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 25 • Number 1 • Winter 1998

Special Issue on the Analysis of Missssing Values and Alternative Regression Procedures

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

Table of Contents

I Don’t Like My Data

Note from the Guest Editor about the Special Issue of *MLRV*

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

This Special Issue of *MLRV* was conceived during the 1998 AERA meeting in San Diego, CA. Isadore Newman and Keith McNeil approached me about guest editing this issue after I had served as the Discussant for a paper session sponsored by the MLR: GLM SIG. The original idea was to include the four papers from that session with my discussion notes as editorial commentary. After accepting this challenge I contacted each of those authors and invited two other papers that I felt would help complete two coherent themes: (1) Analysis of Missing Values and (2) Alternative Regression Procedures. I think you will find the articles enlightening at both the applied and theoretical levels. I can only hope that my comments are equally insightful.

My first general point is that it seems that all data sets have problems, hence the title. “What can be done when these problems arise?” is the central theme to all the articles in this special issue. One problem in particular is that researchers often encounter missing data. In my discussion with many data analysts, the norm seems to be discarding the missing cases. This of course is a loss of information which may bias the results. Another approach involves estimating what the missing value would have been if the subject had actually responded. Of course, the estimation of this replacement value can be biased by many factors. How much bias is created by these two general approaches (i.e., discarding data and imputing missing values) is the underlying theme for the first three articles (Orsak et al.; Mundfrom & Whitcomb; Brockmeier et al.).

The other general problem is that many data sets do not seem to conform to the assumptions of Ordinary Least Squares Regression. Alternative approaches include: (a) transforming the data in some manner or (b) computing parameter estimates in an entirely different manner (i.e., Long; Nevitt & Tam). Some robust methods such as “Trimming” suggest discarding (or downweighting) outliers that may result from a nonnormal error distribution (Nevitt & Tam). It is ironic that purposely deleting values is suggested when assumptions are not met while other researchers are trying to find a way to replace data that is missing.

In this vein of alternative analytic strategies, Kromrey and Hogarty investigate different statistical tests for analyzing the same data without transformation. Thus, even a simple research situation can be approached from several perspectives. The major issue is that different approaches tend to give different interpretations and possibly that is why they remain “alternatives.” This is not to say that alternative methods are somehow inferior, but as researchers we have a tendency to rely on more established methods with which we are familiar.

Concerning the reliance on familiar methods, I feel it necessary to comment on the “controversy” surrounding statistical significance testing. I agree with Joel Levin (e.g., 1993) in that until a better alternative to significance testing is developed researchers should continue its use. In the interpretation of results, however, researchers should also understand and state explicitly the precise meaning and limitations of significance testing. To revamp statistical significance testing, researchers and statisticians alike might benefit from using a confidence interval approach. Moreover, the social science research community should consider the perspective of Ron Serlin (1993) and employ a “range null hypothesis.” Subscribing to the confidence interval approach has particular implications for investigations that compare methodologies and simulation studies. One issue is that methodological researchers should consider is the accuracy of parameter estimates rather than simply investigate Type I error rate and subsequently power. This concern for accuracy is evident in several of the studies in this issue. Yet, one should not overlook Cliff’s (1993) perspective that social science data is typically ordinal in nature and that exact parameter estimates may not be extremely meaningful. One reason that statistical significance testing has been so prevalent in the social sciences has been the scales of measurement issue. To elaborate, when constructing a confidence interval for a parameter estimate for variable measured on an arbitrary scale sometimes the only meaningful value covered (or not covered) by the interval is the null value.

In terms of Monte Carlo studies, statistical hypothesis testing, and therefore investigating whether Type I error rates remain near an expected nominal alpha level, has been the bread-and-butter of simulation researchers. Furthermore, given that statistical hypothesis testing is not going away any time soon, coverage probabilities for confidence intervals should be reported. To elaborate, if a 95% confidence interval is constructed in multiple replications, the confidence interval should cover the population parameter 95% of the time regardless of its value (i.e., whether it is a null or non-null structure). By taking this approach, one can examine the potential biases in: (a) coverage probabilities (i.e., Does the confidence interval cover the population parameter at the specified level?); (b) power (i.e., How often does the confidence interval cover 0 with a non-null structure?); and (c) Type I error rate (i.e., How often does the confidence interval cover 0 with a null structure?).

In summary, as the popular adage goes, "Necessity is the mother of invention." Two notable trends have led to inventions that have increased statistical sophistication among social science researchers but have also resulted in more problematic data sets for most research projects. First of all, research problems, policy analyses, and educational evaluations have increasingly employed a quantitative perspective. This has resulted in more quantitative analyses of "real-life" data. Anyone who has collected their own data in an experiment, but especially those who have collected their own survey data, and those who have analyzed a national data base (e.g., NAEP, NELS) knows that real data have real problems. Secondly, technology has allowed researchers to handle these real data problems but also to view research issues in a more complex manner and subsequently to employ more complex and sophisticated methods.

From my experience in the graduate education of statistics and data analysis, the analogy, "You don't have to be a mechanic to drive a car," has been used to sickening extent. To expand this analogy, researchers are "driving" some very technologically sophisticated machines these days. What happens when there are problems? Today's automobiles are becoming so sophisticated that the "average driver" cannot work on them. (By the way that is not just a coincidence of technology, it is purposeful goal of car manufacturers). Likewise with sophisticated statistical software, the problem is that they will run any data you put in to them and do it very quickly. You will get results; they may just be meaningless. Furthermore, the speed of statistical software has perpetuated a certain level of sloppiness in dealing with quantitative analysis. So the purpose of this special issue is to "look under the hood" of these machines and see what happens if we throw a wrench into it. Sometimes we find that this new machine (i.e., methodology) is just a "souped-up" version of an older model and that it has the same basic problems. Occasionally, we will find that these new machines are true innovations and that either: (a) they have superior performance or (b) they operate in an entirely different manner. Most of all what should be taken from these articles is, "How do these approaches and techniques integrate with what is already known about statistics and data analysis?"

References

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114,* 494-509.

Levin, J. R. (1993). Statistical significance testing from three perspective. *Journal of Experimental Education, 61*, 378-382.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education, 61,* 350-360.

MLRV is the oldest AERA SIG sponsored journal. It has been both a challenge and a pleasure, but I could not have done this alone. First of all, I want to thank former *MLRV* editor Isadore Newman (University of Akron) and current *MLRV* co-editor Keith McNeil (New Mexico State University) for allowing me this opportunity. I also thank the authors for their quick response time in getting their manuscripts to me and their availability when last minute changes were necessary. I also sincerely appreciate the efforts of the most recent former editor (and the person who taught me more about regression than my students care to know), John Pohlmann (Southern Illinois University-Carbondale). His insight in constructing, formatting, and completing this issue was invaluable. I also want to thank the other current co-editor Randall Schumacker (University of North Texas) for his continued support in the area of research in statistics. Special thanks go to former Executive Secretary, Steven Spaner (University of Missouri-St. Louis), and current Executive Secretary, Susan Tracz (California State University-Fresno), for their assistance and guidance in the reproduction and distribution of this Special Issue.

I would like to remind you that I will be the MLR: GLM SIG Program Chair for the 2000 AERA meeting in New Orleans. **START GETTING YOUR PROPOSAL IDEAS TOGETHER NOW. RECRUIT YOUR COLLEAGUES AND STUDENTS**. There is an application form in the back of this issue. I want a BIG SIG in the BIG EASY. I look forward to seeing you in Montreal.

Happy Reading and Regressing,

T. Mark Beasley
St. John's University, New York

Calculating Missing Student Data in Hierarchical Linear Modeling: Uses and Their Effects on School Rankings

Timothy H. Orsak Robert L. Mendro Dash Weerasinghe
Dallas Public Schools

In the age of student accountability, public school systems must find procedures for identifying effective schools, classrooms and teachers that help students continue to excel academically. As a result, researchers have been modeling schools to calculate achievement indicators that will withstand not only statistical review but political criticism. One of the numerous issues encountered in statistical modeling is the management of missing student data. This paper addresses three techniques that elucidate the effects of absent data and highlight consequences on school achievement indicators. The outcomes of each technique are estimated data and School Effectiveness Indices (SEIs). A set of criteria is established from an original data set to determine a baseline to which the analyses will be compared in determining the most appropriate approach in estimating missing data.

Completeness of any data base should be considered a rarity when managing educational data. Numerous factors, not limited to lack of student attendance, data misinterpretation, and mistakes in data entry, all affect the accuracy of any educational database. While incorrect data scores are difficult, if not impossible, to detect, missing scores are readily identifiable. Effective schools within the Dallas Public Schools have been identified by statistical methodologies for several years. Many years of analyses have deduced the accuracy of statistical methods' rankings of schools within the district. Yet these analyses utilized only student data that was complete for both post-test and pre-test years. On average, between 8% and 12% of student data cannot be included in yearly calculations due to at least one year of missing test scores. However, attempts to use all available data while not introducing extraneous trends could more accurately help identify effective schools. In this paper, the question of best estimation of absent post-test data is addressed.

The current problem faced in the computation of school effectiveness rankings relates to missing student test data. How could we effectively rank the school of interest without complete data for its constituents? Several publications have addressed treatment of missing scores in data sets through the use of inference, replacement of missing values with probable values, etc. One example is Sanders and Horn (1993), which implemented a sparse matrix mixed modeling program to predict missing student values. Yet with the typical school district not having the resources to implement such a program, what would be the most effective and efficient method for school analysis? Dallas Public Schools has addressed the missing data issue by not including it in any analysis, thus eliminating possible influences.

The analysis comprised of 5,197 6th grade students who had complete raw data scores for the *Iowa Test of Basic Skills* mathematics and reading tests for years 1995 and 1996 and student characteristics of ethnicity, English proficiency status, census poverty data, census college data, and gender. To analyze the effects of missing data, specific percentages of the post-test scores from the original data set were randomly deleted which produced reduced data sets. The percentages of data deleted in this study were 1%, 2%, 5%, 10%, and 20%. The reduced data sets were then evaluated by *Scientific Software's* HLM2L hierarchical linear modeling software and by Microsoft Excel's Ordinary Least Squares (OLS) software program to produce regression coefficients for each school. The deleted post-test scores were then estimated by HLM (see Bryk & Raudenbush, 1993), by OLS, and by the average post-test score per school. The three new data sets composed of HLM estimates of missing data, OLS estimates of missing data, and average post-test data per school and the original data set (non-deleted scores), were then reprocessed by HLM and school effectiveness indices (SEIs) generated. The SEIs were calculated from HLM as the estimated Bayesian (EB) residuals for the school level intercept rescaled to a mean of 50 and standard deviation of 10. The EB residual reflects the overall achievement of the students within a school. The SEIs from the new data sets were compared to the original data set's SEI scores whereas the estimated post-test scores were compared to the actual scores that were deleted. This process was carried out for three models of varying complexity.

Table 1. Student Characteristic Correlations

	GEN	*LUN*	*BLK*	*HIS*	*LEP*	*INC*	*POV*	*COL*	*R-95*	*M-95*	*M-96*
GEN	1.000										
LUN	-.0122	1.000									
BLK	.0138	.1112	1.000								
HIS	-.0278	.0827	-.6043	1.000							
LEP	.0193	.1390	-.3049	-.1806	1.000						
INC	-.0090	.3407	.2046	.0418	.0215	1.000					
POV	-.0253	.2903	.1530	.0236	.0634	.5804	1.000				
COL	-.0172	.3461	-.0143	.2433	.1412	.6135	.3453	1.000			
R-95	.0951	.2282	.1992	-.0997	.1086	.1863	.1369	.2061	1.000		
M-95	.0169	.1747	.1451	-.0750	.0907	.1682	.1220	.1761	.6112	1.000	
M-96	.0354	.1763	.1303	-.0522	.0966	.1566	.1131	.1901	.5605	.7857	1.000

** GEN is Gender, LUN is Free Lunch Status, BLK represents Black, HIS represents Hispanic, LEP is Limited English Proficient, INC is average block income, POV is percent block poverty, COL is percent block college, R-95 is ITBS Reading for 1995, M-95 is ITBS Mathematics for 1995, M-96 is ITBS Mathematics for 1996.

Table 2. Student Characteristic Summary

	N	*MEAN*	*SD*	*MIN*	*MAX*
GEN	2610	1.54	.50	1	2
LUN	2610	1.28	.45	1	2
BLK	2610	1.50	.5	1	2
HIS	2610	1.74	.44	1	2
LEP	2610	1.92	.28	1	2
INC	2610	28139.44	14488.61	1290	185017.00
POV	2610	74.73	20.88	0	100
COL	2610	9.15	13.12	0	100
R-95	2610	11.91	4.42	1	22
M-95	2610	34.95	8.66	11	54
M-96	2610	37.83	9.23	9	59

** See Table 1 Legend

Investigation and Procedure

This study expands previous studies of HLM to investigate the effects of missing data through the use of HLM models in ranking 118 elementary schools from the Dallas Public Schools at the sixth grade (Webster et al., 1994, 1995; Mendro et al., 1994, 1995; Orsak et al., 1996). Ten school characteristics variables were available for each school. To eliminate undue influences from varying school sizes, the original 5,197 student data set was randomly reduced such that exactly 30 students were included per school. This created a new, reduced data file which contained 2,610 students within 87 schools. Initial analyses for this reduced data set explored OLS and HLM estimates from three models, each more complex than the previous. Then all 5,197 students were used in a fourth analysis. The initial exploratory analysis involved simple data analysis for the reduced data set.

The models used for the prediction of deleted post-test data are as follows. Analyses began with a basic model for prediction and increased in complexity.

The models with no student level variables and no school level variables:

Model 1A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$

Model 1B (OLS):

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{MATH95}_{ik} + r_{ik}$$

The models with two student level variables and no school level variables:

Model 2A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$
$$\beta_{2k} = \gamma_{20} + u_{2k}$$
$$\beta_{3k} = \gamma_{30} + u_{3k}$$

Model 2B (OLS):

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

The basic models with five student level variables and ten school level variables:

Model 3A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{HISPANIC}_{ik} + \beta_{4k}\,\text{BLACK}_{ik} + \beta_{5k}\,\text{GENDER}_{ik} + \beta_{6k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

$p = 0, 1, 2, ..., 6.$

where

W_{1k}	=	School Mobility
W_{2k}	=	School Overcrowdedness
W_{3k}	=	School Average Family Income
W_{4k}	=	School Average Family Education
W_{5k}	=	School Average Family Poverty Index
W_{6k}	=	School Percentage on Free or Reduced Lunch
W_{7k}	=	School Percentage Minority
W_{8k}	=	School Percentage Black
W_{9k}	=	School Percentage Hispanic
W_{10k}	=	School Percentage Limited English Proficient

$\gamma_{00}, \cdots, \gamma_{011}$ = level-2 intercept/slopes to model all β_{0k}s,
$\gamma_{10}, \cdots, \gamma_{111}$ = level-2 intercept/slopes to model all β_{1k}s,
$\gamma_{20}, \cdots, \gamma_{211}$ = level-2 intercept/slopes to model all β_{2k}s,
u_{0k}, u_{1k}, u_{2k} = level-2 random effects for school *k*.

Model 3B (OLS):

$$\text{MATH96}_{ik} = \beta_0 + \beta_1\,\text{CEN-POV}_{ik} + \beta_2\,\text{CEN-COL}_{ik} + \beta_3\,\text{HISPANIC}_{ik} + \beta_4\,\text{BLACK}_{ik} + \beta_5\,\text{GENDER}_{ik} + \beta_6\,\text{MATH95}_{ik} + r_{ik}$$

For this study, the SEIs were calculated only from HLM, two level models. The models used for the calculations were as follows:

Model 1 (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$

Model 2 (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$
$$\beta_{2k} = \gamma_{20} + u_{2k}$$
$$\beta_{3k} = \gamma_{30} + u_{3k}$$

Model 3 (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{HISPANIC}_{ik} + \beta_{4k}\,\text{BLACK}_{ik} + \beta_{5k}\,\text{GENDER}_{ik} + \beta_{6k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

$p = 0, 1, 2, ..., 6.$

The SEI is given by

$$\text{SEI*} = \gamma_{00}\,.$$

Results

The main objective of this study was to determine an acceptable methodology for estimating missing student post-test scores within a school effectiveness analysis. In pursuing the main objective, it was also possible to determine the variability of school ranking based on estimated data. Missing data were estimated by either using HLM estimated values for each school or by OLS estimation within each school for the first two models. Thus, predicted values were not across district but within school. OLS criteria forced district-wide calculations in Model 3B when schools were encountered that where composed of one ethnic group. Correlations were calculated among the actual scores, the two estimated scores, and the average post-test scores per school for each percentage of data estimated. Correlations were also computed among the SEIs for each percentage of data estimated.

Model 1A & 1B

The following tables display the correlations among the original data scores, HLM estimated scores, OLS estimated scores and the school average post-test score, each table reflecting a different percentage of the original data deleted. Also displayed are the correlations among the original SEIs and the SEIs calculated with each of the three estimated data.

Model 1A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$

Model 1B (OLS):

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{MATH95}_{ik} + r_{ik}$$

Model 1 (HLM): SEI CALCULATION

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$

Table 3. 1% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8132	1.0000	
OLS	0.8132	0.9949	1.0000
AVG	0.5224	0.6406	0.6463

Table 4. 1% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9994	1.0000	
OLS--SEI	0.9994	1.0000	1.0000
AVG-SEI	0.9986	0.9988	0.9986

Table 5. 2% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7844	1.0000	
OLS	0.7802	0.9955	1.0000
AVG	0.4673	0.5551	0.5518

Table 6. 2% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9981	1.0000	
OLS--SEI	0.9981	0.9999	1.0000
AVG-SEI	0.9962	0.9974	0.9986

Table 7. 5% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8158	1.0000	
OLS	0.8167	0.9941	1.0000
AVG	0.3710	0.4713	0.4623

Table 8. 5% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9952	1.0000	
OLS--SEI	0.9953	0.9997	1.0000
AVG-SEI	0.9862	0.9927	0.9908

Table 9. 10% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8343	1.0000	
OLS	0.8350	0.9917	1.0000
AVG	0.3893	0.5101	0.4855

Table 10. 10% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9911	1.0000	
OLS--SEI	0.9915	0.9987	1.0000
AVG-SEI	0.9730	0.9875	0.9808

Table 11. 20% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7934	1.0000	
OLS	0.7956	0.9842	1.0000
AVG	0.3452	0.5152	0.4241

Table 12. 20% SEI Correlations

	ACT--SEI	*HLM-SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9794	1.0000	
OLS--SEI	0.9812	0.9928	1.0000
AVG-SEI	0.9405	0.9755	0.9480

The first HLM model examined, Model 1A, used MATH95 to predict MATH96 at the first level with no school-level conditioning variables. Tables 3, 5, 7, 9, and 11 show the correlations among the actual, HLM estimated, OLS estimated and average post-test scores which were 1%, 2%, 5%, 10% and 20% deleted. Note that as the percentage of data deleted increased, the correlation between the actual scores and HLM estimated scores ranged from 0.7844 to 0.8343 whereas the correlation between the actual scores and OLS estimated scores ranged from 0.7802 to 0.8350. The weakest correlations existed between the actual scores and the average school post-test values with a range of 0.3452 to 0.5224. No noticeable pattern existed between the HLM and OLS estimated score correlations to the percentage of data estimated. It was obvious that the HLM and OLS models produced nearly identical results as their estimated values were correlated at a minimal value of 0.9917. Also note that as the percentage of data estimated increased, HLM estimated values were more highly correlated to the average post-test score than the OLS estimated scores, an indication of HLMs shrinkage to the overall mean.

Tables 4, 6, 8, 10, and 12 indicate correlations of SEIs using data from the three estimation sources. As the percentage of estimated data increased, all correlations decreased. In this basic model, it was interesting to note OLS estimated data results had slightly higher correlations with the original SEIs in comparison to HLM estimated data, with the greatest difference at the 20% level (0.9812 versus 0.9794). Note that even the average school value produced correlations within the range of 0.9405 to 0.9986 depending on percentage of missing data.

Now the question of “which is best” in terms of prediction must be decided. Clearly, HLM produced estimates more closely related to the original data than OLS, but not so clear was why the SEIs of OLS were more closely related to the original data than HLM. Light will hopefully be shed on this situation as models become more complex.

Model 2

This next analysis introduced CEN-COL and CEN-POV into the previous model for the prediction of MATH96. CEN-COL represents the percentage of households within the student’s block who attended college. CEN-POV represents the percentage of households who fall below the poverty level.

Model 2A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$
$$\beta_{2k} = \gamma_{20} + u_{2k}$$
$$\beta_{3k} = \gamma_{30} + u_{3k}$$

Model 2B (OLS):

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Model 2 (HLM): SEI CALCULATIONS

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{CEN-POV}_{ik} + \beta_{2k}\,\text{CEN-COL}_{ik} + \beta_{3k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{0k} = \gamma_{00} + u_{0k}$$
$$\beta_{1k} = \gamma_{10} + u_{1k}$$
$$\beta_{2k} = \gamma_{20} + u_{2k}$$
$$\beta_{3k} = \gamma_{30} + u_{3k}$$

Table 13. 1% Predicted Data Correlations.

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8195	1.0000	
OLS	0.8079	0.9653	1.0000
AVG	0.5224	0.6392	0.5804

Table 14. 1% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9985	1.0000	
OLS--SEI	0.9992	0.9977	1.0000
AVG-SEI	0.9984	1.0000	0.9976

Table 15. 2% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7845	1.0000	
OLS	0.7771	0.9702	1.0000
AVG	0.4673	0.5536	0.5259

Table 16. 2% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9980	1.0000	
OLS--SEI	0.9977	0.9995	1.0000
AVG-SEI	0.9958	0.9967	0.9957

Table 17. 5% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8076	1.0000	
OLS	0.8058	0.9669	1.0000
AVG	0.3710	0.4763	0.4537

Table 18. 5% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9948	1.0000	
OLS--SEI	0.9946	0.9988	1.0000
AVG-SEI	0.9843	0.9915	0.9887

Table 19. 10% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8267	1.0000	
OLS	0.8140	0.9274	1.0000
AVG	0.3893	0.5186	0.4474

Table 20. 10% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9894	1.0000	
OLS--SEI	0.9854	0.9902	1.0000
AVG-SEI	0.9708	0.9866	0.9692

Table 21. 20% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7872	1.0000	
OLS	0.7540	0.9172	1.0000
AVG	0.3452	0.5169	0.3969

Table 22. 20% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9748	1.0000	
OLS--SEI	0.9185	0.9447	1.0000
AVG-SEI	0.9343	0.9703	0.8906

Tables 13, 15, 17, 19, and 21 include correlations between the actual, HLM estimated, OLS estimated and average post-test scores for the indicated percentage of data estimated. As the percentage of estimated data increased, the correlations range from 0.7845 to 0.8267 for HLM estimates and 0.7540 to 0.8140 for OLS estimates. In all percentages, HLM estimates were more correlated with the actual data than the OLS estimates, although the differences were extremely slight in one case (0.0018 difference). Again the weakest correlations were between the actual score and the average school post-test value with a range of 0.3452 to 0.5224. It can be noted that as the percentage of estimated data increases, the difference in correlations between HLM and OLS also increased.

Tables 14, 16, 18, 20, and 22 reflect the correlations of SEIs. Once more, as the percentage of estimated data increased, the correlations of SEIs decreased. HLM generated SEIs more correlated with the original SEIs than did OLS, which is in contrast to the first model. The greatest divergence occurred at the 20% level with a difference of 0.0563 while all others were of smaller deviations. Over more, the SEIs from average post-test scores correlated much lower than the estimates within a range of 0.9984 to 0.9343.

The "which is best" decision leans more clearly toward HLM in this particular model.

The third model analyzed included MATH95, CEN-COL, CEN-POV with the new variables of GEN, HIS, BLK, (where GEN represents student gender, HIS represents a Hispanic student and BLK represents a black student) to model MATH96. Ten school conditioning variables were also included in the HLM analysis at the school level. At this point difficulties were encountered in the OLS program in that numerous schools had populations of strictly one ethnic composition; thus it failed to generate estimates. HLM circumvented this predicament by generating estimates for all schools. OLS estimates were now generated across all schools, thus eliminating the problems encountered within schools.

Model 3

Model 3A denotes a true, two-level, hierarchical model with conditioning variables at the second level. This model was compared to the OLS Model 3B where OLS did not adjust for conditioning variables.

Model 3A (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\ \text{CEN-POV}_{ik} + \beta_{2k}\ \text{CEN-COL}_{ik} + \beta_{3k}\ \text{HISPANIC}_{ik} + \beta_{4k}\ \text{BLACK}_{ik} + \beta_{5k}\ \text{GENDER}_{ik} + \beta_{6k}\ \text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

p = 0, 1, 2, . . . , 6.

Model 3B (OLS):

$$\text{MATH96}_{ik} = \beta_0 + \beta_1\ \text{CEN-POV}_{ik} + \beta_2\ \text{CEN-COL}_{ik} + \beta_3\ \text{HISPANIC}_{ik} + \beta_4\ \text{BLACK}_{ik} + \beta_5\ \text{GENDER}_{ik} + \beta_6\ \text{MATH95}_{ik} + r_{ik}$$

Model 3 (HLM): SEI CALCULATION

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\ \text{CEN-POV}_{ik} + \beta_{2k}\ \text{CEN-COL}_{ik} + \beta_{3k}\ \text{HISPANIC}_{ik} + \beta_{4k}\ \text{BLACK}_{ik} + \beta_{5k}\ \text{GENDER}_{ik} + \beta_{6k}\ \text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

p = 0, 1, 2, . . . , 6.

Table 23. 1% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7731	1.0000	
OLS	0.7573	0.9683	1.0000
AVG	0.5224	0.5959	0.4746

Table 24. 1% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9903	1.0000	
OLS--SEI	0.9915	0.9470	1.0000
AVG-SEI	0.9842	0.9873	0.9883

Table 25. 2% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7466	1.0000	
OLS	0.7108	0.9613	1.0000
AVG	0.4673	0.5352	0.3865

Table 26. 2% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9832	1.0000	
OLS--SEI	0.9818	0.9802	1.0000
AVG-SEI	0.9720	0.9709	0.9701

Table 27. 5% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7811	1.0000	
OLS	0.7619	0.9548	1.0000
AVG	0.3710	0.4595	0.3068

Table 28. 5% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9818	1.0000	
OLS--SEI	0.9767	0.9812	1.0000
AVG-SEI	0.9731	0.9915	0.9887

Table 29. 10% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.8182	1.0000	
OLS	0.8075	0.9455	1.0000
AVG	0.3893	0.5064	0.3818

Table 30. 10% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9776	1.0000	
OLS--SEI	0.9710	0.9718	1.0000
AVG-SEI	0.9620	0.9648	0.9592

Table 31. 20% Predicted Data Correlations

	ACTUAL	*HLM*	*OLS*
ACTUAL	1.0000		
HLM	0.7779	1.0000	
OLS	0.7684	0.9316	1.0000
AVG	0.3452	0.5018	0.3190

Table 32. 20% SEI Correlations

	ACT--SEI	*HLM--SEI*	*OLS--SEI*
ACT--SEI	1.0000		
HLM--SEI	0.9503	1.0000	
OLS--SEI	0.9114	0.9447	1.0000
AVG-SEI	0.9175	0.9532	0.9154

Tables 23, 25, 27, 29, and 31 show correlations between the actual, HLM estimated, OLS estimated and average post-test scores for the indicated percentage of data estimated. As the percentage of estimated data increased, the correlations range from 0.7466 to 0.8182 for HLM estimates and 0.7108 to 0.8075 for OLS estimates. In all percentages, HLM estimates were more correlated with the actual data than the OLS estimates. Note again that as the model increased in complexity with the inclusion of more student variables and the addition of school level variables, the correlations decreased in comparison to previous models for the identical level of data estimated.

