Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model

MLRV

Volume 29 • Number 1 • Spring 2003

Table of Contents

Multiple Linear Regression Viewpoints

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: General Linear Model through the **University of Alabama at Birmingham**.

Subscription and SIG membership information can be obtained from:
Jeffrey B. Hecht, MLR:GLM/SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854.
jbhecht@niu.edu

MLRV abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook*.

Some Graphical Methods for Interpreting Interactions in Logistic and OLS Regression

Peter L. Flom **Sheila M. Strauss**
National Development and Research Institutes, Inc.

In statistical models involving one dependent variable (DV) and two or more independent variables (IVs), an interaction occurs when the effect of one IV on the DV is different at different levels of another IV. The existence of an interaction makes interpretation of the model more complicated, but failing to include important interactions in the model can give misleading results. In this paper, we describe how visually examining interactions between two IVs in ordinary least squares regression and in logistic regression can aid comprehension of the interaction, and we present a tool to make such examination easier.

In modeling the relationship between a dependent variable (DV), Y, and a set of independent variables (IVs), X_1, X_2, ….X_k, if the DV is continuous and we are using ordinary least squares regression we have:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \qquad (1)$$

When the DV is dichotomous, OLS regression is inappropriate; probably the most common alternative is using logistic regression (Hosmer & Lemeshow, 2000) where we have

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + ...\beta_k x_k + \varepsilon \qquad (2)$$

where $\pi(x)$ is the conditional mean of Y given X. Given that Y is dichotomous, this is the same as the probability that Y = 1 (assuming that Y is coded 0, 1), or the probability of a 'success' if Y is coded 'failure/success'. The portion before the equals sign is known as the logit. Equivalently, we can model $\pi(x)$ directly as

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + ...\beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + ...\beta_k x_k}} \qquad (3)$$

Models (1.1) and (1.2), and, more generally, any of the class of generalized linear models (McCullagh & Nelder, 1989), assume (among other things) that the effect of each X_i (i = 1, 2, …, k) on Y is the same, regardless of the value of the other X_j (j = 1, 2, …, k; $i \neq j$); that is, that there is no interaction. We may suspect that this is not the case. Earlier research may have found interactive effects; or we may have other substantive reasons for suspecting interactions. For example, if we are examining the likelihood of being HIV positive based on a person's sex and sexual identity, we would include an interaction between the two, since the effect of being homosexual is greater for males than for females. This is usually done by adding an interaction term to the equation. While there are many possible ways to construct such a term, the most usual, and simplest, is to multiply the two IVs that we think may be involved by each other, and add that term (Harrell, 2001). For now, let us suppose (for simplicity) that our model contains only two independent variables: X_1 and X_2. Set $X_3 = X_1X_2$ and add it to the model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad , \qquad (4)$$

which allows for an interaction.

While many researchers recognize the importance of evaluating interactions, interpreting interactions is often difficult and sometimes counterintuitive. This can be so for at least two reasons. First, because the parameter for interaction is multiplied by both IVs, what appears to be a small coefficient can actually be highly meaningful, especially if one of the IVs is large. Second, when the signs of the parameters on the interaction term and the IVs are not all the same, the effect of a change in one of the IVs on the DV may not be readily apparent.

The use of graphics can facilitate the interpretation and presentation of interactions; indeed, graphics can facilitate interpretation of the results of logistic regression even in the absence of interactions (Long, 1997). While many guides to model building, variable selection, and significance testing are available (e.g. Harrell (2001), this is not the case for graphical methods of interpreting interactions. In this paper, we therefore present some graphical methods for displaying and interpreting interactions between two independent variables, at least one of which is continuous. We provide methods for both continuous and dichotomous DVs. These graphs allow the user to examine the effect of a change in the IVs on the DV directly, without any computations. While these methods are not novel, they are under-utilized, and we are unaware of any source, which describes all of the methods described below, or their application specifically to interaction. These methods all involve plotting curves; for ordinary least squares regression, each of these curves is derived from a relatively simple equation; for logistic regression, the curves are considerably more complex. Graphs of this type have two types of uses: First, they may aid one's own analysis of data. Second, they may allow easier presentation of these findings to others.

Data

Data for this paper are drawn from the Drug Use and HIV Risk Among Youth (DUHRAY) project. DUHRAY involved a probability survey of 18-24 year old household-recruited youth in Bushwick, a low-income minority neighborhood in Brooklyn, New York with a population of approximately 100,000 in 1995. It sampled two groups of Bushwick-resident young adults: (1) a population-representative multistage household probability sample; (2) a targeted sample (Watters & Biernacki, 1989) of youth who use heroin, cocaine, crack, or inject drugs. Details of the sampling plan are available elsewhere (Flom et al., 2001).

Graphical Methods

The best method of interpreting interactions depends on the nature of the variables involved in the model. Specifically, it depends on whether the DV and IVs are dichotomous or continuous. In this section, we first present methods for a variety of types of models with two IVs, where at least one of the IVs is continuous. When both IVs are dichotomous, graphical methods are not necessary, and crosstabulations can be very useful. This is so because, in a 2 x 2 crosstabulation, it is relatively straightforward to determine main effects and interactions by hand calculation. We then present some possible extensions to cases where there are more than two IVs.

Continuous DV, One Continuous and One Dichotomous IV

In a regression model having a single DV and a single IV, a scatterplot of the IV and the DV is often useful. When we add a dichotomous IV to the model, we can make a scatterplot with two lines, one for each level of the dichotomous IV.

Example 1: In DUHRAY, we created a variable for peer objection to drug use (DROBJ), based on a factor analysis of five questions, each of which asked what proportion of your friends would object if you used a particular drug. The five drugs were marijuana, cocaine, heroin, crack, and injected drugs. We also asked about recalled childhood misbehavior, using a scale based on one devised by Windle (1993). We then modeled peer objection to drug use as a function of childhood misbehavior and sex (1 for male, 2 for female), using ordinary least squares regression, and including an interaction term. The estimated equation was

$$\textbf{DROBJ = -0.14 - 0.0038 Win + 0.91 SEX - 0.036 SEX*Win} \qquad (5)$$

This yields Figure 1, from which it can be seen that, while objection to drug use decreases as reported childhood misbehavior increases, it does so faster for women than for men. Also, while women, on average, reported more objection to drug use than men did, (the mean for men was -0.10, *SD* = 0.97; for women mean = 0.15, *SD* = 1.00) the opposite was true when there was a lot of reported misbehavior. If there were no interaction, the lines would be parallel.

Dichotomous DV, One Continuous and One Dichotomous IV

Example 2: If the DV is dichotomous, we simply use the results of logistic regression rather than OLS regression to create a graphical display. For example, we modeled using hard drugs (HD) (cocaine, heroin, crack, and/or injected drugs) in the last year (yes=1, no=0) as a function of sex (male = 1, female = 2) and childhood misbehavior (win). The estimated equation was:

$$prob(HD) = \frac{e^{.354-.247*sex+.351*win+.0638*sex*win}}{1+e^{.354-.247*sex+.351*win+.0638*sex*win}} \qquad (6)$$

Figure 1: Objection to drugs as a function of sex and childhood misbehavior

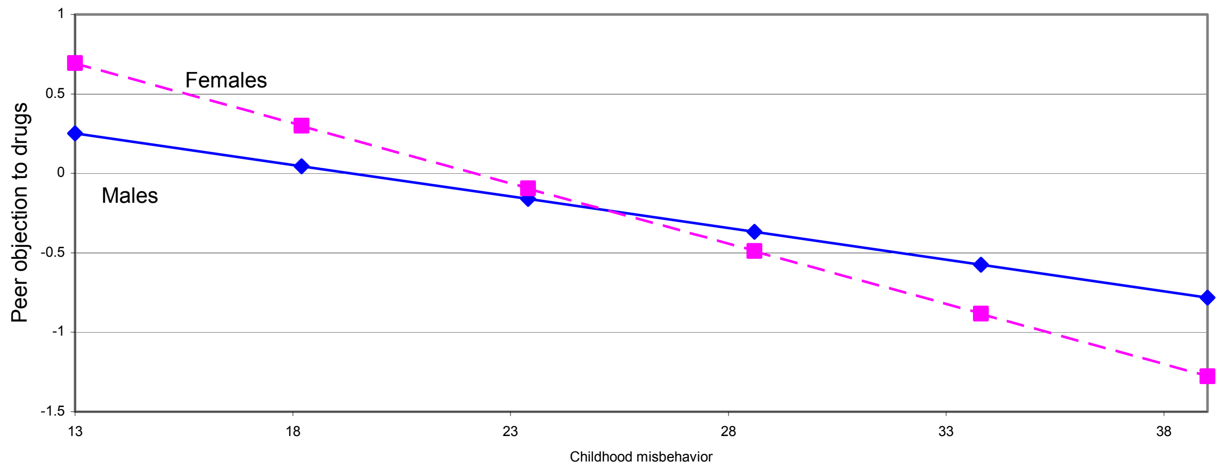


Figure 2: Probability of hard drug use as function of Windle, and sex

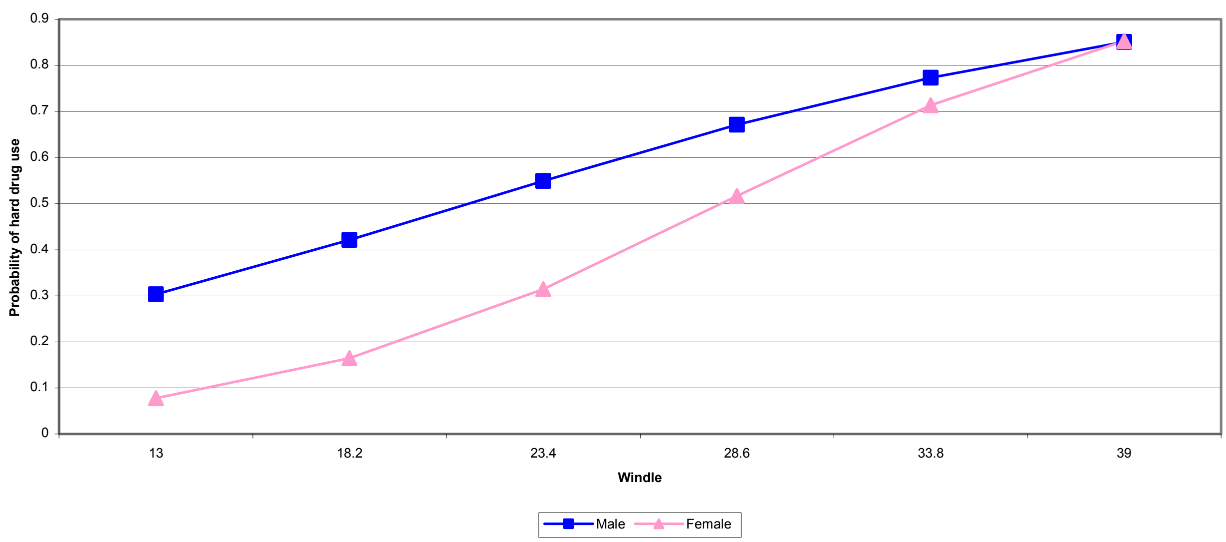


Figure 3: Objection to drugs as function of age (X) and Windle

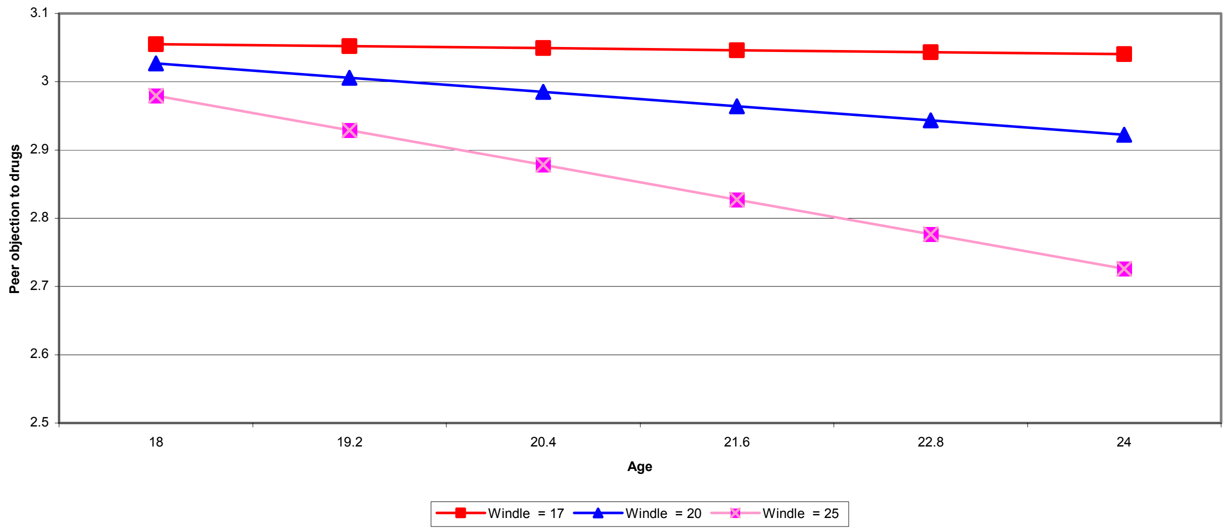
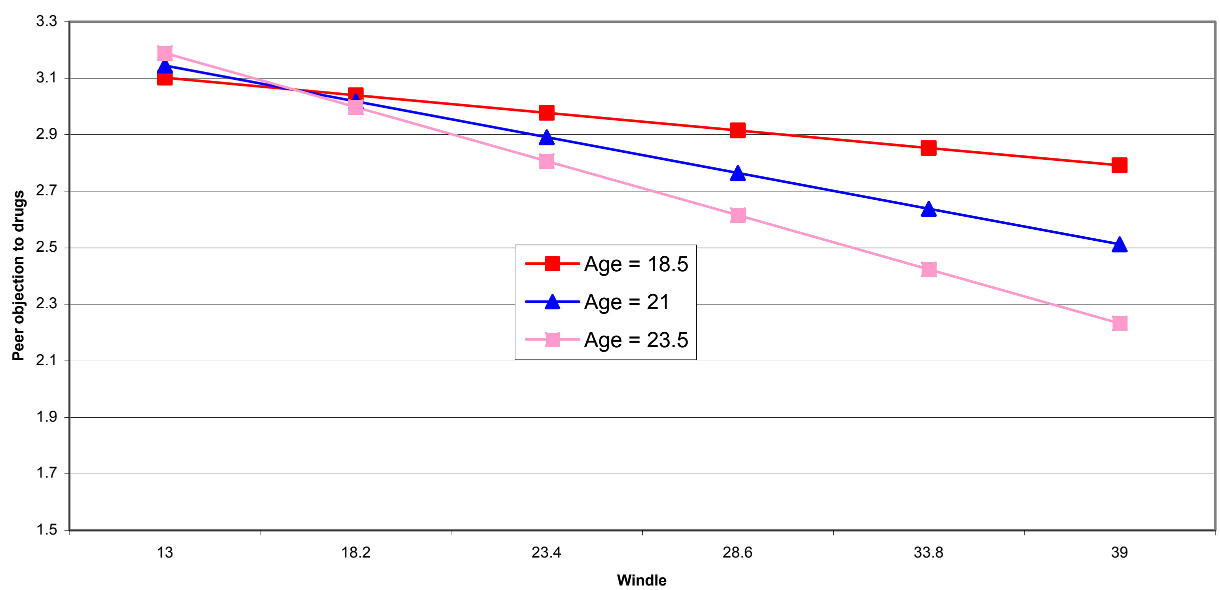


Figure 4: Objection to drugs as function of and Windle (X) and age



This yields Figure 2. If there were no interaction, the two curves would be parallel. As is, we can see signs of a moderate interaction. Although the likelihood of having used hard drugs increases with childhood misbehavior for both males and females, it increases faster for females.

Continuous DV, Two continuous IVs

When both IVs are continuous, there is no statistical reason for deciding which IV to put on the X-axis. There may be substantive reasons for choosing one, but we may need two plots to get a full sense of the interaction. One plot will have one IV on the X axis; the other plot will have the other. Also, in this case, we need to use more than two lines on the scatterplot. The exact number of lines depends on the distribution of each variable, but three is often a good compromise between comprehensibility and completeness. More lines can clutter the page, and fewer lines give an incomplete picture of the changes in the relationship. We need to pick representative values of each IV. One possible set of choices (used below) is the 25^{th}, 50^{th}, and the 75^{th} percentiles.

For example, we modeled peer objection to drug use as a function of age and childhood misbehavior. The estimated equation was:

$$\textbf{DROBJ} = \mathbf{-1.48 + .12*AGE + 1.24*Windle - .063*AGE*Windle} \qquad (7)$$

First, we let age be the X variable, and pick the three values of Windle at the 25^{th}, 50^{th}, and 75^{th} percentiles. This yields Figure 3 from which it can be seen that, while objection to drug use decreases as age increases, it does so much faster for those who reported more childhood misbehavior.

On the other hand, if we let Windle be the X variable, and choose values of age at the 25^{th}, 50^{th}, and 75^{th} percentiles, we get Figure 4, from which it can be seen that the relationship between childhood misbehavior and peer objection to drug use is stronger for older subjects.

Dichotomous DV, Two Continuous IVs

If the DV is dichotomous, we again modify the above procedure by using logistic regression. We modeled the probability of having used hard drugs in the last year (HD), by age and childhood misbehavior. The estimated equation was:

$$prob\left(HD\right) = \frac{e^{5.33-.41*age-.44*win+.027*age*win}}{1+e^{5.33-.41*age-.44*win+.027*age*win}} \qquad (8)$$

With age on the X axis, this yields Figure 5. This indicates that the relationship between age and drug use is stronger for subjects with higher levels of childhood misbehavior (because the slope of the line for Windle = 25 is greater than that for lower values of Windle). If there were no interaction, the lines would be parallel.

Similarly, if we put misbehavior on the X axis, and make separate lines for different ages scores, we get Figure 6. This implies that the relationship between childhood misbehavior and peer objection to drug use is stronger for older subjects.

Discussion and Conclusions

In this paper, we have presented a tool for displaying the effects of interactions involving two independent variables. While this tool is not innovative, making it more widely known and more easily implemented will, we believe, increase understanding about the nature of an interaction. In addition, it has the potential to clarify how changes in various parameters in logistic and ordinary least squares regression affect the relationship between a dependent variable and two independent variables.

While most statistical software (including SPSS, S-Plus, R, or SAS-Graph would allows production of charts similar to those produced in this paper, the graphics presented here were developed in Microsoft EXCEL. Because this software is readily available, allows the user to adjust figures and immediately see the results, and requires no programming skill, utilizing the graphics approach developed here with Microsoft EXCEL will be possible for many data analysts. It should be noted that Excel was not used to calculate the equations, but only to plot the results.

Figure 5: Probability of hard drug use as function of age (X), and Windle

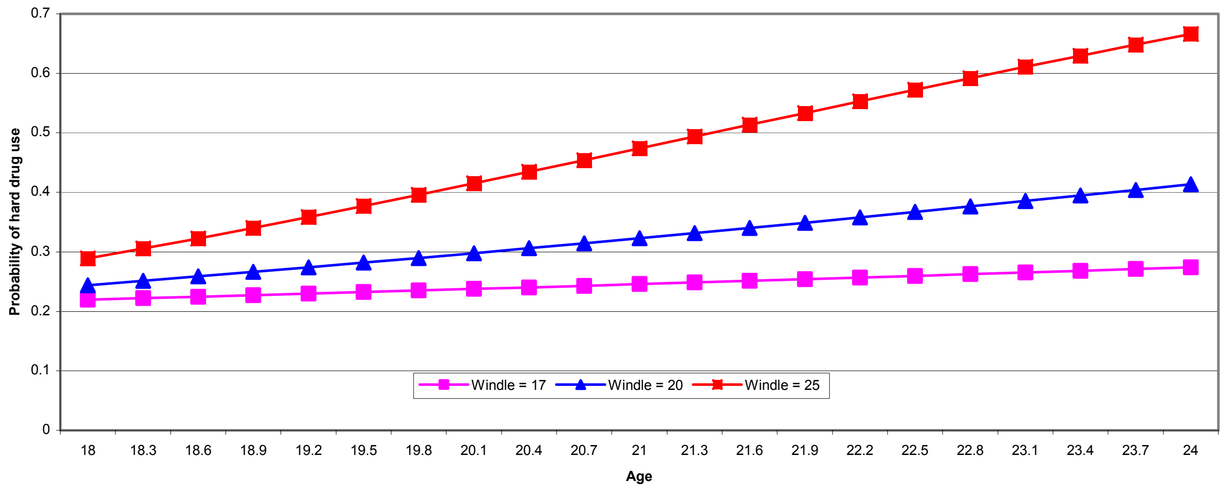
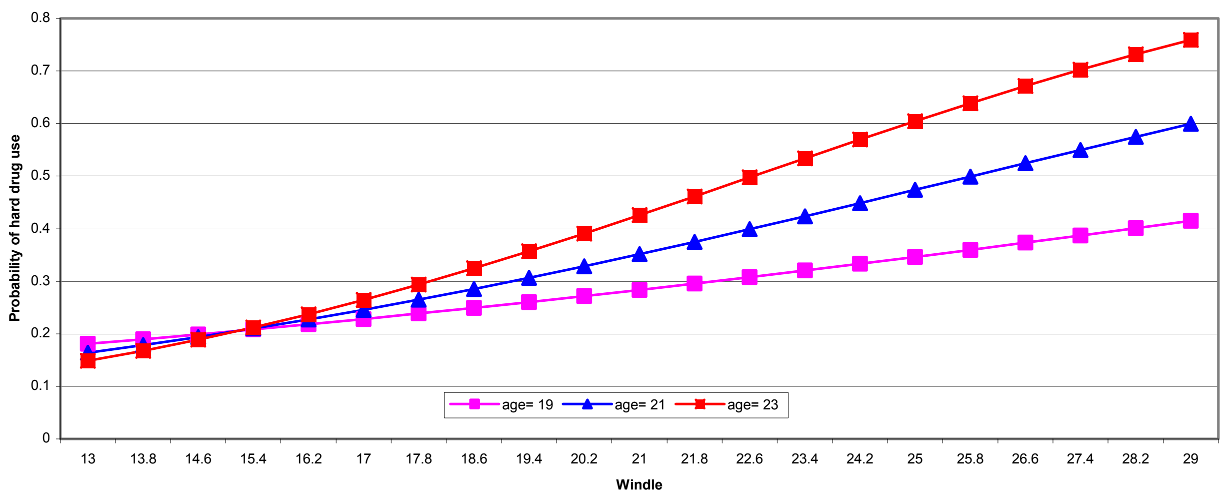


Figure 6: Probability of hard drug use as function of age and Windle(X)



References

Flom, P. L., Friedman, S. R., Kottiri, S. R., Neaigus, A., Curtis, R., Des Jarlais, D. C., Sandoval, M., & Zenilman, J. M. (2001). Stigmatized drug use, sexual partner concurrency, and other sexual risk network and behavioral characteristics of 18-24 Year old youth in a high-risk neighborhood. *Sexually Transmitted Diseases*, *28*, 598-607.

