies, Mean Substitution seemed to be the worst method for estimating regression parameters. Both Deterministic and Stochastic Mean Substitution procedures tended to underestimate ρ^2 raising interpretative issues similar to those concerning using Listwise Deletion. As previously mentioned, replacing values with the mean reduces the average leverage (i.e., correlation) and the variance (i.e., average squared deviation) so that less variance is available to be shared. These problems worsen as the percentage of data missing above the mean increases.

The results for the regression-based imputation procedures create an unusual situation. Both Deterministic and Stochastic Simple Regression imputation approaches typically resulted in the underestimation of ρ^2. One may interpret this from the perspective that since the relationship of *Y* to the missing data was not included in estimating a replacement value, not all relevant information was included. By contrast, both Multiple Regression approaches overestimated ρ^2. One perspective on this is that by including the *Y* relationship to the missing data one increases the likelihood of capitalizing on chance relationships. Furthermore, Deterministic Regression approaches have been reported to "overfit" the data because missing scores are predicted without error (Allison, 1987; Little, 1992). Thus, it would seem that Stochastic Multiple Regression would tend to reduce the amount of overestimation. Although this is not always the case in these results, the Stochastic Multiple Regression procedure performed the best in terms of estimating ρ^2.

These "which is best" results carried over to the estimation of standardized regression coefficients for the most part. In general, increasing amopunts of missing data on X_1 resulted in an increasing underestimation of β_1 for most methods. Also for the standardized regression coefficients, there seems to a "compensatory" process for most of the missing data approaches. That is, where β_1 was *underestimated* β_2 tended to be *overestimated* and vice versa. This compensatory process should be used for aiding interpretation. That is, one should consider the *degree* of over and under estimation in context with which variables have missing data and with what other variables the missing data is correlated.

Summary of Results for Estimating ρ^2 from Brockmeier et al.

Method	Estimation of ρ^2
Listwise Deletion	**Underestimates** with equivalent predictors (Table 4). With a dominant predictor, estimation is better (Table 5).
Pairwise Deletion	**Underestimates**
Deterministic Mean Substitution	**Underestimates**
Deterministic Simple Regression	**Underestimates**
Deterministic Multiple Regression	**Overestimates**
Stochastic Mean Substitution	**Underestimates**
Stochastic Simple Regression	**Underestimates**
Stochastic Multiple Regression	**Overestimates**

It is interesting that these researchers reported that the relative effectiveness of the missing data treatments in this study with systematically missing data were similar to their results obtained with randomly missing data (e.g., Brockmeier, Kromrey, & Hines, 1995, 1996). As they aptly note, however, these results may be due to the particular covariance structures used in this investigation (*p*. 34). Thus, deliberations over whether it is reasonable to assume that data are missing at random may be inconsequential in terms of estimating replacement values. However, I suspect that the efficacy of procedures to handle missing data is complex and depends on (a) the relationships among the criterion variables and predictors, (b) the predictor intercorrelation/covariance matrix, and (c) whether any relationships to data being missing are strong. Furthermore, these issues will become more complicated with more than two predictors.

Address correspondence to:
T. Mark Beasley
School of Education
St. John's University
8000 Utopia Parkway
Jamaica, NY 11439
E-Mail: beasleyt@stjohns.edu

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological Methodology*. San Francisco: Jossey Bass.

Brockmeier, L. L., Kromrey, J. D., & Hines, C. V. (1995, April). *Effective missing data treatments for the multiple regression analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Brockmeier, L. L., Kromrey, J. D., & Hines, C. V. (1996, April). *Missing data treatments for nonrandomly missing data and the multiple regression analysis*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Dayton, C. M. (1970). *The design of educational experiments*. New York: McGraw-Hill.

Frane, J. W. (1976). Some simple procedures for handling missing data in multivariate analysis. *Psychometrika*, *41*, 409-415.

Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics*, *27*, 783-823.

Little, R. J. A. (1992). Regression with missing *X*'s: A review. *Journal of the American Statistical Association*, *87*, 1227-1237.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley and Sons.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21-26.