Tables 24, 26, 28, 30, and 32 reflect the correlations of SEIs. For the most part, as the percentage of estimated data increased, the correlations of SEIs decreased. HLM generated SEIs more correlated with the original SEIs than did OLS. The greatest divergence occurred at the 20% level with a difference of 0.0389 while all others were of smaller deviations. Moreover, the SEIs from average post-test scores correlated much lower than the estimates within a range of 0.9842 to 0.9175. Tests of correlations indicate all were significant.

These three models indicate that HLM is more suitable for estimating missing data than OLS or the average school score. This advantage must be gained by HLM's adjustments for school trends in comparison to overall trends for student scores. Investigations into HLM's ability to predict continued with repeated deletion estimations on the original data set.

The next phase of the investigation focused on twenty-five repeated deletion trials for each percentage of estimated data. The original 5,197 students were used in the computation of SEIs using only HLM estimates of the missing data. The SEIs generated by the twenty-five trials were compared individually to the original SEI and then the average of the twenty-five trials was compared to the original SEI for the complete data set. The model for this comparison was:

Model 4 (HLM):

Level 1:

$$\text{MATH96}_{ik} = \beta_{0k} + \beta_{1k}\,\text{LEP}_{ik} + \beta_{2k}\,\text{HISPANIC}_{ik} + \beta_{3k}\,\text{BLACK}_{ik} + \beta_{4k}\,\text{GENDER}_{ik} + \beta_{5k}\,\text{MATH95}_{ik} + r_{ik}$$

Level 2:

$$\beta_{pk} = \gamma_{p0} + \sum_{k=1}^{10} \gamma_{pk} W_{kj} + u_{pk}$$

$$p = 0, 1, 2, ..., 5$$

Table 33. SEI Correlations with Actual SEI

	AVG(25) vs. ACTUAL	*MAX Corr.*	*MIN Corr.*
1 %	0.9998	0.9989	0.9978
2 %	0.9998	0.9985	0.9977
5 %	0.9996	0.9966	0.9936
10 %	0.9994	0.9937	0.9867
20 %	0.9983	0.9837	0.9735

Table 33 denotes the correlations between the original SEI for the complete data set and the average of the SEIs for twenty-five trials, the maximum correlation between the original SEI and the individual trials as well as the minimum correlation between the individual trials and the original SEIs. The obvious main observation was as the percentage of data increases, the correlation between the actual SEI and estimated data SEI also decreased. Although the correlations remain quite high, an analysis of the ranks of the SEIs revealed changes of up to ten places in rank.

Conclusions

Several observations appear relevant based on this study. First, and perhaps most important, HLM estimates and OLS estimates are both similar to the original data up to approximately the 10% level whereas HLM estimates are more accurate to the original for greater percentages. This highlights the advantage of implementing HLM in educational data analysis when a greater percentage of data is missing. Second, SEIs with HLM estimates of missing data and OLS estimates of missing data are highly correlated when up to 10% of data is estimated for a relatively simple model without school level conditioning variables. This allows a choice of which method to choose for estimating missing data. Differences emerge as estimation models became more complex. The contradicting observation to the previous point is that HLM was able to generate estimates when full rank was not achieved within schools. For example, when students were all of one ethnicity within a school, OLS estimations failed for within school estimation. The alternative was to carry out OLS estimations across schools but it sacrifices potentially useful within-school information.

Future analyses are planned to formulate a test statistic that determines when the deviations of estimated scores from the actual scores are significant, and the deviations of school ranks from actual ranks are significant, along with investigations into the rank changes about their respective quartiles.

References

Bryk, A. S., & Raudenbush, S. W. (1993). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

Mendro, R. L., Webster, W. J., Bembry, K. L., & Orsak, T. H. (1994, October). *Applications of Hierarchical Linear Models in Identifying Effective Schools.* Paper presented at the Annual Meeting of the Rocky Mountain Educational Research Association, Tempe, AZ.

Mendro, R.L., Webster, W. J., Bembry, K. L., & Orsak, T. H., (1995, April). *An Application of Hierarchical Linear Modeling in Determining School Effectiveness*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Orsak, T. H., Mendro, R.L., Webster, W. J., & Weerasinghe, Dash, (1996, April). *Empirical Difficulties in Using Hierarchical Linear Models for School*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Sanders, W. L., & Horn, S. P. (1993). *The Tennessee Value-Added Assessment System (TVAAS): Mixed Model Methodology in Educational Assessment*. Knoxville, TN: University of Tennessee.

Webster, W. J., Mendro, R. L., Bembry, K. L., & Orsak, T. H. (1994, October). *Alternative Methodologies for Identifying Effective Schools*. Paper presented at the Annual Meeting of the Rocky Mountain Educational Research Association, Tempe, AZ.

Webster, W. J., Mendro, R. L., Bembry, K. L., & Orsak, T. H. (1995, April). *Alternative Methodologies for Identifying Effective Schools*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Imputing Missing Values: The Effect on the Accuracy of Classification

Daniel J. Mundfrom, University of Northern Colorado
Alan Whitcomb, Inver Hills Community College

Data from patient records were used to classify cardiac patients as to whether they are likely or unlikely to experience a subsequent morbid event after admission to a hospital. Both a linear discriminant function and a logistic regression equation were developed using a set of nine predictor variables which were chosen on the basis of their correlations with the likelihood of a subsequent morbid event. Once the models were obtained, artificially-generated missing values were replaced with imputed values using mean substitution, regression imputation and hot-deck imputation techniques. The effect on the accuracy of the predictions using models with imputed values was determined by comparing the re-classifications using imputed data with the actual occurrence or non-occurrence of a subsequent morbid event. Mean substitution and hot-deck imputation performed slightly better than regression imputation in this application regardless of whether or not the predictor variable whose values were being imputed was categorical or numerical.

Statistical modeling techniques have been widely used for many years to predict a particular outcome using information from a group of variables which are related to the outcome of interest. That outcome could be a continuous variable such as an achievement test score or a categorical variable such as whether or not an individual graduated from a particular graduate program. When the outcome of interest is continuous, the appropriate statistical procedures would generally be a multiple regression analysis or an analysis of variance. When the outcome of interest is dichotomous, the analysis reduces to classifying an individual into one of two or more groups depending on the observed values of a set of predictor variables, and the appropriate procedure to use is either a discriminant analysis or logistic regression.

One situation in which statistical modeling, with a dichotomous outcome variable, could be used for classification involves the decision that rural hospitals must make when a cardiac patient arrives at the hospital. Rural hospitals frequently cannot afford all the latest technological equipment that their larger urban counterparts have available. One possible decision would be to automatically send all cardiac patients on to the urban hospital. This decision has obvious benefits, but also has at least two drawbacks. One drawback is that some patients will be sent who could have been cared for sufficiently in the local hospital. This decision requires needless expense for the patient that could have been avoided. A second drawback is that every patient transported away from the local hospital takes with him/her revenue that could have been spent locally that would help the local hospital maintain economic viability.

Another decision that could be made is to keep all patients and care for them locally. While this decision keeps the revenue “at home,” it may not be in the best interest of every patient in terms of providing them with the necessary care. The desire to balance the patients’ needs for having the best possible care and the hospitals’ needs to maintain their economic vitality forms the framework for this research.

One way to try and balance these needs is to reduce the number of unnecessary patient transportations from rural hospitals to tertiary care facilities. Technology to assist the rural physician in more accurately predicting which cardiac patients are likely to experience a morbid event and which are not can reduce the number of patient transportations. Since cardiovascular disease is the leading cause of death in the United States, and its prevalence is highest in rural areas where the latest advancements in providing necessary care may not be available, predicting likely candidates for a subsequent morbid event would be a valuable asset for the rural physician. Coronary Care Units (CCUs) have proven to be extremely effective in preventing death from certain cardiac events, but the cost of these units normally limits their presence to tertiary care facilities. Moreover, predicting which cardiac patients are likely to experience a serious morbid event has proven difficult, with only about 25% of patients in a CCU suffering life-threatening events during their stay. The ability to make an accurate prediction would increase the economic viability of the rural hospital and also reduce the financial burden for the patient, without having a negative impact on the adequacy of the care the patient received.

Statistical models, based on patient data collected at the time of the initial hospital visit, could be useful for making cardiac morbidity predictions. However, it is not always possible to obtain a measurement on every variable of interest in real data situations. Missing data values have plagued statisticians for years in their attempts to obtain useful, accurate summaries and predictions. Missing data is an even greater concern when decisions must be made about the appropriateness of the care a patient should receive. From a methodological perspective, missing values either reduce the number of available cases for analysis or introduce bias into the estimation and/or prediction process. Neither scenario is desirable.

The purpose of this research was twofold. The first objective was to develop a statistical model that could be used to predict which cardiac patients are likely to experience a subsequent morbid event. The model that was developed was based upon the complete-case data of actual cardiac patients.

The second objective was to examine the effects on the accuracy of the model's predictions when imputed values from three different imputation techniques were substituted for artificially-generated missing data. Of particular interest was determining which of the techniques would have the smallest detrimental effect on the accuracy of predictions when using imputed values.

Background

The initial phase of this research was to select a suitable model for predicting a morbid event in cardiac patients. Five morbid events were identified and defined as follows: development of sustained ventricular tachycardia (a very rapid heartbeat), ventricular fibrillation (a rapid, quiver-like heartbeat that is incapable of producing a pulse), cardiogenic shock (essentially pump failure--inability of the heart to move blood), development of myocardial infarction or extension of infarction (commonly referred to as a heart attack), and bradycardia of less than 45 beats per minute (a very slow heartbeat). Identifying potential predictor variables that may be indicators of one or more of these events was the next step.

Although prevention of fibrillation has long been recognized as desirable (Lown, Fakhro, Hood, & Thorn, 1967), defining specific electrical parameters heralding fibrillation has not been easy (Campbell, Murray, & Julian, 1981). Like ventricular fibrillation, predicting the development or worsening of pump failure has also been difficult. Nonetheless, numerous studies now exist that have attempted to accurately define the clinical predictors associated with a poor prognosis in CCU patients. Parameters as varied as age, hypertension, diabetes, length of stay in the CCU (Gheorghiade, et al., 1987), ST and T wave changes (Severi, et al., 1988; Bell, Montarello, & Steele, 1990), sex, anterior infarction, hypotension at admission, ventricular tachyarrhythmias, diabetes, Killip class III and IV (De Martini, et al., 1990), previous myocardial infarction (Nishi, et al., 1992), and serum urea (Marik, Lipman, Eidelman, & Erskine, 1990) have all been shown to have short-term prognostic significance.

Assuming that a set of suitable variables for predicting a morbid event can be identified, the problem of missing data must still be addressed. In many real-life situations, one or more of the individual cases will have incomplete data. In this application, one or more of the signs necessary for optimally predicting a morbid event may be unavailable. Perhaps a measurement goes unrecorded, a test is not available to be run, or the results of a test are inadvertently lost. Most standard statistical techniques build their models using only those cases which have a complete set of data values. If the value for even one variable is missing, the entire set of measurements for that individual is excluded from the model-building process. Complete-case analyses are often used because of simplicity of analysis and for comparability resulting from using a common sample for all calculations (Little & Rubin, 1987). However, the loss of potentially useful information in the data which is discarded is undesirable.

Another problem occurs if, after the model has been obtained, one or more of the values required to use the model are unavailable. The optimal model is constructed based on the assumption that data will be available for each variable included in the model. A regression coefficient is calculated for each variable, so its contribution to the prediction of the outcome variable is appropriately weighted. If even one value is missing for an individual, the optimal model cannot be used appropriately and if it is used anyway, the resulting prediction may be suspect. The problem of missing data can be overcome by deleting cases with missing values or by replacing missing values with an imputed value. Imputed values are generally obtained from the existing data and there are a variety of techniques available for imputation, each having different properties that make them more or less useful in any particular situation (see Buck, 1960; Affifi & Elasahoff, 1966, Haitovsky, 1968; Hartley & Hocking, 1971; Chan & Dunn, 1972; Rubin, 1976; Little & Rubin, 1987; Rubin, 1991; Rindskopf, 1992; van Buuren & van Ruckevorsel, 1992; Kromrey & Hines, 1994; Roth & Switzer, 1995).

Among the most commonly recommended missing data treatments are listwise and pairwise deletion, mean substitution, regression imputation, hot-deck imputation, and the EM algorithm (Little & Rubin, 1987). The selection of which of these procedures to examine in this research involves several considerations. First of all, the deletion techniques were deemed inappropriate, since in this phase of this research, a model is not being constructed, but rather a previously built model will be used to classify an individual. Deleting the case would result in the lack of such classification for that individual, an outcome that is unacceptable in this situation. Hence, only the imputation techniques received further consideration. Accuracy of classification is the primary issue, but the principal purpose of this research is to compare accuracy rates, so this characteristic was not used to select techniques for consideration. Ease of use is also a primary factor. In the context of predicting a morbid event, the desire was to use an imputation technique that would not require the physician to perform a complicated or time-consuming task in order to make a decision regarding a particular patient. The need to keep this part of the process simple therefore became the primary criteria for selecting an imputation technique.

Another consideration was the task at hand. The value being predicted was the likelihood of a morbid event. Because this is basically a classification problem (classifying an individual into one of two groups; likely to have a subsequent morbid event or unlikely to have one), an imputation technique with optimal properties in discrimination was desired. Chan and Dunn (1972) reported that mean substitution and the principal components method outperformed other techniques for classification. Kim and Curry (1977) and Raymond and Roberts (1987) report that regression imputation has the desirable property of minimizing the variability in the imputed values. Hot-deck imputation is frequently used in practice because of its intuitive appeal (Roth & Switzer, 1995), but little research regarding its accuracy has been done. Rubin (1991) lists several desirable properties of the EM algorithm that seem to indicate it as the procedure of choice in many situations, especially with large samples. All four of these imputation techniques have desirable characteristics.

From this list of four techniques, the EM algorithm, although highly regarded for many reasons, was deemed to be too complex to have a reasonable expectation of use by a physician in practice. Consequently, the techniques chosen to be examined in this research were mean substitution, regression imputation, and hot-deck imputation. Recognizing that this decision is subjective and may not necessarily be optimal, it still seemed reasonable that due to the relative simplicity of using these procedures, that if any were found to be sufficiently accurate, it would have a high expectation for use in practice.

Method

The archival data used in this research were obtained from patient records for a sample of 99 cardiac patients who had been admitted over a three-year period to a Cardiac Care Unit or a Cardiac Monitored Care Unit (MCU) in an urban University-affiliated hospital after suffering a morbid event for which data existed on a list of 29 variables which had been identified as potential predictors of a subsequent morbid event after suffering an initial such event. Patients who had undergone surgery in the six month period prior to admittance to the CCU/MCU or who were on mechanical breathing support were excluded from the sample. In this sample, 38 individuals experienced at least one subsequent morbid event in the hospital after being admitted.

This list of variables included the continuous variables: height, weight, age, systolic blood pressure, diastolic blood pressure, hematocrit, serum potassium level, serum creatine level, white blood cell count, respiration rate, and heart rate, and the categorical variables: sex, current myocardial infarction, evidence of anterior infarction, atrial arrhythmia, ventricular arrhythmia, S-T depression, diabetes, previous infarction, smoking, rales greater than 1/3 up, presence of heart sound S3, syncope, ventricular ectopics, use of aspirin in treatment, and use of Class I, II, III, or IV drugs. This initial list of potential predictors was reduced from 29 to 9 based upon their correlations with the occurrence or non-occurrence of one or more of the five morbid events ($|r| > .1643$). The final group of nine predictors included sex, age, weight, systolic blood pressure, white blood cell count, ventricular arrhythmia (an indicator of abnormal heart rhythm; measured as present or absent), syncope (an indicator of poorly oxygenated blood; measured as poor or not poor), heart sound S3 (an indicator of heart valve insufficiency; measured as sufficient or not), and use of aspirin (measured as used in treatment or not).

Once these predictors were identified, a linear discriminant function, based on those nine predictors, was created for classifying patients as likely to experience a subsequent morbid event or not likely to experience such an event. Similarly, a logistic regression equation using the same set of nine predictors was generated for the same classification purpose. The coefficients for both the linear discriminant functions and the logistic regression analysis are presented in Table 1. The number of correct classifications for each model was determined by comparing classifications resulting from use of the statistical model with the actual occurrence or non-occurrence of subsequent morbid events.

Table 1. Coefficients of Predictor Variables in Linear Discriminant Functions and Logistic Regression Analysis

Coefficients	LDF for Group 1	LDF for Group 2	Logistic Regression Analysis
Constant	-31.4430	-32.2958	0.6513
Sex	2.1466	0.8996	1.4111
Age	0.2862	0.3154	-0.0288
Weight	0.2306	0.2119	0.0204
SBP	0.1435	0.1387	0.0043
WBCC	0.4197	0.5052	-0.0877
VA	1.6556	2.3389	-0.7457
Syncope	0.9860	1.2777	-0.1887
S3	3.8438	4.7070	-0.7198
Aspirin	2.6690	1.7695	1.1287

Note: LDF = linear discriminant function; Group 1 contains individuals who did not have a subsequent morbid event; Group 2 contains individuals who did have a subsequent morbid event; SBP = systolic blood pressure; WBCC = white blood cell count; VA = ventricular arrhythmia.

To investigate the effect of different techniques for imputing missing values, values for one predictor at a time were deleted for each of the 99 patients and replaced with an imputed value. After replacing the original value with an imputed value, the number of correct re-classifications using the original discriminant function and the logistic regression model were calculated. In turn, this process was repeated for each of the three imputation techniques and for eight of the nine predictor variables. (It was decided that the variable sex is unlikely to ever be unknown in this context, so replacing the actual value of the sex variable with an imputed value seemed unnecessary.) The number of correct re-classifications, using imputed values in both the discriminant analysis and the logistic regression analysis, were then compared to the number of correct classifications using the original data.

For the mean imputation technique, imputed values for a particular variable were obtained by calculating the mean value for that predictor using all 99 patients' records. Using a single variable at a time for imputation, the original values of that variable were replaced with the mean value of that variable in each of the individuals' records. The other eight predictors were left unchanged and the individual was re-classified into one of the two groups. The value for each of the other predictors, excluding sex, was replaced with its mean value in the same way, each time using the original data values for the other predictors, and each individual was re-classified.

Using the regression imputation technique, imputed values for each predictor were calculated for the patients by building a regression equation involving the other eight predictors. Imputed values for the variables which were measured on a continuous scale (e.g., age) were determined using multiple linear regression analysis. For the dichotomous predictors (e.g., ventricular arrhythmia), a logistic regression analysis was used to build the model for prediction. The coefficients used to generate the imputed values for each of the eight predictors (again, excluding sex) are presented in Table 2.

Using the hot-deck imputation technique, an imputed value for one predictor was obtained for each patient by randomly selecting (with replacement) a value from that variable's original set of 99 values. However, since the randomly selected values would vary from one selection to another, so would the number of correct re-classifications. Consequently, the estimate of the accuracy of prediction would be too reliant on the particular value selected. To ensure that this estimate was less dependent upon the particular value that was randomly selected to be used as the imputed value, 1000 repetitions were run for each variable to obtain an average number of correct re-classifications for each of the eight predictors in both the discriminant analysis and the logistic regression analysis.

Results

Using the linear discriminant function with nine predictor variables, 78 of the 99 individuals in the sample were correctly classified into the two groups: likely to experience a subsequent morbid event and unlikely to experience such an event. With the logistic regression analysis, 80 of the 99 individuals were correctly classified.

Overall, the results obtained by using the imputation techniques and comparing the re-classifications, as determined by the discriminant function and the logistic regression equation, with the actual group membership was encouraging. In general some, but not much, accuracy is lost when an original data value is replaced by an imputed value. The re-classification results are presented in Table 3.

Table 2. Coefficients of Predictor Variables Used in Regression Imputation

	Response Variable							
Coefficients	Age (MLR)	Weight (MLR)	SBP (MLR)	WBCC (MLR)	VA (LR)	Syncope (LR)	S3 (LR)	Aspirin (LR)
Constant	56.005	82.124	105.324	8.302	0.334	2.463	-1.523	0.062
Sex	-0.481	-7.363	9.060	0.975	-0.247	0.820	1.330	0.407
Age		-0.404	0.417	0.030	-0.016	-0.037	-0.008	-0.020
Weight	-0.209		0.262	0.038	-0.001	-0.001	0.019	-0.011
SBP	0.106	0.129		-0.029	0.013	0.008	0.014	0.012
WBCC	0.311	0.759	-1.165		0.018	-0.004	-0.074	0.127
VA	2.853	0.164	-8.705	-0.289		-0.473	-1.291	-0.454
Syncope	6.973	-0.100	-7.218	-0.061	-0.445		2.298	0.585
S3	1.057	-5.948	-9.849	1.732	-1.036	2.151		1.869
Aspirin	3.589	3.131	-9.036	-1.181	-0.322	0.528	2.260	

Note: SBP = systolic blood pressure; WBCC = white blood cell count; VA = ventricular arrhythmia; MLR = multiple linear regression; LR = logistic regression

For the discriminant analysis, the variable syncope was least affected by imputation, with 77, 78, and 76.7 individuals being correctly re-classified using mean, regression, and hot-deck techniques, respectively. (Recall, that the number of correct re-classifications using the hot-deck technique are averages of 1000 replications.) Mean substitution appeared to do slightly better than the other two techniques on most variables, particularly ventricular arrhythmia and heart sound S3. Overall, the average number of correctly re-classified individuals, averaging over all eight variables, was very similar for the three imputation techniques with mean substitution having an average number of correct re-classifications of 74.4, only slightly better than hot-deck imputation (73.1) and the regression method (72.0).

For the logistic regression analysis, the variable syncope was again the least affected by the imputation with numbers of correctly re-classified individuals of 78, 77, and 77.7, and systolic blood pressure, which correctly classified 77, 77, and 75.4 individuals also relatively unaffected by imputation. Overall, for the logistic regression analysis, mean substitution was again fairly consistent from variable to variable, although the hot-deck technique, with an average number of correct re-classifications of 73.9, was slightly better than mean substitution (73.5), and regression imputation (72.3).

Discussion

The first phase of this research produced a linear discriminant function and a logistic regression model for classifying individuals as either likely or unlikely to have a subsequent morbid cardiac event after having first experienced an initial such event. Using a set of nine predictor variables, the discriminant function correctly classified 78 of the 99 individuals in the sample, while the logistic regression model classified 80 of the 99 individuals correctly. These numbers are not as high as we would have liked. However, given the relatively small sample size and the large number of variables that needed to be reduced to a manageable size, these results were the best that could be achieved. Given the results of previous research that identified potential predictors and the large number and variety of variables identified in that literature, it should not be surprising, perhaps, that any particular group of variables does not perform exceptionally well in predicting the outcome of interest. Using either the discriminant function or the logistic regression model described above, the three imputation techniques, mean substitution, regression imputation, and hot-deck imputation, were compared to determine the extent to which replacing original data values with imputed values affected the number of correctly classified individuals. Overall, using an imputation technique to replace missing values in this application appeared to produce results which are comparable to those obtained using the actual data. Mean substitution was comparable to the hot-deck technique in the logistic regression analysis and slightly better than the other two techniques in the discriminant analysis. This result was somewhat surprising because of the general lack of trust that researchers appear to have in mean substitution for imputation. It was also somewhat satisfying, since mean substitution is a relatively easy technique to use and does not require sophisticated calculations, thus increasing the probability that it might actually be used in practice.

Table 3. Numbers Of Correct Re-Classifications For Each Predictor Variable And Each Imputation Technique For Discriminant Analysis and Logistic Regression (n=99)

	Discriminant Analysis Imputation Technique			Logistic Regression Imputation Technique		
Variable Imputed	Mean	Regression	Hot-Deck*	Mean	Regression	Hot-Deck*
Age	74	76	72.4	75	77	72.4
Weight	74	72	72.4	73	73	71.4
Systolic Blood Pressure	74	74	73.8	77	77	75.4
White Blood Cell Count	71	71	70.2	72	70	70.1
Ventricular Arrhythmia	74	65	69.9	68	63	70.6
Syncope	77	78	76.7	78	77	77.7
Heart Sound S3	78	67	74.1	73	69	76.6
Aspirin	73	73	75.5	72	72	76.8
Mean of all predictors	74.4	72.0	73.1	73.5	72.3	73.9

* Values in this column represent the average number of correctly re-classified individuals for 1000 repetitions of a hot-deck imputation.

Perhaps it should not have been that surprising as well, since over 25 years ago, Chan and Dunn (1972) identified mean substitution as a preferred technique for imputation with discriminant analysis. One of the main criticisms of mean substitution is the fact that its use underestimates the variability in the variable being imputed. Regression imputation, on the other hand, does not have this same limitation, but these results indicate that regression imputation did not perform as well as either of the other two techniques in either the discriminant analysis or the logistic regression analysis, although the differences were not large. It was also a little surprising to observe that the regression technique did not perform better than the other two techniques, since this method is generally considered to be somewhat better in the sense that it incorporates other information about the individual in calculating the imputed value. This discrepancy might be explained by the fact that in classification, we are less concerned with predicting a specific value for an individual than we are with predicting that individual's group membership. Within each group are a variety of individuals who may possess a wide range of actual values on the criterion variable, which is much different from attempting to predict a specific outcome value (as is the case in multiple regression). Overall, it would appear that either mean substitution or hot-deck imputation would perform credibly in this application. Because mean substitution is easier to use than the hot-deck procedure, it would appear to be the better choice for practice.

There are, of course, limitations to this research. First, our sample was relatively small for the number of predictors used. Larger samples with different predictors would likely produce at least a somewhat different discriminant function and/or logistic regression model. With different models, and different data, it is very likely that the number of correctly re-classified individuals would vary somewhat. With the relatively small differences among the imputation techniques, and between the two classification procedures, even slight differences in the re-classification results could lead to different conclusions than these. Second, not all, nor necessarily even the best, imputation procedures were examined in this research. Choosing different techniques to investigate may also lead to different conclusions. Third, we only imputed a single variable at a time. If two or more variables had been imputed for a single individual the effect on classification may have been different. Fourth, it is possible that interactions among the predictor variables could be important as well, and how imputation may alter those interactions is not addressed here. Finally, it is uncertain how much our results are a function of the particular context, i.e., morbid cardiac events, within which we conducted this research, and how much would generally apply to other research scenarios. A different situation in which the predictor variables are very highly related to the outcomes of interest, resulting in extremely high numbers of correctly classified individuals in the original data, may be affected differently by imputation than was the case here.

At any rate, these results seem to indicate that using imputed values to replace missing values in classification models which have been previously derived from complete-case data can be a useful technique for making predictions we would otherwise be unable to make without re-calculating the models by leaving out the variables on which no data is available or having a series of models for use, each with a different combination of observed variables used as predictors. The ability to make such classifications with comparable accuracy, using a simple imputation technique such as mean substitution, would appear to be quite useful. By replacing missing values with the mean, thus being able to classify individuals who were previously unclassifiable using the same model, and to do so with a level of accuracy that is comparable to what would have been obtained had the values not been missing is a valuable tool. Furthermore the results were comparable regardless of whether the predictor variable being imputed was numerical or categorical.

Partial funding to support this research was provided by the Arkansas Science and Technology Authority and the Graduate School of the University of Northern Colorado. All correspondence concerning this research should be directed to Daniel J. Mundfrom, Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO 80639.

References

Affifi, A. A., & Elasahoff, R. M. (1966). Missing observations in multivariate statistics I: Review of the literature, *Journal of the American Statistical Association*, *61*, 595-604.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of the Royal Statistical Society*, *B22*, 302-306.

Campbell, R. W. F., Murray, A., & Julian, D. G. (1981). Ventricular arrhythmias in the first 12 hours of acute myocardial infarction: Natural history study, *British Heart Journal*, *46*, 351-357.