Harrell, F. E. (2001). *Regression modeling strategies*. New York: Springer-Verlag.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: John Wiley and Sons.

Long, J. S. (1997). *Regression models of categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Boca Raton, FL: Chapman & Hall.

Watters, J., & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, *6*, 416-430.

Windle, M. (1993). A retrospective measure of childhood behavior problems and its use in predicting adolescent problem behaviors. *Journal of Studies on Alcohol*, *54*, 422-431.

Author Notes: This work was supported by NIDA grants R01 DA10411 and P30 DA 11041.

The figures in this paper were all produced in Excel. In each workbook, the worksheet contains a top section, where the user may modify the parameters, ranges of values, and so on, a bottom section, which the user should not edit, and one or two charts, which are linked to the sheet, but which may be further edited by the user. The Excel file are available at the following website:

Send correspondence to: Peter L. Flom, National Development and Research Institutes, Inc.
71 W. 23rd St., 8th Floor
New York, NY 10010
Email: flom@ndri.org

Nonrandomly Missing Data in Multiple Regression Analysis: An Empirical Comparison of Ten Missing Data Treatments

Lantry L. Brockmeier Florida A & M University
Jeffrey D. Kromrey University of South Florida
Kristine Y. Hogarty University of South Florida

This study investigated the effectiveness of ten missing data treatments within the context of a two-predictor multiple regression analysis with nonrandomly missing data. Five distinct types of missing data treatments were examined: deletion (both listwise and pairwise methods), deterministic imputation (with imputations based on the sample mean, simple regression and multiple regression), stochastic imputation (mean, simple regression and multiple regression), maximum likelihood estimation (ML) and multiple imputation (MI). Design factors included in the study were sample size, total proportion of missing data, and the proportion of missing data occurring in the upper stratum of each predictor. The success of each method was evaluated based on the sample estimate of R^2 and each standardized regression coefficient. Results suggest that the stochastic multiple regression imputation procedure evidenced the best performance in providing unbiased estimates of the parameters of interest. Deterministic imputation approaches and the stochastic mean imputation approach resulted in large amounts of bias in the estimates.

It is not uncommon for missing data to occur on one or more variables within an empirical investigation. Missing data may adversely affect data analyses, interpretations and conclusions. Collins, Schafer, and Kam (2001) indicate that missing data may potentially bias parameter estimates, inflate Type I and Type II error rates, and influence the performance of confidence bands. Further, because a loss of data is almost always associated with a loss of information, concerns arise with regard to reductions of statistical power. Unfortunately, researchers' recommendations for managing missing data are not in complete agreement resulting in conceptual difficulties and computational challenges (Guertin, 1968; Beale & Little, 1975; Gleason & Staelin, 1975; Frane, 1976; Kim & Curry, 1977; Santos 1981; Basilevsky, Sabourin, Hum, & Anderson, 1985; Raymond & Roberts, 1987; Schafer & Graham, 2002). Many studies that have examined missing data treatments are not comparable due to the various methods used, the stratification categories (number of variables, sample size, proportion of missing data, and degree of multicollinearity), and the criteria that measure effectiveness of the missing data treatment (Anderson, Basilevsky, & Hum, 1983). Further, Schafer and Graham (2002) argue that the treatment of missing values cannot be properly evaluated apart from the modeling, estimation, or testing procedure in which it is rooted. Before proceeding with an examination of the extant literature, a consideration of some missing data terminology is warranted.

Missing Data Terminology

Contemporary discussion of missing data and their treatment can often be confusing and at times may appear somewhat counterintuitive. For example, the term ignorable, introduced by Little and Rubin (1987) was not intended to convey a message that a particular aspect of missing data could be ignored, but rather under what circumstances the missing data mechanism is ignorable. Additionally, when one speaks of data missing at random, these words should not convey the notion that the missingness is derived from a random process external or unrelated to other variables under study (Collins et al., 2001).

According to Heitjan & Rubin (1991) missing data can take many forms, and missing values are part of a more general concept of *coarsened data*. This general category of missing values results when data are grouped, aggregated, rounded, censored, or truncated, resulting in a partial loss of information. The major classifications of missing data mechanisms can be best explained by the relationship among the variables under investigation. Rubin (1987) identified three general processes that can produce missing data. First, data that are missing purely due to chance are considered to represent data that are missing completely at random (MCAR). Specifically, data are missing completely at random if the probability of a missing response is completely independent of all other measured or unmeasured characteristics under examination. Accordingly, analyses of data of this nature will result in unbiased estimates of the population parameters under investigation. Second, data that are classified as missing at random (MAR), do not depend on the missing value itself, but may depend on other variables that are measured for all participants under study. Lastly, and most problematic statistically, are data missing not at random (MNAR). This type of missingness, also referred to as *nonignorable* missing data, is directly related to the value that would have been observed for a particular variable. A commonly encountered situation, in which data would be classified as MNAR, arises when respondents in a certain income or age strata fail to provide responses to questions of this nature.

In much of the research that has been previously conducted, the key assumption has been that data are missing at random. If data are randomly missing and the percentage of missing data is not too large, researchers are advised that any missing data treatment is effective. However, this assumption of randomly missing data is tenuous in many instances. Although a few procedures have been presented to test the assumption of randomly missing data (e.g., Cohen & Cohen, 1975, 1983; Tabachinick & Fidell, 1983), Kromrey and Hines (1994) assert that this assumption is rarely tested and that the applied researcher is hard pressed to find guidance if data are missing nonrandomly.

Missing Data Treatments

Applied researchers often employ deletion or deterministic imputation procedures to manage missing data rather than choosing from among other available missing data treatments. The former procedures employ a deletion process, utilizing cases with complete data (Glasser, 1964; Haitovsky, 1968). Listwise deletion discards all cases with incomplete information, whereas pairwise deletion constructs a correlation matrix utilizing all pairs of complete data. Deterministic imputation procedures (e.g., mean substitution, simple regression, or multiple regression) provide estimates of the missing values (Santos, 1981; Kalton & Kasprzyk, 1982). In all deterministic approaches to imputation, the residual (error) term is set to zero in the estimation equation. In contrast, stochastic imputation includes a random value for the residual in the estimation equation. Some empirical evidence suggests that the stochastic imputation procedures are superior to the deterministic approaches (Santos, 1981; Kalton & Kasprzyk, 1982; Jinn & Sedransk, 1989; Keawkungal & Benson, 1989; Brockmeier, Hines, & Kromrey, 1993).

Kromrey and Hines (1994) examined the effectiveness of the deletion and deterministic imputation procedures in regression analysis with nonrandomly missing data in the context of missing data on one of two predictor variables. Systematically missing data were produced by generating 60% of the missingness above the mean value of the variable in each simulated sample. These researchers concluded that with moderate amounts of missing data, the deletion procedures yielded results similar to those obtained without missing data. Further, the deterministic imputation procedures evidenced poor performance when compared to the deletion procedures.

Within the context of a two-predictor multiple regression analysis with nonrandomly missing data, Brockmeier, Kromrey, and Hines (2000) investigated the effectiveness of eight missing data treatments on the sample estimate of R^2 and each standardized regression coefficient. These researchers varied the overall proportion of missing data in each sample, as well as the proportion of the missingness that occurred in values greater than the sample mean. The results suggested that the stochastic multiple regression imputation procedure provided the best treatment of missing data.

New Approaches to Missing Data

Traditionally, researchers have not utilized maximum likelihood estimation (ML), multiple imputation (MI), or the aforementioned stochastic imputation procedures. Kromrey (1989) and Brockmeier (1992) indicated that these methods are not typically found in the journals of applied researchers, with much of the scholarly work on maximum likelihood estimation and multiple imputation appearing in technical statistical journals. Little (1992) stated that maximum likelihood estimation is infrequently utilized due to the lack of software and the mathematical complexity of the computations. More recently, however, maximum likelihood estimation has been included in some structural equation modeling software and statistical packages. Additionally, Gregorich (1999) has created a maximum likelihood estimation program using SAS IML that is currently available and freely distributed to SAS users for noncommercial purposes. This program employs the Expectation-Maximization (EM) algorithm to estimate the maximum likelihood covariance matrix and mean vector in the presence of missing data. This maximum likelihood approach assumes that data are missing completely at random or missing at random. While a decade ago few stand-alone programs existed for employing multiple imputation as a missing data treatment, ML and MI are now becoming more popular with the implementation of these procedures in free and commercial software (Schafer & Graham, 2002). For example, a recent release of SAS (version 8.2) introduced a multiple imputation procedure.

Rubin (1996) indicated that the ultimate goal of multiple imputation is to provide statistically valid inferences in applied contexts where researchers employ different analyses and models and when there is no one accepted reason for the missing data. The multiple imputation procedure replaces each missing value with a set of plausible values that represents the degree of uncertainty about the correct value to impute. One might view this approach as an enhancement over simple imputation methods that fail to reflect the uncertainty about the predictions of the missing values, often resulting in point estimates of a variety of parameters that are not statistically valid in any generality.

Multiple imputation inference involves three distinct phases (Schafer, 1997). First, missing data are filled in *m* times to generate complete data sets. Second, the *m* complete data sets are analyzed using standard statistical analyses. Finally, the results from the *m* complete data sets are combined to produce inferential information. For a recent review of MI procedures, see Sinharay, Stern, and Russell (2001).

Evidence of the effectiveness of maximum likelihood estimation and multiple imputation, as missing data treatments has been somewhat limited in the past (for a recent exception, see Collins et al., 2001). However, with the innovations in the software described above, interest and enthusiasm for these alternative methods appears to be growing. A recent issue of *Psychological Methods* (2001) devoted a special section to issues surrounding missing data, with a specific focus on multiple imputation and maximum likelihood estimation.

Purpose

The purpose of this study was to investigate the effectiveness of ten missing data treatments within the context of a two-predictor multiple regression analysis with nonrandomly missing data. Further, the study investigated whether sample size, proportion of missing data occurring in the upper stratum on each predictor, and the total percentage of missing data affected the effectiveness of the ten missing data treatments. The success of each method was evaluated based on the sample estimate of R^2 and each standardized regression coefficient. Five distinct types of missing data treatments were examined: deletion (both listwise and pairwise methods), deterministic imputation (with imputations based on the sample mean, simple regression and multiple regression), stochastic imputation (mean, simple regression and multiple regression), maximum likelihood estimation and multiple imputation.

Method

This research was a Monte Carlo study designed to simulate multiple regression analyses in the presence of missing data. The use of simulation methods allows the control and manipulation of research design factors and the incorporation of sampling error into the analyses. Observations for each sample were generated under known population conditions and missing data were created in each sample.

Data Source

Data for this investigation were simulated to model the correlational structure observed in a sample of responses to an instrument designed to measure teachers' reported perceptions of computers and integration of technology in their classrooms (Hogarty, Lang, & Kromrey, in press). These field data were composite scores based on selected subscales from the instrument. Each subscale contained items measured on either a 5-point Likert scale ranging from *strongly disagree* to *strongly agree* or a 5-point frequency of use scale ranging from *not at all* to *every day*. Two correlation matrices from these data were selected for use as population templates in this Monte Carlo study (Table 1). The first matrix had correlations between variables that ranged from 0.33 to 0.61 with an R^2 of 0.50. In contrast, the second matrix presented lower correlations (ranging from 0.12 to 0.49) and an R^2 value of 0.25.

Table 1. *Correlation Matrices Used for Simulation*

Higher Correlated Data Set			Lower Correlated Data Set		
	Y	X_1		Y	X_1
X_1	0.53		X_1	0.49	
X_2	0.61	0.33	X_2	0.16	0.12

Experimental Design

The study employed a 2 X 3 X 4 X 6 experimental design. The factorial design included two between-subjects factors (population correlation matrix and sample size, N = 50, 100, and 200) and two within-subjects factors (proportion of missing data in the upper stratum of each variable and percentage of missing data). The four proportions of systematically missing data in the upper stratum were 0.60, 0.70, 0.80, and 0.90. The six conditions of missing data generated were 10%, 20%, 30%, 40%, 50%, and 60%. In addition, the complete samples with no missing data were analyzed. Missing data were distributed equally across both predictor variables (i.e., with 20% missing data, 10% of the observations presented missing data on each of the regressors).

The pseudopopulations were not manipulated within the experiment, but were selected to obtain the desired correlational patterns in each data set. The sample sizes, missing data structures and missing data treatments were chosen to replicate and extend the earlier work of Kromrey and Hines (1994) and Brockmeier et al. (2000).

Conduct of the Monte Carlo Study

This research was conducted using SAS/IML version 8.2. Conditions for the study were run under Windows 98. Normally distributed random variables were generated using the RANNOR random number generator in SAS. A different seed value for the random number generator was used in each execution of the program and the program code was verified by hand-checking results from benchmark data sets. For each condition examined in the Monte Carlo study, 5000 samples were simulated. The use of 5000 samples provides adequate precision for this investigation. For example, the use of 5000 samples provides a maximum 95% confidence interval width around an observed proportion that is ± 0.014 (Robey & Barcikowski, 1992).

Within each sample, missing data were created by setting to missing the generated values of one of the two regressors variables, under the constraint that no observation may have both regressors missing. The target percentage of missing data was evenly divided between the two regressors (i.e., half of the missingness occurred on each regressor). In addition, the percentage of the missingness that was imposed on observations in the upper stratum was controlled by dividing each sample into two strata for each regressor, then randomly selecting the correct number of observations from each stratum. The proportion of data missing in the upper stratum was altered to create increasing degrees of distortion in the observed data. The probability of a missing value was established to be proportional to the value of the variable. For the majority of conditions, the upper stratum was defined as observations above the sample median value of the predictor variable. If an insufficient number of observations were available above the median (e.g., with 90% missing and 60% of these occurring above the median) the upper stratum was defined as the top 60% of the distribution.

In each generated sample, the 24 missing data conditions were independently imposed (six total percentages of missing data crossed with four proportions of missing data in the upper stratum), allowing the two missing data factors to be treated as within-subjects factors in the research design. In addition, each complete sample with no missing data was analyzed to provide a reference for the evaluation of the missing data treatments. Each sample was analyzed by computing the regression equation (obtaining the sample estimates of each standardized regression weight and the sample value of R^2) after applying each of the ten missing data treatments.

Missing Data Treatments

The missing data treatments examined in this study included deletion, deterministic imputation, stochastic imputation and model-based approaches. The treatment of missing data was considered with respect to two predictor variables, X_1 and X_2 (with missing data on one but not both predictors), and a single criterion variable, Y_1 (with no missing data). The first method, listwise deletion, necessitated the deletion of any observations in the sample with missing values for either of the predictor variables. The resulting ‘complete’ data set for each sample was used in the calculation of parameter estimates (R^2 and the standardized regression weights). For the second method, pairwise deletion, a correlation matrix was constructed based on all of the available data for each pair of variables. The resulting correlation matrix was subsequently analyzed to obtain the regression equation for each sample.

Three imputation techniques using deterministic methods were employed, a mean imputation approach and both simple and multiple regression imputation. For the mean imputation procedure, missing values were imputed using the sample mean value for each regressor variable. In contrast, deterministic multiple regression imputation for each sample was initiated by calculating a prediction equation based on the data available and deriving a predicted value for the variable with missing data based on the values of the other two variables. For example, to predict missing values for X_1, the sample regression equation for predicting X_1 based on both X_2 and Y was estimated. The resulting equation was then used to compute predicted values for X_1 and those values were imputed to replace the missing data. The same process was followed for simple regression imputation, but only a single predictor variable was used (i.e., either X_2 or Y was used to predict the missing values for X_2). In each sample, the variable evidencing the stronger correlation was used as the predictor.

The three stochastic techniques examined mirrored the deterministic methods described above. However, as stated earlier, stochastic techniques differ from deterministic methods with regard to the incorporation of a random error term. For stochastic mean imputation, the imputation begins with the mean value in the sample for the particular variable with missing data, but each imputation requires the addition of a randomly selected normal deviate drawn from a distribution with $\mu = 0$ and σ^2 equal to the sample variance of the variable. The same steps were followed when employing the stochastic simple and multiple regression imputation methods as those described above for the parallel deterministic approaches with the exception of the addition of a random residual term. This residual was randomly drawn from a distribution with $\mu = 0$ and σ^2 equal to the mean squared residual from the regression that was used to derive the prediction equation.

The multiple imputation approach is an extension of the multiple regression imputation procedure. This approach replaces each missing value with a set of plausible values that represent the degree of uncertainty about the correct values to be imputed. For this study, 10 imputations were employed for the multiple imputation procedure. Rubin (1987) indicated that the efficiency of an estimate based on *m* imputations is approximately:

$$\left(1 + \gamma/m\right)^{-1};$$

where γ is the fraction of missing information for the variable being estimated. For example, with 50% missing information, *m* = 5 imputations provides an estimate of efficiency of approximately 91%, whereas *m* = 10 imputations increases the estimated efficiency to 95%.

As described earlier, first, missing data were filled in ten times to generate ten complete data sets. This phase was accomplished by first conducting a multiple regression analysis on the sample data to derive an initial set of parameter estimates. These estimates, along with the obtained covariance matrix, where then used to generate a sample of ten sets of parameter estimates. These parameter estimates were applied individually to the sample data and the predicted values imputed, resulting in ten distinct samples with complete data. The ten complete data sets were then analyzed sequentially using multiple regression analysis. Finally, the obtained R^2 values and standardized regression weights from the ten regression analyses were averaged and the resultant values were used to assess the performance of this procedure.

The final procedure examined was maximum likelihood estimation via the EM algorithm. This is an iterative approach that seeks parameter values that maximize the likelihood of the observed data. The well-known EM algorithm (Dempster, Laird, & Rubin, 1977) consists of an estimation step (or E-step) that predicts the missing values based upon estimates of the parameters, and a maximization step (or M-step) that revises the estimates of the parameters. The E and M steps are repeated until the parameter values do not change appreciably from one cycle to the next. For an excellent overview of the theory underlying this approach, both with and without missing data see Rubin (1987).

Statistical Analysis

The relative effectiveness of the missing data treatments was evaluated in terms of statistical bias and standard errors for the estimates of the standardized regression weights and the sample value of R^2. The results were analyzed by computing the effect sizes obtained from the missing data treatment conditions relative to the complete sample condition (i.e., 0% missing data). That is,

$$\delta_{ijk} = \frac{\overline{\theta}_{ijk} - \overline{\theta}_{0k}}{\hat{\sigma}_{\theta_{0k}}}$$

where δ_{ijk} = the effect size for missing data treatment *i*, under missing data condition *j* and sample size *k*,

$\overline{\theta}_{ijk}$ = the mean value of the parameter estimate over the 5000 samples,

$\overline{\theta}_{0k}$ = the mean value of the parameter estimate over the 5000 samples of size *k* with no missing data, and

$\hat{\sigma}_{\theta_{0k}}$ = the standard deviation of the parameter estimates across the 5000 samples of size *k* with no missing data.

In addition, standard deviation ratios were calculated to assess the sampling variability of each missing data treatment for the sample estimate of R^2 and standardized regression coefficients relative to the complete sample condition:

$$SD_{Ratio_{ijk}} = \frac{\hat{\sigma}_{\theta_{ijk}}}{\hat{\sigma}_{\theta_{0k}}}$$

where $SD_{Ratio_{ijk}}$ = the ratio for missing data treatment *i*, under missing data condition *j* and sample size *k*,

$\hat{\sigma}_{\theta_{ijk}}$ = the standard deviation of the parameter estimates obtained across the 5000 samples with missing data treatment *i*, under missing data condition *j* and sample size *k*,

$\hat{\sigma}_{\theta_{0k}}$ = the standard deviation of the parameter estimates across the 5000 samples of size *k* with no missing data.