Chan, L. S., & Dunn, O. J. (1972). The treatment of missing values in discriminant analysis I: The sampling experiment, *Journal of the American Statistical Association*, *67*, 473-477.

De Martini, M., Valentini, R., Cesana, B., Massari, F. M., Lettino, M., Pupilella, T., Ambrosini, F., Eriano, G., La Marchesina, U., & Lotto, A. (1990). Early and late prognosis in acute myocardial infarction: A retrospective study in patients admitted to the coronary care unit in the past 10 years, *Italian Journal of Cardiology*, *20*, 215-226.

Gheorghiade, M., Anderson, J., Rosman, D. G., Lakier, J., Velardo, B., Goldberg, D., Friedman, A., Schultz, L., Tilley, B., & Goldstein, S. (1987). Risk identification at the time of admission to coronary care unit in patients with suspected myocardial infarction, *American Heart Journal*, *116*, 1212-1217.

Haitovsky, Y. (1968). Missing data in regression analysis, *Journal of the Royal Statistical Society*, *B30*, 67-81.

Hartley, H. O., & Hocking, R. R. (1971). The analysis of incomplete data, *Biometrics*, *27*, 783-808.

Kim, J. O., & Curry, J. (1977). The treatment of missing data in multivariate analysis, *Sociological Methods and Research*, *6*, 215-241.

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*, 573-593.

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

Lown, B., Fakhro, A. M., Hood, W. B., & Thorn, G. W. (1967). The coronary care unit: New perspectives and directions, *Journal of the American Medical Association*, *199*, 188-198.

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.

Rindskopf, D. (1992). A general approach to categorical data analysis with missing data, using generalized linear models with composite links, *Psychometrika*, *57*, 29-42.

Roth, P. L., & Switzer III, F. S. (1995). A Monte Carlo analysis of missing data techniques in a HRM setting, *Journal of Management*, *21*, 1003-1023.

Rubin, D. B. (1976). Inference and missing data, *Biometrika*, *63*, 581-592.

Rubin, D. B. (1991). EM and beyond, *Psychometrika*, *56*, 241-254.

Van Buuren, S., & van Ruckevorsel, J. L. A. (1992). Imputation of missing categorical data by maximizing internal consistency, *Psychometrika*, *57*, 567-580.

Systematically Missing Data and Multiple Regression Analysis: An Empirical Comparison of Deletion and Imputation Techniques

Lantry L. Brockmeier, Florida Department of Education
Jeffrey D. Kromrey, University of South Florida
Constance V. Hines, University of South Florida

The purpose of this study was to investigate, within the context of a two-predictor multiple regression analysis with systematically missing data, the effectiveness of eight missing data treatments on the sample estimate of R^2 and each standardized regression coefficient. Furthermore, the study investigated whether sample size, proportion of systematically missing data above the mean of the regressor, and the percentage of missing data affected the effectiveness of the eight missing data treatments. One thousand samples of size 50, 100, and 200 were generated per data set. The percentages of missing data were 0%, 10%, 20%, 30%, 40%, 50%, and 60%, occurring either on one regressor or across both regressors. The proportions of missing data that were above the mean value of the regressors were 0.60, 0.70, 0.80, or 0.90. The data were analyzed by computing effect sizes obtained from the missing data treatment conditions relative to the complete sample condition (i.e., 0% missing data). The results suggest that the stochastic multiple regression imputation technique was the most effective treatment of the missing data. Listwise and pairwise deletion approaches were less effective than stochastic multiple regression imputation but were superior to the other techniques examined.

Empirical investigations frequently have missing data on one or more variables. Researchers have long recognized that missing data within a study may be detrimental to any subsequent data analyses, interpretations, and conclusions. Unfortunately, researchers' recommendations (Guertin, 1968; Beale & Little, 1975; Gleason & Staelin, 1975; Frane, 1976; Kim & Curry, 1977; Santos 1981; Basilevsky et al., 1985; Raymond & Roberts, 1987) for managing missing data are not in complete agreement. Anderson, Basilevsky, and Hum (cited in Rossi, Wright, & Anderson, 1983) observed that the results of many research studies on missing data treatments are not comparable due to the method used, stratification categories (number of variables, sample size, proportion of missing data, and degree of multicollinearity), and the criteria that measure effectiveness. Kromrey and Hines (1991) stated that in the multiple regression context, the criteria should be the accuracy of the sample estimate of R^2 (coefficient of determination) and the regression coefficients.

While numerous missing data treatments are available for use by the applied researcher to manage missing data, researchers have characteristically utilized only two classes of procedures. Applied researchers typically employ the deletion procedures or the deterministic imputation procedures. The deletion procedures only utilize 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. With the deterministic imputation procedures (Santos, 1981; Kalton & Kasprzyk, 1982), the applied researcher employs a statistical procedure (e.g., mean substitution, simple regression, or multiple regression) to estimate the missing values. The residual (error) term in the equation for this estimation is set to zero.

Stochastic imputation is not typically utilized by applied researchers. However, evidence suggests that stochastic imputation procedures might be a viable alternative in the treatment of missing data (Santos, 1981; Kalton & Kasprzyk, 1982; Jinn & Sedransk, 1989; Keawkungal & Benson, 1989; Brockmeier, Hines, & Kromrey, 1993; Brockmeier, Kromrey, & Hines, 1994, 1995, and 1996). As with the deterministic imputation procedures, a statistical procedure is employed to estimate the missing values. The residual term when employing stochastic imputation is a randomly appended value in the estimation equation instead of zero as occurs with deterministic imputation (an example of SAS code that provides stochastic imputation is included in Appendix A).

Applied researchers also do not usually employ maximum likelihood estimation and multiple imputation for managing missing data. Scholarly work on maximum likelihood estimation and multiple imputation are found in the technical statistical journals, not usually in the journals of applied researchers (Kromrey, 1989; Brockmeier, 1992). Maximum likelihood estimation is infrequently utilized due to the lack of software and mathematical complexity (Little, 1992).

In most of the previously conducted research, the key assumption is that data are missing at random. Researchers are often advised that if data are randomly missing, and the percentage of missing data is not too large, then any missing data treatment is effective. This assumption of randomly missing data is tenuous in many cases. Cohen and Cohen (1975, 1983) and Tabachnick and Fidell (1983) describe procedures to test the assumption of randomly missing data. Kromrey and Hines (1994) elucidate that the assumption of randomly missing data is rarely tested and that the applied researcher is hard pressed to find guidance if data are missing systematically.

Kromrey and Hines (1994) examined the effectiveness of the deletion procedures and deterministic imputation procedures with systematically missing data in the context of missing data on one of two predictor variables. The authors stated that with moderate amounts of missing data, the deletion procedures yielded results similar to those results obtained without missing data. Kromrey and Hines indicated that the deterministic imputation procedures generally did not work well when compared to the results obtained with complete data.

Brockmeier, Kromrey, and Hines (1996) investigated, within the context of a two-predictor multiple regression analysis with systematically missing data, the effectiveness of eight missing data treatments on the sample estimate of R^2 and each standardized regression coefficient. The stochastic multiple regression imputation technique was effective with as much as 60% of the data missing. With smaller proportions of missing data, both the listwise and pairwise deletion approaches were also effective in estimating R^2 and the regression weights.

The present study extends the previous work of Kromrey and Hines (1994) and Brockmeier et al. (1996). First, the study continues to investigate the effectiveness of the stochastic and deterministic imputation procedures and the deletion procedures with systematically missing data in data sets with different correlations between variables. Second, the number of levels of systematically missing data was increased to be more representative of authentic data sets.

Purpose

The purpose of this study was to investigate, within the context of a two-predictor multiple regression analysis with systematically missing data, the effectiveness of eight missing data treatments on the sample estimate of R^2 and each standardized regression coefficient. The study also examined whether the proportion of systematically missing data above the mean of each independent variable affected the effectiveness of the eight missing data treatments. Three types of missing data treatments were examined: deletion, deterministic imputation, and stochastic imputation. The missing data treatments examined in this study were: (a) listwise deletion, (b) pairwise deletion, (c) deterministic mean substitution, (d) deterministic simple regression, (e) deterministic multiple regression, (f) stochastic mean substitution, (g) stochastic simple regression, and (h) stochastic multiple regression.

Method

Data Source

Data selected for this investigation were chosen from the work of Skaalvik and Rankin (1995). Skaalvik and Rankin examined the relationship between math and verbal achievement and measures of motivation. One data set consisted of correlations between the measures of mathematics achievement, self-perceived ability to learn mathematics, and mathematics intrinsic motivation for grade six students. The second data set consisted of correlations between the same three measures, but for grade nine students. In the data obtained from the older students, the correlations between variables were higher.

Table 1. Summary Descriptive Statistics for the Population on the Grade Six Data

			Correlations	
	Mean	SD	(X_1)	(X_2)
(Y) Mathematics Achievement	12.2672	4.6380	0.33	0.25
(X_1) Mathematics Self-perceived Ability	13.1075	2.1900	- -	0.58
(X_2) Mathematics Intrinsic Motivation	50.3355	14.5836		

Table 2. Summary Descriptive Statistics for the Population on the Grade Nine Data

			Correlations	
	Mean	SD	(X_1)	(X_2)
(Y) Mathematics Achievement	9.3039	4.4363	0.58	0.59
(X_1) Mathematics Self-perceived Ability	12.2064	2.7616	- -	0.70
(X_2) Mathematics Intrinsic Motivation	45.0580	17.5328		

Table 3. Regression Models in the Study as Computed on the Population.

Data set	Dependent Variable	Independent Variables	Beta	R^2
Grade Six	Y	X_1	0.2779	0.1152
Data		X_2	0.0917	
Grade Nine	Y	X_1	0.3239	0.4027
Data		X_2	0.3638	

SAS/IML was employed to generate multivariate normal random variables given the correlation between variables and the mean and standard deviation of each variable. Tables 1 and 2 present the means and standard deviations of each variable and the correlation between variables by data set. Table 3 presents the regression model for each data set.

Experimental Design

The study employed a 2 x 3 x 4 x 13 x 8 experimental design. The design included two between-subjects variables (pseudopopulation and sample size) and three within-subjects variables (proportion of missing data above the mean, percentage of missing data, and missing data treatment). One thousand samples of size 50, 100, and 200 were generated per data set. The four proportions of systematically missing data above the mean of each independent variable were 0.60, 0.70, 0.80, and 0.90. The 13 percentages of missing data generated by predictor variable (X_1,X_2) were (0%,0%), (10%,0%), (20%,0%), (30%,0%), (40%,0%), (50%,0%), (60%,0%), (10%,10%), (20%,10%), (20%,20%), (30%,20%), (40%,20%), and (30%,30%). The eight missing data treatments examined were listwise deletion, pairwise deletion, deterministic mean substitution, deterministic simple regression, deterministic multiple regression, stochastic mean substitution, stochastic simple regression, and stochastic multiple regression.

The pseudopopulations were not manipulated within the experiment, but were generated to obtain the desired correlational differences between variables in each data set. The sample sizes and missing data treatments were chosen to replicate the earlier work of Kromrey and Hines (1994) and Brockmeier et al. (1996). The percentages of missing data were chosen to be representative of the research of Kromrey and Hines (1994) and Brockmeier et al. (1993, 1994, 1995, and 1996). The proportion of systematically missing data above the mean of the regressors was altered to create increasing degrees of distortion in the observed data. The probability of a missing value was established as proportional to the value of the variable. Kromrey and Hines (1994) indicated that this process reduces the variance and exaggerates the skewness in the observed distribution, and that the value of the observed mean is altered by the asymmetry.

Statistical Analysis

The dependent variables analyzed were the sample estimate of R^2 and the standardized regression coefficients. The data 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).

Results

To conserve space, the results are presented as effect sizes representing the difference between the mean value of the sample statistic (R^2 or standardized regression weight) and the mean value obtained from the complete data condition. This difference in means was then divided by the standard deviation of the statistic obtained in the complete data condition. For more complete results, the raw means and standard deviations are available from the first author.

Sample estimates were considered to be reasonably unbiased and to present few practical problems to applied researchers if the absolute value of the effect size was less than 0.3 (Kromrey & Hines, 1991). The criterion of 0.3 was chosen because the regression coefficients and the sample estimate of R^2 are both subject to substantive interpretation and tests of statistical significance.

Effects of Missing Data on Sample Estimate of R^2

Effect sizes for the estimation of R^2 are presented in Tables 4 and 5. These data reveal that stochastic multiple regression generated fewer effect sizes greater than the criterion of 0.3 than any other missing data treatment. Stochastic multiple regression generated effect sizes greater than the criterion 4.3% (12 of 280 effect sizes) of the time. Of the 12 effect sizes greater than the criterion, 11 effect sizes occurred when the percentage of missing data was 60%. Ten of the twelve effect sizes greater than the criterion occurred for the grade six data set and the sample size of 50. Two other cases occurred with the sample size of 200 when the proportion of missing data above the mean was 0.90 and the proportion of missing data was 60%.

Pairwise deletion produced effect sizes greater than the criterion 8.9% (25 of 280 effect sizes) of the time, more than twice as frequently as that provided by stochastic multiple regression. Effect sizes greater than 0.3 occurred 19 of 25 times when the proportion of missing data above the mean was 0.80 or 0.90 and 22 of 25 times when the percentage of missing data was 50% or 60%. Listwise deletion yielded effect sizes greater than the criterion 22.5% (63 of 280 effect sizes) of the time. Across both data sets and sample sizes, 58 of 63 effect sizes greater than the criterion occurred when the percentage of missing data was 50% or 60%. The other five effect sizes greater than the criterion occurred when the percentage of missing data was 40%.

Notably worse performance was observed for the other missing data treatments. Deterministic simple regression, deterministic mean substitution, deterministic multiple regression, stochastic simple regression, and stochastic mean substitution yielded effect sizes greater than the criterion from 45.7% (128 of 280 effect sizes) to 89.6% (251 of 280 effect sizes) of the time. For each of these missing data treatments, effect sizes greater than the criterion occurred about equally across the two data sets.

Effects of Missing Data on the First Standardized Regression Coefficient (X_1)

Tables 6 and 7 report the effect sizes for the first standardized regression coefficient. Stochastic multiple regression, listwise deletion, and pairwise deletion yielded the fewest effects sizes greater than the criterion for this coefficient. Stochastic multiple regression generated effect sizes greater than the criterion 9.6% (27 of 280 effect sizes) of the time. Eighteen of these 27 conditions occurred when the percentage of missing data was 60%, and 24 of the 27 conditions were when the proportion of missing data above the mean was 0.80 or 0.90. Listwise deletion generated effects sizes greater than the criterion 11.4% (32 of 280 effect sizes) of the time, only slightly more frequently than that of stochastic multiple regression. Twenty of the 32 conditions were those with 60% missing data, and 28 of the 32 conditions were those in which the proportion of missing data above the mean was 0.80 or 0.90. Pairwise deletion produced effect sizes greater than the criterion 18.2% (51 of 280 effect sizes) of the time.

As with the estimation of R^2, the effectiveness of deterministic simple regression, deterministic multiple regression, deterministic mean substitution, stochastic simple regression, and stochastic mean substitution were notably lower than that of stochastic multiple regression and the two deletion procedure. These techniques produced effect sizes greater than the criterion from 40.4% (113 of 280 effect sizes) to 85.7% (240 of 280 effect sizes) of the time.

Effects of Missing Data on the Second Standardized Regression Coefficient (X_2)

Examination of Tables 8 and 9 reveals that listwise deletion yielded no effect sizes greater than the criterion of 0.3 with the lower correlated data set (i.e., grade six data) and only six effect sizes (2.1%) greater than the criterion with the higher correlated data set (i.e., grade nine data). Five of these six effect sizes occurred when the proportion of missing data above the mean was 0.80, or 0.90 and the percentage of missing data was 50% or 60%. Stochastic multiple regression yielded effect sizes greater than the criterion 9.3% (26 of 280 effect sizes) of the time. These effect sizes occurred 23 of 26 times when the proportion of missing data above the mean was 0.80 or 0.90, and 18 of 26 times when the percentage of missing data was 60%. Deterministic simple regression yielded effect sizes greater than the criterion 16.4% (46 of 280 effect sizes) of the time and pairwise deletion yielded effect sizes greater than the criterion 22.1% (62 of 280 effect sizes) of the time.

Deterministic multiple regression, stochastic simple regression, deterministic mean substitution, and stochastic mean substitution yielded effect sizes greater than the criterion from 52.5% (147 of 280 effect sizes) to 78.9% (221 of 280 effect sizes) of the time.

Discussion

In the context of the current study, a two-predictor multiple regression analysis with systematically missing data, the results suggest large differences in the effectiveness of the eight missing data treatments. Stochastic multiple regression performed the best of the missing data treatments in yielding fewer sample estimates of R^2 that differed from the complete sample condition. Pairwise deletion, the second best performer of the missing data treatments, produced biased sample estimates of R^2 more than twice as frequently as stochastic multiple regression. The effect sizes greater than 0.3 that were generated by pairwise deletion occurred when the percentage of missing data was 50% or 60%. Listwise deletion yielded five times more effect sizes greater than the criterion when compared to stochastic multiple regression. Deterministic simple regression was the next most effective missing data treatment, but its performance was notably worse than those of the three best treatments. Deterministic mean substitution, deterministic multiple regression, stochastic simple regression, and stochastic mean substitution produced sample estimates of R^2 that differed from the complete sample condition at least 45.7% of the time. These four missing data treatments were simply ineffective in producing unbiased sample estimates of R^2.

Similar results were obtained for the estimation of regression weights in the presence of missing data. Stochastic multiple regression was the most effective at producing unbiased estimates of the first standardized regression coefficient but listwise deletion was more effective at generating unbiased estimates of the second. The conditions under which the estimates of the standardized regression coefficients differed from the complete sample condition, were the most extreme conditions examined (i.e., when the proportion of missing data above the mean was 0.80 or 0.90 and the percentage of missing data was 50% or 60%).

Table 4. Effect Sizes of the Sample Estimate of R-Square for the Grade Six Data by Sample Size, Proportion of Missing Data High, Percentage of Missing Data, and Missing Data Treatment

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
50	.60	10%,0%	0.0316	0.0208	-0.0907	-0.0589	0.0983	-0.1480	-0.0947	0.0396
	.60	20%,0%	0.0783	0.0621	-0.1671	-0.1264	0.2198	-0.2556	-0.2054	0.0859
	.60	30%,0%	-0.1561	0.1199	-0.2322	-0.1832	0.3925	-0.3452	-0.2935	0.1189
	.60	40%,0%	-0.2511	0.1898	-0.2929	-0.2453	0.6098	-0.4063	-0.3719	0.1968
	.60	50%,0%	-0.4747	0.4125	-0.2998	-0.2634	1.0192	-0.4466	-0.3964	0.3747
	.60	60%,0%	-0.4832	0.3935	-0.4123	-0.3612	1.3667	-0.5001	-0.4876	0.4105
	.70	10%,0%	0.0000[a]							
	.70	20%,0%	0.0784	0.0668	-0.1600	-0.1170	0.2249	-0.2446	-0.1997	0.0777
	.70	30%,0%	0.1236	0.0823	-0.2546	-0.1935	0.3641	-0.3703	-0.3092	0.1337
	.70	40%,0%	0.1661	0.1353	-0.3076	-0.2492	0.5669	-0.4037	-0.3723	0.1773
	.70	50%,0%	0.2621	0.2209	-0.3693	-0.3023	0.8991	-0.4738	-0.4355	0.2724
	.70	60%,0%	0.3924	0.3424	-0.4257	-0.3704	1.3806	-0.5150	-0.4813	0.3850
	.80	10%,0%	0.0371	0.0265	-0.0887	-0.0623	0.0992	-0.1429	-0.1134	0.0259
	.80	20%,0%	0.0883	0.0735	-0.1570	-0.1262	0.2187	-0.2491	-0.2075	0.0918
	.80	30%,0%	0.1338	0.1102	-0.2311	-0.1854	0.3775	-0.3595	-0.3102	0.1176
	.80	40%,0%	0.1881	0.1584	-0.2976	-0.2389	0.6096	-0.4139	-0.3694	0.2013
	.80	50%,0%	0.1121	0.0933	-0.4178	-0.3574	0.7386	-0.4775	-0.4638	0.1630
	.80	60%,0%	0.0386	0.0647	-0.4949	-0.4464	1.0187	-0.5253	-0.4954	0.2159
	.90	10%,0%	0.0000[a]							
	.90	20%,0%	0.0660	0.0516	-0.1835	-0.1292	0.2271	-0.2854	-0.2069	0.0662
	.90	30%,0%	0.0687	0.0655	-0.2598	-0.1997	0.3557	-0.3791	-0.2914	0.1200
	.90	40%,0%	0.0784	0.0699	-0.3426	-0.2674	0.5478	-0.4427	-0.3767	0.1917
	.90	50%,0%	-0.0758	0.0095	-0.4391	-0.3913	0.6428	-0.4879	-0.4804	0.1107
	.90	60%,0%	-0.1736	-0.0573	-0.5101	-0.4648	0.7441	-0.5477	-0.5358	0.1111
50	.60	10%,10%	0.0668	0.0642	-0.1032	-0.0696	0.1614	-0.1823	-0.1231	0.0835
	.60	20%,10%	0.1825	0.1211	-0.1579	-0.1188	0.3004	-0.3043	-0.2057	0.1367
	.60	20%,20%	0.2525	0.1839	-0.1693	-0.1282	0.4036	-0.3397	-0.2312	0.2196
	.60	30%,20%	0.3912	0.2644	-0.2219	-0.1551	0.6706	-0.4229	-0.3362	0.2982
	.60	40%,20%	0.5757	0.3297	-0.3145	-0.2581	0.9462	-0.5471	-0.4105	0.4146
	.60	30%,30%	0.5596	0.3378	-0.2431	-0.1908	0.8362	-0.4589	-0.3260	0.4152
	.70	10%,10%	0.0000 [a]							
	.70	20%,10%	0.1397	0.0870	-0.1857	-0.1337	0.2858	-0.3247	-0.2365	0.1523
	.70	20%,20%	0.2503	0.1796	-0.1624	-0.1311	0.3908	-0.3195	-0.2189	0.1999
	.70	30%,20%	0.2837	0.2257	-0.2306	-0.1848	0.5699	-0.4496	-0.3327	0.2444
	.70	40%,20%	0.4986	0.2942	-0.3018	-0.2588	0.8634	-0.5471	-0.3759	0.3573
	.70	30%,30%	0.4928	0.3182	-0.2355	-0.1877	0.8204	-0.4691	-0.3174	0.4053
	.80	10%,10%	0.0784	0.0695	-0.0902	-0.0653	0.1530	-0.1703	-0.1163	0.0852
	.80	20%,10%	0.1690	0.1218	-0.1598	-0.1303	0.2893	-0.2934	-0.2253	0.1338
	.80	20%,20%	0.2363	0.1857	-0.1555	-0.1145	0.4178	-0.3264	-0.2300	0.2212
	.80	30%,20%	0.2973	0.2596	-0.2221	-0.1983	0.5685	-0.4153	-0.3207	0.2411
	.80	40%,20%	0.3558	0.2919	-0.2950	-0.2453	0.8117	-0.4976	-0.3692	0.3531
	.80	30%,30%	0.3642	0.3446	-0.2068	-0.1789	0.7251	-0.4288	-0.3142	0.3056
	.90	10%,10%	0.0000 [a]							
	.90	20%,10%	0.1532	0.0992	-0.1730	-0.1274	0.2958	-0.3016	-0.2181	0.1364
	.90	20%,20%	0.1790	0.1734	-0.1590	-0.1323	0.3715	-0.3300	-0.2472	0.1918
	.90	30%,20%	0.1710	0.1761	-0.2527	-0.2224	0.5001	-0.4448	-0.3353	0.2168
	.90	40%,20%	0.2523	0.2907	-0.2792	-0.2577	0.7274	-0.4997	-0.4064	0.3049
	.90	30%,30%	0.1893	0.3533	-0.2068	-0.2088	0.5802	-0.4602	-0.3393	0.2635

(Table continues)

Table 4 (continued).

Sample Size	Proportion High	Percentage of Missing Data	L	P	MS	SR	MR	SMS	SSR	SMR
200	.60	10%,0%	0.0122	0.0076	-0.1719	-0.1177	0.1385	-0.3146	-0.2137	0.0106
	.60	20%,0%	0.0311	0.0208	-0.3309	-0.2425	0.3107	-0.5390	-0.4345	0.0510
	.60	30%,0%	0.0535	0.0449	-0.4685	-0.3578	0.5391	-0.7191	-0.6083	0.0404
	.60	40%,0%	0.0854	0.0675	-0.6004	-0.4819	0.8254	-0.8479	-0.7762	0.0852
	.60	50%,0%	0.1278	0.0915	-0.7154	-0.5910	1.2319	-0.9612	-0.8658	0.1529
	.60	60%,0%	0.2099	0.1652	-0.8206	-0.7157	1.8387	-1.0430	-0.9857	0.1984
	.70	10%,0%	0.0217	0.0177	-0.1676	-0.1148	0.1492	-0.3077	-0.2213	0.0290
	.70	20%,0%	0.0205	0.0131	-0.3309	-0.2410	0.3044	-0.5452	-0.4208	0.0136
	.70	30%,0%	0.0162	0.0145	-0.4806	-0.3733	0.5032	-0.7247	-0.6202	0.0464
	.70	40%,0%	0.0313	0.0242	-0.6153	-0.4833	0.8179	-0.8423	-0.7503	0.0797
	.70	50%,0%	-0.0770	-0.0329	-0.7593	-0.6325	1.1078	-0.9856	-0.8926	0.0664
	.70	60%,0%	-0.2324	-0.1434	-0.8993	-0.7893	1.4826	-1.0682	-1.0277	0.0194
	.80	10%,0%	0.0096	0.0105	-0.1719	-0.1204	0.1405	-0.3063	-0.2187	0.0194
	.80	20%,0%	0.0080	0.0015	-0.3437	-0.2508	0.2952	-0.5598	-0.4387	0.0160
	.80	30%,0%	-0.0028	0.0025	-0.4917	-0.3764	0.5189	-0.7447	-0.6108	0.0545
	.80	40%,0%	-0.1251	-0.0911	-0.6667	-0.5236	0.7137	-0.8937	-0.7924	0.0294
	.80	50%,0%	-0.3413	-0.2211	-0.8305	-0.6904	0.9385	-0.9975	-0.9341	-0.0146
	.80	60%,0%	-0.9049	-0.5265	-1.0072	-0.9076	0.8731	-1.1066	-1.0576	-0.2344
	.90	10%,0%	0.0007	-0.0017	-0.1834	-0.1292	0.1301	-0.2913	-0.2390	-0.0043
	.90	20%,0%	-0.0107	-0.0127	-0.3565	-0.2580	0.2983	-0.5573	-0.4564	0.0230
	.90	30%,0%	-0.1395	-0.0984	-0.5468	-0.4031	0.4478	-0.7640	-0.6320	0.0022
	.90	40%,0%	-0.3295	-0.2164	-0.7229	-0.5716	0.6058	-0.9102	-0.7991	-0.0423
	.90	50%,0%	-0.7429	-0.4629	-0.9122	-0.7765	0.6343	-1.0366	-0.9678	-0.2109
	.90	60%,0%	-1.2608	-0.7036	-1.0424	-0.9421	0.4194	-1.1131	-1.0620	-0.4166
200	.60	10%,10%	0.0611	0.0371	-0.1853	-0.1331	0.1785	-0.3434	-0.2513	0.0329
	.60	20%,10%	0.0328	0.0243	-0.3686	-0.2744	0.3197	-0.6670	-0.4680	0.0377
	.60	20%,20%	0.1028	0.0841	-0.3538	-0.2689	0.3921	-0.6758	-0.4771	0.0719
	.60	30%,20%	0.2269	0.1357	-0.4737	-0.3677	0.6743	-0.9117	-0.6555	0.1561
	.60	40%,20%	0.2765	0.1301	-0.6176	-0.5077	0.9548	-1.1043	-0.8155	0.1773
	.60	30%,30%	0.1862	0.1358	-0.5197	-0.4165	0.7010	-1.0358	-0.7265	0.1476
	.70	10%,10%	0.0055	0.0256	-0.1961	-0.1302	0.1784	-0.3738	-0.2467	0.0367
	.70	20%,10%	0.0667	0.0494	-0.3443	-0.2614	0.3434	-0.6321	-0.4696	0.0545
	.70	20%,20%	0.0820	0.0872	-0.3432	-0.2720	0.3748	-0.7092	-0.4867	0.0772
	.70	30%,20%	0.0697	0.0769	-0.5058	-0.3863	0.6182	-0.9386	-0.6511	0.1252
	.70	40%,20%	-0.0032	0.1072	-0.6300	-0.5120	0.8961	-1.1317	-0.8237	0.1396
	.70	30%,30%	0.1016	0.1564	-0.4803	-0.3931	0.7031	-1.0259	-0.6758	0.1637
	.80	10%,10%	0.0220	0.0297	-0.1884	-0.1304	0.1771	-0.3683	-0.2478	0.0358
	.80	20%,10%	0.0245	0.0504	-0.3382	-0.2447	0.3537	-0.6297	-0.4484	0.0542
	.80	20%,20%	0.0220	0.0646	-0.3642	-0.2751	0.3878	-0.7061	-0.4768	0.1053
	.80	30%,20%	-0.0646	0.0892	-0.4791	-0.3778	0.5804	-0.8920	-0.6548	0.1042
	.80	40%,20%	-0.1855	0.0639	-0.6158	-0.5293	0.7564	-1.1101	-0.8279	0.0724
	.80	30%,30%	-0.1998	0.1923	-0.4574	-0.3909	0.6485	-0.9951	-0.6928	0.1460
	.90	10%,10%	0.0331	0.0450	-0.1746	-0.1235	0.1853	-0.3604	-0.2477	0.0462
	.90	20%,10%	-0.0064	0.0525	-0.3296	-0.2340	0.3582	-0.6149	-0.4397	0.0710
	.90	20%,20%	-0.1236	0.0321	-0.3820	-0.2909	0.3498	-0.7376	-0.5033	0.0500
	.90	30%,20%	-0.2198	0.0610	-0.4850	-0.3993	0.5346	-0.9192	-0.6615	0.0620
	.90	40%,20%	-0.5107	0.1100	-0.5785	-0.5146	0.7081	-1.0827	-0.8029	0.0728
	.90	30%,30%	-0.5501	0.2709	-0.4148	-0.3865	0.5613	-0.9770	-0.6972	0.1042

Note. L: listwise deletion, P: pairwise deletion, MS: mean substitution, SR: simple regression, MR: multiple regression, SMS: stochastic mean substitution, SSR: stochastic simple regression, SMR: stochastic multiple regression. [a] Data were not computed for this combination of sample size, proportion of missing data high, and percentage of missing data.