Results

The results of this study are presented in terms of the effect sizes and the inflation of sampling error associated with the 10 missing data treatments. To save space, only summary statistics are reported, but complete tables are available from the first author.

Effect Sizes for Missing Data Treatments

The distributions of the obtained effect sizes for the estimation of R^2 for all conditions examined in this study are presented in Figures 1 and 2 for the high and low correlation matrices, respectively. Similar patterns of results were seen for both population matrices, although the biases associated with some of the missing data treatments (MDTs) were more extreme in the high correlation matrix. Four particularly poor MDTs are evident in these figures: the three deterministic imputation procedures and the stochastic mean imputation approach. These methods led to large biases in sample R^2 across most of the conditions examined, with the mean imputation approaches yielding underestimates of R^2 and the deterministic regression procedures yielding overestimates. The other MDTs provided relatively unbiased estimates of R^2 across the majority of conditions examined.

Figures 3 and 4 present the distributions of effect sizes associated with the estimation of the standardized regression weight for X_1 in the conditions simulated for the high and low correlation matrices, respectively. For this regression weight, the three deterministic imputation procedures and the stochastic mean imputation procedure showed relatively extreme bias in the conditions simulated in the low correlation matrix (Figure 4). For the high correlation matrix (Figure 3), although the pattern of over and under estimation remained the same, substantially less bias was evident, especially for the deterministic mean procedure.

Figures 5 and 6 present the distributions of effect sizes associated with the estimation of the standardized regression weight for X_2. A notably different pattern was observed for this regression weight. In the high correlation matrix (see Figure 5), the same four procedures evidenced the bias pattern

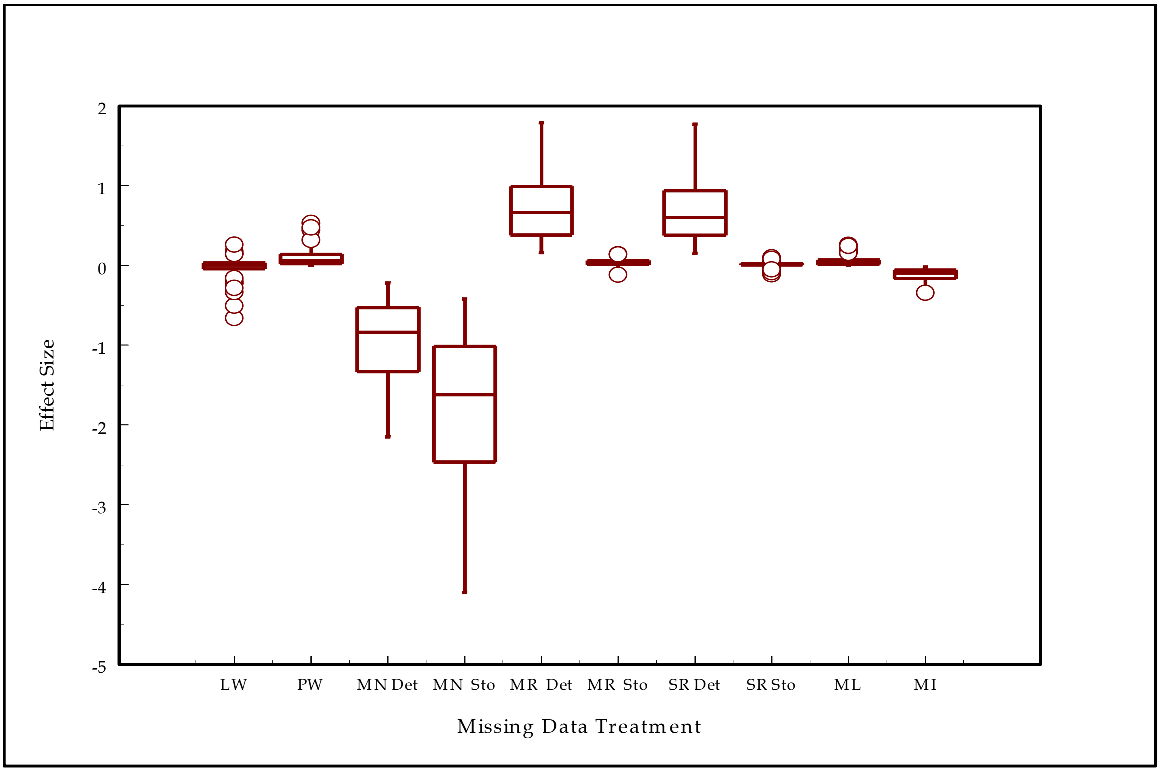


Figure 1. *Distributions of Effect Sizes for Estimates of* R^2 *in High Correlation Matrix.*

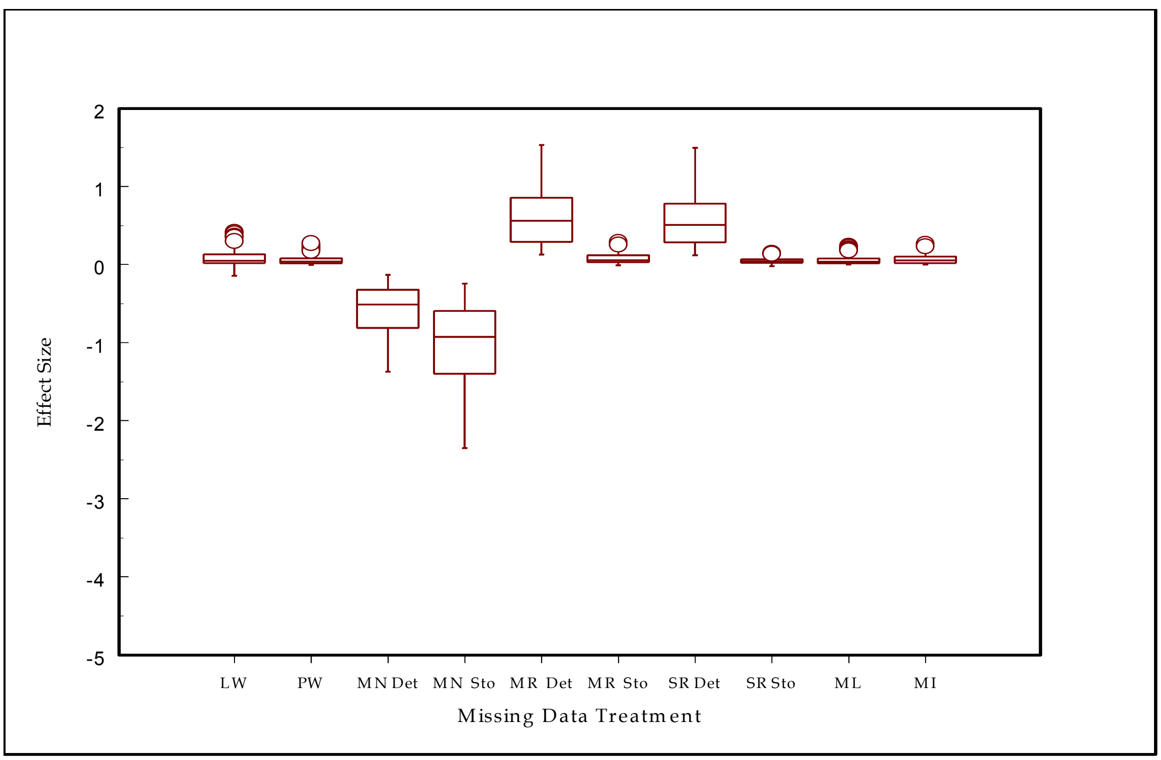


Figure 2. *Distributions of Effect Sizes for Estimates of* R^2 *in Low Correlation Matrix.*

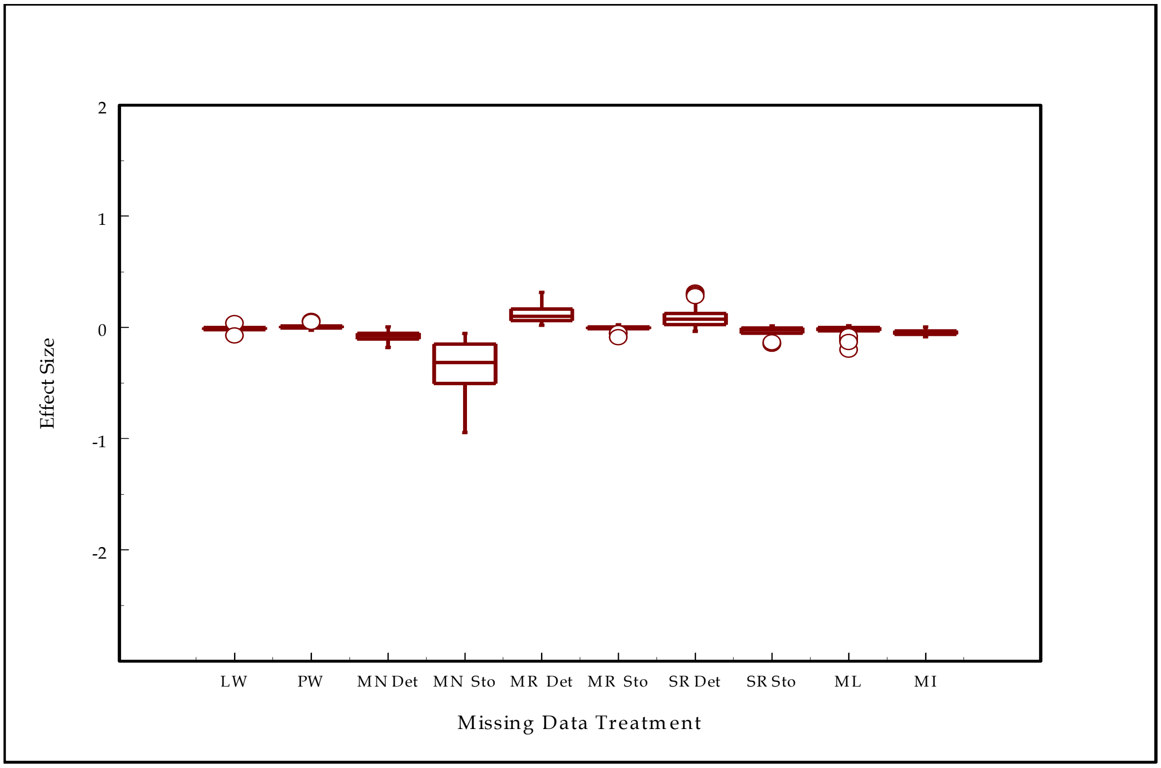


Figure 3. *Distributions of Effect Sizes for Estimates of* β_1 *in High Correlation Matrix.*

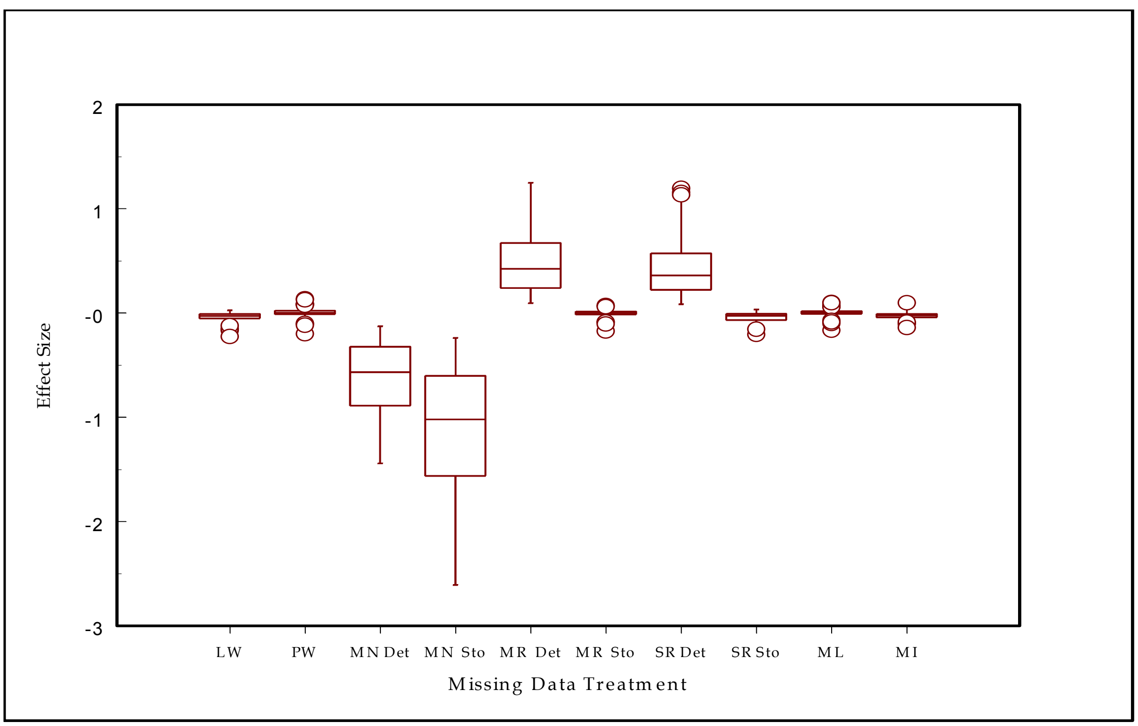


Figure 4. *Distributions of Effect Sizes for Estimates of* β_1 *in Low Correlation Matrix.*

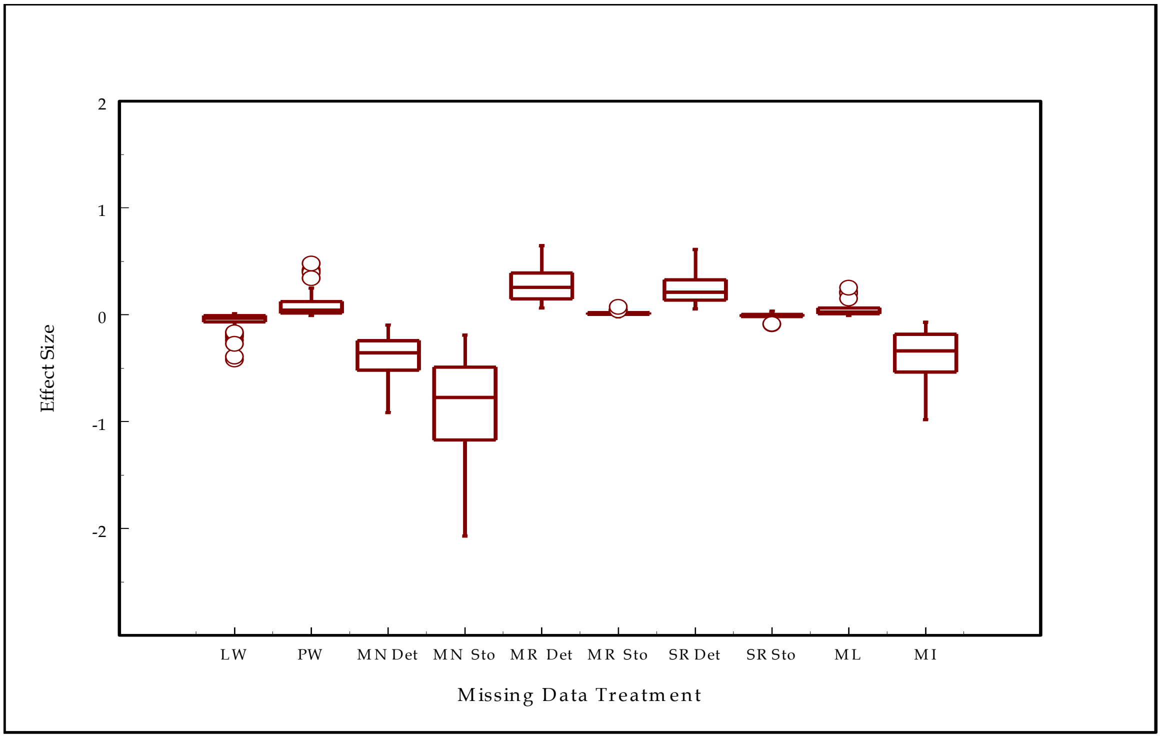


Figure 5. *Distributions of Effect Sizes for Estimates of* β_2 *in High Correlation Matrix.*

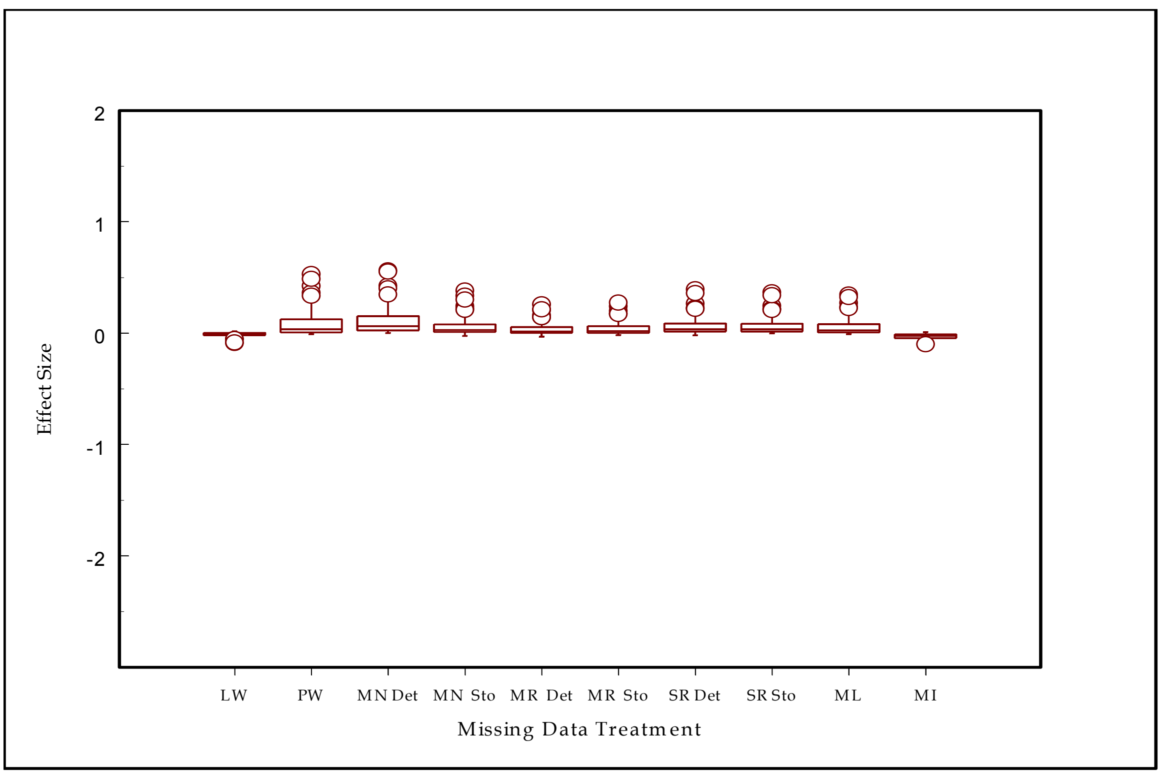


Figure 6. *Distributions of Effect Sizes for Estimates of* β_2 *in Low Correlation Matrix.*

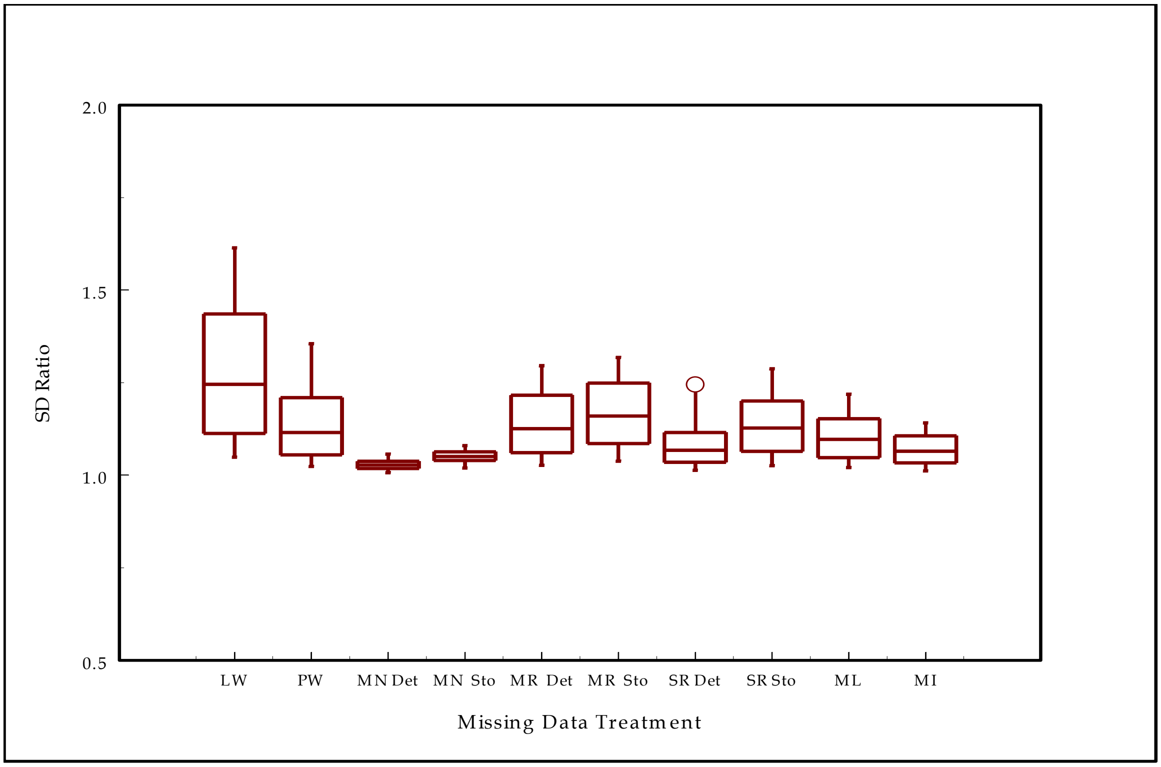


Figure 7. *Distributions of Standard Deviation Ratios for Estimates of R^2 in High Correlation Matrix.*