Table 5. Effect Sizes of the Sample Estimate of R-Square for the Grade Nine Data by Sample Size, Proportion of Missing Data High, Percentage of Missing Data, and Missing Data Treatment

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
50	.60	10%,0%	0.0225	0.0251	-0.0867	-0.0461	0.0853	-0.1487	-0.1048	0.0342
	.60	20%,0%	0.0311	0.0392	-0.1723	-0.1075	0.1765	-0.2539	-0.1796	0.0516
	.60	30%,0%	0.0409	0.0611	-0.2397	-0.1681	0.2707	-0.3402	-0.2793	0.0759
	.60	40%,0%	0.0902	0.1202	-0.2979	-0.2228	0.4520	-0.3884	-0.3377	0.1198
	.60	50%,0%	0.1163	0.1573	-0.3452	-0.2735	0.6156	-0.4425	-0.3922	0.1396
	.60	60%,0%	0.1533	0.2641	-0.3889	-0.3278	0.9518	-0.4677	-0.4338	0.2299
	.70	10%,0%	0.0000[a]							
	.70	20%,0%	0.0491	0.0536	-0.1556	-0.0942	0.1860	-0.2470	-0.1734	0.0680
	.70	30%,0%	0.0073	0.0573	-0.2565	-0.1666	0.2877	-0.3525	-0.2797	0.0700
	.70	40%,0%	0.0201	0.0847	-0.3104	-0.2167	0.4530	-0.4112	-0.3285	0.1248
	.70	50%,0%	-0.1186	0.0684	-0.3800	-0.2913	0.6189	-0.4453	-0.4065	0.1581
	.70	60%,0%	-0.2708	0.1164	-0.4250	-0.3486	0.8867	-0.4721	-0.4497	0.2016
	.80	10%,0%	0.0214	0.0153	-0.0943	-0.0557	0.0741	-0.1616	-0.1069	0.0204
	.80	20%,0%	0.0275	0.0364	-0.1701	-0.1155	0.1630	-0.2574	-0.1880	0.0406
	.80	30%,0%	-0.0367	0.0396	-0.2602	-0.1695	0.2874	-0.3340	-0.2730	0.0777
	.80	40%,0%	-0.0753	0.0540	-0.3180	-0.2291	0.4322	-0.4130	-0.3257	0.0982
	.80	50%,0%	-0.3059	0.0200	-0.3943	-0.3083	0.5711	-0.4511	-0.4013	0.1279
	.80	60%,0%	-0.7960	-0.0397	-0.4639	-0.3988	0.6800	-0.4945	-0.4799	0.1295
	.90	10%,0%	0.0000[a]							
	.90	20%,0%	-0.0216	0.0147	-0.1856	-0.1184	0.1515	-0.2670	-0.1921	0.0231
	.90	30%,0%	-0.0858	0.0330	-0.2579	-0.1572	0.3041	-0.3371	-0.2515	0.0931
	.90	40%,0%	-0.2914	-0.0302	-0.3506	-0.2535	0.3725	-0.4294	-0.3481	0.0777
	.90	50%,0%	-0.8275	-0.1089	-0.4434	-0.3601	0.4342	-0.4771	-0.4249	0.0116
	.90	60%,0%	-1.1856	-0.1301	-0.4798	-0.4160	0.4876	-0.5019	-0.4587	0.0060
50	.60	10%,10%	0.0700	0.0448	-0.2013	-0.1172	0.1554	-0.4308	-0.2247	0.0462
	.60	20%,10%	0.0496	0.0735	-0.3044	-0.1697	0.2504	-0.6226	-0.3165	0.0819
	.60	20%,20%	0.1256	0.1120	-0.3882	-0.2450	0.3524	-0.8244	-0.4480	0.1249
	.60	30%,20%	0.1449	0.1592	-0.4423	-0.2709	0.4822	-0.9791	-0.5223	0.1756
	.60	40%,20%	0.1463	0.2344	-0.5050	-0.3232	0.6690	-1.1814	-0.5983	0.2030
	.60	30%,30%	0.1251	0.2282	-0.5653	-0.3628	0.6297	-1.2934	-0.6529	0.1897
	.70	10%,10%	0.0000[a]							
	.70	20%,10%	0.0273	0.0617	-0.3027	-0.1696	0.2483	-0.6288	-0.3257	0.0763
	.70	20%,20%	0.0084	0.0933	-0.4020	-0.2569	0.3142	-0.8763	-0.4526	0.0916
	.70	30%,20%	0.0209	0.1798	-0.4328	-0.2825	0.4748	-1.0079	-0.5355	0.1528
	.70	40%,20%	-0.0484	0.2653	-0.4817	-0.3476	0.6329	-1.1538	-0.6445	0.1784
	.70	30%,30%	-0.1473	0.2947	-0.5058	-0.3842	0.5906	-1.2845	-0.6659	0.1824
	.80	10%,10%	0.0284	0.0566	-0.1967	-0.1096	0.1696	-0.3867	-0.2047	0.0642
	.80	20%,10%	-0.0058	0.0724	-0.2800	-0.1613	0.2547	-0.5874	-0.3124	0.0821
	.80	20%,20%	-0.0316	0.1193	-0.3727	-0.2447	0.3329	-0.7880	-0.4605	0.1053
	.80	30%,20%	-0.1470	0.1989	-0.4295	-0.2992	0.4689	-0.9975	-0.5439	0.1307
	.80	40%,20%	-0.4512	0.2800	-0.4574	-0.3741	0.5466	-1.1321	-0.6342	0.1260
	.80	30%,30%	-0.3978	0.3997	-0.4380	-0.3653	0.5564	-1.2396	-0.6626	0.1472
	.90	10%,10%	0.0000[a]							
	.90	20%,10%	0.0186	0.0817	-0.2877	-0.1684	0.2684	-0.5675	-0.2973	0.0994
	.90	20%,20%	-0.2366	0.1093	-0.3848	-0.2596	0.3070	-0.8182	-0.4560	0.0650
	.90	30%,20%	-0.4642	0.2436	-0.3805	-0.3183	0.4009	-1.0173	-0.5446	0.0867
	.90	40%,20%	-0.8097	0.4404	-0.3643	-0.3655	0.5384	-1.0630	-0.6417	0.1400
	.90	30%,30%	-1.1361	0.5863	-0.3586	-0.4131	0.4015	-1.1560	-0.6953	0.0379

(Table continues)

Table 5. (continued)

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
200	.60	10%,0%	0.0088	0.0048	-0.1783	-0.1012	0.1045	-0.3026	-0.1862	0.0052
	.60	20%,0%	0.0002	0.0103	-0.3282	-0.1957	0.2402	-0.5146	-0.3456	0.0177
	.60	30%,0%	0.0097	0.0239	-0.4479	-0.2919	0.4056	-0.6369	-0.4907	0.0380
	.60	40%,0%	0.0186	0.0436	-0.5449	-0.3828	0.6306	-0.7417	-0.6105	0.0566
	.60	50%,0%	-0.0320	0.0282	-0.6507	-0.4940	0.8763	-0.8155	-0.7259	0.0587
	.60	60%,0%	-0.1513	-0.0102	-0.7457	-0.6056	1.1857	-0.8851	-0.8131	0.0477
	.70	10%,0%	0.0019	0.0056	-0.1749	-0.0973	0.1081	-0.3034	-0.1865	0.0196
	.70	20%,0%	-0.0124	0.0109	-0.3253	-0.1934	0.2427	-0.4975	-0.3404	0.0404
	.70	30%,0%	-0.0251	0.0092	-0.4496	-0.2940	0.3976	-0.6545	-0.5041	0.0210
	.70	40%,0%	-0.0733	-0.0052	-0.5608	-0.4005	0.5881	-0.7467	-0.6275	0.0429
	.70	50%,0%	-0.3215	-0.0582	-0.6721	-0.4968	0.8593	-0.8279	-0.7227	0.0349
	.70	60%,0%	-0.8557	-0.2373	-0.7895	-0.6454	1.0227	-0.8982	-0.8270	-0.0467
	.80	10%,0%	0.0039	0.0106	-0.1668	-0.0932	0.1096	-0.2970	-0.1729	0.0134
	.80	20%,0%	-0.0349	-0.0092	-0.3392	-0.2080	0.2237	-0.5224	-0.3620	0.0053
	.80	30%,0%	-0.1256	-0.0395	-0.4774	-0.3124	0.3670	-0.6779	-0.5095	0.0061
	.80	40%,0%	-0.3246	-0.1024	-0.6094	-0.4343	0.5377	-0.7736	-0.6435	-0.0098
	.80	50%,0%	-0.8434	-0.2470	-0.7304	-0.5588	0.6957	-0.8526	-0.7613	-0.0601
	.80	60%,0%	-2.0911	-0.5009	-0.8518	-0.7193	0.6845	-0.9211	-0.8607	-0.2185
	.90	10%,0%	-0.0289	-0.0073	-0.1865	-0.1018	0.0997	-0.3090	-0.1918	-0.0056
	.90	20%,0%	-0.1028	-0.0308	-0.3538	-0.2023	0.2222	-0.5253	-0.3524	0.0112
	.90	30%,0%	-0.2402	-0.0737	-0.4972	-0.3155	0.3668	-0.6818	-0.5065	-0.0135
	.90	40%,0%	-0.6653	-0.2103	-0.6480	-0.4472	0.4927	-0.7823	-0.6499	-0.0238
	.90	50%,0%	-1.6310	-0.4491	-0.7934	-0.6178	0.5080	-0.8865	-0.7948	-0.1651
	.90	60%,0%	-2.7342	-0.6396	-0.8767	-0.7458	0.3644	-0.9206	-0.8572	-0.3129
200	.60	10%,10%	0.0187	0.0128	-0.4179	-0.2292	0.2164	-0.8489	-0.4477	0.0163
	.60	20%,10%	0.0020	0.0179	-0.5787	-0.3259	0.3239	-1.1697	-0.6206	0.0095
	.60	20%,20%	0.0358	0.0486	-0.7713	-0.4349	0.4769	-1.6598	-0.8437	0.0446
	.60	30%,20%	0.0030	0.0558	-0.9093	-0.5542	0.5719	-2.0329	-1.0384	0.0427
	.60	40%,20%	-0.0592	0.0870	-1.0251	-0.6417	0.7648	-2.3233	-1.1413	0.0669
	.60	30%,30%	-0.0962	0.0888	-1.0873	-0.6697	0.7398	-2.5399	-1.2424	0.0702
	.70	10%,10%	-0.0213	0.0060	-0.4260	-0.2272	0.2181	-0.8397	-0.4321	0.0285
	.70	20%,10%	-0.0475	0.0182	-0.5691	-0.3189	0.3272	-1.1807	-0.5993	0.0191
	.70	20%,20%	-0.0619	0.0536	-0.7487	-0.4352	0.4585	-1.6632	-0.8401	0.0390
	.70	30%,20%	-0.1976	0.0770	-0.8965	-0.5570	0.5701	-2.0138	-1.0381	0.0432
	.70	40%,20%	-0.4626	0.1337	-0.9598	-0.6517	0.7462	-2.2493	-1.1657	0.0460
	.70	30%,30%	-0.5203	0.1825	-1.0096	-0.6996	0.6919	-2.4983	-1.2857	0.0315
	.80	10%,10%	0.0059	0.0247	-0.3989	-0.2235	0.2207	-0.8156	-0.4411	0.0215
	.80	20%,10%	-0.0921	0.0230	-0.5512	-0.3223	0.3234	-1.1501	-0.6135	0.0142
	.80	20%,20%	-0.2632	0.0396	-0.7615	-0.4726	0.4187	-1.6917	-0.8555	0.0175
	.80	30%,20%	-0.5341	0.1117	-0.8442	-0.5744	0.5377	-1.9998	-1.0314	0.0033
	.80	40%,20%	-1.1578	0.2250	-0.8749	-0.6776	0.6482	-2.2202	-1.1802	-0.0127
	.80	30%,30%	-1.2698	0.3677	-0.8887	-0.7154	0.6104	-2.4096	-1.2646	-0.0324
	.90	10%,10%	-0.0501	0.0226	-0.3953	-0.2185	0.2207	-0.8145	-0.4262	0.0125
	.90	20%,10%	-0.1844	0.0225	-0.5542	-0.3351	0.3148	-1.1721	-0.6347	0.0088
	.90	20%,20%	-0.4234	0.1062	-0.6870	-0.4523	0.4384	-1.6081	-0.8709	0.0212
	.90	30%,20%	-0.9301	0.2304	-0.7203	-0.5570	0.5346	-1.9020	-1.0174	0.0190
	.90	40%,20%	-2.2154	0.4767	-0.7293	-0.6999	0.5528	-2.0722	-1.2006	-0.0876
	.90	30%,30%	-2.4545	0.7049	-0.7391	-0.7797	0.4242	-2.3257	-1.3174	-0.2027

Note. L: listwise deletion, P: pairwise deletion, MS: mean substitution, SR: simple regression, MR: multiple regression, SMS: stochastic mean substitution, SSR: stochastic simple regression, SMR: stochastic multiple regression. [a] Data were not computed for this combination of sample size, proportion of missing data high, and percentage of missing data.

Table 6. Effect Sizes of the First Standardized Regression Coefficient (X_1) for the Grade Six Data by Sample Size, Proportion of Missing Data High, Percentage of Missing Data, and Missing Data Treatment

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
50	.60	10%,0%	0.0012	-0.0044	-0.1720	-0.0568	0.1136	-0.2985	-0.1595	-0.0102
	.60	20%,0%	-0.0267	-0.0149	-0.3430	-0.1404	0.2216	-0.5528	-0.3496	-0.0176
	.60	30%,0%	0.0005	0.0145	-0.4704	-0.1749	0.4172	-0.7577	-0.5079	-0.0180
	.60	40%,0%	-0.0223	-0.0044	-0.6214	-0.2630	0.5878	-0.9338	-0.6869	-0.0272
	.60	50%,0%	0.0313	0.1356	-0.6839	-0.2745	0.8829	-1.0613	-0.8026	0.0260
	.60	60%,0%	-0.0985	-0.1456	-0.9517	-0.4406	0.9745	-1.2245	-1.0789	-0.0852
	.70	10%,0%	0.0000[a]							
	.70	20%,0%	-0.0080	-0.0131	-0.3377	-0.1233	0.2387	-0.5294	-0.3496	-0.0316
	.70	30%,0%	-0.0070	-0.0228	-0.4929	-0.1791	0.4096	-0.7757	-0.5105	0.0232
	.70	40%,0%	-0.0431	-0.0885	-0.6643	-0.2723	0.5596	-0.9277	-0.6781	-0.0416
	.70	50%,0%	-0.1064	-0.1629	-0.8285	-0.3909	0.7015	-1.1133	-0.9174	-0.1010
	.70	60%,0%	-0.2444	-0.3544	-1.0307	-0.5754	0.7853	-1.2830	-1.0695	-0.2024
	.80	10%,0%	0.0026	0.0065	-0.1661	-0.0545	0.1169	-0.2899	-0.1781	-0.0106
	.80	20%,0%	0.0059	0.0251	-0.3120	-0.1082	0.2638	-0.5291	-0.3140	0.0419
	.80	30%,0%	-0.0341	-0.0443	-0.5007	-0.2031	0.3717	-0.7903	-0.5333	-0.0391
	.80	40%,0%	-0.0509	-0.1029	-0.6720	-0.2792	0.5582	-0.9583	-0.6945	-0.0547
	.80	50%,0%	-0.2707	-0.4184	-0.9573	-0.5299	0.4825	-1.1988	-0.9860	-0.2289
	.80	60%,0%	-0.5365	-0.7498	-1.1769	-0.8084	0.3357	-1.3907	-1.2023	-0.4654
	.90	10%,0%	0.0000[a]							
	.90	20%,0%	-0.0162	-0.0300	-0.3642	-0.1277	0.2514	-0.5852	-0.3245	-0.0211
	.90	30%,0%	-0.0779	-0.1114	-0.5571	-0.2397	0.3422	-0.8213	-0.5237	-0.0284
	.90	40%,0%	-0.1450	-0.2564	-0.7552	-0.3605	0.4262	-0.9872	-0.7263	-0.0849
	.90	50%,0%	-0.4779	-0.7050	-1.0750	-0.7049	0.1883	-1.2344	-1.0965	-0.4430
	.90	60%,0%	-0.6693	-0.9448	-1.2350	-0.9045	0.0246	-1.4042	-1.2092	-0.5652
50	.60	10%,10%	-0.0276	-0.0010	-0.1257	-0.0715	0.0688	-0.2054	-0.1498	-0.0097
	.60	20%,10%	0.0102	0.0056	-0.2540	-0.1134	0.2180	-0.4110	-0.3006	-0.0391
	.60	20%,20%	0.0192	0.0263	-0.2022	-0.1054	0.2110	-0.3688	-0.2482	0.0421
	.60	30%,20%	0.0759	0.0563	-0.3260	-0.1090	0.4804	-0.5101	-0.4405	0.0496
	.60	40%,20%	-0.1082	-0.0058	-0.5047	-0.3017	0.4553	-0.7507	-0.6444	-0.0749
	.60	30%,30%	-0.0195	0.0131	-0.2916	-0.1518	0.4147	-0.4324	-0.4033	0.0317
	.70	10%,10%	0.0000[a]							
	.70	20%,10%	-0.0196	-0.0098	-0.2780	-0.1164	0.2270	-0.4556	-0.3081	0.0424
	.70	20%,20%	-0.0441	-0.0424	-0.2440	-0.1665	0.1239	-0.3977	-0.3140	-0.0433
	.70	30%,20%	-0.0452	-0.0672	-0.3814	-0.2144	0.3148	-0.5753	-0.4739	-0.0351
	.70	40%,20%	-0.0484	-0.0747	-0.5118	-0.2800	0.5424	-0.7560	-0.5953	-0.0352
	.70	30%,30%	-0.1038	-0.0965	-0.3326	-0.2185	0.2554	-0.5031	-0.4090	-0.0422
	.80	10%,10%	0.0004	0.0096	-0.1103	-0.0521	0.0907	-0.1795	-0.1526	0.0066
	.80	20%,10%	-0.0332	-0.0097	-0.2773	-0.1524	0.1745	-0.4397	-0.3415	-0.0467
	.80	20%,20%	-0.0051	-0.0107	-0.2182	-0.1177	0.1943	-0.3601	-0.2807	0.0000
	.80	30%,20%	-0.0707	-0.0772	-0.3910	-0.2375	0.2887	-0.5579	-0.4493	-0.0633
	.80	40%,20%	-0.0808	-0.1362	-0.5433	-0.3263	0.4422	-0.7481	-0.6321	-0.0250
	.80	30%,30%	-0.0406	-0.1167	-0.3296	-0.1932	0.3173	-0.4785	-0.3871	-0.0252
	.90	10%,10%	0.0000[a]							
	.90	20%,10%	-0.0147	-0.0541	-0.3141	-0.1434	0.1980	-0.4698	-0.3287	-0.0263
	.90	20%,20%	-0.0304	-0.0590	-0.2548	-0.1452	0.1665	-0.3917	-0.2802	-0.0052
	.90	30%,20%	-0.1275	-0.2289	-0.4788	-0.3416	0.1706	-0.6425	-0.5376	-0.1246
	.90	40%,20%	-0.1347	-0.2766	-0.6011	-0.4046	0.3446	-0.8020	-0.7091	-0.1223
	.90	30%,30%	-0.1005	-0.2021	-0.3868	-0.2528	0.2170	-0.5601	-0.4502	-0.0826

(Table continues)

Table 6 (continued).

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
	.60	10%,0%	-0.0024	-0.0079	-0.3493	-0.1208	0.2364	-0.6485	-0.3426	-0.0118
	.60	20%,0%	-0.0150	-0.0178	-0.6859	-0.2554	0.5237	-1.1479	-0.7291	0.0159
	.60	30%,0%	-0.0060	-0.0103	-0.9836	-0.3761	0.8989	-1.5784	-1.0902	-0.0142
	.60	40%,0%	-0.0386	-0.0511	-1.3066	-0.5348	1.3247	-1.9465	-1.4660	-0.0244
	.60	50%,0%	-0.0449	-0.0861	-1.5866	-0.6788	1.8978	-2.2854	-1.7637	-0.0132
	.60	60%,0%	-0.0667	-0.1007	-1.8626	-0.8456	2.6981	-2.5775	-2.1450	-0.0451
	.70	10%,0%	0.0082	0.0090	-0.3416	-0.1105	0.2580	-0.6381	-0.3466	0.0181
	.70	20%,0%	-0.0168	-0.0356	-0.6924	-0.2575	0.5126	-1.1663	-0.7034	-0.0366
	.70	30%,0%	-0.0532	-0.0742	-1.0205	-0.4155	0.8363	-1.5909	-1.1172	-0.0291
	.70	40%,0%	-0.0691	-0.1241	-1.3360	-0.5598	1.2866	-1.9270	-1.4423	-0.0455
	.70	50%,0%	-0.2148	-0.3498	-1.7047	-0.8221	1.6594	-2.3532	-1.8427	-0.1771
200	.70	60%,0%	-0.4562	-0.6802	-2.0781	-1.1678	2.0551	-2.6978	-2.3204	-0.3695
	.80	10%,0%	-0.0081	-0.0141	-0.3592	-0.1268	0.2372	-0.6448	-0.3481	0.0021
	.80	20%,0%	-0.0295	-0.0465	-0.7059	-0.2692	0.5049	-1.1889	-0.7335	-0.0327
	.80	30%,0%	-0.0651	-0.1095	-1.0481	-0.4266	0.8465	-1.6300	-1.1141	-0.0289
	.80	40%,0%	-0.1934	-0.3260	-1.4417	-0.6676	1.1220	-2.0502	-1.5470	-0.1539
	.80	50%,0%	-0.4498	-0.7127	-1.8801	-1.0156	1.3559	-2.4127	-1.9950	-0.3324
	.80	60%,0%	-1.1026	-1.5660	-2.4345	-1.7067	0.9128	-2.8839	-2.5623	-0.9232
	.90	10%,0%	-0.0239	-0.0309	-0.3753	-0.1414	0.2202	-0.6274	-0.3832	-0.0369
	.90	20%,0%	-0.0533	-0.0810	-0.7399	-0.2892	0.4990	-1.2032	-0.7708	-0.0346
	.90	30%,0%	-0.1399	-0.2739	-1.1567	-0.4894	0.7471	-1.7049	-1.1481	-0.1237
	.90	40%,0%	-0.3769	-0.6126	-1.6041	-0.8239	0.9016	-2.1277	-1.5764	-0.2992
	.90	50%,0%	-0.8213	-1.2375	-2.1183	-1.3391	0.7883	-2.5702	-2.1538	-0.7129
	.90	60%,0%	-1.4187	-1.9514	-2.5564	-1.9316	0.1251	-2.9109	-2.5651	-1.2907
	.60	10%,10%	-0.0076	-0.0273	-0.2662	-0.1427	0.1588	-0.4268	-0.2954	-0.0461
	.60	20%,10%	-0.0691	-0.0201	-0.5707	-0.2865	0.4356	-0.9292	-0.6667	-0.0334
	.60	20%,20%	-0.0401	0.0107	-0.4282	-0.2612	0.3935	-0.6895	-0.5788	-0.0324
	.60	30%,20%	-0.1472	-0.3800	-0.8904	-0.5913	0.5228	-1.2245	-1.0570	-0.1864
	.60	40%,20%	-0.0463	-0.0871	-0.9904	-0.5830	1.2253	-1.5132	-1.3127	-0.0534
	.60	30%,30%	-0.0350	-0.0326	-0.6026	-0.3983	0.7406	-0.9836	-0.8860	0.0107
	.70	10%,10%	-0.0116	0.0104	-0.2291	-0.1095	0.1907	-0.4174	-0.2824	-0.0011
	.70	20%,10%	-0.0254	-0.0340	-0.5706	-0.2732	0.4555	-0.9354	-0.6628	-0.0285
	.70	20%,20%	-0.0398	-0.0449	-0.4555	-0.2900	0.3424	-0.7431	-0.6232	-0.0591
	.70	30%,20%	-0.0242	-0.0973	-0.7573	-0.4103	0.7805	-1.1709	-0.9298	-0.0051
	.70	40%,20%	-0.1286	-0.1618	-1.0144	-0.6027	1.1845	-1.5522	-1.2859	-0.0497
200	.70	30%,30%	-0.0376	-0.1142	-0.6147	-0.4193	0.6948	-0.9938	-0.8498	-0.0062
	.80	10%,10%	-0.0045	-0.0089	-0.2493	-0.1222	0.1807	-0.4264	-0.2956	0.0201
	.80	20%,10%	-0.0173	-0.0393	-0.5732	-0.2610	0.4621	-0.9272	-0.6617	-0.0262
	.80	20%,20%	-0.0648	-0.0647	-0.4747	-0.2986	0.3614	-0.7499	-0.5965	-0.0215
	.80	30%,20%	-0.1105	-0.1964	-0.7899	-0.5016	0.6142	-1.1580	-0.9883	-0.1073
	.80	40%,20%	-0.2315	-0.4575	-1.1488	-0.8111	0.8179	-1.6120	-1.4101	-0.2634
	.80	30%,30%	-0.1317	-0.2697	-0.6731	-0.4963	0.5411	-1.0438	-0.9152	-0.0878
	.90	10%,10%	-0.0099	-0.0201	-0.2517	-0.1332	0.1634	-0.4268	-0.3138	-0.0124
	.90	20%,10%	-0.0178	-0.0709	-0.5915	-0.2690	0.4517	-0.9443	-0.6785	-0.0211
	.90	20%,20%	-0.1087	-0.1995	-0.5628	-0.3578	0.2949	-0.8185	-0.6421	-0.0986
	.90	30%,20%	-0.1472	-0.3800	-0.8904	-0.5913	0.5228	-1.2245	-1.0570	-0.1864
	.90	40%,20%	-0.3494	-0.6288	-1.2274	-0.8975	0.6420	-1.6996	-1.4292	-0.3022
	.90	30%,30%	-0.2251	-0.3989	-0.7368	-0.5438	0.3995	-1.0763	-0.9203	-0.1514

Note. L: listwise deletion, P: pairwise deletion, MS: mean substitution, SR: simple regression, MR: multiple regression, SMS: stochastic mean substitution, SSR: stochastic simple regression, SMR: stochastic multiple regression. [a] Data were not computed for this combination of sample size, proportion of missing data high, and percentage of missing data.

Table 7. Effect Sizes of the First Standardized Regression Coefficient (X_1) for the Grade Nine Data by Sample Size, Proportion of Missing Data High, Percentage of Missing Data, and Missing Data Treatment

Sample	Proportion	Percentage of								
Size	High	Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
	.60	10%,0%	0.0131	0.0138	-0.2761	-0.0435	0.1677	-0.4660	-0.2285	0.0260
	.60	20%,0%	-0.0069	-0.0002	-0.5497	-0.1125	0.3482	-0.8228	-0.4123	-0.0031
	.60	30%,0%	-0.0266	-0.0102	-0.7773	-0.1911	0.5422	-1.1236	-0.6547	-0.0279
	.60	40%,0%	-0.0239	0.0465	-0.9492	-0.2495	0.8340	-1.3272	-0.8824	0.0273
	.60	50%,0%	-0.0388	-0.0085	-1.1361	-0.3244	1.1149	-1.4995	-1.0769	-0.0604
	.60	60%,0%	-0.0723	-0.0380	-1.3218	-0.4155	1.5024	-1.6637	-1.2814	-0.0526
	.70	10%,0%	0.0000[a]							
	.70	20%,0%	0.0284	0.0304	-0.5080	-0.0831	0.3717	-0.8079	-0.4053	0.0392
	.70	30%,0%	-0.0196	-0.0188	-0.7966	-0.1900	0.5558	-1.1378	-0.6636	-0.0034
	.70	40%,0%	-0.0446	-0.0553	-1.0025	-0.2595	0.8175	-1.3677	-0.8424	-0.0025
	.70	50%,0%	-0.1664	-0.2596	-1.2420	-0.4201	1.0326	-1.5566	-1.1250	-0.0438
50	.70	60%,0%	-0.3181	-0.5020	-1.4609	-0.6084	1.1746	-1.7353	-1.3966	-0.1958
	.80	10%,0%	-0.0061	-0.0009	-0.2921	-0.0618	0.1515	-0.5015	-0.2177	0.0050
	.80	20%,0%	-0.0220	0.0004	-0.5428	-0.1302	0.3285	-0.8282	-0.4162	-0.0089
	.80	30%,0%	-0.0564	-0.0927	-0.8301	-0.2136	0.5299	-1.1139	-0.6663	-0.0264
	.80	40%,0%	-0.1076	-0.1508	-1.0368	-0.3080	0.7639	-1.3904	-0.8640	-0.0654
	.80	50%,0%	-0.2864	-0.4574	-1.3029	-0.5171	0.8512	-1.5948	-1.1485	-0.1718
	.80	60%,0%	-0.6482	-1.0008	-1.5957	-0.8856	0.6884	-1.7984	-1.5403	-0.4545
	.90	10%,0%	0.0000[a]							
	.90	20%,0%	-0.0391	-0.0525	-0.5862	-0.1436	0.3034	-0.8684	-0.4290	-0.0420
	.90	30%,0%	-0.0843	-0.1549	-0.8566	-0.2322	0.5082	-1.1624	-0.6265	-0.0387
	.90	40%,0%	-0.2574	-0.4258	-1.1611	-0.4290	0.5785	-1.4491	-0.9656	-0.1694
	.90	50%,0%	-0.6277	-0.9750	-1.5154	-0.8141	0.4012	-1.7413	-1.3351	-0.5021
	.90	60%,0%	-0.9072	-1.3876	-1.7052	-1.0967	0.1642	-1.8526	-1.5675	-0.7215
	.60	10%,10%	-0.0167	-0.0111	0.0071	-0.0830	-0.0321	-0.0368	-0.0657	-0.0329
	.60	20%,10%	-0.0585	-0.0176	-0.2451	-0.1419	0.1405	-0.4377	-0.2721	-0.0332
	.60	20%,20%	0.0263	0.0393	0.0736	-0.0937	-0.0012	0.0328	-0.0739	0.0004
	.60	30%,20%	0.0056	-0.0501	-0.1705	-0.1944	0.1610	-0.3495	-0.3575	0.0008
	.60	40%,20%	0.0014	-0.0246	-0.3223	-0.2321	0.4735	-0.6160	-0.5426	0.0020
	.60	30%,30%	-0.0209	-0.0561	0.0781	-0.1236	0.0127	-0.0960	-0.1276	0.0276
	.70	10%,10%	0.0000 [a]							
	.70	20%,10%	0.0142	0.0119	-0.2160	-0.0875	0.2001	-0.3841	-0.2726	0.0106
	.70	20%,20%	-0.0096	0.0018	0.0312	-0.0960	0.0115	-0.0434	-0.1024	0.0099
	.70	30%,20%	0.0104	-0.0478	-0.1809	-0.1724	0.2233	-0.3746	-0.3020	0.0210
	.70	40%,20%	-0.0260	-0.0540	-0.3364	-0.2202	0.4652	-0.6414	-0.5655	0.0335
50	.70	30%,30%	-0.0120	0.0145	0.1166	-0.0389	0.1088	-0.0751	-0.0763	0.0631
	.80	10%,10%	-0.0182	-0.0037	0.0118	-0.0627	-0.0109	0.0019	-0.0481	0.0107
	.80	20%,10%	-0.0018	-0.0455	-0.2525	-0.1097	0.1746	-0.3834	-0.2408	0.0205
	.80	20%,20%	-0.0245	-0.0246	0.0214	-0.1025	0.0049	-0.0211	-0.1372	0.0168
	.80	30%,20%	-0.0696	-0.0678	-0.1855	-0.1777	0.2010	-0.3594	-0.3051	0.0040
	.80	40%,20%	-0.2091	-0.2155	-0.4195	-0.2713	0.3371	-0.7196	-0.5772	-0.1347
	.80	30%,30%	0.0052	0.0504	0.1140	0.0963	0.2580	-0.0983	-0.0281	0.1995
	.90	10%,10%	0.0000 [a]							
	.90	20%,10%	-0.0293	-0.0121	-0.2421	-0.1208	0.1769	-0.4030	-0.2358	0.0065
	.90	20%,20%	-0.0740	-0.0548	0.0054	-0.1116	-0.0186	-0.0571	-0.1310	-0.0056
	.90	30%,20%	-0.1215	-0.1658	-0.2454	-0.1790	0.2026	-0.4005	-0.3605	-0.0507
	.90	40%,20%	-0.2424	-0.2973	-0.5198	-0.2412	0.3572	-0.7747	-0.5342	-0.0876
	.90	30%,30%	-0.1600	0.0632	0.0476	0.2112	0.4088	-0.1360	0.0208	0.2139

(Table continues)

Table 7 (continued).

Sample Size	Proportion High	Percentage of Missing Data	L	P	MS	SR	MR	SMS	SSR	SMR
200	.60	10%,0%	-0.0059	0.0058	-0.5746	-0.1129	0.3145	-0.9948	-0.4220	-0.0079
	.60	20%,0%	-0.0050	-0.0077	-1.0785	-0.2210	0.6971	-1.7258	-0.8331	-0.0077
	.60	30%,0%	-0.0014	0.0017	-1.4967	-0.3332	1.1722	-2.1964	-1.2497	0.0141
	.60	40%,0%	-0.0025	-0.0054	-1.8614	-0.4439	1.7640	-2.6200	-1.6561	-0.0088
	.60	50%,0%	-0.0570	-0.0904	-2.2536	-0.6120	2.4624	-2.9886	-2.1129	-0.0478
	.60	60%,0%	-0.1967	-0.2914	-2.6257	-0.8497	3.2920	-3.3139	-2.5705	-0.1423
	.70	10%,0%	-0.0085	-0.0073	-0.5847	-0.1126	0.3132	-1.0183	-0.4299	0.0184
	.70	20%,0%	-0.0104	-0.0209	-1.0906	-0.2245	0.6951	-1.7000	-0.8298	0.0310
	.70	30%,0%	-0.0392	-0.0592	-1.5230	-0.3612	1.1261	-2.2686	-1.3120	-0.0329
	.70	40%,0%	-0.0872	-0.1358	-1.9199	-0.5111	1.6731	-2.6624	-1.7214	-0.0254
	.70	50%,0%	-0.1748	-0.3873	-2.3535	-0.6947	2.3401	-3.0534	-2.1461	-0.1409
	.70	60%,0%	-0.6356	-1.0727	-2.8479	-1.1908	2.6882	-3.4274	-2.6916	-0.4917
	.80	10%,0%	0.0007	-0.0031	-0.5689	-0.1066	0.3104	-0.9997	-0.3974	0.0024
	.80	20%,0%	-0.0531	-0.0663	-1.1196	-0.2635	0.6511	-1.7626	-0.8812	-0.0332
	.80	30%,0%	-0.1036	-0.1754	-1.5988	-0.4110	1.0708	-2.3388	-1.3308	-0.0744
	.80	40%,0%	-0.2615	-0.4227	-2.0909	-0.6512	1.5276	-2.7648	-1.8055	-0.1649
	.80	50%,0%	-0.5722	-0.9932	-2.5808	-1.0196	1.8545	-3.1429	-2.3695	-0.4254
	.80	60%,0%	-1.3159	-2.1213	-3.1616	-1.7865	1.5394	-3.5579	-2.9700	-1.1140
	.90	10%,0%	-0.0192	-0.0389	-0.6123	-0.1235	0.2950	-1.0266	-0.4517	-0.0291
	.90	20%,0%	-0.0558	-0.1307	-1.1708	-0.2622	0.6396	-1.7808	-0.8723	-0.0313
	.90	30%,0%	-0.1541	-0.3028	-1.6812	-0.4483	1.0310	-2.3726	-1.3366	-0.1400
	.90	40%,0%	-0.4150	-0.7812	-2.2408	-0.7720	1.3360	-2.8235	-1.8840	-0.2634
	.90	50%,0%	-1.0185	-1.7119	-2.8500	-1.4062	1.2405	-3.3215	-2.5826	-0.7847
	.90	60%,0%	-1.6916	-2.5967	-3.2668	-2.0845	0.6154	-3.6220	-3.0131	-1.3553
200	.60	10%,10%	-0.0153	-0.0192	0.0140	-0.1286	-0.0303	-0.0150	-0.1199	-0.0088
	.60	20%,10%	-0.0062	-0.0183	-0.4451	-0.2292	0.3439	-0.7511	-0.5325	-0.0276
	.60	20%,20%	0.0450	0.0401	0.1221	-0.1779	0.0079	-0.0008	-0.1831	0.0672
	.60	30%,20%	-0.0303	-0.0489	-0.3077	-0.3702	0.3935	-0.7175	-0.6172	-0.0498
	.60	40%,20%	-0.0317	-0.0111	-0.5894	-0.4639	1.0407	-1.1742	-0.9943	0.0050
	.60	30%,30%	-0.0600	0.0352	0.2277	-0.2829	-0.0416	-0.0990	-0.3074	0.0165
	.70	10%,10%	-0.0267	0.0099	0.0568	-0.1101	-0.0225	0.0097	-0.0922	-0.0136
	.70	20%,10%	-0.0655	-0.0593	-0.4760	-0.2643	0.3126	-0.7974	-0.5326	-0.0478
	.70	20%,20%	0.0127	-0.0364	0.0758	-0.2150	-0.0361	-0.0562	-0.1661	0.0105
	.70	30%,20%	-0.0727	-0.1136	-0.3258	-0.3511	0.3846	-0.7030	-0.5781	-0.0514
	.70	40%,20%	-0.0533	-0.2082	-0.6500	-0.4562	0.9607	-1.2431	-1.0012	-0.0477
	.70	30%,30%	-0.0469	-0.0687	0.1982	-0.2206	0.0112	-0.1498	-0.2307	0.0293
	.80	10%,10%	-0.0195	-0.0077	0.0368	-0.1232	-0.0316	0.0223	-0.1156	0.0090
	.80	20%,10%	-0.0304	-0.1077	-0.5108	-0.2581	0.3194	-0.8438	-0.5193	-0.0096
	.80	20%,20%	-0.0510	-0.0602	0.0617	-0.2113	-0.0326	-0.0721	-0.1423	0.0122
	.80	30%,20%	-0.1112	-0.2063	-0.3669	-0.3324	0.3808	-0.7406	-0.6186	-0.0639
	.80	40%,20%	-0.2377	-0.3691	-0.7356	-0.4675	0.8549	-1.3129	-1.0400	-0.0818
	.80	30%,30%	-0.1329	-0.0045	0.2236	0.0361	0.2921	-0.1193	-0.1260	0.1663
	.90	10%,10%	-0.0213	-0.0384	0.0092	-0.1194	-0.0329	-0.0180	-0.0835	-0.0058
	.90	20%,10%	-0.0778	-0.1175	-0.5054	-0.2746	0.2986	-0.8310	-0.5356	-0.0591
	.90	20%,20%	-0.0382	-0.0888	0.0582	-0.1679	0.0074	-0.0578	-0.1896	0.0381
	.90	30%,20%	-0.1486	-0.2569	-0.3614	-0.2784	0.3943	-0.7047	-0.6067	-0.0091
	.90	40%,20%	-0.4503	-0.4708	-0.8584	-0.4037	0.7856	-1.3911	-0.9864	-0.1467
	.90	30%,30%	-0.1995	0.1407	0.2294	0.4789	0.8062	-0.1592	0.1240	0.4457

Note. L: listwise deletion, P: pairwise deletion, MS: mean substitution, SR: simple regression, MR: multiple regression, SMS: stochastic mean substitution, SSR: stochastic simple regression, SMR: stochastic multiple regression. [a] Data were not computed for this combination of sample size, proportion of missing data high, and percentage of missing data.

Table 8. Effect Sizes of the Second Standardized Regression Coefficient (X_2) for the Grade Six Data by Sample Size, Proportion of Missing Data High, Percentage of Missing Data, and Missing Data Treatment

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
50	.60	10%,0%	-0.0139	0.0017	0.1389	0.0020	-0.0924	0.2376	0.0862	0.0042
	.60	20%,0%	0.0091	0.0024	0.2698	0.0125	-0.1937	0.4280	0.1902	0.0114
	.60	30%,0%	-0.0059	-0.0212	0.3694	-0.0090	-0.3561	0.5677	0.2726	0.0079
	.60	40%,0%	0.0105	-0.0035	0.4800	0.0172	-0.4947	0.6770	0.3788	0.0348
	.60	50%,0%	0.0471	-0.0904	0.5504	-0.0213	-0.7293	0.7632	0.4405	-0.0038
	.60	60%,0%	-0.0507	0.0758	0.6927	0.0485	-0.8601	0.8292	0.5919	0.0451
	.70	10%,0%	0.0000[a]							
	.70	20%,0%	-0.0088	0.0045	0.2687	0.0038	-0.2021	0.4175	0.1898	0.0152
	.70	30%,0%	-0.0340	0.0059	0.3859	-0.0030	-0.3524	0.5842	0.2797	-0.0204
	.70	40%,0%	-0.0263	0.0583	0.5147	0.0266	-0.4751	0.6878	0.3795	0.0352
	.70	50%,0%	-0.0484	0.1251	0.6321	0.0848	-0.5864	0.7874	0.5304	0.0880
	.70	60%,0%	-0.0501	0.2589	0.7420	0.1897	-0.6592	0.8533	0.6117	0.1746
	.80	10%,0%	-0.0078	-0.0081	0.1344	-0.0042	-0.0984	0.2389	0.0955	0.0020
	.80	20%,0%	0.0049	-0.0196	0.2571	-0.0076	-0.2187	0.4168	0.1664	-0.0263
	.80	30%,0%	0.0054	0.0264	0.3977	0.0178	-0.3194	0.5909	0.2901	0.0234
	.80	40%,0%	-0.0399	0.0708	0.5208	0.0352	-0.4686	0.7045	0.3900	0.0450
	.80	50%,0%	-0.0317	0.2851	0.6885	0.1891	-0.4293	0.8092	0.5714	0.1674
	.80	60%,0%	-0.0419	0.5436	0.8148	0.3989	-0.2797	0.8846	0.7168	0.3616
	.90	10%,0%	0.0000[a]							
	.90	20%,0%	-0.0215	0.0134	0.2845	0.0013	-0.2156	0.4408	0.1699	0.0011
	.90	30%,0%	-0.0137	0.0797	0.4377	0.0511	-0.2892	0.6114	0.2960	0.0242
	.90	40%,0%	-0.0518	0.1852	0.5749	0.1078	-0.3594	0.7208	0.4136	0.0786
	.90	50%,0%	-0.0414	0.4991	0.7591	0.3482	-0.1705	0.8405	0.6465	0.3243
	.90	60%,0%	-0.0178	0.6760	0.8450	0.4993	-0.0250	0.8909	0.7314	0.4388
50	.60	10%,10%	0.0063	-0.0057	0.0924	-0.0071	-0.0509	0.1207	0.0553	0.0042
	.60	20%,10%	-0.0067	-0.0061	0.2052	-0.0225	-0.1777	0.2791	0.1318	0.0231
	.60	20%,20%	-0.0022	-0.0102	0.1756	-0.0366	-0.1567	0.2245	0.0612	-0.0290
	.60	30%,20%	-0.0968	-0.0466	0.2779	-0.0928	-0.4033	0.3109	0.1596	-0.0547
	.60	40%,20%	0.0452	0.0210	0.4346	0.0296	-0.3726	0.4698	0.2932	0.0628
	.60	30%,30%	-0.0715	0.0208	0.2786	-0.0608	-0.3405	0.2387	0.1258	-0.0199
	.70	10%,10%	0.0000 [a]							
	.70	20%,10%	-0.0267	-0.0016	0.2179	-0.0197	-0.1930	0.3086	0.1312	-0.0307
	.70	20%,20%	0.0557	0.0558	0.2177	0.0188	-0.0735	0.2744	0.1299	0.0510
	.70	30%,20%	-0.0281	0.0800	0.3423	0.0154	-0.2480	0.3549	0.2085	0.0395
	.70	40%,20%	0.0131	0.1266	0.4754	0.0308	-0.4221	0.4777	0.2893	0.0616
	.70	30%,30%	0.0405	0.1809	0.3519	0.0259	-0.1465	0.2947	0.1626	0.0706
	.80	10%,10%	-0.0183	-0.0190	0.0747	-0.0284	-0.0809	0.1129	0.0546	-0.0166
	.80	20%,10%	0.0559	0.0224	0.2386	0.0198	-0.1301	0.3313	0.1735	0.0539
	.80	20%,20%	0.0003	0.0413	0.2106	-0.0082	-0.1246	0.2443	0.1027	0.0219
	.80	30%,20%	0.0226	0.1217	0.3707	0.0461	-0.2087	0.4052	0.2033	0.0670
	.80	40%,20%	-0.0400	0.2192	0.5177	0.1006	-0.3357	0.5238	0.3402	0.0631
	.80	30%,30%	-0.0713	0.2523	0.3741	0.0302	-0.1940	0.3293	0.1506	0.0543
	.90	10%,10%	0.0000 [a]							
	.90	20%,10%	-0.0109	0.0578	0.2689	0.0117	-0.1549	0.3491	0.1642	0.0228
	.90	20%,20%	0.0068	0.0839	0.2430	0.0018	-0.1252	0.2691	0.0987	0.0025
	.90	30%,20%	0.0536	0.2924	0.4704	0.1596	-0.0854	0.4643	0.2868	0.1406
	.90	40%,20%	-0.0325	0.4460	0.6205	0.2296	-0.1736	0.5922	0.4434	0.1641
	.90	30%,30%	-0.0160	0.4876	0.4949	0.1568	-0.0452	0.4475	0.2599	0.1542

(Table continues)

Table 8 (continued).

Sample Size	Proportion High	Percentage of Missing Data	L	P	MS	SR	MR	SMS	SSR	SMR
200	.60	10%,0%	-0.0028	0.0046	0.2971	0.0007	-0.2066	0.5362	0.1954	0.0067
	.60	20%,0%	-0.0052	0.0070	0.5678	0.0050	-0.4635	0.9275	0.4176	-0.0106
	.60	30%,0%	-0.0115	0.0059	0.8084	0.0044	-0.7889	1.2448	0.6226	0.0019
	.60	40%,0%	-0.0091	0.0300	1.0467	0.0189	-1.1847	1.4824	0.8335	0.0184
	.60	50%,0%	-0.0078	0.0698	1.2588	0.0343	-1.6969	1.6760	1.0200	0.0188
	.60	60%,0%	0.0013	0.0938	1.4472	0.0602	-2.4132	1.8178	1.2420	0.0667
	.70	10%,0%	-0.0114	-0.0056	0.2923	-0.0055	-0.2195	0.5364	0.1988	-0.0116
	.70	20%,0%	-0.0027	0.0257	0.5784	0.0129	-0.4488	0.9440	0.4052	0.0237
	.70	30%,0%	0.0159	0.0544	0.8387	0.0355	-0.7412	1.2591	0.6467	0.0224
	.70	40%,0%	-0.0244	0.0986	1.0792	0.0533	-1.1363	1.4888	0.8392	0.0458
	.70	50%,0%	-0.0002	0.2752	1.3371	0.1671	-1.4808	1.7161	1.0843	0.1586
	.70	60%,0%	0.0289	0.5387	1.5737	0.3601	-1.8281	1.8576	1.3785	0.3252
	.80	10%,0%	-0.0039	0.0085	0.3019	0.0048	-0.2058	0.5323	0.1973	-0.0002
	.80	20%,0%	-0.0033	0.0291	0.5831	0.0180	-0.4472	0.9576	0.4199	0.0191
	.80	30%,0%	-0.0030	0.0943	0.8666	0.0569	-0.7286	1.2774	0.6556	0.0433
	.80	40%,0%	-0.0054	0.2540	1.1536	0.1508	-0.9928	1.5423	0.9177	0.1348
	.80	50%,0%	-0.0067	0.5399	1.4415	0.3360	-1.2168	1.7501	1.1931	0.2804
	.80	60%,0%	0.0183	1.1692	1.7557	0.8534	-0.8070	1.9244	1.5678	0.7728
	.90	10%,0%	0.0060	0.0196	0.3131	0.0139	-0.1958	0.5278	0.2180	0.0199
	.90	20%,0%	0.0011	0.0646	0.6169	0.0407	-0.4314	0.9768	0.4507	0.0332
	.90	30%,0%	-0.0523	0.2013	0.9323	0.1029	-0.6638	1.3189	0.6814	0.0989
	.90	40%,0%	-0.0034	0.4658	1.2585	0.2846	-0.8040	1.5911	0.9534	0.2516
	.90	50%,0%	0.0013	0.9400	1.5926	0.6396	-0.6905	1.8147	1.3333	0.5924
	.90	60%,0%	-0.0053	1.4292	1.8129	1.0991	-0.1098	1.9343	1.6115	1.0380
200	.60	10%,10%	0.0240	0.0363	0.2214	-0.0134	-0.1127	0.3119	0.0934	0.0496
	.60	20%,10%	0.0369	0.0041	0.4605	-0.0114	-0.3777	0.6491	0.3036	0.0179
	.60	20%,20%	0.0270	-0.0152	0.3501	-0.0684	-0.3185	0.4139	0.1773	0.0215
	.60	30%,20%	0.0234	0.0224	0.6316	-0.0647	-0.6901	0.7296	0.3816	0.0335
	.60	40%,20%	0.0770	0.1036	0.8961	0.0012	-1.0638	0.9816	0.6257	0.0605
	.60	30%,30%	-0.0051	0.0406	0.5648	-0.1027	-0.6371	0.5013	0.2709	-0.0231
	.70	10%,10%	-0.0233	-0.0180	0.1681	-0.0497	-0.1506	0.2663	0.0881	-0.0008
	.70	20%,10%	0.0146	0.0366	0.4772	-0.0142	-0.3812	0.6705	0.3065	0.0263
	.70	20%,20%	0.0500	0.0631	0.4074	-0.0332	-0.2642	0.4553	0.2353	0.0618
	.70	30%,20%	-0.0505	0.1075	0.6758	-0.0392	-0.6798	0.7631	0.3959	-0.0001
	.70	40%,20%	-0.0493	0.2037	0.9315	0.0348	-1.0381	0.9864	0.6108	0.0322
	.70	30%,30%	0.0123	0.2103	0.6536	-0.0440	-0.5581	0.5467	0.2868	0.0328
	.80	10%,10%	-0.0116	0.0103	0.2009	-0.0331	-0.1386	0.2943	0.1004	-0.0245
	.80	20%,10%	-0.0038	0.0524	0.4978	-0.0093	-0.3773	0.6942	0.3241	0.0313
	.80	20%,20%	0.0393	0.1005	0.4409	-0.0165	-0.2745	0.4890	0.2116	0.0424
	.80	30%,20%	0.0433	0.2726	0.7761	0.0899	-0.4809	0.8178	0.4760	0.1319
	.80	40%,20%	0.0884	0.5830	1.1276	0.2856	-0.6430	1.1165	0.7532	0.2661
	.80	30%,30%	0.0201	0.5173	0.8051	0.0986	-0.3557	0.6931	0.4006	0.1534
	.90	10%,10%	0.0067	0.0337	0.2162	-0.0184	-0.1151	0.2998	0.1203	0.0143
	.90	20%,10%	-0.0186	0.0949	0.5313	0.0093	-0.3598	0.7144	0.3503	0.0371
	.90	20%,20%	-0.0399	0.2331	0.5309	0.0394	-0.2268	0.5420	0.2459	0.0832
	.90	30%,20%	-0.0064	0.5093	0.9166	0.2011	-0.3722	0.9053	0.5646	0.2111
	.90	40%,20%	0.0128	0.9071	1.2911	0.4507	-0.4007	1.2430	0.8498	0.3781
	.90	30%,30%	-0.0405	0.8685	0.9817	0.2386	-0.1616	0.7968	0.4493	0.2552

Note. L: listwise deletion, P: pairwise deletion, MS: mean substitution, SR: simple regression, MR: multiple regression, SMS: stochastic mean substitution, SSR: stochastic simple regression, SMR: stochastic multiple regression. [a] Data were not computed for this combination of sample size, proportion of missing data high, and percentage of missing data.