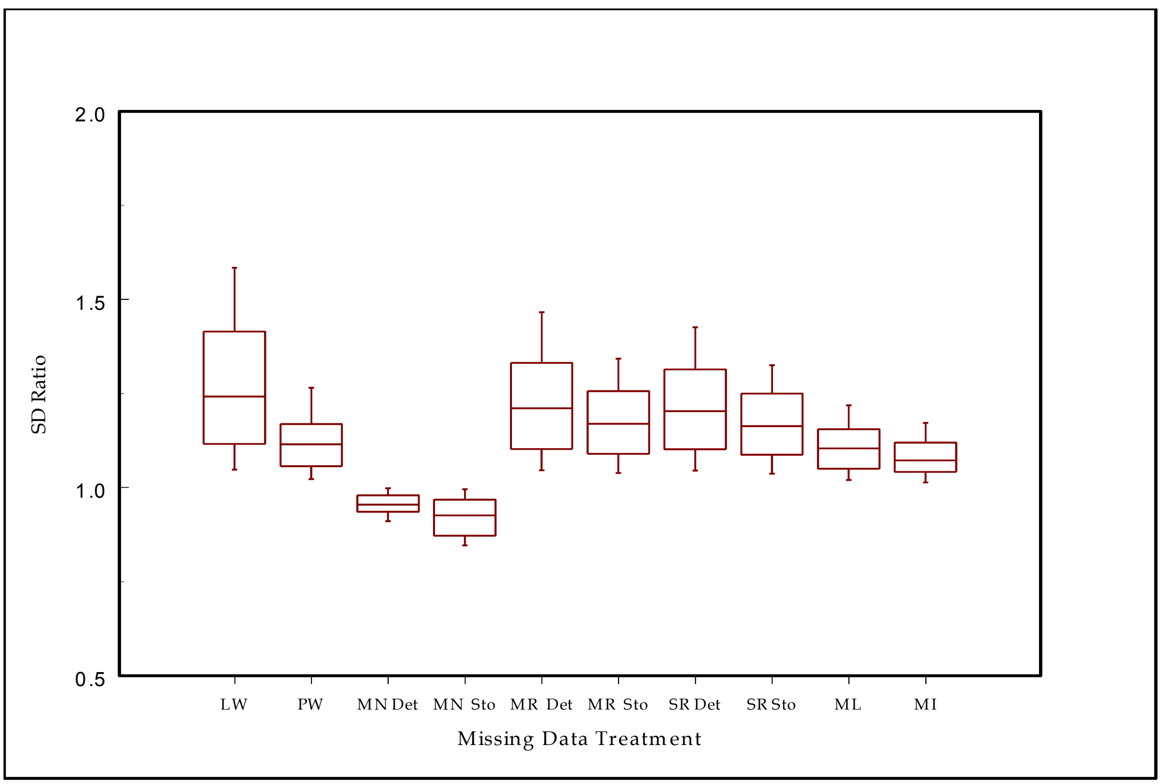


Figure 8. *Distributions of Standard Deviation Ratios for Estimates of R^2 in Low Correlation Matrix.*

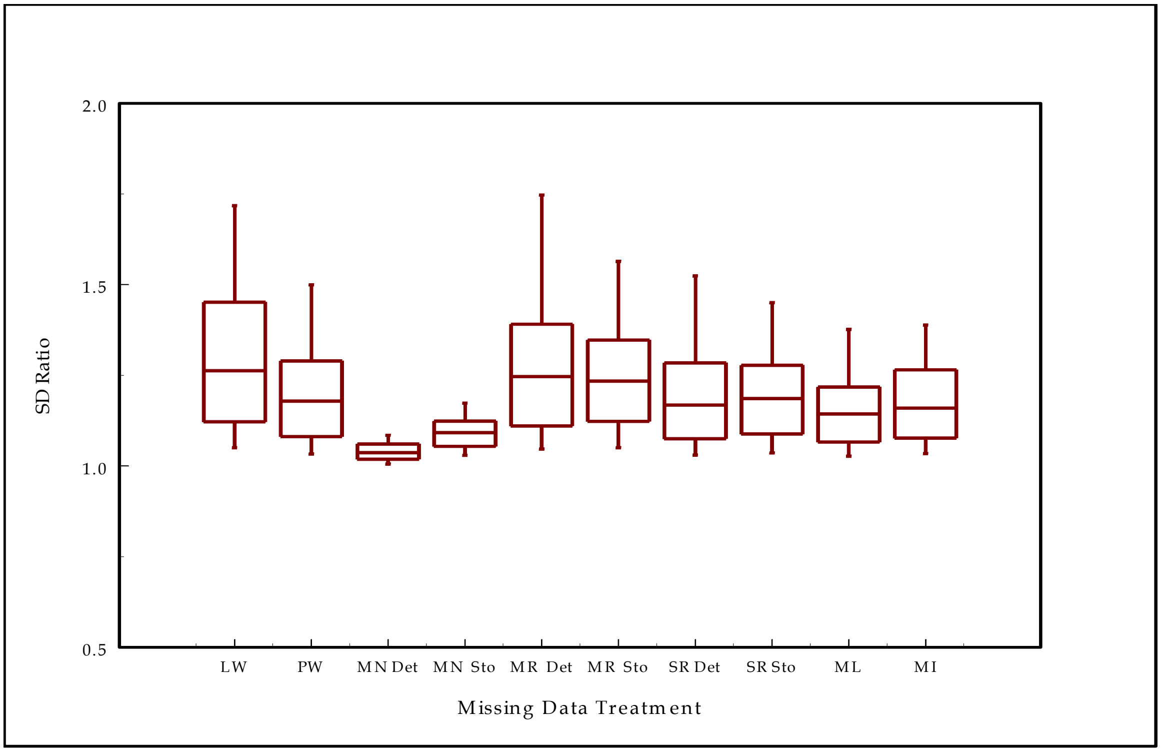


Figure 9. *Distributions of Standard Deviation Ratios for Estimates of* β_1 *in High Correlation Matrix.*

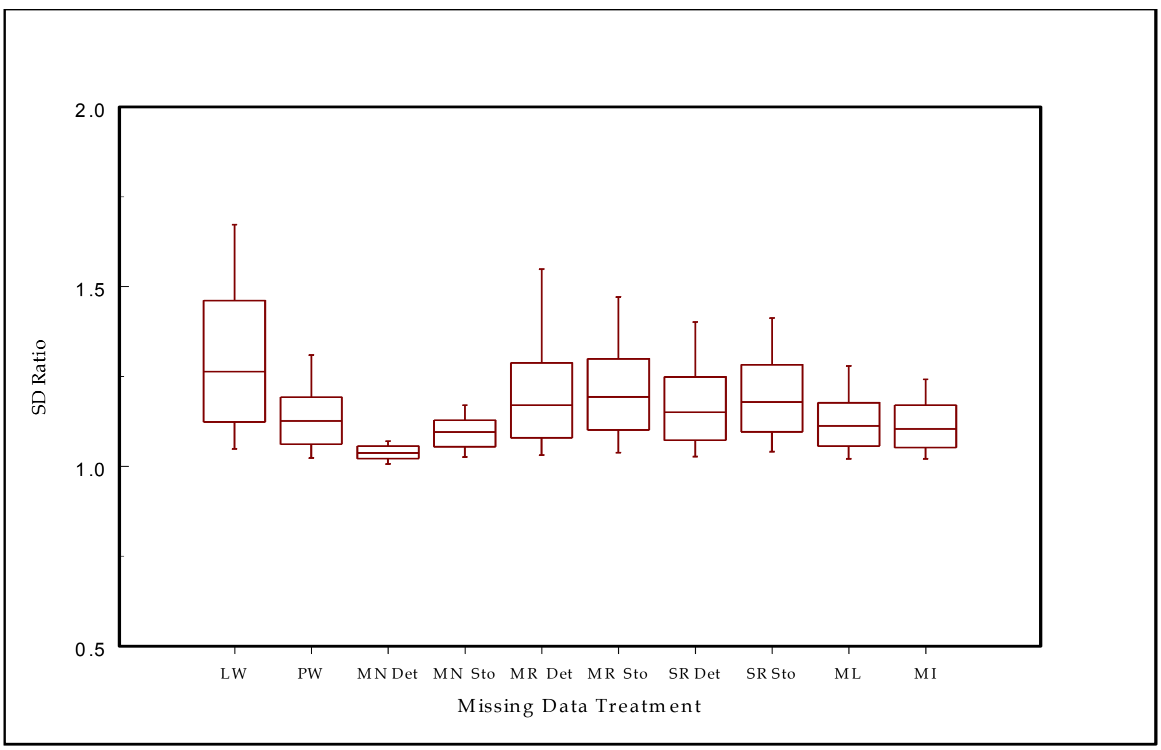


Figure 10. *Distributions of Standard Deviation Ratios for Estimates of* β_1 *in Low Correlation Matrix.*

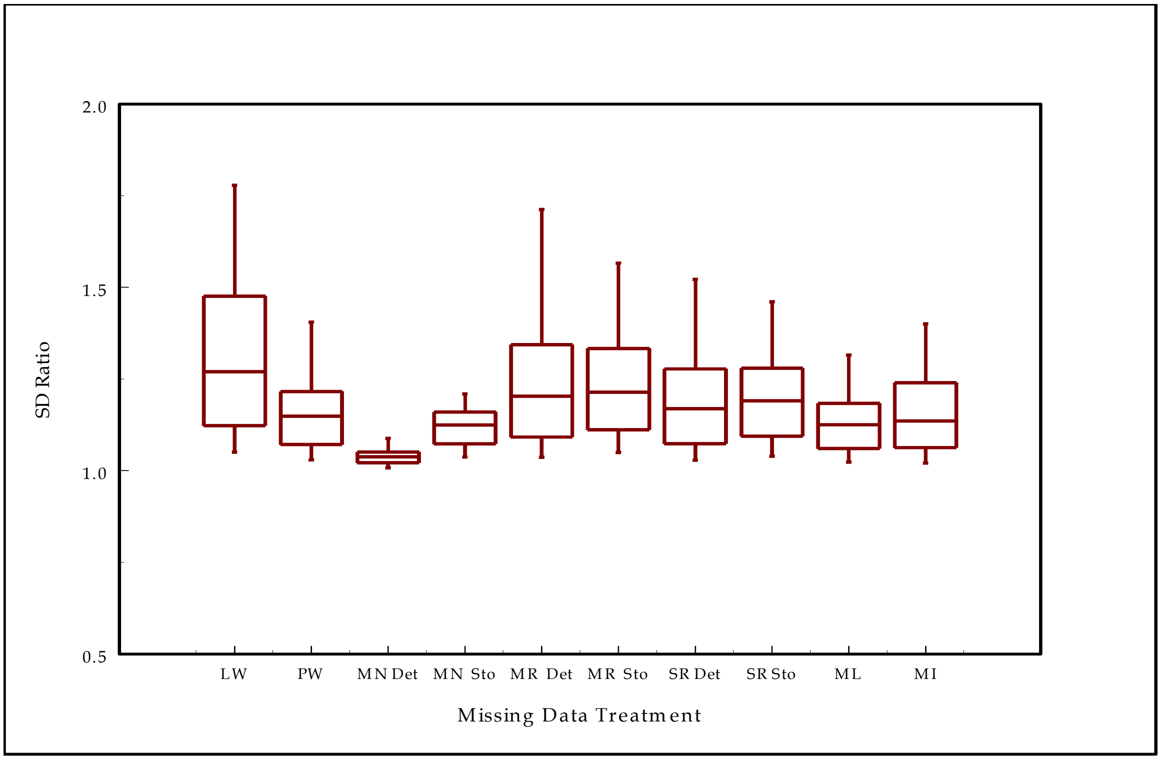


Figure 11. *Distributions of Standard Deviation Ratios for Estimates of* β_2 *in High Correlation Matrix.*

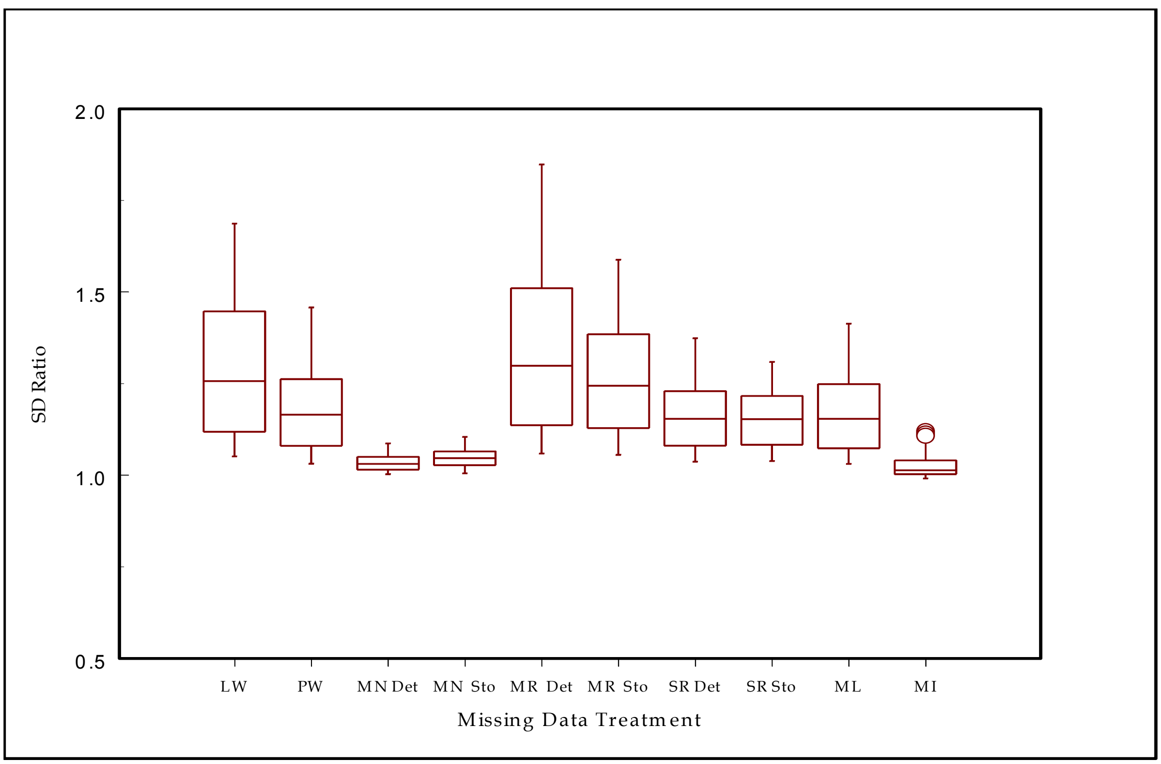


Figure 12. *Distributions of Standard Deviation Ratios for Estimates of* β_2 *in Low Correlation Matrix.*

seen before, but the MI procedure also revealed substantial negative bias. In the low correlation matrix (Figure 6), for most conditions very little bias was evident in any of the methods. When bias was present, the bias was consistently in a positive direction (overestimation of the regression weight).

To aid in the interpretation of the distributions of effect sizes observed, the differences in mean effect sizes were examined across each of the design factors in the study. Table 2 presents the mean effect sizes across conditions examined in each of the two population correlation matrices. On average, the two deletion procedures provided relatively unbiased estimates of all parameters (with the average effect size not exceeding 0.10). For the three deterministic procedures, substantial bias was evident in the estimates for both matrices, with the exception of the estimation of β_1 in the high correlation matrix and β_2 in the low correlation matrix. The simple regression, multiple regression, and stochastic imputation procedures performed well on average, with the average bias not exceeding 0.08, and the MI procedure performed well with the exception of the estimation of β_2 in the high correlation matrix, a condition that led to substantial negative bias. Finally, the maximum likelihood (EM) approach provided reasonably unbiased estimates of all parameters across the conditions.

Table 3 presents the patterns of effect sizes by sample size. Neither deletion procedure was appreciably influenced by the sample size, with the average bias ranging from -0.02 to 0.10 across the parameters and conditions examined. In contrast, the deterministic imputation procedures and the stochastic mean imputation procedure showed greater bias with larger samples, often doubling in magnitude as sample size increased from 50 to 200. The remaining stochastic imputation procedures and the EM approach showed minimal bias across sample sizes examined, with the exception of the MI procedure in the estimation of the second regression weight.

Table 4 presents the pattern of mean effect sizes by the percentage of missing data. The deletion procedures, deterministic imputation procedures and the stochastic mean imputation procedure all evidenced greater degrees of bias with larger percentages of missing data. In the most extreme condition (60% missing data), the stochastic mean imputation procedure yielded an effect size of –2.31 for the estimation of R^2. The stochastic regression procedures (both simple and multiple) also evidenced greater bias with more missing data, but the increase was quite small (reaching only as large as 0.07 for simple regression and 0.14 for multiple regression). The bias observed with the MI procedure in the estimation of the second regression weight is also evident in this table, with increasing bias accompanying greater proportions of missing data. Finally, the EM approach showed a small increase in bias with greater amounts of missing data (with mean effect size reaching as high as 0.16 in the estimation of R^2).

Table 5 presents the pattern of mean effect sizes by the proportion of missing data occurring in the upper stratum of each regressor. Surprisingly, for most of the MDTs, the estimation bias was not strongly related to this factor. Exceptions were pairwise deletion, which showed larger bias in the estimation of R^2 and β_2 with increasing proportions of missing data in the upper stratum, and the maximum likelihood procedure, which showed greater bias in the estimation of β_2.

As a final approach to guide interpretation of the success of the missing data treatments, the proportions of conditions in which the effect size was less than 0.30 in absolute value were computed (see Table 6). Effect sizes this small are considered to present little or no practical problem to researchers (Kromrey & Hines, 1991). The criterion of 0.30 was chosen (rather than Cohen's 0.50, a medium effect size) because the regression coefficients and the sample estimate of R^2 are subject to both substantive interpretation and tests of statistical significance. The best performance, overall, in terms of providing relatively unbiased estimates of the three parameters of interest was the stochastic multiple regression imputation procedure. With this method, none of the conditions examined yielded effect sizes greater than 0.30. Closely following the stochastic multiple regression approach were the stochastic simple regression approach and the maximum likelihood (EM) approach. Both of these methods provided effect sizes less than 0.30 for all conditions except for two, both of which involved the estimation of the second regression weight. The deletion procedures also performed relatively well, providing relatively unbiased estimates in more than 90% of the conditions examined. The deterministic imputation procedures and the stochastic mean imputation procedures produced very poor performance in this analysis. For example, none of these procedures produced effect sizes less than 0.30 for the estimates of R^2 in more than 24% of the conditions.

Table 2. *Effect Sizes for Missing Data Treatments by Population Correlation Matrix*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	R Matrix	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	High	-0.02	0.10	-0.95	0.70	0.73	-1.80	0.01	0.04	-0.12	0.05
	Low	0.09	0.07	-0.59	0.57	0.61	-1.04	0.05	0.08	0.07	0.06
β_1	High	-0.01	0.01	-0.08	0.09	0.12	-0.35	-0.03	-0.01	-0.05	-0.02
	Low	-0.03	0.01	-0.62	0.42	0.48	-1.12	-0.04	0.00	-0.02	0.00
β_2	High	-0.06	0.09	-0.40	0.25	0.28	-0.87	-0.01	0.01	-0.38	0.05
	Low	-0.01	0.08	0.11	0.06	0.04	0.06	0.06	0.04	-0.03	0.06

Table 3. *Effect Sizes for Missing Data Treatments by Sample Size*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	N	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	50	0.10	0.10	-0.53	0.44	0.51	-0.96	0.04	0.09	0.01	0.08
	100	0.02	0.08	-0.73	0.60	0.63	-1.36	0.03	0.05	-0.02	0.05
	200	-0.02	0.07	-1.04	0.86	0.86	-1.94	0.03	0.03	-0.05	0.04
β_1	50	-0.02	0.02	-0.24	0.14	0.22	-0.50	-0.06	0.01	-0.05	0.01
	100	-0.02	0.01	-0.33	0.25	0.29	-0.70	-0.03	0.00	-0.03	-0.01
	200	-0.02	-0.01	-0.48	0.39	0.40	-1.01	-0.02	-0.01	-0.03	-0.02
β_2	50	-0.02	0.07	-0.10	0.10	0.12	-0.28	0.01	0.03	-0.14	0.04
	100	-0.04	0.08	-0.14	0.15	0.15	-0.39	0.03	0.02	-0.20	0.05
	200	-0.04	0.10	-0.19	0.22	0.20	-0.54	0.05	0.03	-0.27	0.06

Table 4. *Effect Sizes for Missing Data Treatments by Percentage of Missing Data*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	Percent	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	10	0.01	0.01	-0.23	0.18	0.18	-0.44	0.01	0.02	-0.01	0.01
	20	0.02	0.02	-0.44	0.34	0.35	-0.84	0.02	0.03	-0.02	0.02
	30	0.02	0.04	-0.68	0.53	0.55	-1.26	0.03	0.05	-0.03	0.03
	40	0.01	0.07	-0.89	0.71	0.74	-1.64	0.03	0.06	-0.03	0.05
	50	-0.01	0.13	-1.12	0.90	0.95	-2.02	0.02	0.06	-0.04	0.08
	60	0.16	0.23	-1.24	1.15	1.25	-2.31	0.07	0.14	-0.01	0.16
β_1	10	0.00	0.00	-0.10	0.08	0.08	-0.21	0.00	0.00	-0.02	0.00
	20	-0.01	0.00	-0.20	0.15	0.16	-0.40	-0.01	0.00	-0.03	0.00
	30	-0.02	0.00	-0.31	0.22	0.24	-0.63	-0.02	-0.01	-0.04	-0.01
	40	-0.03	-0.01	-0.41	0.29	0.32	-0.85	-0.04	-0.01	-0.04	-0.02
	50	-0.05	0.00	-0.52	0.35	0.42	-1.09	-0.07	-0.02	-0.04	-0.03
	60	-0.01	0.05	-0.55	0.45	0.57	-1.26	-0.06	0.02	-0.05	0.01
β_2	10	0.00	0.01	-0.06	0.04	0.04	-0.12	0.01	0.00	-0.05	0.01
	20	0.00	0.02	-0.11	0.09	0.08	-0.24	0.01	0.01	-0.10	0.01
	30	-0.02	0.04	-0.15	0.13	0.13	-0.36	0.02	0.02	-0.17	0.03
	40	-0.04	0.09	-0.17	0.18	0.18	-0.46	0.04	0.03	-0.23	0.05
	50	-0.10	0.17	-0.17	0.23	0.23	-0.55	0.05	0.05	-0.32	0.10
	60	-0.05	0.19	-0.20	0.26	0.28	-0.68	0.04	0.06	-0.35	0.12

Table 5. *Effect Sizes for Missing Data Treatments by Percentage of Missingness in the Upper Stratum*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	Percent	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	60	0.07	0.06	-0.76	0.64	0.68	-1.42	0.03	0.06	-0.04	0.05
	70	0.07	0.07	-0.76	0.64	0.68	-1.42	0.04	0.06	-0.03	0.05
	80	0.03	0.09	-0.77	0.63	0.67	-1.42	0.03	0.06	-0.02	0.06
	90	-0.03	0.12	-0.77	0.62	0.66	-1.42	0.03	0.05	-0.01	0.07
β_1	60	-0.01	0.02	-0.34	0.27	0.31	-0.73	-0.02	0.01	-0.05	0.01
	70	-0.01	0.01	-0.34	0.27	0.31	-0.73	-0.03	0.01	-0.04	0.01
	80	-0.03	0.00	-0.35	0.25	0.30	-0.74	-0.04	-0.01	-0.04	-0.01
	90	-0.04	-0.01	-0.36	0.24	0.28	-0.75	-0.05	-0.02	-0.02	-0.04
β_2	60	-0.01	0.02	-0.19	0.13	0.14	-0.45	0.00	0.01	-0.20	0.01
	70	-0.02	0.05	-0.17	0.14	0.15	-0.43	0.01	0.02	-0.20	0.03
	80	-0.04	0.10	-0.14	0.16	0.16	-0.40	0.03	0.03	-0.20	0.06
	90	-0.07	0.18	-0.07	0.19	0.18	-0.33	0.06	0.06	-0.22	0.11

Table 6. *Number and Proportions of Conditions with Effect Sizes Less than 0.30 in Absolute Value*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic		Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	*N*	136	140	24	35	35	5	144	144	143	144
	%	94%	97%	17%	24%	24%	3%	100%	100%	99%	100%
β_1	*N*	144	144	88	97	92	40	144	144	144	144
	%	100%	100%	61%	67%	64%	28%	100%	100%	100%	100%
β_2	*N*	142	134	89	120	111	77	142	144	106	142
	%	99%	93%	62%	83%	77%	53%	99%	100%	74%	99%

Table 7. *Standard Deviation Ratios for Missing Data Treatments by Population Correlation Matrix*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	R Matrix	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	High	1.28	1.13	1.03	1.08	1.14	1.05	1.13	1.17	1.07	1.10
	Low	1.27	1.12	0.96	1.21	1.22	0.92	1.17	1.18	1.08	1.11
β_1	High	1.30	1.20	1.04	1.20	1.28	1.09	1.20	1.25	1.18	1.15
	Low	1.30	1.13	1.04	1.17	1.20	1.09	1.20	1.21	1.11	1.12
β_2	High	1.31	1.16	1.04	1.19	1.24	1.12	1.20	1.23	1.16	1.13
	Low	1.30	1.19	1.03	1.16	1.35	1.05	1.15	1.27	1.03	1.17

Table 8. *Standard Deviation Ratios for Missing Data Treatments by Sample Size*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	N	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	50	1.28	1.14	1.00	1.16	1.19	0.99	1.16	1.18	1.07	1.11
	100	1.27	1.12	0.99	1.14	1.18	0.99	1.15	1.17	1.08	1.10
	200	1.28	1.12	0.99	1.14	1.18	0.98	1.15	1.17	1.08	1.10
β_1	50	1.33	1.19	1.04	1.22	1.29	1.09	1.23	1.26	1.15	1.16
	100	1.29	1.16	1.04	1.18	1.23	1.09	1.19	1.22	1.14	1.13
	200	1.28	1.15	1.04	1.15	1.21	1.10	1.17	1.21	1.14	1.12
β_2	50	1.33	1.20	1.04	1.19	1.34	1.09	1.19	1.28	1.11	1.18
	100	1.29	1.16	1.04	1.17	1.28	1.08	1.17	1.24	1.09	1.15
	200	1.28	1.15	1.03	1.17	1.26	1.08	1.17	1.23	1.08	1.14

Table 9. *Standard Deviation Ratios for Missing Data Treatments by Percentage of Missing Data*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	Percent	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	10	1.06	1.03	1.00	1.04	1.04	1.01	1.04	1.05	1.02	1.03
	20	1.12	1.06	1.00	1.07	1.08	1.01	1.08	1.09	1.04	1.05
	30	1.20	1.09	1.00	1.11	1.14	1.00	1.12	1.14	1.06	1.08
	40	1.29	1.13	0.99	1.16	1.19	0.99	1.17	1.19	1.08	1.11
	50	1.43	1.19	0.98	1.22	1.28	0.96	1.23	1.25	1.11	1.15
	60	1.57	1.25	0.98	1.29	1.36	0.95	1.28	1.32	1.14	1.20
β_1	10	1.06	1.04	1.01	1.04	1.05	1.03	1.05	1.06	1.03	1.03
	20	1.12	1.07	1.02	1.08	1.10	1.06	1.09	1.11	1.06	1.06
	30	1.21	1.12	1.03	1.13	1.17	1.08	1.15	1.18	1.11	1.11
	40	1.31	1.18	1.05	1.19	1.24	1.11	1.21	1.25	1.16	1.15
	50	1.46	1.25	1.06	1.28	1.37	1.13	1.29	1.34	1.22	1.21
	60	1.63	1.33	1.07	1.38	1.51	1.14	1.38	1.44	1.28	1.27
β_2	10	1.06	1.04	1.01	1.04	1.06	1.03	1.05	1.06	1.02	1.04
	20	1.12	1.08	1.02	1.08	1.12	1.05	1.09	1.12	1.03	1.07
	30	1.21	1.13	1.03	1.13	1.20	1.08	1.14	1.20	1.06	1.12
	40	1.31	1.18	1.04	1.19	1.29	1.10	1.19	1.27	1.09	1.16
	50	1.47	1.26	1.05	1.27	1.45	1.11	1.26	1.37	1.14	1.23
	60	1.64	1.35	1.07	1.36	1.63	1.13	1.33	1.48	1.20	1.30

Table 10. *Standard Deviation Ratios for Missing Data Treatments by Percentage of Missingness in the Upper Stratum*

		Deletion		Deterministic Imputation			Stochastic Imputation				
Statistic	Percent	Listwise	Pairwise	Mean	SR	MR	Mean	SR	MR	MI	ML
R^2	60	1.27	1.12	0.99	1.15	1.18	0.99	1.15	1.17	1.08	1.10
	70	1.28	1.12	0.99	1.15	1.18	0.99	1.15	1.17	1.08	1.10
	80	1.28	1.12	0.99	1.15	1.18	0.99	1.15	1.17	1.08	1.10
	90	1.28	1.13	0.99	1.15	1.18	0.98	1.15	1.17	1.07	1.10
β_1	60	1.30	1.16	1.04	1.18	1.24	1.09	1.20	1.23	1.14	1.14
	70	1.30	1.16	1.04	1.18	1.24	1.09	1.20	1.23	1.14	1.14
	80	1.30	1.17	1.04	1.18	1.24	1.09	1.20	1.23	1.14	1.14
	90	1.30	1.17	1.04	1.18	1.24	1.09	1.20	1.23	1.14	1.14
β_2	60	1.30	1.17	1.04	1.18	1.29	1.08	1.18	1.25	1.09	1.15
	70	1.30	1.18	1.04	1.18	1.29	1.09	1.18	1.25	1.09	1.15
	80	1.30	1.17	1.04	1.18	1.29	1.08	1.18	1.25	1.09	1.15
	90	1.31	1.17	1.03	1.18	1.30	1.08	1.18	1.25	1.09	1.15

Sampling Error Inflation

The distributions of the standard deviation ratios for R^2 in samples from high and low correlation matrices are presented in Figures 7 and 8, respectively. The least inflation of sampling error was seen with the mean imputation procedures (both deterministic and stochastic), while the largest increase was seen with the listwise deletion approach. With the exception of some conditions observed with the listwise deletion procedure, the increase in sampling error did not exceed 50% across any of the conditions examined. Somewhat greater increases in sampling error were evident for the regression weights (see Figures 9 and 10 for the estimation of β_1, and Figures 11 and 12 for the estimation of β_2).

The average inflation of sampling error was similar across the two population matrices (see Table 7) and across the sample sizes examined (see Table 8), with slightly less inflation observed with larger samples. In contrast, the inflation was directly related to the proportion of missing data (see Table 9). For example, all of the MDTs evidenced minimal sampling error inflation when only 10% of the data were missing (for these conditions, the greatest increase in sampling error across parameters and across methods was only 6%). However, as the proportion of missing data increased, the sampling error also increased for all of the MDTs with the exception of the mean imputation procedures in the estimation of R^2. For both the stochastic and deterministic mean imputation approaches, the sampling error decreased as the proportion of missing data increased. Finally, sampling error inflation was not associated with the proportion of missing data that occurred in the upper stratum of each regressor (see Table 10).

Discussion and Conclusions

Researchers in many fields are often confronted with challenges regarding appropriate methods for dealing with missing data. This issue continues to be a pervasive concern for applied researchers conducting inquiries across a multitude of research contexts. If missing data are ignored or improperly handled, resulting parameter estimates are likely to be biased, inferences distorted and conclusions unsubstantiated. Our investigation was designed to inform the treatment of nonrandomly missing data across a variety of commonly encountered situations in the conduct of multiple regression analysis. In this vein, we examined the influence of various configurations of missing data across two predictor variables. We studied two distinct correlational structures (both low and high correlations among variables) and three levels of sample size (small, medium and large). Missing data were simulated by varying the proportion of missingness in the upper stratum of each variable (i.e., above the median in most instances) and the total percentage of missingness.

The influence of the degree of relationship between the variables under investigation was negligible for the deletion procedures, but rather substantial for the deterministic procedures. When the influence of sample size was examined, the deletion procedures appeared relatively unaffected, yet sample size evidenced a dramatic effect on the deterministic procedures and the stochastic mean imputation procedures, with larger sample sizes resulting in rather substantially biased estimates. For most of the conditions examined, the influence of the proportion of missing data in the upper stratum had little effect on most of the missing data treatment methods. As expected, the most problematic conditions appeared to be those in which the proportion of missing data in the upper stratum was high, coupled with a large percentage of missing data.

The influence of missing data on sample estimates of R^2 varied considerably among the missing data techniques across the conditions examined. In many instances, nonbiased parameter estimates were evidenced for the two stochastic regression approaches and the maximum likelihood estimation method. Both pairwise and listwise deletion methods and the multiple imputation approach were nearly as effective, yielding relatively few conditions presenting some degree of bias. In contrast, the stochastic mean substitution approaches and all three deterministic approaches evidenced considerable bias, suggesting that these methods should be avoided when faced with missing data of this nature. Similar patterns were observed for these four missing data treatments for both the high and low correlation matrices, however, the bias associated with these procedures was more pronounced when the variables were more highly correlated.

If we turn our attention to the estimation of the regression parameters, we find that both deletion procedures, the stochastic regression approaches and the two model-based methods (i.e., multiple imputation and maximum likelihood estimation) consistently yielded unbiased estimates of the standardized regression weight for X_1. Once again, the three deterministic procedures and the stochastic mean imputation procedure showed a similar pattern of bias across both the high and low correlation matrices, but for these analyses, bias was considerably less pronounced for the high correlation matrix conditions.

Of interest is the notably different pattern inherent in the bias associated with the estimation of the standardized regression weight for X_2. For the high correlation matrix conditions, the same four procedures evidenced extreme bias, but surprisingly the multiple imputation procedure exhibited substantial negative bias. In contrast, multiple imputation evidenced very little bias for the lower correlation condition across the various treatments and conditions.

When we consider the results from this examination of ten missing data treatments under various conditions of missing data on two predictors, we find no convincing evidence to recommend any of the deterministic procedures or the stochastic mean imputation approach as valid approaches to dealing with issues of missing data. On the contrary, the results reported here provide substantial evidence that these methods should be avoided. However, these results suggest that the use of stochastic regression methods, multiple imputation techniques and maximum likelihood estimation warrant careful consideration under many circumstances. Further, the simple deletion methods performed relatively well in many conditions. Although the stochastic multiple regression imputation approach performed the best over all of the conditions examined, the simpler deletion methods may be sufficient in many circumstances.

The results of this study need to be considered in the context of the limitations of this research. First, the findings are somewhat narrowly focused on a two-predictor multiple regression model. The success of these methods in more complex situations, such as factor analysis or path analysis remains uncertain, although recent work suggests promise for the maximum likelihood approach in structural equation modeling (Enders, 2001). Clearly, additional research is called for to examine these methods, as well as others, in multiple contexts under supplementary conditions. Second, the structure of the missing data was such that the missingness was balanced across the two predictors, and the missingness was never allowed to occur across both predictors for any observation. Additionally, the behavior of the missing data techniques was explored for only two correlational structures. The nature and structure of the simulated data sets may additionally limit the ability to generalize the results beyond the conditions examined. Further, the simulated data used in this study were multivariate normal. The success of these missing data treatments with nonnormal distributions will require additional research. Finally, the focus of this work was on bias and sampling error of R^2 and standardized regression weights. The impact of missing data and their treatment on Type I error rates, statistical power, and the accuracy of confidence intervals constructed around parameter estimates requires additional investigation.

References

Anderson, A. B., Basilevsky, A., & Hum, D. P. (1983). Missing data. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 415-494). New York: Academic Press.

Basilevsky, A., Sabourin, D., Hum, D., & Anderson, A. (1985). Missing data estimators in the general linear model: An evaluation of simulated data as an experimental design. *Communications in Statistics*, *14*, 371 394.

Beale, E. M. L., & Little R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, *37*, 129 145.

Brockmeier. L. L. (1992). *Missing data treatments: An empirical investigation of stochastic imputation, deterministic imputation, and the deletion procedures*. Unpublished doctoral dissertation, University of South Florida, Tampa.

Brockmeier, L. L., Hines, C. V., & Kromrey, J. D. (1993, April). *Missing data treatments for multiple regression: Stochastic imputation, deterministic imputation, and the deletion procedures*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Brockmeier, L. L., Kromrey, J. D., & Hines, C. V. (2000, November). *Multiple regression with nonrandomly missing data: A comparison of missing data treatments*. Paper presented at the annual meeting of the Florida Educational Research Association, Tallahassee, FL.

Brockmeier, L. L., Kromrey, J. D., Hogarty, K. Y. (2001, November). *A comparison of ten missing data treatments for nonrandomly missing data in the multiple regression analysis*. Paper presented at the annual meeting of the Florida Educational Research Association, Marco Island, FL.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, L. M., Schafer, J. L., & Kam, C. M (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330-351.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *39*, 1-38.

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*, 352–370.

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

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, *59*, 834 844.

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing observations. *Psychometrika*, *40*, 229 252.

Guertin, W. H. (1968). Comparison of three methods of handling missing observations. *Psychological Reports*, *22*, 896.

Gregorich, S. (1999). EM_COVAR.SAS [computer program]. San Francisco: University of California at San Francisco.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society*, *Series B*, *30*, 67 82.

Heitjan, D. F. & Rubin, D.B. (1991). Ignorability and coarse data. *Annals of Statistics*, *12*, 2244-2253.
Hogarty, K. Y., Lang, T. R., & Kromrey, J. D. (in press). Another look at technology use in classrooms: The development and validation of an instrument to measure teachers' perceptions. *Educational and Psychological Measurement*.
Jinn, J. H., & Sedransk, J. (1989). Effect on secondary data analysis of common imputation methods. In C. C. Clogg (Ed.), Sociological Methodology (pp. 213-241). Oxford: Basil Blackwell.
Kalton, G., & Kasprzyk, D. (1982). Imputing for missing survey responses. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 22-33.
Keawkungal, J., & Benson, J. (1989, March). *The effect of missing data on confirmatory factor analysis models: A Monte Carlo comparison of two regression imputation methods*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
Kim, J., & Curry J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, *6*, 215 240.
Kromrey, J. D. (1989). *Toward practical advice on the treatment of missing data*. Unpublished doctoral dissertation, University of South Florida, Tampa.
Kromrey, J. D., & Hines, C. V. (1990, November). *Nonrandomly missing data in multiple regression: An empirical examination of the effectiveness of common missing data treatments*. Paper presented at the meeting of the Florida Educational Research Association, Deerfield Beach, FL.
Kromrey, J. D., & Hines, C. V. (1991, February). *Randomly missing data in multiple regression: An empirical comparison of common missing data treatments*. Paper presented at the meeting of the Eastern Educational Research Association, Boston, MA.
Kromrey, J. D., & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing data treatments. *Educational and Psychological Measurement*, *54*(*3*), 573-593.
Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, *87*, 1227 1237.
Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.
Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, *47*, 13 26.
Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473-489.
Santos, R. (1981). Effects of imputation on regression coefficients. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 141-145.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman and Hall.
Schafer, J. L. & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.
Sinharay, S., Stern, H.S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*, 317-329.
Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics*. New York: Harper & Row.

Send correspondence to: Lantry L. Brockmeier
Department of Educational Leadership and Human Services, 308 GEC-B
Florida A & M University, Tallahassee, FL 32307-4900
Email: lantry.brockmeier@famu.edu

An Empirical Comparison of Regression Analysis Strategies with Discrete Ordinal Variables

Jeffrey D. Kromrey **Gianna Rendina-Gobioff**
University of South Florida

The Type I error control and statistical power of three tests of regression models incorporating discrete ordinal variables were compared in a Monte Carlo study. Samples were generated from populations measured on discrete ordinal variables representing 5-point and 7-point response scales. Each sample was analyzed using ordinary least squares regression, ordinal multiple regression and cumulative logit models. Factors examined in the Monte Carlo study were the population effect size, number of regressor variables, level of regressor intercorrelation, population distribution shape and sample size. Results suggest that the logistic regression approach evidenced poor Type I error control with small samples or with large numbers of regressors. In contrast, both the ordinary least squares approach and the ordinal multiple regression approach evidenced good Type I error control across the majority of conditions examined. Further, the power differences between these approaches were negligible.

A common response format for the measurement of many variables in the social sciences is a forced-choice, ordinal scale. For example, data may be collected using a Likert scale in which respondents indicate the extent of their agreement or disagreement with a stimulus. Similarly, ordinal rating scales may be used to obtain ratings by participants regarding their perceptions of the frequency or intensity of a target phenomenon.