Table 9. Effect Sizes of the Second Standardized Regression Coefficient (X_2) for the Grade Nine Data by Sample Size, Proportion of Missing Data High, Percentage of Missing Data, and Missing Data Treatment

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
50	.60	10%,0%	-0.0102	-0.0109	0.2560	-0.0081	-0.1518	0.4244	0.1564	-0.0170
	.60	20%,0%	-0.0046	-0.0029	0.5015	-0.0036	-0.3235	0.7445	0.2811	-0.0001
	.60	30%,0%	0.0093	0.0039	0.6975	0.0138	-0.5094	0.9792	0.4478	0.0209
	.60	40%,0%	0.0122	-0.0446	0.8459	0.0153	-0.7739	1.1385	0.6058	-0.0124
	.60	50%,0%	0.0162	-0.0072	0.9927	0.0249	-1.0548	1.2406	0.7405	0.0389
	.60	60%,0%	0.0095	0.0210	1.1344	0.0558	-1.4003	1.3468	0.8791	0.0613
	.70	10%,0%	0.0000[a]							
	.70	20%,0%	-0.0152	-0.0230	0.4730	-0.0224	-0.3374	0.7366	0.2783	-0.0264
	.70	30%,0%	-0.0320	0.0104	0.7058	0.0179	-0.5155	0.9848	0.4565	0.0010
	.70	40%,0%	-0.0448	0.0473	0.8866	0.0279	-0.7585	1.1552	0.5805	0.0145
	.70	50%,0%	-0.0257	0.2213	1.0749	0.1233	-0.9571	1.2768	0.7852	0.0714
	.70	60%,0%	-0.0555	0.4345	1.2301	0.2503	-1.0851	1.3780	0.9843	0.2012
	.80	10%,0%	0.0038	-0.0011	0.2666	0.0040	-0.1408	0.4531	0.1468	-0.0049
	.80	20%,0%	0.0263	0.0029	0.5022	0.0175	-0.3002	0.7523	0.2875	0.0110
	.80	30%,0%	-0.0017	0.0850	0.7486	0.0448	-0.4822	0.9885	0.4642	0.0306
	.80	40%,0%	-0.0041	0.1322	0.9227	0.0741	-0.7058	1.1752	0.6053	0.0596
	.80	50%,0%	0.0088	0.3904	1.1232	0.2164	-0.7886	1.3087	0.8088	0.1736
	.80	60%,0%	-0.0147	0.8394	1.3081	0.5185	-0.6355	1.4013	1.1093	0.4172
	.90	10%,0%	0.0000[a]							
	.90	20%,0%	0.0031	0.0403	0.5292	0.0261	-0.2857	0.7767	0.2961	0.0318
	.90	30%,0%	-0.0124	0.1487	0.7818	0.0710	-0.4527	1.0299	0.4460	0.0547
	.90	40%,0%	0.0040	0.3548	1.0091	0.1819	-0.5471	1.2124	0.6851	0.1501
	.90	50%,0%	-0.0048	0.8044	1.2515	0.4952	-0.3803	1.3649	0.9722	0.4345
	.90	60%,0%	0.0178	1.1298	1.3618	0.7310	-0.1692	1.4149	1.1406	0.6196
50	.60	10%,10%	0.0346	0.0143	-0.0246	-0.0363	0.0483	-0.0481	-0.0769	0.0310
	.60	20%,10%	0.0220	0.0172	0.2056	-0.0317	-0.1093	0.2673	0.0707	0.0358
	.60	20%,20%	-0.0151	-0.0270	-0.0735	-0.1452	0.0410	-0.1932	-0.1992	0.0050
	.60	30%,20%	0.0232	0.0721	0.2156	-0.0910	-0.1092	0.1526	0.0295	0.0149
	.60	40%,20%	-0.0249	0.0740	0.4108	-0.0985	-0.3910	0.3248	0.1587	0.0184
	.60	30%,30%	-0.0786	0.0878	0.0050	-0.2333	0.0325	-0.2048	-0.2826	-0.0337
	.70	10%,10%	0.0000 [a]							
	.70	20%,10%	-0.0422	-0.0104	0.1876	-0.0835	-0.1681	0.2354	0.0639	-0.0074
	.70	20%,20%	-0.0154	0.0043	-0.0258	-0.1455	0.0162	-0.1406	-0.1743	-0.0119
	.70	30%,20%	-0.0254	0.0877	0.2441	-0.1154	-0.1753	0.1605	-0.0353	-0.0044
	.70	40%,20%	-0.0093	0.1543	0.4558	-0.1055	-0.3841	0.3615	0.1710	-0.0216
	.70	30%,30%	-0.0420	0.1554	0.0704	-0.2799	-0.0285	-0.1726	-0.2933	-0.0404
	.80	10%,10%	0.0085	0.0104	-0.0278	-0.0501	0.0348	-0.0711	-0.0774	-0.0012
	.80	20%,10%	-0.0205	0.0550	0.2432	-0.0559	-0.1371	0.2617	0.0415	-0.0108
	.80	20%,20%	-0.0026	0.0563	0.0128	-0.1300	0.0317	-0.1114	-0.1463	-0.0068
	.80	30%,20%	0.0069	0.1437	0.2730	-0.1008	-0.1446	0.1752	-0.0228	-0.0057
	.80	40%,20%	0.0570	0.3941	0.5727	-0.0330	-0.2679	0.4610	0.2017	0.1275
	.80	30%,30%	-0.0576	0.2376	0.1389	-0.3629	-0.1545	-0.0977	-0.3082	-0.1621
	.90	10%,10%	0.0000 [a]							
	.90	20%,10%	0.0070	0.0381	0.2371	-0.0415	-0.1275	0.2947	0.0607	0.0164
	.90	20%,20%	-0.0190	0.1084	0.0515	-0.1151	0.0494	-0.0628	-0.1364	-0.0073
	.90	30%,20%	0.0135	0.3288	0.4028	-0.0737	-0.1394	0.2432	0.0596	0.0542
	.90	40%,20%	0.0800	0.6490	0.7457	0.0215	-0.2071	0.5936	0.2276	0.1504
	.90	30%,30%	-0.0161	0.4937	0.3155	-0.3691	-0.2361	0.0613	-0.2707	-0.1300

(Table continues)

Table 9 (continued).

Sample Size	Proportion High	Percentage of Missing Data	**L**	**P**	**MS**	**SR**	**MR**	**SMS**	**SSR**	**SMR**
	.60	10%,0%	-0.0024	-0.0079	-0.3493	-0.1208	0.2364	-0.6485	-0.3426	-0.0118
	.60	20%,0%	-0.0150	-0.0178	-0.6859	-0.2554	0.5237	-1.1479	-0.7291	0.0159
	.60	30%,0%	-0.0060	-0.0103	-0.9836	-0.3761	0.8989	-1.5784	-1.0902	-0.0142
	.60	40%,0%	-0.0386	-0.0511	-1.3066	-0.5348	1.3247	-1.9465	-1.4660	-0.0244
	.60	50%,0%	-0.0449	-0.0861	-1.5866	-0.6788	1.8978	-2.2854	-1.7637	-0.0132
	.60	60%,0%	-0.0667	-0.1007	-1.8626	-0.8456	2.6981	-2.5775	-2.1450	-0.0451
	.70	10%,0%	0.0082	0.0090	-0.3416	-0.1105	0.2580	-0.6381	-0.3466	0.0181
	.70	20%,0%	-0.0168	-0.0356	-0.6924	-0.2575	0.5126	-1.1663	-0.7034	-0.0366
	.70	30%,0%	-0.0532	-0.0742	-1.0205	-0.4155	0.8363	-1.5909	-1.1172	-0.0291
	.70	40%,0%	-0.0691	-0.1241	-1.3360	-0.5598	1.2866	-1.9270	-1.4423	-0.0455
	.70	50%,0%	-0.2148	-0.3498	-1.7047	-0.8221	1.6594	-2.3532	-1.8427	-0.1771
200	.70	60%,0%	-0.4562	-0.6802	-2.0781	-1.1678	2.0551	-2.6978	-2.3204	-0.3695
	.80	10%,0%	-0.0081	-0.0141	-0.3592	-0.1268	0.2372	-0.6448	-0.3481	0.0021
	.80	20%,0%	-0.0295	-0.0465	-0.7059	-0.2692	0.5049	-1.1889	-0.7335	-0.0327
	.80	30%,0%	-0.0651	-0.1095	-1.0481	-0.4266	0.8465	-1.6300	-1.1141	-0.0289
	.80	40%,0%	-0.1934	-0.3260	-1.4417	-0.6676	1.1220	-2.0502	-1.5470	-0.1539
	.80	50%,0%	-0.4498	-0.7127	-1.8801	-1.0156	1.3559	-2.4127	-1.9950	-0.3324
	.80	60%,0%	-1.1026	-1.5660	-2.4345	-1.7067	0.9128	-2.8839	-2.5623	-0.9232
	.90	10%,0%	-0.0239	-0.0309	-0.3753	-0.1414	0.2202	-0.6274	-0.3832	-0.0369
	.90	20%,0%	-0.0533	-0.0810	-0.7399	-0.2892	0.4990	-1.2032	-0.7708	-0.0346
	.90	30%,0%	-0.1399	-0.2739	-1.1567	-0.4894	0.7471	-1.7049	-1.1481	-0.1237
	.90	40%,0%	-0.3769	-0.6126	-1.6041	-0.8239	0.9016	-2.1277	-1.5764	-0.2992
	.90	50%,0%	-0.8213	-1.2375	-2.1183	-1.3391	0.7883	-2.5702	-2.1538	-0.7129
	.90	60%,0%	-1.4187	-1.9514	-2.5564	-1.9316	0.1251	-2.9109	-2.5651	-1.2907
	.60	10%,10%	0.0133	0.0175	-0.0709	-0.1096	0.0631	-0.1591	-0.1607	0.0090
	.60	20%,10%	-0.0045	0.0073	0.3931	-0.1224	-0.3153	0.5089	0.1294	0.0106
	.60	20%,20%	-0.0333	-0.0289	-0.1298	-0.2920	0.0644	-0.3571	-0.3576	-0.0701
	.60	30%,20%	0.0373	0.0562	0.3626	-0.2232	-0.3344	0.2393	-0.0582	0.0489
	.60	40%,20%	-0.0176	0.0275	0.7287	-0.2418	-0.9908	0.5868	0.2374	-0.0199
	.60	30%,30%	-0.0313	-0.0069	-0.0605	-0.4376	0.1228	-0.5388	-0.5061	-0.0285
	.70	10%,10%	0.0054	-0.0107	-0.1086	-0.1249	0.0583	-0.1767	-0.1793	0.0240
	.70	20%,10%	0.0454	0.0597	0.4441	-0.0790	-0.2769	0.5523	0.1508	0.0373
	.70	20%,20%	-0.0016	0.0640	-0.0517	-0.2487	0.1044	-0.2829	-0.3695	-0.0104
	.70	30%,20%	0.0176	0.1579	0.4193	-0.2353	-0.3238	0.2648	-0.0869	0.0486
	.70	40%,20%	-0.0220	0.3173	0.8853	-0.2271	-0.8923	0.7274	0.2561	0.0404
200	.70	30%,30%	-0.0283	0.2274	0.0963	-0.4826	0.0825	-0.3996	-0.5788	-0.0307
	.80	10%,10%	0.0260	0.0154	-0.0754	-0.1125	0.0666	-0.1681	-0.1647	-0.0073
	.80	20%,10%	0.0187	0.1182	0.5010	-0.0840	-0.2827	0.6158	0.1320	-0.0002
	.80	20%,20%	-0.0131	0.0991	-0.0159	-0.2654	0.0865	-0.2617	-0.3938	-0.0143
	.80	30%,20%	-0.0023	0.3146	0.5455	-0.2465	-0.3268	0.3502	-0.0262	0.0445
	.80	40%,20%	0.0268	0.6331	1.0894	-0.1832	-0.7925	0.8487	0.3169	0.0704
	.80	30%,30%	0.0422	0.4105	0.2529	-0.6799	-0.1822	-0.2856	-0.6122	-0.1551
	.90	10%,10%	0.0102	0.0506	-0.0350	-0.1108	0.0696	-0.1260	-0.1817	-0.0004
	.90	20%,10%	0.0188	0.1378	0.5098	-0.0721	-0.2642	0.6068	0.1357	0.0481
	.90	20%,20%	-0.0196	0.1895	0.0761	-0.2913	0.0579	-0.1941	-0.3504	-0.0410
	.90	30%,20%	0.0647	0.5042	0.6985	-0.2582	-0.3129	0.4448	-0.0044	0.0151
	.90	40%,20%	0.1155	1.0825	1.3977	-0.1361	-0.6577	1.0901	0.3462	0.1869
	.90	30%,30%	0.0228	0.7197	0.4850	-1.0172	-0.6737	-0.1019	-0.7715	-0.4352

Note. L: listwise deletion, P: pairwise deletion, MS: mean substitution, SR: simple regression, MR: multiple regression, SMS: stochastic mean substitution, SSR: stochastic simple regression, SMR: stochastic multiple regression. [a] Data were not computed for this combination of sample size, proportion of missing data high, and percentage of missing data.

Pairwise deletion was the third most effective missing data treatment estimating the regression weights for the first data set, but was less effective than deterministic simple regression for the second data set (in which higher zero-order correlations were present). Deterministic mean substitution, deterministic multiple regression, stochastic simple regression, and stochastic mean substitution were generally ineffective in generating unbiased estimates of the regression coefficients.

Overall, these results suggest that applied researchers can be reasonably confident in utilizing stochastic multiple regression and the deletion procedures to generate unbiased parameter estimates of the standardized regression coefficients even when the proportion of missing data is as high as 50%. As the proportion of missing data increases, and as the probability of missingness becomes more highly related to the values of the regressors, however, we can be more confident in employing the stochastic multiple regression or listwise deletion procedures than the pairwise deletion technique.

The relative effectiveness of the missing data treatments in this study with systematically missing data were similar to the results obtained with randomly missing data in previous studies (Brockmeier, Hines, & Kromrey, 1993; Brockmeier, Kromrey, & Hines, 1994; Brockmeier, Kromrey, & Hines, 1995). Stochastic multiple regression and pairwise deletion were the most effective procedures in estimating the sample estimate of R^2. Listwise deletion was the next closest procedure in yielding parameter estimates that did not differ from the complete sample condition. The pattern of effectiveness for the missing data treatments also was similar for the standardized regression coefficients. Across the studies, deterministic mean substitution, deterministic simple regression, deterministic multiple regression, and stochastic mean substitution generally did not perform well in producing unbiased estimates of the sample estimate of R^2 and standardized regression coefficients.

Similarly, the effectiveness of the missing data treatments with systematically missing data in this study and in Kromrey and Hines (1994) was consistent. The deletion procedures were more effective than the deterministic imputation procedures in generating unbiased estimates of the sample estimate of R^2 and standardized regression coefficients. Pairwise deletion was more effective than listwise deletion for the estimation of R^2, but listwise deletion was more effective than pairwise deletion in estimating the regression weights.

Finally, the effectiveness of the missing data treatments with systematically missing data in this study and in Brockmeier et al. (1996) was also congruent. Stochastic multiple regression and pairwise deletion were the most effective missing data treatments for estimating R^2, with listwise deletion being the third most effective method. Stochastic multiple regression and listwise deletion were the most effective missing data treatments for estimating the regression weights in both studies, with pairwise deletion being the third most effective. Across the studies, deterministic mean substitution, deterministic multiple regression, and stochastic mean substitution did not perform well in generating unbiased estimates of the sample estimate of R^2 and standardized regression coefficients.

Three limitations should be considered when interpreting the results of the present investigation. First, generalizability of the results to other data sets is a limitation. The data sets were selected based on the type of data and the correlational differences between variables in each data set. The data sets were not randomly selected from all possible data sets. Second, the outcomes are limited to a two-predictor regression model. The outcomes of regression models with additional predictor variables require examination. Finally, variations in the missing data mechanism need further investigation.

Given these limitations, however, the consistency of the results across several years of research suggest that the choice of a missing data treatment is an important one for researchers. Many of the procedures yield large degrees of bias in the resulting sample estimates, even in the presence of small proportions of missing data. In contrast, stochastic multiple regression and the deletion procedures appear to maintain the integrity of the data matrix and provide relatively unbiased estimates in the presence of large proportions of missing data.

Address correspondence to:
Lantry L. Brockmeier
Florida Department of Education,
614 Turlington Building
325 Gaines Street
Tallahassee, FL 32399-0400
E-Mail: brockml@mail.doe.state.fl.us

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. (1994, April). *Effectiveness of imputation procedures and deletion procedures on the multiple regression analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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

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

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

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

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.

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

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 21-32.

Jinn, J. H., & Sedransk, J. (1989). Effect on secondary data analysis of common imputation methods. In 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.

Kim, J., & Curry J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, *6*, 215-240.

Kromrey, J. (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.

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.

Santos, R. (1981). Effects of imputation on regression coefficients. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 141-145.

Skaalvik, E. M., & Rankin, R. J. (1995). A test of the internal/external frame of reference model at different levels of math and verbal self-perception. *American Educational Research Journal*, *32*(1), 161-184.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics*. New York: Harper & Row.

Appendix A
Example SAS Code for Three Stochastic Imputation Procedures

```
*--------------------------------------------------------*;
* STOCHASTIC MEAN SUBSTITUTION*;
*--------------------------------------------------------*;;

PROC MEANS DATA=SAM50MD NOPRINT;
   VAR X1 X2;
   BY SAMPLE;
   OUTPUT OUT=SMS MEAN=MX1 MX2
                STD=STDX1 STDX2;

DATA SMSUB; MERGE SAM50MD SMS;
   BY SAMPLE;
   IF X1=. OR X2=.  THEN DO;
      RX1=STDX1*RANNOR(0);
      RX2=STDX2*RANNOR(0);
      IF X1=. THEN X1=MX1+RX1;
      IF X2=. THEN X2=MX2+RX2;
      OUTPUT;
   END;
   ELSE OUTPUT;
   KEEP SAMPLE Y X1 X2;

PROC MEANS DATA=SMSUB NOPRINT;
   VAR Y X1 X2 ;
   BY SAMPLE;
  OUTPUT OUT=SMES STD=STDY STDX1 STDX2;

DATA SMSUBS; SET SMES;
   BY SAMPLE;
   KEEP SAMPLE STDY STDX1 STDX2;
PROC REG DATA=SMSUB OUTEST=SMEAS
             NOPRINT;
  MODEL Y=X1 X2/SELECTION=RSQUARE
             START=2 B;
  BY SAMPLE;

DATA SMEANSUB; MERGE SMSUBS SMEAS;
  BY SAMPLE;
  STBX1=X1*(STDX1/STDY);
  STBX2=X2*(STDX2/STDY);
  KEEP SAMPLE _RSQ_ STBX1 STBX2;
  RENAME _RSQ_=SMSRSQ STBX1=SMSX1
                                           STBX2=SMSX2;
```

```
*--------------------------------------------------------*;
* STOCHASTIC SIMPLE REGRESSION *;
*--------------------------------------------------------*;
PROC REG DATA=SAM50MD OUTEST=A
             NOPRINT;
  MODEL X1=X2/SELECTION=RSQUARE
                   START=1 B;
   BY SAMPLE;
   OUTPUT OUT=SSR1 P=X1PV1;
DATA A1; SET A;
   BY SAMPLE;
   KEEP SAMPLE _RMSE_;
   RENAME _RMSE_=RMSE1;
DATA SSIM;  MERGE SAM50MD A1 SSR1;
             * B1 SSR2;
   BY SAMPLE;
   IF X1=. OR X2=. THEN DO;
      RX1=RMSE1*RANNOR(0);
*      RX2=RMSE2*RANNOR(0);
      IF X1=. AND X1PV1 NE . THEN
X1=X1PV1+RX1;
      OUTPUT;
      END;
   ELSE OUTPUT;
   KEEP SAMPLE Y X1 X2;
PROC MEANS DATA=SSIM NOPRINT;
   VAR Y X1 X2;
   BY SAMPLE;
   OUTPUT OUT=SSIMP
                   STD=STDY STDX1
STDX2;
DATA F2;  SET SSIMP;
  BY SAMPLE;
  KEEP SAMPLE STDY STDX1 STDX2;
PROC REG DATA=SSIM OUTEST=SSIMPL
NOPRINT;
  MODEL Y=X1 X2/SELECTION=RSQUARE
START=2 B;
  BY SAMPLE;
DATA SSIMR;  SET SSIMPL;
  BY SAMPLE;
  KEEP SAMPLE _RSQ_ X1 X2;
DATA SSIMREG;  MERGE F2 SSIMR;
  BY SAMPLE;
  STBX1=X1*(STDX1/STDY);
  STBX2=X2*(STDX2/STDY);
  KEEP SAMPLE _RSQ_ STBX1 STBX2;
  RENAME _RSQ_=SSRRSQ STBX1=SSRX1
STBX2=SSRX2;
```

```
*-----------------------------------------------------------------*;
* STOCHASTIC  MULTIPLE  REGRESSION  *;
*-----------------------------------------------------------------*;
PROC REG DATA=SAM50MD OUTEST=AB
     NOPRINT;
  MODEL X1=Y X2/SELECTION=RSQUARE
START=2 B;
   BY SAMPLE;
   OUTPUT OUT=SMR1 P=X1PV1;
DATA AB1; SET AB;
   BY SAMPLE;
   KEEP SAMPLE _RMSE_;
   RENAME _RMSE_=RMSE1;
PROC REG DATA=SAM50MD OUTEST=AC
NOPRINT;
  MODEL X2=Y X1/SELECTION=RSQUARE
START=2 B;
   BY SAMPLE;
   OUTPUT OUT=SMR2 P=X2PV1;

DATA AC1; SET AC;
   BY SAMPLE;
   KEEP SAMPLE _RMSE_;
   RENAME _RMSE_=RMSE2;

DATA GR1A; MERGE SMR1 AB1 SMR2 AC1;
   BY SAMPLE;
DATA SMUL; MERGE SAM50MD GR1A;
    BY SAMPLE;
     IF X1=. OR X2=. THEN DO;
        RX1=RMSE1*RANNOR(0);
        RX2=RMSE2*RANNOR(0);
        IF X1=. AND X1PV1 NE .
            THEN X1=X1PV1+RX1;
        IF X2=. AND X2PV1 NE . THEN
                X2=X2PV1+RX2;
        OUTPUT;
        END;
   ELSE OUTPUT;
   KEEP SAMPLE Y X1 X2;

PROC MEANS DATA=SMUL NOPRINT;
   VAR Y X1 X2;
   BY SAMPLE;
   OUTPUT OUT=SMULT
                STD=STDY STDX1 STDX2;
DATA F3;  SET SMULT;
  BY SAMPLE;
  KEEP SAMPLE STDY STDX1 STDX2;
PROC REG DATA=SMUL OUTEST=SMULTI
NOPRINT;
  MODEL Y=X1 X2/SELECTION=RSQUARE
START=2 B;
  BY SAMPLE;
DATA SMULTR;  SET SMULTI;
  BY SAMPLE;
  KEEP SAMPLE _RSQ_ X1 X2;
DATA SMREGRES;  MERGE F3 SMULTR;
  BY SAMPLE;
  STBX1=X1*(STDX1/STDY);
  STBX2=X2*(STDX2/STDY);
  KEEP SAMPLE _RSQ_ STBX1 STBX2;
  RENAME _RSQ_=SMRRSQ STBX1=SMRX1
           STBX2=SMRX2;
```

Comments on the Analysis of Data with Missing Values

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

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

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

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

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

By contrast, the authors report that the "which is best" decision leaned more clearly to HLM for the Model 2 analysis. However, this may not be attributable to the HLM approach. Rather it may be due to EB estimation procedure. That is, Models 2 and 2A do not have "nested" or hierarchical structures. Thus, the difference in the results for these random effects models may be due to the superiority of the EB estimators over OLS for the added predictor variables (i.e., Percent block poverty (POV) and Percent block college (COL)), both of which are likely to be skewed. Thus, only the Model 3 results are convincing in demonstrating a definite advantage of the HLM approach with EB estimation. The possibility remains, however, that this advantage could potentially dissipate if interaction terms are created so that the OLS regression could model the hierarchical structure of the data. Therefore, OLS regression should still be considered as a viable method for estimating missing data. Fortunately, both Mundrom and Whitcomb (*pp*. 13-19) as well as Brockmeier et al. (*pp*. 20-39) also investigate the properties of regression-based imputation procedures.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

Descriptive and Inferential Aspects of Ordinal Multiple Regression

Jeffrey D. Long
St. John's University

This paper discusses the ordinal multiple regression (OMR) method of Cliff (1994, 1996) and the development of a confidence interval (CI) for population regression weights. First, the OMR methodology is presented along with a discussion of the differences between OMR and least squares multiple regression (LSMR). Next, it is shown how a confidence interval (CI) for a population predictor weight can be derived. The OMR CI is based on an estimated standard deviation of a weight derived from a fixed effects model. Finally, the sampling properties of the OMR CI are discussed. It is pointed out that the OMR CI is more robust than the LSMR CI to predictor correlations and violations of assumptions. The OMR CI is recommended when a researcher wants to consider only ordinal information in multivariate prediction, and/or when predictor correlations are moderate to high, and/or when the assumptions of fixed effects LSMR are violated.

The first purpose of this paper is to introduce applied researchers to a type of ordinal multiple regression (OMR) due to Cliff (1994, 1996). The method is analogous to least squares multiple regression (LSMR) in that regression weights are obtained that optimally combine information on the predictors. OMR can be used as a descriptive method to determine the ability of predictors to predict the order on a criterion.

The second purpose of this paper is to move beyond description and extend Cliff's work to include a confidence interval (CI) for the OMR weights. The CI may be used to make inferences about the size of the OMR weights in the population. To enhance accessibility, examples of computational formulas are provided throughout.

Ordinal Multiple Regression Methodology

There are a number of reasons why OMR might be of interest to the applied researcher (for an extended discussion, see Cliff, 1996). Briefly, (1) OMR is based on operations appropriate for ordinal data, (2) it is suitable for answering ordinal prediction questions, (3) it is relatively unaffected by violations of parametric assumptions, (4) it can accommodate non-linear but monotonic relationships, (5) and its results are invariant under monotonic transformation.

OMR is based on Kendall's tau, t_{jk}, an ordinal correlation coefficient (Kendall, 1970). t_{jk} is an index of the amount of (dis)agreement between two sets of rankings. This definition is best illustrated through the use of dominance scores (Cliff, 1993; Kendall, 1970; Long, in press). A dominance score is an index of the rank order of a pair of raw scores on a variable. The dominance score is defined for observations i, h on variable y, as,

$$d_{ihy} = \text{sign}(y_i - y_h) \tag{1}$$

where $d_{ihy} = +1$ when $y_i > y_h$ (the two scores are in ascending rank order), $d_{ihy} = -1$ when $y_i < y_h$ (the two scores are in descending rank order), and $d_{ihy} = 0$ when $y_i = y_h$ (the scores are tied). Letting n represent the number of subjects, there are $n(n - 1)$ possible pairings of y_i and y_h. The dominance scores for $i < h$ can be computed by multiplying the dominance scores for $i > h$ by (-1). Therefore, only $n(n - 1)/2$ of the $n(n - 1)$ dominance scores for each variable are unique. Note that dominance scores are appropriate for ordinal level data because they index the relations
<, >, = (Stevens, 1959).

To illustrate the computation of dominance scores, consider two hypothetical variables, a criterion variable, y, and a predictor x_1 (both sorted on the criterion),

y:	4,	6,	8,	9,	13
x_1:	10.5,	11,	15,	11,	30

The dominance scores for these two variables are obtained with equation (1). Starting with $i = 1$ and $h = 2$, $d_{12y} = -1$ because $4 < 6$, and $d_{121} = -1$ because $10.5 < 11$. Increasing to $h = 3$ we have $d_{13y} = -1$ (because $4 < 8$) and $d_{131} = -1$ (because $10.5 < 15$), and so on. The reader is invited to check that the dominance scores for the two variables when $i < h$ are

d_{ihy}: -1, -1, -1, -1, -1, -1, -1, -1, -1, -1
d_{ih1}: -1, -1, -1, -1, -1, 0, -1, +1, -1, -1 .