The level of measurement of this type of variable and the appropriate statistical techniques for the analysis of such data have been the subject of debate for many years. Stevens (1951) defined four levels of measurement (nominal, ordinal, interval and ratio) and the appropriateness of their use within statistical analyses. The subsequently published literature presents conflicting views about the use of parametric statistics with such discrete ordinal variables. Critics of Stevens have proposed that ordinal data can be treated as interval data when parametric statistics are employed (e.g., Borgatta & Bohrnstedt, 1980; Gaito, 1980). In contrast, some researchers have proposed alternative analyses that are purported to more appropriately represent the discrete ordinal nature of such data (Agresti & Finlay, 1997; Cliff, 1994, 1996; Long, 1999), while others have recommended a rescaling of such data to provide a closer approximation to interval-level measurement (Harwell & Gatti, 2001). The long-standing debate about how to treat ordinal variables in statistical analyses is fueled by the need within the social sciences to accurately answer research questions in which measures are obtained using a discrete ordinal response scale.

The purpose of this study was to investigate available analysis options for ordinal level data in the context of multiple regression analysis, and to empirically compare the performance of these analysis options. Multiple regression was the focus of this study because it has proven to be a useful, general purpose tool for data analysis encompassing the analysis of variance and analysis of covariance as special cases. As stated above, the literature reflects disagreement about appropriate analyses of discrete ordinal data. Therefore, one purpose of our work is to raise researchers' awareness of multiple regression analysis options with discrete ordinal variables. Further, the literature base is lacking empirical studies that investigate factors of the research context (such as effect size, number of regressors, sample size, and level of measurement) that may affect the performance of the analysis options. Therefore, a second purpose was to compare the performance of three analysis options with various manipulations of factors under controlled conditions.

Analysis Options with Discrete Ordinal Variables

Through the years, ordinal data have been analyzed using a variety of models, including ordinary least squares regression, variants of logistic regression analysis, and techniques developed specifically for ordinal-level analyses. Each of these methods will be briefly described.

Ordinary Least Squares. Ordinary least squares regression (OLS) analysis essentially ignores the discreteness and the ordinal level of measurement of variables, treating all values as if they represented a continuous interval-level measure. The well-known general linear model in matrix form is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon} \ ,$$

where $\boldsymbol{y}$ is an n x 1 vector of values for the dependent variable,
$\boldsymbol{X}$ is an n x $(k + 1)$ matrix of observations on the k regressor variables augmented with a unit vector to provide an intercept,
$\boldsymbol{b}$ is a $(k + 1)$ x1 vector of regression coefficients, and
$\boldsymbol{\varepsilon}$ is an n x 1 vector of residuals.

The OLS regression coefficients are obtained as:

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \;.$$

Hypothesis tests associated with OLS regression are typically of the form $H_0 : \beta_i = 0$, where β_i is the parameter associated with one of the elements of the vector $\boldsymbol{b}$. This test is conducted by obtaining the ratio of the parameter estimate to its standard error and comparing that ratio to a t distribution with df = n – k – 1. A simultaneous test that all of the β_i parameters are equal to zero (equivalent to the test that the population squared multiple correlation coefficient is equal to zero) is obtained by calculating the ratio

$$F = \frac{R^2 / k}{\left(1 - R^2\right) / \left(n - k - 1\right)}$$

and comparing the result to an F distribution with df = k, n – k – 1.

Ordinal Multiple Regression. Cliff (1994) presented an ordinal multiple regression strategy (OMR) and Long (1999) provided a method for estimating confidence intervals around the parameters of this model. The OMR model was developed to predict the ordinal information (that which Cliff refers to as “dominance”) in a criterion variable, based upon the observed ordinal information among the predictor variables. That is, OMR predicts the dominance scores on the criterion variable (d_{ihy}) by weighting the observed dominance scores on the regressor variables (d_{ihx}). The dominance scores are given as

$$d_{ihx} = sign\left(x_i - x_h\right),$$

taking the value of 1 if the i^{th} observation has a higher “score” on X than does the h^{th} observation, a value of –1 if the score on X is lower, and a value of zero if the two observations are tied on X. The weights for the OMR model are obtained as

$$\boldsymbol{w} = \boldsymbol{T}_{\boldsymbol{x}}^{-1} \boldsymbol{t}_{\boldsymbol{y}}$$

where $\boldsymbol{w}$ is a k X 1 vector of regression weights,
$\boldsymbol{T}_{\boldsymbol{x}}$ is a k X k matrix of Kendall tau-a correlation coefficients among the regressors, and
$\boldsymbol{t}_{\boldsymbol{y}}$ is a k X 1 vector of tau-a correlations between the regressors and the criterion variable.

Confidence intervals and hypothesis tests about the OMR weights ($\boldsymbol{w}$) are conducted using the standard normal (z) distribution. That is, the confidence interval for the j^{th} weight, described by Long (1999) is $w_j \pm Z_{(1-\alpha/2)} \hat{\sigma}_{wj}$ and a test of the null hypothesis that the population weight is zero is conducted by dividing the sample weight by its standard error and comparing the ratio to a critical value of Z.

Logistic Regression. A cumulative logit model for ordinal criterion variables (LR) was described by Agresti and Finlay (1997). The LR analysis models the probability of a value on the dependent variable being in the j^{th} category or lower, that is, modeling $P\left(y \le j\right)$ based on the values of predictor variables. In logit form, the model is

$$\log\left(\frac{P\left(y \le j\right)}{P\left(y > j\right)}\right) = \alpha_j + \beta X$$

where α_j is the intercept for the j^{th} category of the ordinal response, and
β is a k X 1 vector of regression weights for the predictors.

Parameter estimates for the LR models are usually obtained using maximum likelihood methods. Inference in the cumulative logit model is conducted using Wald tests for the individual regression parameters (ratios of the estimates to their standard errors), and likelihood ratio tests to obtain a simultaneous test that all parameters are equal to zero.

Method

This research was a Monte Carlo study in which random samples were generated under known and controlled population conditions. In the Monte Carlo study, samples were generated from populations of discrete ordinal variables. In each sample the data were analyzed using the OLS, OMR, and cumulative logit regression approaches.

The Monte Carlo study included six factors in the design. These factors were (a) the true population effect size of the individual regressors (populations were simulated with effect sizes corresponding to small, medium and large values of Cohen's effect size f^2, as well as a null condition (effect size of zero), (b) number of regressor variables (with k = 2, 5 and 10 regessors), (c) correlation between regressor variables (with r_{12} = .10 and .30), (d) sample sizes (with n = 5*k, 10*k and 100*k), (e) level of measurement of variables (discrete, ordinal variables were investigated as 5-option response scales, and 7-option response scales), and (f) population distribution shape (conditions were simulated in which the discrete ordinal variables were uniform, unimodal symmetric, and unimodal skewed).

Generation of Pseudo-populations. Because the population parameters corresponding to regression weights in the models being investigated for discrete ordinal data cannot be determined analytically, the simulation study was conducted by generating a pseudo-population for each condition, then drawing random samples with replacement from these pseudo-populations. The pseudo-populations consisted of 10,000 observations randomly generated from the corresponding population (level of discreteness of the measurements, distribution shape and degree of relationship between the regressors and the criterion variable).

The data were generated by transforming uniform random variates obtained from the RANUNI function in SAS, using a modification of the technique described by Bradley and Fleisher (1994), and operationalized by Ferron, Yi, and Kromrey (1997). In this method, a population correlation matrix, R, based on discrete ordinal variables is constructed by an iterative process in which large simulated samples (n = 100,000) are generated from an approximation of R, $(\tilde{R})$. The observed correlation matrix obtained from this large sample $(\hat{R})$ is compared elementwise to R, and the residuals $(R-\hat{R})$ are used to adjust the generating matrix $\tilde{R}$. This sequence of large sample generation, matrix estimation, and adjustment of $\tilde{R}$ continues until the process converges. The resulting matrix, $\tilde{R}$, is used as a template to generate correlated discrete data for the Monte Carlo study.

Each analysis model was calculated on all observations in these pseudo-populations providing values that served as the corresponding population parameters. The Monte Carlo simulation was then conducted by randomly sampling observations, with replacement, from these pseudo-populations.

The research was conducted using SAS/IML version 8.1. Conditions for the study were run under Windows 98. Discrete random variables were generated using the RANUNI function of SAS. A different seed value was used in each execution of the program and the program code was verified by hand-checking results from benchmark datasets.

For each condition investigated, 5,000 samples were generated. The use of 5,000 samples provides adequate precision for the investigation of the sampling behavior of these statistics. For example, 5,000 samples provides a maximum 95% confidence interval width around an observed proportion that is $\pm$ .014 (Robey & Barcikowski, 1992).

The relative performance of the analysis strategies was evaluated by a comparison of the Type I error control and statistical power of the tests of regression coefficients. Estimates of Type I error control and statistical power were obtained by conducting the hypothesis tests associated with each parameter of each analytical model. For the OLS and LR models, the simultaneous test that all regression parameters is equal to zero was also conducted. The proportion of rejections of hypothesis tests for conditions in which

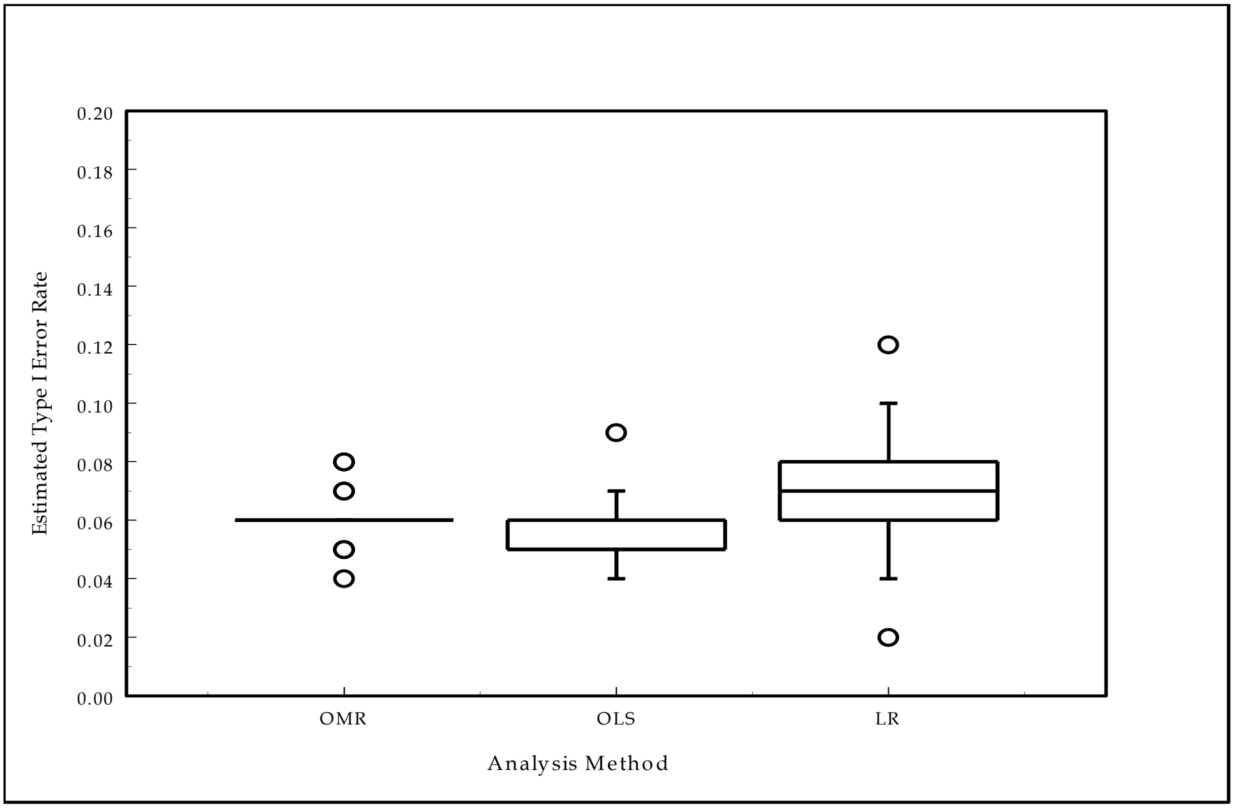


FIGURE 1: Distribution of Type I Error Rate Estimates for Tests of Individual Regression Coefficients

the population parameter is zero provided an estimate of Type I error control. Similarly, the proportion of rejections of such tests for conditions in which the parameter is not zero provided an estimate of statistical power.

Results

Type I Error Control

Box-and-whisker plots in Figure 1 depict the distribution of Type I error rate estimates for tests of individual regression weights across the conditions examined in the research. The figure clearly portrays a higher typical Type I error rate for the LR analysis method, relative to the OMR and OLS methods. The tails of the plot for the LR method range from .02 to .12, with half ranging from .06 to .08. The OMR method exhibits the most consistent Type I estimates, with the majority of estimates at .06. The OLS method is similar to the OMR with tails ranging from .04 to .07, with half ranging from .05 to .06.

Estimates of Type I error rates for the tests of individual regression weights in samples from the unimodal symmetric population are presented in Table 1. Overall the 5-point scale and 7-point scale exhibited few differences across the three analysis methods for all conditions. With two regressors and small sample size, the Type I error rate for the LR analysis method was lower than the OMR and OLS analysis methods. For example, with a 5-point scale, k=2, r_{12}=.30, and n=10, the estimated Type I error rate for OMR was .07, for OLS was .05, and for LR was .04. However, with more than 2 regressors and small to medium sample sizes, the Type I error rate was higher for LR when compared to OLS and OMR. For example, with a 5-point scale, k=10, r_{12}=.10, and n=50, the estimated Type I error rate for LR (.09) was higher than OMR (.05) and OLS (.05) . The OMR method had slightly higher Type I error rates with 2 regressors when compared to its performance with 5 and 10 regressors.

The analysis methods provided similar results across models with five and ten regressors, with LR consistently having higher Type I error rates than OMR or OLS. The OMR and OLS methods for small, medium, and larger sample sizes exhibited similar Type I error rates, which ranged from .05 to .07. In contrast the LR analysis method for small and medium sample sizes exhibited a higher Type I error rate,

Table 1. Type I Error Rate Estimates for Tests of Individual Regression Weights with Unimodal Symmetric Populations

5-point Scale			Analysis Method		
k	r_{12}	N	OMR	OLS	LR
2	.10	10	0.08	0.06	0.04
		20	0.07	0.05	0.06
		200	0.06	0.06	0.06
	.30	10	0.07	0.05	0.04
		20	0.06	0.05	0.06
		200	0.06	0.06	0.06
5	.10	25	0.06	0.05	0.08
		50	0.06	0.06	0.07
		500	0.06	0.06	0.06
	.30	25	0.06	0.05	0.09
		50	0.06	0.06	0.07
		500	0.06	0.06	0.06
10	.10	50	0.05	0.05	0.09
		100	0.06	0.06	0.08
		1000	0.06	0.06	0.07
	.30	50	0.05	0.05	0.10
		100	0.06	0.06	0.08
		1000	0.06	0.07	0.07
7-point Scale			Analysis Method		
k	r_{12}	N	OMR	OLS	LR
2	.10	10	0.08	0.05	0.05
		20	0.06	0.05	0.06
		200	0.06	0.05	0.06
	.30	10	0.07	0.05	0.05
		20	0.06	0.05	0.06
		200	0.06	0.06	0.05
5	.10	25	0.06	0.06	0.09
		50	0.06	0.05	0.07
		500	0.06	0.06	0.06
	.30	25	0.06	0.05	0.09
		50	0.06	0.06	0.08
		500	0.05	0.05	0.05
10	.10	50	0.06	0.05	0.05
		100	0.06	0.06	0.07
		1000	0.05	0.05	0.05
	.30	50	0.06	0.06	0.10
		100	0.06	0.06	0.08
		1000	0.07	0.06	0.07

which ranged from .06 to .10. For the small sample conditions, the LR method provided slightly higher Type I error rates with 10 regressors than with 5 regressors (although it remained liberal with small samples in both conditions). For example, with a 7-point scale, k=10, r_{12}=.30, and n=50, the estimated Type I error rate for LR was .10. In contrast, with k = 5, r_{12}=.30, and n = 25, the Type I error rate estimate for LR dropped to .09. The LR method evidenced better Type I error control as the sample size increased. For example, with a 7-point scale, k=5, r_{12}=.30, the LR method was .09 with the small sample, .08 with the medium sample, and .05 with the large sample.

The Type I error rate estimates for tests of individual regression weights in selected conditions across distribution shapes are presented in Table 2. The three distribution shapes (unimodal symmetric, uniform,

Table 2. Type I Error Rates Estimates for Tests of Individual Regression Weights Across Distribution Shapes.

5-point Scale			Symmetric			Uniform			Skewed		
k	r_{12}	N	OMR	OLS	LR	OMR	OLS	LR	OMR	OLS	LR
2	.10	10	0.08	0.06	0.04	0.07	0.04	0.05	0.08	0.05	0.02
		20	0.07	0.05	0.06	0.07	0.05	0.07	0.07	0.05	0.05
	.30	10	0.07	0.05	0.04	0.07	0.06	0.05	0.08	0.05	0.02
		20	0.06	0.05	0.06	0.06	0.05	0.06	0.07	0.05	0.05
5	.10	25	0.06	0.05	0.08	0.06	0.06	0.09	0.06	0.05	0.08
		50	0.06	0.06	0.07	0.06	0.05	0.07	0.06	0.05	0.07
	.30	25	0.06	0.05	0.09	0.06	0.05	0.09	0.06	0.06	0.07
		50	0.06	0.06	0.07	0.07	0.06	0.08	0.06	0.05	0.06
10	.10	50	0.05	0.05	0.09	0.05	0.05	0.09	0.06	0.05	0.09
		100	0.06	0.06	0.08	0.05	0.05	0.07	0.05	0.05	0.07
	.30	50	0.05	0.05	0.10	0.05	0.05	0.09	0.04	0.09	0.12
		100	0.06	0.06	0.08	0.05	0.05	0.07	0.05	0.05	0.07
7-point Scale			Symmetric			Uniform			Skewed		
k	r_{12}	N	OMR	OLS	LR	OMR	OLS	LR	OMR	OLS	LR
2	.10	10	0.08	0.05	0.05	0.08	0.05	0.07	0.07	0.04	0.04
		20	0.06	0.05	0.06	0.06	0.05	0.07	0.06	0.05	0.06
	.30	10	0.07	0.05	0.05	0.08	0.05	0.07	0.07	0.06	0.04
		20	0.06	0.05	0.06	0.06	0.05	0.06	0.07	0.05	0.06
5	.10	25	0.06	0.06	0.09	0.05	0.05	0.09	0.06	0.05	0.08
		50	0.06	0.05	0.07	0.06	0.05	0.07	0.06	0.06	0.07
	.30	25	0.06	0.05	0.09	0.05	0.05	0.09	0.06	0.06	0.08
		50	0.06	0.06	0.08	0.06	0.05	0.07	0.06	0.05	0.07
10	.10	50	0.06	0.05	0.05	0.05	0.05	0.09	0.06	0.06	0.10
		100	0.06	0.06	0.07	0.06	0.06	0.07	0.06	0.06	0.08
	.30	50	0.06	0.06	0.10	0.05	0.05	0.10	0.05	0.09	0.12
		100	0.06	0.06	0.08	0.06	0.06	0.08	0.06	0.06	0.07

and skewed) yielded similar Type I error rates for all three analysis methods. For example, with the 7-point scale, k=5, r_{12}=.10, and n=25, the estimated Type I error rate ranged from .05 to .06 across the distribution shapes for the OMR and OLS analysis method. Similarly, for the LR analysis method the estimated Type I error rate ranged from .08 to .09 across the distribution shapes.

Figure 2 contains box-and-whisker plots depicting the distribution of Type I error rate estimates for simultaneous tests of regression coefficients. The LR method portrays a wide range of estimates, .00 to .18, with half ranging from .00 to .09. The OLS method portrays a much smaller range of estimates, .04 to .07, with half ranging from .05 to .06. Thus the OLS method exhibits a much smaller type I error rate for simultaneous tests of regression coefficients across all research conditions.

Table 3 presents the estimates of Type I error rates for the simultaneous test of all regression weights when samples were drawn from the unimodal symmetric populations. As evident in this table, the LR method evidenced substantially greater variability in Type I error control for this test than did the OLS method. Specifically, the LR method was very conservative in conditions with both k = 2 and k = 5 and was liberal with k = 10, with Type I error rates reaching as high as .14 (with the 7-point scale, k = 10, r_{12}=.3, and n=50). The OLS method maintained Type I error rates near the nominal level across these conditions, with larger estimates in the larger sample size and 10k conditions. The Type I error rates for these tests across distribution shapes (Table 4) indicated a slight increase in Type I error rates for the OLS method when samples were drawn from the 7 point, 10 regressors, symmetric and skewed distributions, relative to the rates obtained from the uniform distribution. However, these higher rates remained very close to the nominal alpha level and did not exceed .09.