To obtain the dominance scores for $i > h$ we simply multiply these scores by (-1).

It is convenient to list the dominance scores in matrices with rows representing the i index and columns representing the h index. Table 1 shows the $\mathbf{d}_{ihy}$ and $\mathbf{d}_{ih1}$ matrices for the variables. The upper triangle of these matrices represents the case when $i < h$ and the lower triangle the case when $i > h$. The diagonals represent the case when $i = h$.

Table 1. Dominance scores and $\mathbf{t}_{ih1y}$ matrix for two hypothetical variables, y and x_1.

	$\mathbf{d}_{ihy}$					$\mathbf{d}_{ih1}$					$\mathbf{t}_{ih1y}$					
	h					h					h					
i	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	$\sum t_{ih1y}$
1	0	-1	-1	-1	-1	0	-1	-1	-1	-1	0	+1	+1	+1	+1	4
2	+1	0	-1	-1	-1	+1	0	-1	0	-1	+1	0	+1	0	+1	3
3	+1	+1	0	-1	-1	+1	+1	0	+1	-1	+1	+1	0	-1	+1	2
4	+1	+1	+1	0	-1	+1	0	-1	0	-1	+1	0	-1	0	+1	1
5	+1	+1	+1	+1	0	+1	+1	+1	+1	0	+1	+1	+1	+1	0	4

t_{jk} is the proportional agreement between the dominance scores on variable j and k. The index of dominance score agreement is t_{ihjk}, defined as the product of the corresponding dominance scores on variables j and k,

$$t_{ihjk} = d_{ihj}\, d_{ihk} \qquad (2)$$

where t_{ihjk} is +1 when dominance scores have the same sign (the rank order of the pair is the same on both variables), -1 when dominance scores have different signs (rank order of the pair is different), and 0 when there is a tie on either variable.

The third matrix in Table 1, $\mathbf{t}_{ih1y}$, contains the t_{ih1y} computed with equation (2). The last column of Table 1 contains the $\sum_h t_{ih1y}$, which are the sum of the elements in each row of $\mathbf{t}_{ihxy}$. The $\sum_h t_{ih1y}$ are used to calculate t_{1y} and the estimated variance.

t_{jk} is defined as the sum of the t_{ihjk} divided by their total number, $n(n-1)$, or

$$t_{jk} = \frac{\sum_i \sum_h t_{ihjk}}{n(n-1)} \, . \qquad (3)$$

Calculating t_{1y} with our data,

$$\sum_i \sum_h t_{ih1y} = 4 + 3 + 2 + 1 + 4 = 14$$

and $$t_{1y} = \frac{\sum_i \sum_h t_{ih1y}}{n(n-1)} = 14/20 = 0.70$$

Equation (3) shows that t_{jk} is the proportion of paired rank order agreement on variable j and variable k. This form of tau is known as “tau-a” (Kendall & Gibbons, 1991).

OMR deals with the case of one criterion variable, y, and p predictor variables, $x_1,\ldots, x_p$. Let us define $\mathbf{T}_x$ as the p by p matrix of tau correlations among the predictors, and $\mathbf{t}_y$ as the p by 1 vector of correlations between the predictors and y (the tau validities). OMR weights are derived by using tau correlations in place of Pearson correlations in the familiar LSMR equation,

$$\mathbf{w} = \mathbf{T}_x^{-1}\, \mathbf{t}_y \qquad (4)$$

The **w** vector contains the weights used to combine the dominance scores of the predictors to predict the dominance scores of the criterion.

Let us define d_{ihy} as the dominance scores on the criterion and d_{ihj} as the dominance scores on predictor x_j. Then $\hat{d}_{ihy}$ is the predicted criterion dominance scores obtained by solving $\hat{d}_{ihy} = \sum_j d_{ihj} w_j$. It can be shown (see Cliff, 1994) that equation (4) yields weights that optimize the ordinal loss function,

$$Q = \frac{\sum_i \sum_h d_{ihy}\,[\text{sign}\,(\hat{d}_{ihy})]}{n(n-1)}, \qquad (5)$$

where sign(·) is the same as equation (1), and takes on the values, -1, 0, +1. Consistent with permissible ordinal operations, Q indexes the agreement between the criterion dominance scores and the *sign* of the predicted dominance scores.

Differences Between OMR and LSMR

The loss function of OMR is markedly different than the loss function of LSMR. Q defines optimal prediction in terms of an ordinal criterion rather than a least squares criterion. Because of this, there are some important differences between OMR and LSMR.

Interpretation of the weights. An important issue in multiple regression is the interpretation of the weights. As in LSMR, the properties of the OMR weights are most clear when the predictors are not correlated. When the predictor correlations are zero and there are no ties, $\mathbf{T}_x = \mathbf{T}_x^{-1} = \mathbf{I}$, and equation (4) shows that the OMR weights are equal to the tau validities.

When the predictor correlations are greater than zero, the OMR weights have a more ambiguous interpretation. This is also true of LSMR weights (see Cliff, 1987). However, there are a number of reasons why LSMR weights are more interpretable than OMR weights. One reason is that the LSMR weights always have an explicit algebraic definition in relation to the criterion. The LSMR function, $Y_i = \Sigma_j b_j X_{ji} + b_0 + e_i$, indicates that the LSMR weights can always be interpreted as the constants applied to the predictors to literally construct (or deconstruct) the raw scores of the criterion (allowing for error).

The OMR function is much more ambiguous in its specification of the relationship between the weights and the criterion. In fact, an algebraic formula expressing the criterion in terms of the weighted predictors is not possible. The reason is that the ordinal loss function, Q, uses the sign of the weighted predictor dominance scores, sign($\hat{d}_{ihy}$). Because the sign($\hat{d}_{ihy}$) take only the values –1, 0, +1, the criterion dominance scores can not be literally decomposed into a weighted sum of predictor dominance scores. This would be possible if $\hat{d}_{ihy}$ were used in the loss function, but doing so would violate the logic of ordinal analysis. Therefore, we must rely on a verbal description of the functional relationship between the weights and the criterion: the OMR weights are the constants that when applied to the predictor dominance scores best predict order on the criterion, "best" meaning that Q is optimal.

Another reason why LSMR weights may be more interpretable is that OMR weights can not be used to study partial relationships (also called structural relationships). What is deceptive here is that the mathematical equations of the OMR weights seem to suggest ordinal counterparts of some common LSMR coefficients used to examine partial relationships. Consider the OMR case involving two untied correlated predictors, x_1 and x_2. Solving equation (4) for the first weight gives

$$w_1 = \frac{t_{1y} - t_{2y} t_{12}}{1 - t_{12}^2} \qquad (6)$$

where the t_{ky} are the tau correlations between x_k and the criterion, y (the tau validity of x_k and y), and t_{12} is the tau correlation between x_1 and x_2. It is tempting to call equation (6) the "partial OMR weight" because it is identical in form to the partial LSMR weight. It is even more tempting to take the square root of the denominator and call this "semipartial tau," or multiply the denominator by $(1 - t_{2y}^2)$ and take the square root and call this "partial tau." However, these terms are misleading because they do not necessarily represent the same types of relationships as their LSMR counterparts (Cliff, 1996).

Recall that least squares analysis of partial relationships involves the analysis of residuals (Cohen & Cohen, 1983). Residual analysis based on dominance scores leads to interpretation problems. When dominance scores are used with least squares methods, their residuals can take on any real value. This violates the very nature of dominance scores as indices of the order relations <, >, =. Unlike interval/ratio scale scores, it makes no sense to express dominance scores as a function of a predictable component and an error component. More importantly, the residuals of dominance scores do not behave like residuals of raw interval/ratio scores. The value of the partial tau correlation for example, can be different than the value of the partial Pearson correlation in situations where they should be equal (Cliff, 1996; Nelson & Yang, 1988; Somers, 1974).[1] The same inconsistencies can occur in the values of the OMR and LSMR weights (Kim, 1975; Reynolds, 1974).

The above discussion makes it clear that when the predictor correlations are greater than zero, the OMR weights should be interpreted only as practical devices for predicting the order on the criterion. The OMR weights cannot be interpreted in any causal or explanatory sense. The size of an OMR weight represents the relative importance of a predictor to the overall prediction system. A variable with a large weight has a relatively large influence in prediction, and a variable with a small weight has a smaller influence in prediction. Substantive interpretations beyond this have little or no justification.

The limited interpretability of the OMR weights should not be viewed as a flaw. The interpretation of the OMR weights is true to the ambiguity of ordinal data. Ordinal variables carry less information than interval/ratio variables. Therefore, functional relationships among ordinal variables cannot be as precisely specified as among interval/ratio variables. This fact is reflected in the narrower interpretation of the OMR weights. LSMR holds a strong temptation for over-interpretation when ordinal data are analyzed. Numbers derived from ordinal-level variables do not know they are only ordinal (see Lord, 1953), and LSMR computer programs will always produce estimates of partial coefficients. This does not mean that structural interpretations are justified.

Relative magnitude of correlations and weights. The relative sizes of the OMR and LSMR weights are a result of the relative sizes of Pearson and tau correlations. Pearson and tau correlations can be quite different for the same data. It can be shown that under bivariate normality (Kendall & Gibbons, 1991),

$$\text{tau} = \frac{2}{\pi}\sin^{-1}\rho. \qquad (7)$$

This nonlinear relationship means that, over the range of values $0 < |\rho| < 1$, tau can be as much as two-thirds the size of rho.[2]

When predictor intercorrelations are zero, the standardized LSMR weights, b^*_j, are equal to the

Pearson validities, and the OMR weights, w_j, are equal to the tau validities. Assuming multivariate normality, equation (7) indicates that $|b^*_j| > |w_j|$, as long as the Pearson validities are not equal to zero or unity. This relationship also holds for the w_j and the unstandardized LSMR weights, b_j, as long as the predictor variances are greater than unity. When predictor variances are greater than unity, covariances are greater in absolute value than Pearson correlations. It follows that $|b_j| > |b^*_j|$ and, by transitivity, $|b_j| > |w_j|$. When predictor correlations exit, the situation becomes much more complicated, but the same relationships hold.

When some or all of the bivariate relationships are not normal, the above relationships do not necessarily hold. Under non-normality, tau correlations can have absolute values greater than Pearson correlations (Long & Cliff, 1997) and LSMR weights can be smaller in absolute value than OMR weights. Experience has shown that unless the distributions are extremely non-normal the $|b_j|$ are often larger than the $|w_j|$. However, this inequality appears to hold less frequently for the $|b^*_j|$ and the $|w_j|$.

The indeterminacy of the OMR weights. Unlike LSMR, the set of OMR weights obtained by equation (4) are optimal in terms of predicting the order on the criterion, but not unique. It is theoretically possible to find another set of weights that produce an equal number of signed agreements between d_{ihy} and $\hat{d}_{ihy}$. The reason is that equation (4) yields unique weights only when the $\hat{d}_{ihy}$ are normally distributed. Dominance scores almost always produce $\hat{d}_{ihy}$ values that are non-normal because of the influence of common patterns (see Cliff, 1994). Therefore, the equation (4) weights are not the only coefficients that optimize Q for a set of data.

In defense of the OMR weights, LSMR weights can also be indeterminate in many applied situations. Consider the two-group discriminant analysis, which is equivalent to LSMR with a dichotomous criterion. In this case, the LSMR weights maximize the probability of correctly classifying subjects into groups only when the assumptions of normality and equal variances are met (Cliff, 1987). Since these assumptions tend to be violated in applied research (Hill & Dixon, 1982; Micceri, 1989; Pearson & Please, 1975; Sawilowsky & Blair, 1992; Wilcox, 1990), it is often possible to find other sets of weights that are equally effective in classifying subjects for the data at hand.

Prediction and the number of predictors. In contrast to LSMR, prediction does not necessarily improve with additional variables in OMR. Because dominance scores have only three possible values (-1, 0, +1), there are a finite number of possible patterns of dominance relations across predictors. The result is that Q is not necessarily more optimal with the addition of more predictors. In fact, two predictors predict no better than one in the sense of optimizing Q. In the $p = 2$ case, the OMR weights will still indicate the relative importance of the predictors but Q will not be more optimal than in the $p = 1$ case. With more than two predictors, the addition of another does improve prediction (see Cliff, 1994).

Having discussed some of the differences between the OMR and LSMR weights, we now move on to the development of a CI for the OMR population weights. The CI is based on a standard deviation (*SD*) of an OMR weight derived under a fixed effects regression model.

Confidence Interval for the OMR Weights

OMR was completely descriptive in its original presentation (Cliff, 1994). Descriptive methods have been very valuable to the field of psychology (e.g. exploratory factor analysis). However, applied researchers usually want to go beyond description and make inferences about population parameters. In multiple regression, it is often of interest to determine the magnitude of a population regression weight and whether its value is significantly different than zero (Cohen & Cohen, 1983). For OMR, these inferences can be achieved by computing a CI for the OMR weights.

Let us start with the standard form of a CI for a single population weight, π_i. The CI is defined as

$$w_j \pm z_{\alpha/2}\, \hat{\sigma}_{w_j} \tag{8}$$

where w_j is a sample weight from equation (4), $z_{\alpha/2}$ is the appropriate critical value from the standard normal distribution ($z_{\alpha/2} = 1.96$ for the 95% CI), and $\widehat{\sigma}_{w_j}$ is the estimated *SD* of w_j (i.e. the standard error). Since w_j and $z_{\alpha/2}$ are readily available, constructing the CI becomes the problem of computing $\widehat{\sigma}_{w_j}$.

In order to compute $\widehat{\sigma}_{w_j}$ we assume a fixed effects regression model, in which values of the predictors are determined or "fixed" by the researcher. The fixed effects regression model is adopted for a number of reasons. First, there are no known methods for deriving an OMR CI under the random effects model. Second, though random effects data are the most common in psychology, fixed effects methods are often used for analyses because they require fewer assumptions. A number of textbooks for applied researchers assume the fixed effects model in their development of inferential methods (e.g. Cohen & Cohen, 1983; Hays, 1988), and fixed effects inferential methods are used in the multiple regression modules of common computer programs like SPSS (Norusis/SPSS, 1990). Third, the fixed effects model seems most appropriate with OMR because the predictor values are “fixed” to –1, 0, or +1 by the dominance score transformation. The random effects model is usually used only when the predictors take on a wider range of values (Cliff, 1987). Finally, some applied research seems more amenable to the fixed effects model. There are a number of instances where predictor levels appear to be constructions of the researcher rather than random samples from a population of levels as implied by the random effects model.

Under the fixed effects model in OMR the elements of the matrix of predictor tau correlations, $\mathbf{T}_x$, are constants, and so are the elements of $\mathbf{T}_x^{-1}$. Under these assumptions, any sample weight, w_j, can be viewed as a linear combination,

$$w_j = \sum_{k=1}^{p} t^*_{jk}\, t_{ky} \qquad (9)$$

where t_{jk}* is an element from the $\mathbf{T}_x^{-1}$ matrix (t_{jk} is the tau correlation between x_j and x_k), and t_{ky} is the tau correlation between x_k and the criterion, y (the tau validity of x_k and y).

When a sample weight is defined as a linear combination, then the variance of the weight can be obtained by computing the variance of a linear combination (Hays, 1988). In this context, the formula for the variance of a linear combination, say w_j, is

$$\sigma^2_{w_j} = \sum_k (t^*_{jk})^2 \,\mathrm{var}(t_{ky}) + 2\sum_{k<m} t^*_{jk}, t^*_{jm} \,\mathrm{cov}(t_{ky}, t_{my}) \qquad (10)$$

In this equation, var(t_{ky}) is the variance of the tau validity between x_k and y, and cov(t_{ky},t_{my}) is the covariance between the two respective tau validities. The task here is to compute estimates of these two quantities. Then the square root of equation (10) can be used to construct the CI.

The estimated variance of t_{ky}. To estimate the variance of a tau validity (t_{ky}), we will use a biased but consistent estimate of the variance of bivariate tau (Cliff & Charlin, 1991). Consider the tau correlation between x_1 and y, t_{1y}. The consistent estimate of the variance of t_{1v} is

$$\mathrm{Est}[\mathrm{var}(t_{1y})] = \frac{4(n-2)s^2_{t_{i.1y}} + 2s^2_{t_{ih1y}}}{n(n-1)} \qquad (11)$$

In this equation, $s^2_{t_{i.1y}}$ is the variance of the $t_{i.1y}$, formally defined as

$$s^2_{t_{i.1y}} = \frac{\sum_h (t_{i.1y} - t_{1y})^2}{(n-1)} \qquad (12)$$

The $t_{i.1y}$ are the sum of the column elements of the $\mathbf{t}_{ih1y}$ matrix divided by $(n-1)$(see Table 1),

$$t_{i.1y} = \frac{\sum_h t_{ih1y}}{(n-1)} \qquad (13)$$

$s^2_{t_{ih1y}}$ is the variance of the t_{ih1y}, formally defined as

$$s^2_{t_{ih1y}} = \frac{\sum_i \sum_h (t^2_{ih1y} - n(n-1)t^2_{1y})}{n(n-1)-1} \qquad (14)$$

The first term in the numerator of equation (14) is the sum of the squared t_{ih1y} produced by equation (2). The last term in the numerator of equation (14) is the square of the tau validity between x_1 and y. Long and Cliff (1997) found that the CI for bivariate tau based on this consistent estimate of the variance performed well under a wide number of conditions.

To compute the estimated variance of t_{y1} for our data, we must solve equation (11) which involves computing $s^2_{t_{i.1y}}$ and $s^2_{t_{ih1y}}$. First let us consider $s^2_{t_{i.1y}}$, the variance of the $t_{i.1y}$. According to equation (13), the $t_{i.1y}$ are computed by dividing each $\sum_h t_{ih1y}$ by $(n - 1)$. Performing this operation on the last column of Table 1 we obtain the elements of the $\mathbf{t}_{i.1y}$ vector,

$$\mathbf{t}_{i.1y}{}' = \begin{bmatrix} \frac{4}{4} & \frac{3}{4} & \frac{2}{4} & \frac{1}{4} & \frac{4}{4} \end{bmatrix} = [1 \;\; .75 \;\; .50 \;\; .25 \;\; 1]\,.$$

Using the $t_{i.1y}$ and recalling that $t_{1y} = .7$, equation (12) yields

$$s^2_{t_{i.1y}} = \frac{[(1-.7)^2 + (.75-.7)^2 + (.5-.7)^2 + (.25-.7)^2 + (1-.7)^2]}{4}$$
$$= .1188.$$

To derive $s^2_{t_{ih1y}}$, equation (14) indicates that we must compute $\sum_i \sum_h t^2_{ih1y}$, the sum of the squared elements of $\mathbf{t}_{ih1y}$. The reader should verify that there are 18 non-zero elements in $\mathbf{t}_{ih1y}$. Therefore, $\sum_i \sum_h t^2_{ih1y}$ = 18 and equation (14) yields

$$s^2_{t_{ih1y}} = \frac{18 - (5)(4)(.7^2)}{(5)(4)-1} = .4316\,.$$

Finally, substituting all the elements in equation (11), the estimated variance is

$$\mathrm{Est}[\mathrm{var}(t_{1y})] = \frac{4(n-2)s^2_{t_{i.1y}} + 2s^2_{t_{ih1y}}}{n(n-1)}$$

$$= \frac{4(3)(.1188) + 2(.4316)}{5(4)} = .1144\,.$$

Covariance between two t_{ky}. To estimate the covariance between two tau validities, the logic of the variance is extended. Consider the case where we have the tau correlation between x_1 and y, t_{1y}, and the tau correlation between x_2 and y, t_{2y}. In this case there are two sets of t_{ihjk}, t_{ih1y} and t_{ih2y}. The estimate of the covariance between t_{1y} and t_{2y} is (Cliff & Charlin, 1991),

$$\text{Est}[\text{cov}(t_{1y},t_{2y})] = \frac{4(n-2)s_{t_{i.1y},t_{i.2y}} + 2s_{t_{ih1y},t_{ih2y}}}{n(n-1)} . \quad (15)$$

$s_{t_{i.1y},t_{i.2y}}$ is the covariance between the $t_{i.1y}$ and $t_{i.2y}$, formally defined as

$$s_{t_{i.1y},t_{i.2y}} = \frac{\sum_i (t_{i.1y} - t_{1y})(t_{i.2y} - t_{2y})}{(n-1)} . \quad (16)$$

$s_{t_{ih1y},t_{ih2y}}$ is the covariance between the t_{ih1y} and t_{ih2y}, formally defined as

$$s_{t_{ih1y},t_{ih2y}} = \frac{\sum_{i \neq} \sum_{h} (t_{ih1y} - t_{1y})(t_{ih2y} - t_{2y})}{n(n-1) - 1} . \quad (17)$$

To illustrate the computation of the covariance between two tau validities, consider the additional variable, x_2,

x_2: 5, 2, 7, 8, 9

Computation of the $\mathbf{d}_{ih2}$ matrix is left to the reader. Table 2 shows the $\mathbf{t}_{ih2y}$ matrix and the $\sum_h t_{ih2y}$. The reader should verify that $t_{2y} = (16 / 20) = .80$, and

$$\mathbf{t}_{i.2y}{}' = \begin{bmatrix} \frac{2}{4} & \frac{2}{4} & \frac{4}{4} & \frac{4}{4} & \frac{4}{4} \end{bmatrix} = [.50 \quad .50 \quad 1 \quad 1 \quad 1].$$

To compute the covariance between t_{1y} and t_{2y} we solve equation (15). $s_{t_{i.1y},t_{i.2y}}$ is the covariance between the $t_{i.jk}$ of both pairs of variables. Working with the elements of the $\mathbf{t}_{i.1y}$ and the $\mathbf{t}_{i.2y}$ vectors and recalling that $t_{1y} = .7$ and $t_{2y} = .8$, equation (16) yields

$$s_{t_{i.1y},t_{i.2y}} = \frac{(1-.7)(.5-.8) + \ldots + (1-.7)(1-.8)}{4} = -.0438$$

$s_{t_{ih1y},t_{ih2y}}$ is the covariance between the t_{ihjk} of both variables. Working with the elements of the $\mathbf{t}_{ih1y}$ and $\mathbf{t}_{ih2y}$ matrices, we use equation (17) to compute

$$s_{t_{ih1y},t_{ih2y}} = \frac{(1-.7)(.5-.8) + \ldots + (1-.7)(1-.8)}{5(4) - 1} . \\ = -.0632$$

Table 2. The $\mathbf{t}_{ih2y}$ matrix for variables y and x_2.

	$\mathbf{t}_{ih2y}$					
	h					
i	1	2	3	4	5	$\sum t_{ih2y}$
1	0	-1	+1	+1	+1	2
2	-1	0	+1	+1	+1	2
3	+1	+1	0	+1	+1	4
4	+1	+1	+1	0	+1	4
5	+1	+1	+1	+1	0	4

Finally, substituting all the elements into equation (15), the estimated covariance between t_{1y} and t_{2y} is

$$\text{Est}[\text{cov}(t_{1y},t_{2y})] = \frac{4(n-2)s_{t_{i.1y},t_{i.2y}} + 2s_{t_{ih1y},t_{ih2y}}}{n(n-1)} \\ = \frac{4(3)(-.0438)+2(-.0632)}{5(4)} = -.0326$$

To compute the estimated variance of an OMR weight, $\widehat{\sigma}^2_{w_j}$, the above equations for the variance and covariance can be used in equation (10). The square root, $\widehat{\sigma}_{w_j}$, can then be used in the CI of equation (8). A FORTRAN program for computing all equations and performing OMR is available from the author.

Sampling Properties of the OMR CI

In a simulation study using multivariate normal data, Long (1998) found the OMR CI performed well in terms of Type I error and coverage, though coverage became more conservative as effect size increased. The results for power were mixed. The LSMR CI had higher power for all of the conditions in which the predictors were not correlated. However, the OMR CI had higher power for almost all the conditions in which the predictors where moderately to highly correlated (r = .3, .5, respectively). This last finding was especially favorable to the OMR CI given that predictor correlations are almost always non-zero and can be quite substantial in applied research (Cohen, 1994; Meehl, 1997).

In addition to a simulation study, Long (1998) analyzed a real data set with both the OMR CI and the LSMR CI. The data set violated the assumptions of the fixed effects LSMR. That is, the conditional distributions of the criterion were not normal, the conditional variances were not equal, and the predictor-criterion relationships were non-linear but monotonic. It was shown that the OMR CI had higher power than the LSMR CI. This finding is especially favorable to the OMR CI given that fixed effects assumptions (e.g. conditional normality) are violated in many applied situations (Hill & Dixon, 1982; Micceri, 1989; Pearson & Please, 1975; Sawilowsky & Blair, 1992; Wilcox, 1990).

Considering the results of the simulation and the real data analysis, Long (1998) recommended the OMR CI for use when predictor correlations are moderate to high, and/or when fixed effects LSMR assumptions are violated.[3] Earlier it was mentioned that because OMR uses only ordinal information, interpretations tend to be narrower than with LSMR. The tradeoff is that the OMR CI is more versatile than the LSMR CI. We now turn to a more detailed discussion of this versatility.

Outliers. The OMR CI can have higher power than the LSMR CI when outliers are present because the former method is more resistant to extreme values. Recall that the OMR CI is based on dominance scores that index rank order relations. Because only rank order is considered, an outlying observation will not unduly influence the OMR CI. A very large outlier, for instance, is simply treated as the largest value in computing the dominance scores and ultimately, the OMR CI. It does not matter if the largest value is close to the next smallest value, or very far above it. In contrast, a single outlier can have a strong adverse influence on the LSMR CI. Outliers tend to inflate standard errors, causing the LSMR CI to be wide and its power to be low (Birkes & Dodge, 1993).

Non-linearity. The OMR CI can have higher power when predictors have a non-linear but monotonic relationship with the criterion because it is unaffected by such relationships. In both linear and non-linear monotonic relationships, the rank order on the criterion and the predictor is continuously increasing or decreasing. Since rank order is consistent, the dominance scores, and ultimately, the OMR CI are the same for both types of relationships. This means that power and coverage are the same as well. This is not true for the LSMR CI. Non-linear monotonic relationships inflate standard errors and tend to lower the power of the LSMR CI (Birkes & Dodge, 1993).

Parametric assumptions and monotonic transformations. As in the case of outliers and monotonic non-linearity, the OMR CI can have higher power with non-normal conditional distributions and unequal conditional variances. The power superiority of the OMR CI is due again to its use of only ordinal information. To understand this, consider what can happen when normal variables are monotonically transformed.

Assume we compute a LSMR CI on normal, equal conditional variance data, and find the CI to have good coverage and high power. Then we apply a monotonic transformation, such as a power transformation (e.g. $f(x) = x^b$, $b > 0$), that causes the data to be non-normal with unequal conditional variances. These conditions are known to adversely affect the LSMR CI, so we would expect it to have lower power (and perhaps poorer coverage; Birkes & Dodge, 1993; Duncan & Layard, 1973; Wilcox, 1996). Even if the transformation did not cause drastic violations of assumptions, the LSMR CI could be quite different because monotonic transformations change the Pearson correlation structure (Cohen, 1978).

Tau correlations and the OMR CI are invariant under monotonic transformation. In the above situation, the tau correlations and the OMR CI would be exactly the same for the normal data and the transformed non-normal data. It follows that the power (and coverage) of the OMR CI would also be the same. The fact that the properties of the OMR CI remain constant under any monotonic transformation means that the OMR CI is applicable for a wider class of distributions than the LSMR CI.

Monotonic transformations are especially important in applied research because they are commonly used to induce data to meet one or more parametric assumptions, such as normality. The supposition is that some natural process has monotonically transformed the data, resulting in non-normality (Tadikamalla, 1980). The applied researcher attempts to find the transformation that will “undo” the natural transformation and allow the data to meet the desired assumption of normality.