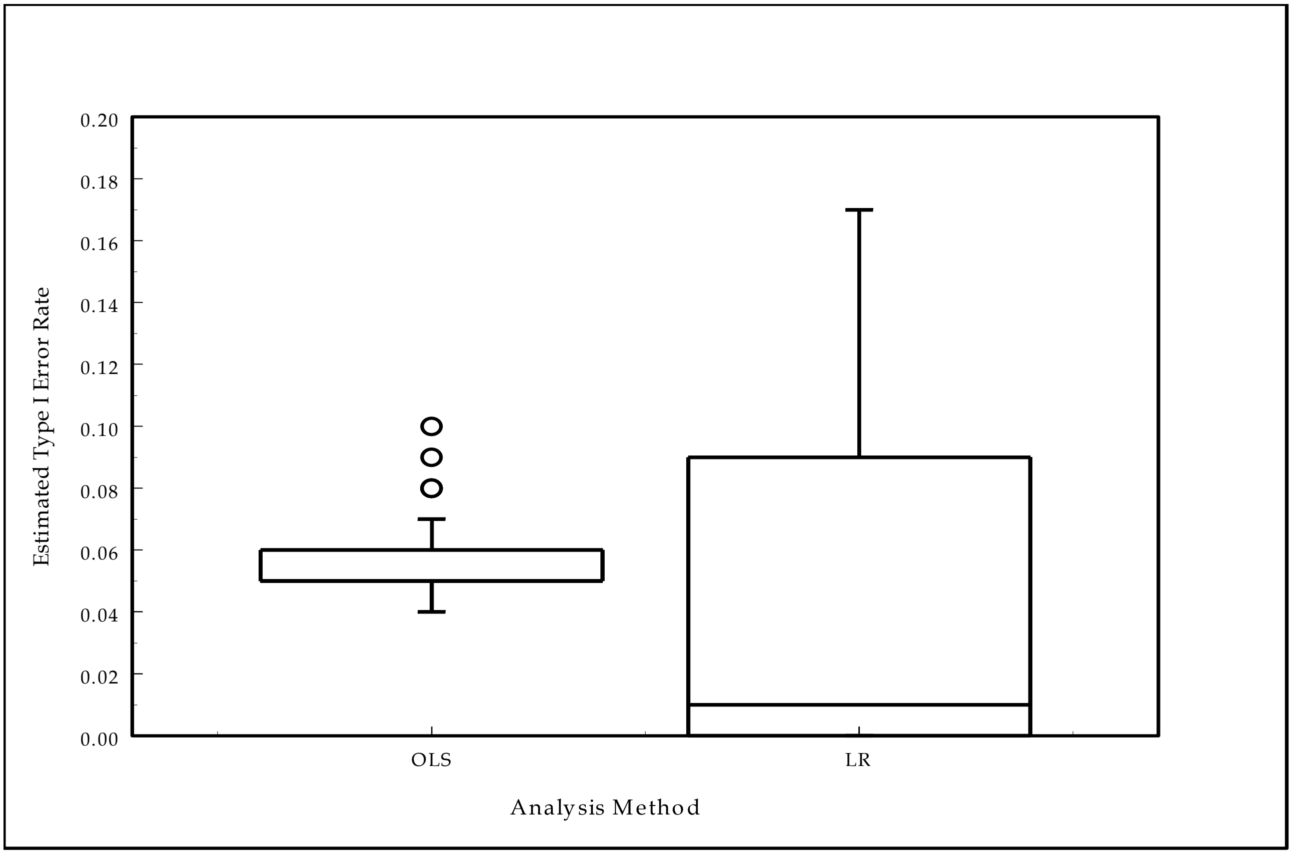


FIGURE 2: Distribution of Type I Error Rate Estimates for Simultaneous Test of Regression Coefficients

Statistical Power

The distribution of estimates of statistical power for the tests of regression weights across the conditions examined in this research are presented as box-and-whisker plots in Figure 3. As is evident in this figure, most of the conditions examined yielded relatively low power and the power differences across analysis methods was not great. However, the LR analysis method provided slight power advantages relative to OMR and OLS analyses (a result that should be expected because of its tendency to be liberal in Type I error control).

Power estimates for the tests of individual regression weights with samples drawn from the three distribution shapes are presented in Tables 5 and 6. Few differences were evident between the analyses of variables with 5 points and those with 7 points. With small effect sizes (f^2 = .02) the power estimates for all analysis methods ranged from .04 to .12. With large effect sizes (f^2 = .35), modest but greater power differences were evident across the analysis methods. With a small number of regressors (k = 2), the OMR method evidenced power advantages over the OLS and LR methods with small samples. With a 5-point scale, f^2 = .35, k = 2, r_{12} = .10, and n = 10, for example, the statistical power of the OMR method was estimated to be .19, while the power estimates of the OLS and LR methods were .15 and .09, respectively (Table 5). These slight power advantages of the OMR method disappeared with larger samples and larger numbers of regressors (conditions in which the LR approach evidenced modest power advantages). The differences in statistical power between tests conducted on samples from unimodal symmetric, uniform, and skewed distributions are negligible. For example, with the 7-point response scale, f^2 = .35, k = 5, r_{12} = .10, and n = 50, the power estimates for the OMR method ranged from .21 for the skewed population to .28 for the uniform population (Table 6). Similarly, the power estimates for OLS ranged only from .28 to .31, and those of the LR method ranged from .32 to .35.

The distribution of power estimates for simultaneous tests of regression coefficients across all research conditions is presented with box-and-whisker plots in Figure 4. The OLS method evidenced a slight power advantage over the LR method. Power estimates for the simultaneous test of all regression

Table 3. Estimates of Type I Error Rate for Simultaneous Tests of All Regression Weights with Unimodal Symmetric Populations

			5-Point Scale Analysis Method		7-Point Scale Analysis Method	
k	r_{12}	N	OLS	LR	OLS	LR
2	.10	10	0.05	0.01	0.05	0.01
		20	0.05	0.00	0.04	0.00
		200	0.06	0.00	0.06	0.00
	.30	10	0.05	0.00	0.05	0.01
		20	0.05	0.00	0.05	0.00
		200	0.05	0.00	0.06	0.00
5	.10	25	0.05	0.01	0.06	0.01
		50	0.06	0.01	0.06	0.01
		500	0.08	0.00	0.07	0.00
	.30	25	0.06	0.01	0.05	0.02
		50	0.05	0.01	0.08	0.01
		500	0.08	0.01	0.05	0.00
10	.10	50	0.06	0.11	0.07	0.13
		100	0.06	0.10	0.06	0.09
		1000	0.08	0.08	0.08	0.08
	.30	50	0.06	0.14	0.07	0.14
		100	0.07	0.10	0.07	0.10
		1000	0.10	0.10	0.08	0.09

weights with samples drawn from unimodal symmetric, uniform, and skewed populations are presented in Tables 7 and 8. With a small effect size (.02) and k = 2 or k = 5, the LR test provided rejection rates less than the nominal alpha level across most conditions. For the OLS test, similarly low power levels were observed except for large samples (and power estimates were still less than .50 for these conditions). As expected, with larger effect sizes the power estimates for both methods increased. The OLS method evidenced substantially greater power than the LR method for k = 2 and k = 5, while the LR method evidenced smaller power advantages for k = 10. A comparison of power estimates across distribution shapes suggests that population shape is not a major influence on the power of either the OLS or the LR method.

Conclusions

Surprisingly few and relatively small differences were evident among the OMR, OLS and LR methods in terms of their Type I error control and statistical power in tests of their respective weights. The OMR approach, while accurately representing the ordinal nature of the discrete response scales and discrete regressor variables, provided neither superior Type I error control nor superior statistical power. In contrast, the LR analysis evidenced Type I error control problems with small sample sizes or large numbers of regressors, conditions in which the Wald tests of the logistic regression weights became liberal. Although this analysis method presents modest power advantages relative to OMR and OLS, such power comes at a cost of relatively tenuous Type I error control. Finally, the surprisingly good performance of the OLS approach suggests that researchers who approach the analysis of discrete ordinal data (such as individual Likert items) with OLS tools should feel no guilt in such a tactic. The Type I error control in tests conducted of the OLS regression weights was as good as that obtained with tests of OMR weights and was superior to tests obtained in the LR context. Further, the statistical power evidenced with OLS was comparable or superior to that of OMR.

Of course, testing hypotheses about weights obtained in these models are only a small part of the inferential machinery applied to discrete ordinal data. Additional research is needed to focus on the relative bias in sample estimates of weights for OMR, OLS, and LR models and on the accuracy of confidence bands constructed around the sample estimated weights. Furthermore, research conducted on models that present a mixture of discrete ordinal variables and continuous variables is needed in order to explore the relative performance of these models in such a context. Finally, the substantive nature of the

Table 4. Type I Error Rate Estimates for Simultaneous Tests of All Regression Weights Across Distribution Shapes.

5-point Scale			Symmetric		Uniform		Skewed	
k	r_{12}	N	OLS	LR	OLS	LR	OLS	LR
2	.10	10	0.05	0.01	0.05	0.00	0.05	0.00
		20	0.05	0.00	0.06	0.00	0.05	0.00
	.30	10	0.05	0.00	0.06	0.01	0.05	0.01
		20	0.05	0.00	0.05	0.00	0.05	0.00
5	.10	25	0.05	0.01	0.06	0.02	0.05	0.01
		50	0.06	0.01	0.06	0.00	0.05	0.01
	.30	25	0.06	0.01	0.06	0.01	0.05	0.01
		50	0.05	0.01	0.06	0.01	0.05	0.01
10	.10	50	0.06	0.11	0.05	0.12	0.05	0.12
		100	0.06	0.10	0.05	0.09	0.06	0.09
	.30	50	0.06	0.14	0.04	0.13	0.04	0.14
		100	0.07	0.10	0.05	0.09	0.05	0.12

7-point Scale			Symmetric		Uniform		Skewed	
k	r_{12}	N	OLS	LR	OLS	LR	OLS	LR
2	.10	10	0.05	0.01	0.05	0.00	0.05	0.00
		20	0.04	0.00	0.05	0.00	0.05	0.00
	.30	10	0.05	0.01	0.05	0.01	0.06	0.01
		20	0.05	0.00	0.05	0.00	0.05	0.00
5	.10	25	0.06	0.01	0.04	0.01	0.04	0.01
		50	0.06	0.01	0.05	0.00	0.06	0.01
	.30	25	.0.05	0.02	0.05	0.01	0.06	0.02
		50	0.08	0.01	0.05	0.00	0.06	0.01
10	.10	50	0.07	0.13	0.05	0.12	0.07	0.15
		100	0.06	0.09	0.06	0.09	0.09	0.12
	.30	50	0.07	0.14	0.05	0.15	0.05	0.17
		100	0.07	0.10	0.05	0.09	0.06	0.13

inferences suggested by these models requires attention from the perspective of psychometrics. That is, the OMR model provides a vehicle for inferences about the ordinal position of observations on the criterion variable, while the LR model estimates the probability of an observation being "in category j or lower" on the criterion variable. In contrast, the OLS model provides a prediction of the criterion value as though it was a continuous variable (i.e., seeking to minimize the sum of squared errors of prediction). In addition (although not analyzed in this research study), methods for determining both the strength and direction of the response have been proposed (Jones & Sobel (2000); Brody & Dietz (1997) . These types of inferences have obvious substantive differences, and the validity of such inferences extends beyond the estimation of Type I error rates and statistical power.

In conclusion, this study has sought to bring increased awareness and clarity to three analysis options for multiple regression with discrete ordinal variables. The social sciences often focus on variables which are measured with ordinal scales (most commonly Likert scales). Unfortunately, the appropriate application of multiple regression with discrete ordinal data has been insufficiently addressed in the literature. There are many reasons for this paucity of treatment of the issue: general debates continue about the treatment of ordinal data with statistics, the breadth of analysis options are spread among numerous sources, and the consequences of analysis choice in terms of Type I error control and statistical power has not been thoroughly investigated. The current study presents three analysis options within the literature and clarifies their differences with the hopes of both increasing the appropriate analysis of discrete ordinal variables and stimulating additional methodological research in this area.

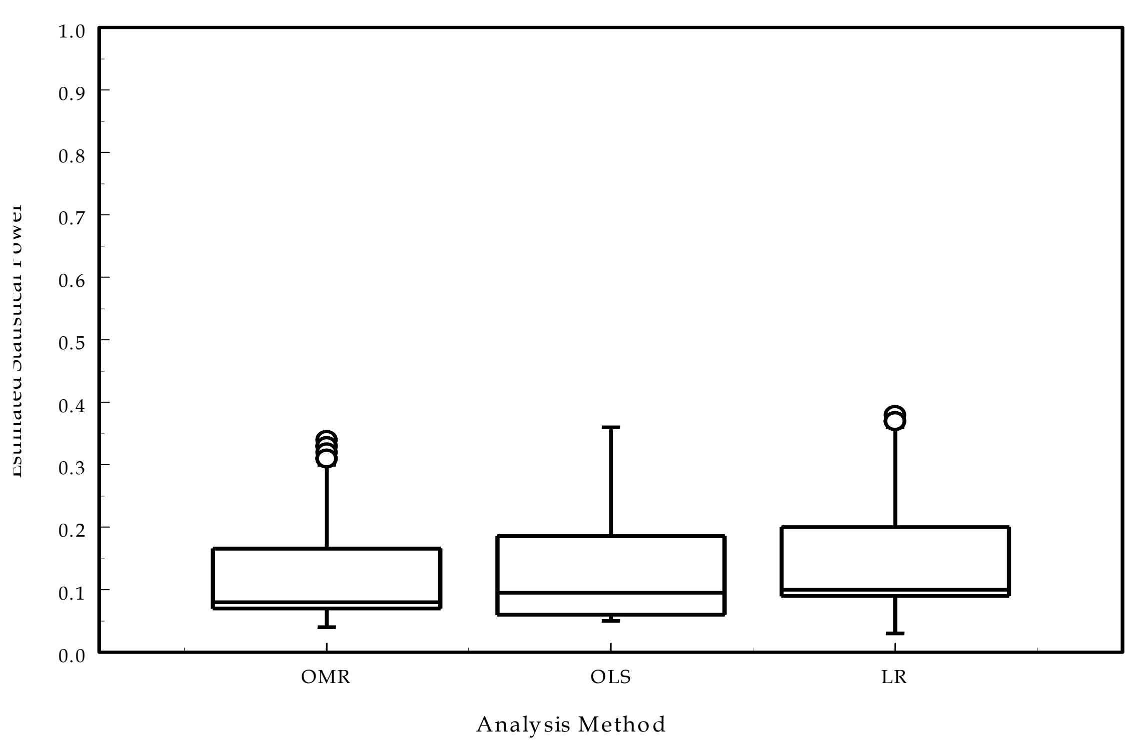


FIGURE 3: Distribution of Statistical Power Estimates for Tests of Individual Regression Coefficients

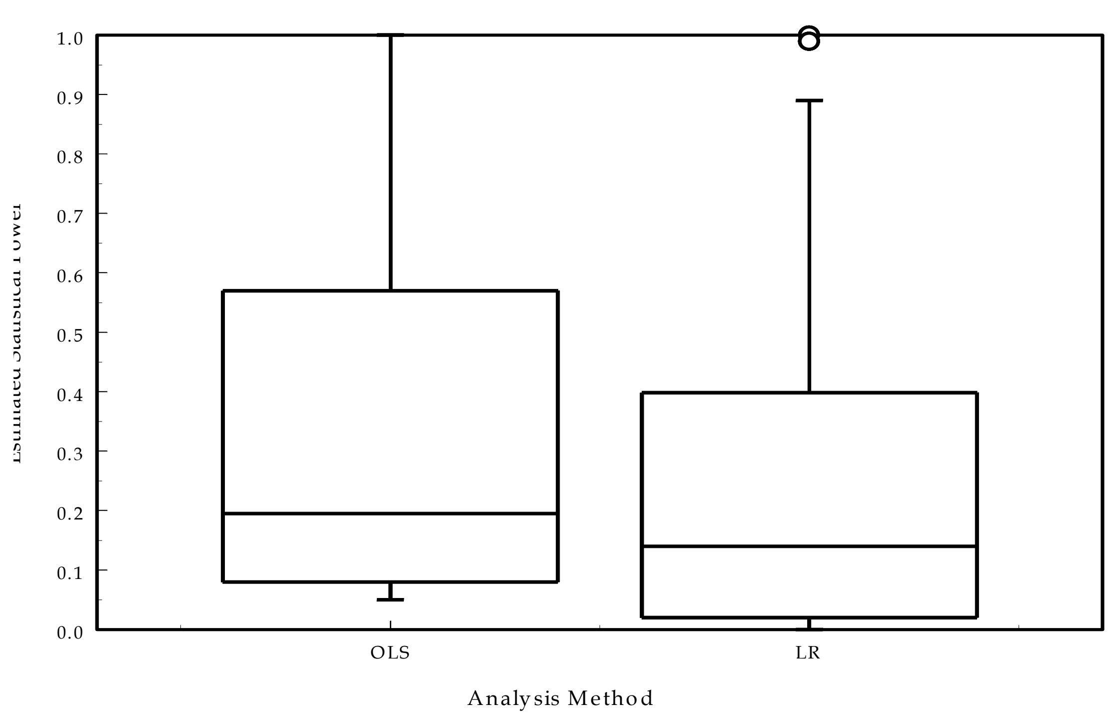


FIGURE 4: Distribution of Statistical Power Estimates for Simultaneous Test of Regression Coefficients

Table 5. Estimates of Statistical Power for Tests of Individual Regression Weights Across Distribution Shapes.

5 Point Likert				Symmetric			Uniform			Skewed		
Effect	k	r_{12}	N	OMR	OLS	LR	OMR	OLS	LR	OMR	OLS	LR
.02	2	.10	10	0.09	0.06	0.04	0.09	0.05	0.04	0.08	0.06	0.03
			20	0.09	0.07	0.09	0.08	0.06	0.08	0.08	0.07	0.07
		.30	10	0.09	0.06	0.04	0.09	0.05	0.03	0.08	0.07	0.03
			20	0.08	0.07	0.08	0.10	0.08	0.10	0.07	0.08	0.08
	5	.10	25	0.07	0.06	0.09	0.07	0.06	0.10	0.06	0.06	0.09
			50	0.08	0.07	0.09	0.08	0.07	0.09	0.07	0.07	0.08
		.30	25	0.06	0.05	0.09	0.07	0.05	0.09	0.05	0.11	0.11
			50	0.11	0.10	0.12	0.06	0.06	0.07	0.05	0.08	0.08
	10	.10	50	0.07	0.06	0.10	0.06	0.06	0.10	0.06	0.06	0.10
			100	0.07	0.07	0.09	0.07	0.06	0.08	0.06	0.07	0.09
		.30	50	0.06	0.05	0.09	0.05	0.05	0.09	0.04	0.10	0.15
			100	0.07	0.06	0.08	0.06	0.06	0.08	0.05	0.07	0.09
.35	2	.10	10	0.19	0.15	0.09	0.18	0.16	0.12	0.18	0.19	0.08
			20	0.34	0.34	0.37	0.31	0.32	0.35	0.28	0.35	0.36
		.30	10	0.19	0.14	0.07	0.17	0.14	0.10	0.14	0.18	0.07
			20	0.30	0.27	0.30	0.33	0.31	0.34	0.25	0.31	0.30
	5	.10	25	0.15	0.16	0.22	0.13	0.14	0.20	0.13	0.19	0.23
			50	0.27	0.30	0.34	0.24	0.26	0.29	0.22	0.33	0.36
		.30	25	0.11	0.11	0.16	0.10	0.10	0.15	0.07	0.19	0.17
			50	0.20	0.20	0.23	0.18	0.18	0.21	0.11	0.23	0.22
	10	.10	50	0.11	0.14	0.19	0.11	0.13	0.19	0.09	0.15	0.19
			100	0.21	0.24	0.28	0.21	0.24	0.28	0.16	0.25	0.27
		.30	50	0.07	0.08	0.14	0.05	0.08	0.14	0.04	0.19	0.19
			100	0.11	0.12	0.15	0.09	0.12	0.17	0.05	0.18	0.17

References

Agresti, A. & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd Ed.). Upper Saddle River, NJ: Prentice-Hall.

Borgatta, E. F. & Bohrnstedt, G. W. (1980). Level of measurement: Once over again. *Sociological Methods & Research*, *9*(2), 147-160.

Bradley, D. R. & Fleisher, C. L. (1994). Generating multivariate data from nonnormal distributions: Mihal and Barrett revisited. *Behavior Research Methods, Instruments, and Computers*, *26*, 156-166.

Brody, C. J. & Dietz, J. (1997). On the dimensionality of two-question format likert attitude scales. *Social Science Research*, *26*, 197-204.

Cliff, N. (1994). Predicting ordinal relations. *British Journal of Mathematical and Statistical Psychology*, *47*, 127-150.

Cliff, N. (1996). *Ordinal Methods for Behavioral Data Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ferron, J., Yi, Q. & Kromrey, J.D. (1997). NNCORR: A SAS/IML program for generating nonnormal correlated data. *Applied Psychological Measurement*, *21*, 64.

Table 6. Estimates of Statistical Power for Tests of Individual Regression Weights Across Distribution Shapes.

7 point Likert				Symmetric			Uniform			Skewed		
Effect	k	r_{12}	N	OMR	OLS	LR	OMR	OLS	LR	OMR	OLS	LR
.02	2	.10	10	0.08	0.06	0.06	0.08	0.05	0.07	0.08	0.06	0.04
			20	0.08	0.07	0.09	0.08	0.07	0.09	0.07	0.07	0.08
		.30	10	0.08	0.05	0.05	0.08	0.06	0.07	0.09	0.07	0.04
			20	0.08	0.07	0.08	0.08	0.06	0.08	0.09	0.08	0.08
	5	.10	25	0.07	0.06	0.10	0.07	0.06	0.10	0.07	0.08	0.11
			50	0.07	0.07	0.09	0.07	0.07	0.09	0.07	0.08	0.09
		.30	25	0.06	0.05	0.09	0.06	0.05	0.09	0.05	0.09	0.11
			50	0.07	0.07	0.09	0.06	0.06	0.08	0.06	0.07	0.10
	10	.10	50	0.06	0.06	0.11	0.06	0.06	0.10	0.06	0.06	0.10
			100	0.08	0.07	0.09	0.07	0.07	0.09	0.07	0.07	0.09
		.30	50	0.06	0.05	0.09	0.05	0.06	0.10	0.04	0.14	0.18
			100	0.06	0.06	0.08	0.06	0.06	0.08	0.04	0.09	0.09
.35	2	.10	10	0.18	0.15	0.12	0.18	0.16	0.17	0.16	0.20	0.10
			20	0.32	0.33	0.37	0.33	0.34	0.38	0.26	0.36	0.37
		.30	10	0.18	0.13	0.09	0.17	0.14	0.14	0.12	0.15	0.10
			20	0.31	0.28	0.31	0.29	0.28	0.32	0.25	0.32	0.32
	5	.10	25	0.14	0.14	0.20	0.14	0.16	0.23	0.12	0.18	0.22
			50	0.26	0.28	0.32	0.28	0.31	0.35	0.21	0.31	0.34
		.30	25	0.12	0.11	0.16	0.11	0.11	0.17	0.07	0.19	0.18
			50	0.20	0.18	0.22	0.19	0.19	0.22	0.13	0.24	0.24
	10	.10	50	0.11	0.13	0.19	0.11	0.13	0.19	0.10	0.16	0.21
			100	0.21	0.23	0.28	0.22	0.25	0.28	0.18	0.26	0.28
		.30	50	0.07	0.08	0.14	0.07	0.09	0.16	0.04	0.19	0.20
			100	0.12	0.13	0.16	0.12	0.13	0.18	0.05	0.19	0.19

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*(*3*), 564-567.

Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*, 105–131.

Jones, B. S. & Sobel, M. E. (2000). Modeling direction and intensity in semantically balanced ordinal scales: An assessment of congressional incumbent approval. *American Journal of Political Science*, *44*(*1*), 174-185.

Long, J. D. (1999). A confidence interval for ordinal multiple regression weights. *Psychological Methods*, *4*(*3*), 315-330.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* 1-49. New York: Wiley.

Robey, R. R. & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.

Table 7. Estimates of Statistical Power for Simultaneous Tests of All Regression Weights Across Distribution Shapes.

5 Point Likert				Symmetric		Uniform		Skewed	
Effect	k	r_{12}	N	OLS	LR	OLS	LR	OLS	LR
.02	2	.10	10	0.07	0.01	0.05	0.01	0.07	0.01
			20	0.08	0.00	0.08	0.00	0.07	0.00
		.30	10	0.06	0.01	0.05	0.01	0.08	0.01
			20	0.07	0.00	0.09	0.00	0.09	0.00
	5	.10	25	0.06	0.01	0.08	0.03	0.08	0.02
			50	0.11	0.01	0.13	0.02	0.12	0.01
		.30	25	0.06	0.02	0.06	0.02	0.10	0.02
			50	0.35	0.18	0.08	0.01	0.13	0.02
	10	.10	50	0.10	0.18	0.08	0.16	0.11	0.17
			100	0.14	0.18	0.13	0.17	0.15	0.17
		.30	50	0.09	0.20	0.06	0.19	0.17	0.19
			100	0.15	0.22	0.10	0.16	0.17	0.23
.35	2	.10	10	0.21	0.02	0.22	0.03	0.28	0.02
			20	0.53	0.04	0.50	0.03	0.53	0.04
		.30	10	0.22	0.03	0.22	0.02	0.26	0.02
			20	0.52	0.04	0.55	0.04	0.54	0.04
	5	.10	25	0.44	0.20	0.42	0.18	0.54	0.24
			50	0.85	0.56	0.80	0.49	0.87	0.58
		.30	25	0.51	0.27	0.38	0.18	0.52	0.22
			50	0.90	0.66	0.84	0.52	0.86	0.52
	10	.10	50	0.74	0.87	0.72	0.84	0.73	0.82
			100	0.99	1.00	0.99	0.99	0.98	0.99
		.30	50	0.69	0.84	0.71	0.89	0.75	0.81
			100	0.98	0.99	0.99	1.00	0.99	0.99

Table 8. Estimates of Statistical Power for Simultaneous Tests of All Regression Weights Across Distribution Shapes.

7 Point Likert				Symmetric		Uniform		Skewed	
Effect	k	r_{12}	N	OLS	LR	OLS	LR	OLS	LR
.02	2	.10	10	0.06	0.00	0.06	0.01	0.07	0.01
			20	0.08	0.00	0.09	0.00	0.09	0.00
		.30	10	0.06	0.01	0.06	0.00	0.08	0.00
			20	0.08	0.00	0.08	0.00	0.12	0.00
	5	.10	25	0.07	0.02	0.07	0.02	0.14	0.03
			50	0.11	0.02	0.10	0.02	0.13	0.02
		.30	25	0.06	0.02	0.08	0.02	0.10	0.03
			50	0.11	0.02	0.09	0.01	0.13	0.02
	10	.10	50	0.08	0.17	0.08	0.19	0.11	0.18
			100	0.13	0.20	0.14	0.20	0.15	0.19
		.30	50	0.06	0.14	0.06	0.18	0.08	0.18
			100	0.10	0.14	0.10	0.16	0.19	0.22
.35	2	.10	10	0.20	0.02	0.22	0.02	0.27	0.02
			20	0.51	0.03	0.52	0.03	0.54	0.04
		.30	10	0.20	0.02	0.23	0.02	0.23	0.03
			20	0.53	0.05	0.51	0.04	0.56	0.05
	5	.10	25	0.40	0.18	0.46	0.21	0.49	0.22
			50	0.81	0.52	0.86	0.58	0.84	0.55
		.30	25	0.45	0.23	0.44	0.21	0.56	0.25
			50	0.85	0.58	0.86	0.55	0.88	0.55
	10	.10	50	0.68	0.82	0.70	0.82	0.76	0.84
			100	0.98	0.99	0.99	1.00	0.98	0.99
		.30	50	0.74	0.88	0.76	0.89	0.82	0.87
			100	0.99	1.00	1.00	1.00	0.99	1.00

Send correspondence to: Jeffrey D. Kromrey
Department of Educational Measurement and Research
University of South Florida
4202 E. Fowler Ave., EDU162, Tampa, FL 33620
Email: Kromrey@tempest.coedu.usf.edu

A Strategy for Addressing the Validity of a Teacher Effectiveness Instrument

Dale Shaw
University of Northern Colorado

Suzanne Young
University of Wyoming

Jay Schaffer
University of Northern Colorado

Daniel Mundfrom
University of Northern Colorado

This paper deals with the validation of an objective teacher effectiveness instrument for rating the classroom effectiveness of college and university teachers. It includes a description of how the instrument was developed and the process by which validity evidence for the instrument was generated and analyzed via regression and factor analyses.

The purpose of this study was to create a research-based teacher evaluation instrument and analyze data gathered with it to obtain validity evidence for its use as a measure of college and university teacher effectiveness. Institutions use such instruments to collect student ratings of teachers for one or more of the following purposes: (1) to provide teachers with feedback for improving their teaching, (2) to provide students with information they may use to select future courses and instructors, and (3) to provide administrators and faculty with a measure of a teacher's effectiveness that may inform their decisions about a faculty member's tenure, promotion, or retention (Marsh & Dunkin, 1992). Typically the validation of such an instrument requires several studies dealing with at least two aspects of validity: (1) to determine the degree to which obtained ratings reflect the true feelings of students, which is important for the first and second purposes above, and (2) to establish the degree to which the items collectively capture, or truly do measure, the construct of teacher effectiveness, which is important to the third purpose. The focus of this paper is on those aspects of validity that surround the instrument's use for the third purpose, that is, to provide a measure of teacher effectiveness.

This paper consists of an explanation of how the instrument was developed and a description of how data were collected and analyzed for validity evidence. First, items were developed that met two or more of the following three criteria: (1) the item is used prevalently in other teacher rating instruments, (2) the item bears a high relationship to the global construct of teacher effectiveness as evidenced in previous research, or (3) it is a key item in a previously developed teacher effectiveness model (i.e., McKay, 1997). In her model of teacher effectiveness, McKay argues that the three most important items to include in a teacher rating instrument are subject matter knowledge, teacher enthusiasm, and communication skills. Second, data were gathered about college and university teachers from former students in an effort to acquire data about teachers from the entire spectrum of teacher effectiveness. Third, these data were submitted to regression and factor analyses. Evidence of the instrument's construct validity could be indicated in several ways, including: (1) high multiple correlation coefficients between a global score and the collection of items or subsets of the items suggested by previous teacher effectiveness models (i.e., McKay, 1997), (2) high factor loadings in the first extracted principle component suggesting that the items provide a common measure of a unitary construct, (3) obtaining a meaningful factor structure consistent with the work of other teacher effectiveness researchers (Marsh, 1991; Marsh & Hocevar, 1984, 1991; Abrami, d'Apollonia, & Rosenfield,1997).

Instrument Development

One hundred twenty-five different items were gleaned from objective teacher effectiveness instruments described in research studies published since 1985. Only items that were demonstrated to be correlates of teacher effectiveness in the studies wherein students provided ratings of teachers were selected for our study. In all we found 44 studies that identified items that were teacher effectiveness correlates. This pool of one hundred twenty-five items was analyzed for duplicates and near-duplicates, and was edited to achieve a uniformity of presentation in style and format. Twenty-five items from this pool were retained for further consideration. We relied heavily on the works of Feldman (1976, 1984, & 1986), Murray (1980), Erdle, Murray, & Rushton (1985) and Marsh (1987) as we sought to assess the adequacy of the twenty-five items to collectively capture the construct of teacher effectiveness. The twenty-five items include all but two of the nineteen instructional rating dimensions that Feldman (1976) identified in his classic teacher effectiveness review study as well as two additional items recommended by Murray (1980). These items are teacher's interest in the course, enthusiasm, subject matter knowledge, breadth of subject coverage, preparation and organization, presentation skills, speaking skills, sensitivity to student achievement, clarity of objectives, value of the course, value of supplementary materials, classroom management, course difficulty including appropriateness of workload, fairness, value and frequency of feedback, openness, encouragement and challenge, respect and friendliness, availability, clear explanations and encouragement of student participation.

A pilot study of the twenty-five-item instrument led us to conclude that, at twenty-five items, the instrument was much too long to be practical. An eleven-item version was developed from the twenty-five-item version by selecting in large part those items that bore the highest relationships with teacher effectiveness while still covering the spectrum of issues captured in the original item pool. The eleven items are presented in Table 1. In the form for administering the items, a 9-interval rating scale from 1 to 9 with anchors 1 (Very Low), 3 (Low), 5 (Average), 7 (High), and 9 (Very High) followed the presentation of each item.

Data for addressing the validity of the instrument were obtained from students in 22 undergraduate and graduate classes who were asked to rate three professors of their choice from whom they had taken a course in the recent past. The students were given a brief training regarding halo effect and leniency effect in ratings and admonished to not succumb to these rater errors as they filled out the instrument. They were also asked to select professors to rate from a variety of points along the teacher effectiveness continuum to the extent that it was possible for them to do so. In a cover sheet, the students were given written instructions regarding the study and an overall or global rating item to be filled out for each instructor that they planned to rate on subsequent rating forms. The global item, that served as the criterion variable in the regression analyses below, was worded "Everything considered, I would rate the instructor's effectiveness" and was rated on the same 1 to 9 scale as the 11 items. In all, 1082 useable cases were obtained from 384 students. These data was submitted to regression and factor analyses in an effort to acquire evidence of the 11-item instrument's validity to measure college and university teacher effectiveness.

Regression analysis

Table 2 presents information about 4 regression models. The first model is the complete model derived from the data collected in this study by regressing the global score onto all eleven items. An R^2 of 0.8918 was obtained for this model indicating that 89% of the variance in the global scores is accounted

Table 1. Instrument Items

Item Name	Actual Wording on the Instrument
1. Subject matter knowledge	The instructor's subject matter knowledge
2. Communication skills	The effectiveness of the instructor's communication skills
3. Enthusiasm	The instructor's enthusiasm for teaching
4. Comfortable atmosphere	The degree to which the instructor created a comfortable learning atmosphere
5. Respectful of students	The degree to which the instructor was respectful of students
6. Warm and friendly	The instructor's warmth and friendliness
7. Motivate & stimulate	The degree to which the instructor was motivating and stimulating
8. Concern for learning	The instructor's genuine concern for student learning
9. Increased interest	The degree to which the course increased my interest in the subject
10. Increased understanding	The degree to which the course increased my understanding of concepts
11. Course organization	The degree to which the course was well organized
Global Item	Everything considered, I would rate the instructor's effectiveness

Table 2. Regression Models

Item	Complete	Feldman	Young/Shaw	McKay
Subject matter knowledge	X*	X		X
Communication skills	X	X	X	X
Enthusiasm	X	X		X
Comfortable atmosphere	X			
Respectful of students	X	X		
Warmth and friendliness	X	X		
Motivate and stimulate	X	X	X	
Concern for learning	X	X	X	
Increased interest	X			
Increased understanding	X		X	
Course organization	X	X	X	
R^2	0.8918	0.8659	0.8788	0.7877

* An X indicates that the item is included in the model.

for by this eleven-item instrument. The multiple correlation coefficient for the global score and the best linear combination of the 11 items is 0.9444 indicating that the global score and the eleven-item instrument score bear a very high relationship to each other. Considering the criteria used to select items for inclusion in the instrument, this is compelling validity evidence. The 11-item instrument does indeed capture the construct of overall teacher rating extremely well.

Additional validity evidence is provided by the Feldman, Young/Shaw, and McKay models presented in Table 2. Of the 11 items in the instrument, eight were among those that Feldman identified as being used prevalently in teacher evaluation instruments at many colleges and universities. To the extent that an item's prevalence of use in other scales serves as a validity criterion for its inclusion in this teacher effectiveness scale, the subset of eight commonly used items identified by Feldman alone accounts for almost 87% (R^2 = 0.8659) of the variance in the global ratings. This provides further substantial evidence of the eleven-item instrument's validity. In a like manner, the Young/Shaw and McKay models offer additional validity evidence. These authors have demonstrated that communication skills, instructor enthusiasm, subject matter knowledge, and ability to motivate and stimulate students are among the most important items to include in a teacher effectiveness instrument (Young & Shaw, 1999 and McKay, 1997). The 5-item subset of Young/Shaw and the 3-item subset of McKay account for 88% and 79% of the variance in global scores, respectively. Regarding the validity of the eleven-item instrument developed in this study, validity is evident in that the instrument contains subsets of items, known to have validity as measures of teacher effectiveness in their own right, that bear high relationships to the global score.

Factor analysis

Factor analysis was used to extract the first principal component from the data in an effort to ascertain the degree to which the eleven-item instrument captures a single, unitary construct. The results are presented in the first column in Table 3. With the single exception of subject matter knowledge that had a moderate loading, loadings are high to very high providing substantial evidence that the eleven-item instrument is indeed capturing a unitary construct of teacher effectiveness. The items were also factored to determine whether the unitary dimension would sub-divide into two or more factors. A five-factor solution, with well-identified factors that is easily interpreted, is presented in Table 3. The single dimension of teacher effectiveness in this study subdivides into 5 factors: instructor's subject knowledge; course organization; instructor communication skills, enthusiasm and ability to motivate; increased student interest and understanding; and instructor's general regard for, and treatment of, students. This sub-division of the overall dimension of teacher effectiveness into two or more (in this case, five) factors closely matches factor structures reported by other teacher effectiveness researchers (Marsh ,1991; Marsh and Hocevar, 1984 and 1991; Abrami, d'Apollonia, & Rosenfield,1997).

Table 3. Factor Analysis Results

Item	First Principal Component	Rotated Five Factor Orthogonal Solution				
		F1	F2	F3	F4	F5
Subject matter knowledge	.553					.937*
Course organization	.726				.847	
Communication skills	.861			.606		
Motivate and stimulate	.891			.617		
Enthusiasm	.835			.736		
Increased interest	.820		.831			
Increased understanding	.815		.819			
Comfortable atmosphere	.868	.724				
Respectful of students	.834	.866				
Warmth and friendliness	.796	.860				
Concern for learning	.870	.647				

* Loadings less than .500 are not reported.

Results and Discussion

Our findings consist of the following two statements: 1) the 11 items capture 89% of the variation in overall teacher ratings indicating that the instrument does indeed capture a very large portion of the variation in teacher ability, and 2) the 11 items have high loadings on a single factor indicating the extent to which the instrument is indeed unidimensional, however, the items do subdivide as expected into five, easily interpreted sub-factors, some of which deal more with the instructor and the others more with course-related matters. These findings provide substantial validity evidence for the eleven-item instrument. In general, the evidence is compelling. Our conclusion is that the instrument indeed appears to capture the construct of teacher effectiveness very well.

This work has resulted in the development of a teacher effectiveness instrument to which is attached a substantial body of validity evidence. This instrument may ultimately prove to be a viable teacher-rating instrument for use in a college or university, however, it is important to point out that its intent is to calibrate teacher effectiveness as a global construct. It may or may not be very useful as a device for providing teachers with itemized student feedback or students with information for their future scheduling. However, of possibly greater value than the creation of a single instrument, is the process by which the instrument was developed and validated. This process may be used again with different or modified item bases or underlying dimensions of teacher effectiveness.

References

Abrami, P., d'Apollonia, S., and Rosenfield, S. (1996). *The dimensionality of student ratings of instruction: What we know and what we do not*. In J. C. Smart (ed.) Higher Education: Handbook of Theory and Research (Vol. 11). New York: Agathon Press.

Erdle, S., Murray, H. & Rushton, J. (1985). Personality, classroom behavior, and college teaching effectiveness: A path analysis. *Journal of Educational Psychology*, *77*, 394-407.

Feldman, K. (1976). The superior college teacher from the student's view. *Research in Higher Education*, *5*, 243-288.

Feldman, K. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, *21*, 45-116.

Feldman, K. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education*, *24*, 139-213.

Marsh, H. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285-296.

Marsh, H. & Dunkin, M (1992). *Students' evaluations of university teaching: A multidimensional approach*. In J. Smart (ed.) Higher education: Handbook of Theory and Research (Vol. 8). New York: Agathon Press.

Marsh, H & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal*, *21*, 341-366.

Marsh, H. & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level and course level. *Teaching and Teacher Education*, *7*, 9-18.

McKay, J. (1997). Examining the elusive essence of the superlative teacher. *Education*, *177*, 3-9.

Murray, H. (1980). *Effective teaching behaviors in the college classroom*. In J. Smart (ed.) Higher education: Handbook of Theory and Research (Vol. 7). New York: Agathon Press.

Young, S. & Shaw, D. (1999). Profiles of effective teaching in higher education. *Journal of Higher Education*, *70*, 670-686.

Send correspondence to: Dale Shaw, Department of Applied Statistics and Reseach Methods
University of Northern Colorado, Greeley, Colorado 80639
Email: dale.shaw@unco.edu

Multiple Linear Regression Viewpoints
Information for Contributors

Multiple Linear Regression Viewpoints (*MLRV*) is a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model (MLR/GLM SIG). It is published twice a year to facilitate communication among professionals who focus their research on the theory, application, or teaching of multiple linear regression models and/or the general linear model. Manuscripts submitted to *MLRV* should conform to the language, style, and format of the Publication Manual of the American Psychological Association (5th ed., 2001). Three copies (two blind) of a doubled spaced manuscript (including equations, footnotes, quotes, and references) of approximately 25 pages in length, a 100 word abstract, and an IBM formatted diskette with the manuscript formatted in WordPerfect or Word should be submitted to one of the editors listed below.

Mathematical and Greek symbols should be clear and concise. All figures and diagrams must be photocopy-ready for publication. Manuscripts will be anonymously peer reviewed by two editorial board members. Author identifying information should appear on the title page of only one submitted manuscript. The review process will take approximately 3 months. A letter acknowledging receipt of the manuscript will be sent to the first author, and upon review, a letter indicating the peer review decision will be sent to the first author. Potential authors are encouraged to contact the editors to discuss ideas for contributions or determine if their manuscript is suitable for publication in *MLRV*.

EDITORS

T. Mark Beasley, Editor *MLRV*
Department of Biostatistics
School of Public Health
343C Ryals Public Health Bldg.
University of Alabama at Birmingham
Birmingham, AL 35294
(205) 975-4957 (voice)
(205) 975-2540 (fax)
mbeasley@uab.edu

Robin K. Henson, Associate Editor
Department of Technology and Cognition
College of Education,
P.O. Box 311337
University of North Texas
Denton, Texas 76203-1337
(940) 369-8385 (voice)
(940) 565-2185 (fax)
rhenson@tac.coe.unt.edu

ORDER INFORMATION
Jeffrey B. Hecht, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854
jbhecht@niu.edu

Check out our website at: http://www.coe.unt.edu/mlrv/

POSTMASTER: Send address changes to:
Jeffrey B. Hecht, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the
AERA Special Interest Group on Multiple Linear Regression: General Linear Model
through the **University of Alabama at Birmingham** and the **Dallas Independent School District**.