One problem with this practice is that the transformed metric may be difficult to interpret (Salthouse, 1985; Emerson & Stoto, 1983). For example, transforming reaction time data (in milliseconds) by the equation $f(\text{ms}) = \text{ms}^{(.835)}$ might induce normality. But it may be difficult to give a substantive meaning to milliseconds raised to the .835th power.

Another problem is that transforming solely to meet parametric assumptions is driven by the sample data. Sample-driven transformations may not be replicable (Games, 1984). The transformation $f(\text{ms}) = \text{ms}^{(.835)}$ may induce normality in one data set whereas $f(\text{ms}) = \text{ms}^{(-.5)}$ may induce normality in another data set. If parametric methods such as the LSMR CI are used, the results based on different transformations will not be comparable. Of course, these problems can be avoided by using the OMR CI. Results obtained with the initial data, whether normal or non-normal, are invariant for any monotonic transformation.

Latent Trait theory. Another reason why monotonic transformations are of interest is that they are basic elements in latent trait theory (Lord & Novick, 1968). Latent trait theory specifies that latent and manifest variables have a non-linear but monotonic functional relationship. Manifest scores can be monotonically transformed to estimate latent scores (e.g. true ability; see Suen, 1990).

Taken seriously, latent trait theory presents a problem for LSMR. If raw scores are analyzed, then LSMR tells us only about the relationships among the manifest variables. The relationships among the latent variables are still unknown and may be very different.[4] If monotonically transformed scores are analyzed, the results will not be comparable to analyses performed on the untransformed scores.

Latent trait theory does not pose the same problem for OMR. Since the OMR CI is invariant under monotonic transformation, it will yield the same limits whether a variable is analyzed in its latent or manifest form. This consistency seems highly desirable because results from different studies can be compared regardless of whether manifest or latent scores were used.

Conclusion

This paper discussed the ordinal multiple regression (OMR) method of Cliff (1994, 1996) and the development of a confidence interval (CI) for population regression weights. The OMR methodology was presented along with a discussion of differences between OMR and least-squares multiple regression (LSMR). Then it was shown how a confidence interval (CI) for a population predictor weight could be derived. The CI is based on an estimated *SD* of a weight derived from a fixed effects model. Finally, it was pointed out that the OMR CI has many favorable sampling properties. Especially important is the fact that the OMR CI is more robust than the LSMR CI to predictor correlations and violations of assumptions. The OMR CI is recommended when a researcher wants to consider only ordinal information in multivariate prediction, and/or when predictor correlations are moderate to high, and/or when the assumptions of fixed effects LSMR are violated.

Footnotes

[1] Alternative strategies for assessing partial relationships with tau exist that avoid some of these inconsistencies. For example, one might examine the tau correlation between x_2 and y while blocking on x_1 (see Korn, 1984).

[2] Applied researchers may feel uncomfortable that tau correlations can be smaller than Pearson correlations for the same data. Some solace is provided by the fact that t_{jk} is more comparable to r_{jk}^2 than r_{jk}. Recall that t_{jk} is a proportional measure of paired rank order agreement. Therefore, t_{jk} is closer in meaning to r_{jk}^2, another proportional measure (proportion of variance).

[3] Note that if fixed effects assumptions are violated, random effects assumptions must also be violated. The random effects model is more stringent requiring multivariate (not just criterion) normality (Cliff, 1987).

[4] It should be noted that methods such as structural equation modeling (SEM) do not solve the dilemma. Though SEM allows latent variables to be specified, the relationships with the manifest variables are still assumed to be linear. See McDonald (1985) for a discussion of non-linear models.

Portions of this manuscript were based on Long (1998). I wish to thank Norman Cliff, John Horn, Rand Wilcox, and Jennie Noll for their very helpful comments on the manuscript.

Correspondence should be directed to:
Jeffrey D. Long, Department of Psychology,
St. John's University, Marillac Hall SB36,
8000 Utopia Parkway, Jamaica, NY 11439.
Email: longj@stjohns.edu.

References

Birkes, D., Dodge, Y. (1993). *Alternative Methods of Regression*. New York: Wiley.

Cliff, N. (1987). *Analyzing Multivariate Data*. New York: Harcourt, Brace, Javanovich.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

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* (Ch. 4). Mahwah, New Jersey: Lawrence Erlbaum.

Cliff, N., & Charlin, V. (1991). Variances and covariances of Kendall's tau and their estimation. *Multivariate Behavioral Research, 26*, 693-707.

Cohen, J. (1978). Partial products are interactions; partialed products are curve components. *Psychological Bulletin, 93*, 549-562.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Duncan, G. T. & Layard, M. W. (1973). A Monte-Carlo study of asymptotically robust tests of correlation coefficients. *Biometrika, 60*, 551-558.

Emerson, J. D., & Stoto, M. A., (1983). Transforming data. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley.

Games, P. A. (1984). Data transformation, power, and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin, 95*, 345-347.

Hays, W. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart, and Winston.

Hill, M. A., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics, 38*, 377-396.

Kendall, M. G. (1970). *Rank Correlation Methods* (4th ed.). London: Charles Griffin.

Kendall, M., & Gibbons, J. D. (1991). *Rank Correlation Methods* (5th ed.). New York: Oxford University Press.

Kim, J. (1975). Multivariate analysis of ordinal variables. *American Journal of Sociology*, *81*, 261-298.

Korn, E. L. (1984). Kendall's tau with a blocking variable. *Biometrics, 40*, 209-214.

Long, J. D. (1998). *A Confidence Interval for Ordinal Multiple Regression Weights*. Manuscript submitted for publication.

Long, J. D. (in press). Kendall's tau (update). In S. Kotz, C. B. Read, & D. L. Banks (Eds.), *Encyclopedia of Statistical Sciences Update, Vol. 3*. New York: John Wiley & Sons.

Long, J. D., & Cliff, N. (1997). Confidence intervals for Kendall's tau. *British Journal of Mathematical and Statistical Psychology*, *50*, 31-41.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 750-751.

Lord, F. M., & Novick, N. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1985). *Factor Analysis and Related Methods*. Hillsdale, NJ: Lawrence Erlbaum.

Meehl, P. E. (1997). The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Muliak, & J. H. Steiger (Eds.) *What If There Were No Significance Tests?* Hillsdale, NJ: Lawrence Erlbaum.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Nelson, P., & Yang, S. (1988). Some properties of Kendall's partial rank correlation coefficient. *Statistics & Probability Letters, 6*, 147-150.

Norusis, M. J., & SPSS Inc. (1990). *SPSS Base System User's Guide*. Chicago: SPSS Inc.

Pearson, E. S, & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*, 223-241.

Reynolds, H. (1974). Ordinal partial correlation and causal inferences. In H. M. Blalock (Ed.), *Measurement in the Social Sciences*. Chicago: Aldine.

Salthouse, T. A. (1985). Speed of behavior and its implications for cognition. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the Psychology of Aging* (2nd ed.). New York: Van Nostrand Reinhold.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*, 352-360.

Somers, R. (1974). Analysis of partial rank correlation measures based on the product-moment model: Part one. *Social Forces, 53*, 229-246.

Stevens, S. S. (1959). Mathematics, measurement, and psychophysics, in S. S. Stevens (Ed.), *Handbook of Experimental Psychology*. New York: Wiley.

Suen, H. K. (1990). *Principles of Test Theories* (pp. 99-115). Hillsdale, NJ: Lawrence Erlbaum.

Tadikamalla, P. R. (1980). On simulating non-normal distributions. *Psychometrika, 45*, 273-280.

Wilcox, R. R. (1990). Comparing the means of two independent groups. *Biometrical Journal, 32*, 771-780.

Wilcox, R. R. (1996). *Statistics for the Social Sciences*. New York: Academic Press.

A Comparison of Robust and Nonparametric Estimators Under the Simple Linear Regression Model

Jonathan Nevitt, University of Maryland, College Park
Hak P. Tam, National Taiwan Normal University

The present study investigates parameter estimation under the simple linear regression model for situations in which the underlying assumptions of ordinary least squares (OLS) estimation are untenable. Classical nonparametric estimation methods are directly compared against some robust estimation methods for conditions in which varying degrees of outliers are present in the observed data. Additionally, estimator performance is considered under conditions in which the normality assumption regarding error distributions is violated. The study addresses the problem via computer simulation methods. The study design includes three sample sizes (n = 10, 30, 50) crossed with five types of error distributions (unit normal, 10% contaminated normal, 30% contaminated normal, lognormal, t-5df). Variance, bias, mean square error, and relative mean square error are used to evaluate estimator performance. Recommendations to applied researchers and direction for further study are considered.

Applied statistics in the social sciences has focused heavily on modeling data via a linear model (Pedhazur, 1997). Under this framework, a model is posited in which it is assumed that a linear combination of predictors is useful in explaining or predicting some random outcome variable of interest. The most basic form of this model, simple linear regression, is the situation in which a single predictor is included in the explanatory model.

The simple linear regression model, in terms of the observed data, may be expressed by the equation: $y_i = \alpha + \beta x_i + \varepsilon_i$, in which y_i is the score for the response measure for the ith individual; x_i is the value of the explanatory variable for the ith individual; α is the Y-intercept, the mean of the population when the value of X is zero; β is the regression coefficient in the population, the slope of the line; ε_i is a random disturbance, or error, for individual i and is computed as the discrepancy between the observed value of Y for a given individual and the predicted value of Y for that subject). Under this model, it is posited that the score for an individual is partitioned into a structural component, $\widehat{y}_i = (\widehat{\alpha} + \widehat{\beta} X_i)$, which is common to all subjects at the same level of X, and a random component (ε_i) which is unique to each individual.

In the simple linear regression model, the population parameters α and β are unknown quantities which are estimated from the sample data. The most widely employed method for estimating these parameters is the method of ordinary least squares (OLS). Under OLS, sample estimates of α and β (denoted $\widehat{\alpha}$ and $\widehat{\beta}$, respectively) are chosen to minimize the sum of the squared errors of prediction, $\sum e_i^2$, where $e_i = y_i - (\widehat{\alpha} + \widehat{\beta} X_i)$ is the sample estimate of ε_i. OLS regression yields estimates for the parameters that have the desirable property of being minimum variance unbiased estimators (Pedhazur, 1997).

Ordinary least squares estimation places certain restrictive assumptions on the random component in the model, the errors of prediction. OLS estimation assumes, among others, that the errors of prediction are normally distributed, with a common error variance at all levels of X [$\varepsilon \sim N(0,\sigma^2)$]. The normality assumption is frequently untenable in practice. Violation of this assumption is often manifested by the presence of outliers in the observed data. Thus data containing outlying values may reflect nonnormal error distributions with heavy tails or normal error distributions containing observations atypical of the usual normal distribution with larger variance than the assumed σ^2 (Draper & Smith, 1981; Hamilton, 1992). It is well demonstrated that outliers in the sample data heavily influence estimates using OLS regression, sometimes even in the presence of one outlier (e.g., Rousseeuw & Leroy, 1987).

It is also recognized that in the presence of normally distributed errors and homoscedasticity, OLS estimation is the method of choice. For situations in which the underlying assumptions of OLS estimation are not tenable, the choice of method for parameter estimation is not clearly defined. Thus, the choice of estimation method under non-ideal conditions has been a long-standing problem for methodological researchers. The history of this problem is lengthy with many alternative estimation methods having been proposed and investigated

(Birkes & Dodge, 1993; Dietz, 1987; Iman & Conover, 1979; Tam, 1996; Theil, 1950; Yale & Forsythe, 1976).

Robust Regression

Alternatives to OLS regression may be regarded as falling into broad classes based upon the approach to the problem of parameter estimation and the assumptions placed upon the model. Robust regression is a general term that encompasses a wide array of estimation methods. In general, robust estimation methods are considered to perform reasonably well if the errors of prediction have a distribution that is not necessarily normal but “close” to normal (Birkes & Dodge, 1993). Thus, these methods have been developed for situations in which symmetric error distributions have heavy tails due to outliers in the observed data (Hamilton, 1992). A common element to these methods is the definition of a loss function on the residuals, which is subject to minimization via differentiation with respect to the slope and *Y*-intercept parameters (Draper & Smith, 1981). Examples of this type of robust estimation are Huber M-estimation, the method of Least Median of Squares, and the method of Least Absolute Deviations (LAD).

The robust LAD estimator is investigated in the present study and so a brief description of the method is mentioned here. LAD was developed by Roger Joseph Boscovich in 1757, nearly 50 years before OLS estimation (see Birkes & Dodge, 1993 for a review and historical citations). In contrast to OLS estimation which defines the loss function on the residuals as $\sum e_i^2$, LAD finds the slope and *Y*-intercept that minimize the sum of the absolute values of the residuals, $\sum |e_i|$. In concept, the LAD estimator is no more complex than the OLS estimator. Some have considered LAD to be simpler than OLS because $|e_i|$ is a more straightforward measure of the size of a residual as compared to e_i^2. Unfortunately, computing LAD estimates is more difficult than computing OLS estimates; there are no exact formulas for LAD estimates and thus algorithmic methods must be employed to calculate them.

Other forms of robust regression involve iterative modification of the sample data, often based upon the residuals from OLS estimation. Examples of this type of robust estimation are Winsorized Regression (Yale & Forsythe, 1976) and regression using data trimming methods (Hamilton, 1992). These methods maintain the assumptions of OLS estimation and employ smoothing techniques to resolve the influence of *Y*-outliers on the estimates of slope and *Y*-intercept. The trimmed least squares estimator (TLS) is computationally similar to a trimmed mean (Hamilton, 1992). Estimates for TLS are computed by deleting cases corresponding to a specified percentage of the largest positive and the largest negative residuals under an initial OLS estimation. After case deletion, OLS estimation is performed on the remaining data to compute the TLS estimates of slope and *Y*-intercept.

Winsorized regression, which can take on several different forms, is used as a method to reduce the effect of *Y*-outliers in the sample by smoothing the observed *Y*-data rather than simply deleting outlying cases (as in TLS). Fundamental to the method is the formulation of an observed response measure as $y_i = \hat{y}_i + e_i$. If an observed response measure is far from the majority of the other *Y*-values (i.e., an outlier), then the residual for that case will tend to be large in absolute value. Winsorization methods modify extreme *Y*-values, in an iterative fashion, by replacing the observed residual for an extreme *Y*-value with the next closest (and smaller) residual in the data set, and then computing new *Y*-values using the formulation for an observed score as presented above. These new *Y*-values are used to compute new slope and intercept estimates for the regression line, and then a new set of residuals is obtained. The process of estimation, obtaining residuals, and data modification is continued for a specified number of iterations.

Variations on Winsorization methods for linear regression are described by Yale and Forsythe (1976) and incorporate techniques for both computing the residuals and for modifying the observed *Y*-data. They note the most common method for obtaining the residuals is to compute the OLS estimates of slope and intercept and form the residuals in the usual manner as $e_i = y_i - (\hat{\alpha} + \hat{\beta} X_i)$. The most straightforward method for smoothing the data is a process in which a specified percentage of the *Y*-data, at each extreme of the ordered residuals, is modified iteratively. Iterations involve computing OLS estimates, obtaining residuals, and then replacing extreme *Y*-values with modified *Y*-values as described above.

Nonparametric Regression

The robust regression methods described above assume normally distributed error terms in the regression model. In distinction, classical nonparametric approaches to linear regression typically employ parameter estimation methods that are regarded as distribution free. Since nonparametric regression procedures are developed without relying on the assumption of normality of error distributions, the only presupposition behind such procedures is that the errors of prediction are independently and identically distributed (i.i.d.) (Dietz, 1989). The assumption that the data are i.i.d. is a considerably weaker assumption as compared to the normality assumption underlying OLS regression and robust regression procedures. Hence nonparametric regression methods are expected to perform well without regard to the nature of the distribution of errors. Several classical nonparametric approaches to linear regression are reviewed by Tam (1996) and are briefly described here.

Many nonparametric procedures are based on using the ranks of the observed data rather than the observed data themselves. An application of rank transformation in the linear regression model was developed by Iman and Conover (1979) and is known as monotonic regression. This technique has been proposed for estimating slope and *Y*-intercept when the data exhibit a nonlinear relationship (i.e., data that exhibit a monotonic increasing or decreasing relationship). Monotonic regression uses the rank ordering of the data as the values for criterion and independent variables in the estimation of slope and *Y*-intercept. Iman and Conover (1979) compared the performance of the rank regression method against OLS, mean isotonic regression, and median isotonic regression and found that for data exhibiting a strictly monotonic increasing or decreasing relationship, monotonic regression shows strong estimator performance. They also note that the procedure fits the monotone non-linear trend in the sample data while robust regression is forced to treat non-linearity in the data as outliers. Therefore, Iman and Conover suggest using monotonic regression for situations of non-linearity but not for cases in which the sample data is contaminated by outliers.

In addition to methods based on ranks, nonparametric procedures have been developed that use the median as a robust measure (rather than means, as in OLS). Theil (1950) considered the geometric formula for the slope of the line between any two data points (say the *i*th and *j*th points) as

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i},$$

where $x_i \neq x_j$. He proposed a robust measure for the slope of the regression line passing through all *n* sample data points by taking the median of all possible pairwise slopes. Conceptually, this method would yield an estimate of slope that is resistant to outliers in the sample data.

Modifications to Thiel's original method for computing the slope of the regression line have been proposed in which each of the pairwise slopes, b_{ij}, are weighted using a weighting scheme. The median of these weighted pairwise slopes is then taken as the slope of the regression line passing through all *n* observations in the sample. Jaeckel (1972) proposed that each slope should be weighted by the *X*-distance between the i^{th} and j^{th} observations (i.e., $w_{ij} = x_j - x_i$). Sievers (1978) and Scholz (1978) suggested the use of $w_{ij} = (j - i)$ as the weighting scheme, which is the number of steps between i^{th} and j^{th} observations. Still another weighting method, as discussed by Birkes and Dodge (1993), uses $w_{ij} = |x_j - x_i|$.

Using medians, several methods for computing the *Y*-intercept have been proposed and investigated. It can be shown that the intercept of the line joining any two data points is given by

$$a_{ij} = \frac{x_j y_i - x_i y_j}{x_j - x_i},\ i < j,\ x_i \neq x_j.$$

Under this formulation, several nonparametric estimators for *Y*-intercept have been proposed. The most obvious one is to take the median of the a_{ij} values.

A different approach that does not require the a_{ij} terms explicitly is to make use of the various nonparametric slope estimators previously mentioned. For some estimator of slope, $\widehat{\beta}$, the term $y_i - \widehat{\beta} X_i$ is computed for each observation, and then the median of these terms is taken for the *Y*-intercept of the regression line passing through all *n* observations. Theil (1950) originally proposed this estimator for the *Y*-intercept of the line using his proposed median of pairwise slopes as $\widehat{\beta}$. A variant of this *Y*-intercept may be formed substituting the modified (weighted) Theil slope estimator as $\widehat{\beta}$.

Yet another approach to estimating the *Y*-intercept is to compute it as the median of all pairwise averages of the $y_i - \widehat{\beta} X_i$ terms. This *Y*-intercept can also be computed using either the original Theil median of pairwise slopes as $\widehat{\beta}$ or using the modified Theil slope as $\widehat{\beta}$. Finally, Conover (1980) proposed estimating the *Y*-intercept using $\widehat{\alpha}$ = median(y_i) - median(x_i), using the Theil median of pairwise slopes as $\widehat{\beta}$. This *Y*-intercept estimate is usually paired with the Theil median of pairwise slopes estimator for $\widehat{\beta}$ in the regression equation.

Tam (1996) reviews two important studies that compare the performance of median based classical nonparametric methods for estimating the slope and *Y*-intercept in linear regression. Hussain and Sprent (1983) present a simulation study in which they compared the OLS regression estimator against the Theil pairwise median and weighted Theil estimators in a study using 100 replications per condition. Hussain and Sprent characterized the data modeled in their study as typical data patterns that might result from contamination due to outliers. Contaminated data sets were generated using a mixture model in which each error term is either a random observation from a unit normal distribution [N(0,1)] or an observation from a normal distribution with a larger variance [N(0,k^2), k >1].

The investigators present results from simulated data sets with the probability, *p*, of drawing data from the N(0,1) distribution fixed between 0.85 and 0.95. Sample sizes of 10 and 30 are presented for the situation in which there are no outliers (*p* = 1.0) and for the condition in which the data contain approximately 10% outliers (k = 9; p = 0.85 for n = 10, p = 0.90 for n = 30). X-values in the Hussain and Sprent study follow an equally spaced, sequential additive series (x_i = 1, 2,..., n). Observed outcome values are generated by the model: $y_i = 2 + x_i + e_i$, in which e_i is a random deviate drawn from the appropriate normal distribution.

Results from Hussain and Sprent (1983) indicate that Theil's method was appreciably better than OLS in the presence of outliers, especially for small sample sizes. Such results pertain especially to the estimation of the Y-intercept term in the linear regression model. Furthermore, their results showed no real advantage of the weighted median estimator as compared to the Theil estimator under their simulated data conditions.

In addition to the work of Hussain and Sprent, findings in Dietz (1987) have contributed substantially to the field of classical nonparametric regression. Dietz estimated and compared the mean square errors (MSE) of the Theil slope and several weighted median slope estimators under a variety of simulated data conditions. Additionally, Dietz examined several nonparametric estimators of Y-intercept. Dietz simulated data according to two sample sizes (20 and 40), three X-designs to generate X-values, and nine error distributions (i.e. standard normal, 6 contaminated normal distributions with various degrees of flatness, heavy-tailed t-distribution with 3 degrees of freedom, and an asymmetric lognormal distribution). Dietz generated 500 data replications per condition.

Findings in Dietz (1987) demonstrated that for normal error distributions, the OLS slope estimator yielded the lowest MSE, while for nonnormal errors the OLS slope estimator had the largest MSE. The weighted median slope estimators showed strong performance under the moderately contaminated data conditions while the Theil unweighted median slope estimator yielded the lowest MSE under the heavily contaminated data conditions. Dietz also reported that the Y-intercept estimator as proposed by Theil (1950) yielded large MSE values and should be avoided in practice.

Alternatives to OLS regression continue to intrigue applied statisticians and methodological researchers. The present study explores the behavior of robust regression and nonparametric approaches to simple linear regression under various situations with respect to contaminated data and nonnormal error distributions. This study provides an extension to previous research in some important areas. As noted by Tam (1996), very little research exists in which classical nonparametric alternatives to linear regression are directly compared against robust regression methods. Additionally, comparisons of alternative regression methods are often presented only within the framework of statistical theory or by examining estimator performance on exemplary data sets (e.g., Birkes & Dodge, 1993). The present study serves to begin addressing the issue of comparing alternatives to OLS regression within the framework of a simulation study.

Method

All programming for the simulation study was developed using GAUSS (Aptech Systems, 1996). In the present study, three levels of sample size (n = 10, 30, 50) were crossed with five types of error distributions (unit normal, contaminated unit normal with 10% Y-outliers, contaminated unit normal with 30% Y-outliers, lognormal, t-5df). For each of the 15 cells in the study, 1000 simulated bivariate data sets were generated. Algorithms for drawing random deviates from contaminated unit normal, lognormal, and t-5df distributions are found in Evans, Hastings, and Peacock (1993).

Data generation methods are conformable to those of Hussain and Sprent (1983). Vectors of random error variates were drawn from the appropriate error distribution. Error vectors for the contaminated normal distributions were mixtures of deviates drawn from a unit normal distribution and from a normal $N(0,k^2)$ distribution with k = 9. It has been demonstrated that drawing deviates from this larger variance normal distribution will result in some (potentially) large Y-outliers (Hussain & Sprent, 1983).

Simulated bivariate data sets consisted of $(\mathbf{X},\mathbf{Y})$ vectors. The vector of X-values was generated to follow an equally spaced, sequential additive series (x_i = 1, 2,..., n). The Y-vector was generated by the model: $y_i = 2 + x_i + e_i$, in which e_i is a random deviate drawn from the appropriate error distribution. Thus, the population parameters underlying the model are $\alpha = 2$ and $\beta = 1$ for Y-intercept and slope, respectively.

For each simulated data set, estimators of slope and Y-intercept were computed. The robust regression estimators considered in this study are LAD, 10% and 20% Winsorized least squares, and 10% TLS. Algorithms for computing the LAD estimator are found in Birkes and Dodge (1993). Winsorization methods for computing residuals and smoothing the Y-data were implemented via the methods described previously, and used five iterations of data smoothing. We conducted pilot studies using Winsorized regression, with results showing very little change in the parameter estimates beyond five iterations of data adjustment. Estimates for the 10% TLS were computed by deleting cases corresponding to the 10% largest positive and the 10% largest negative residuals under an initial OLS estimation. After case deletion, OLS estimation was performed on the remaining observations to compute the TLS estimates.

Table 1. Summary Measures for Estimating Population Slope (β = 1.0).

Estimation	Error Distribution: N(0,1) - 0% contamination			
Method	Variance	Bias	MSE	RMSE
OLS:	0.01115491	0.00707727	0.01120500	0
LAD:	0.01838679	0.00598824	0.01842265	-0.64414576
WIN10:	0.01223615	0.00756652	0.01229340	-0.09713513
WIN20:	0.01299585	0.00830138	0.01306476	-0.16597602
TLS:	0.01646757	0.00737854	0.01652201	-0.47452125
MON:	0.00096072	-0.04701818	0.00317143	0.71696304
Theil:	0.01266696	0.00790564	0.01272946	-0.13605202
Wtd. Theil:	0.01235103	-0.00126754	0.01235263	-0.10242155
	Error Distribution: N(0,1) - 10% contamination			
OLS:	0.11142026	0.01378250	0.11161021	0
LAD:	0.02767390	0.00432905	0.02769264	0.75188074
WIN10:	0.02192931	0.00375534	0.02194342	0.80339239
WIN20:	0.02942458	0.00682076	0.02947111	0.73594615
TLS:	0.01880606	0.00268830	0.01881329	0.83143757
MON:	0.02047459	-0.15438788	0.04431021	0.60299146
Theil:	0.02066901	0.00651707	0.02071149	0.81443018
Wtd. Theil:	0.02018951	-0.00604903	0.02022610	0.81877913
	Error Distribution: N(0,1) - 30% contamination			
OLS:	0.31264452	-0.00547711	0.31267452	0
LAD:	0.06165909	-0.00054303	0.06165939	0.80280009
WIN10:	0.14933177	-0.01329565	0.14950854	0.52183970
WIN20:	0.10990528	-0.00357852	0.10991809	0.64845845
TLS:	0.15258114	-0.01516154	0.15281101	0.51127769
MON:	0.04915750	-0.34893333	0.17091197	0.45338696
Theil:	0.06853707	-0.00716128	0.06858835	0.78063978
Wtd. Theil:	0.09594470	-0.02908675	0.09679074	0.69044252
	Error Distribution: Lognormal			
OLS:	0.05361053	0.00528236	0.05363843	0
LAD:	0.02574529	-0.00334989	0.02575651	0.51981235
WIN10:	0.02661642	-0.00448754	0.02663656	0.50340532
WIN20:	0.02639584	0.00045408	0.02639604	0.50788934
TLS:	0.03613776	-0.00986773	0.03623513	0.32445574
MON:	0.01326078	-0.10921212	0.02518806	0.53041014
Theil:	0.01489242	-0.00264993	0.01489945	0.72222444
Wtd. Theil:	0.01521499	-0.01259078	0.01537352	0.71338612
	Error Distribution: *t*-5*df*			
OLS:	0.01764455	-0.00219277	0.01764936	0
LAD:	0.02363482	0.00078222	0.02363543	-0.33916683
WIN10:	0.01707596	-0.00241412	0.01708179	0.03215808
WIN20:	0.01734658	-0.00224915	0.01735164	0.01686858
TLS:	0.02184069	-0.00112768	0.02184196	-0.23755002
MON:	0.00321083	-0.07123636	0.00828545	0.53055218
Theil:	0.01810661	-0.00002527	0.01810661	-0.02590776
Wtd. Theil:	0.01704933	-0.01184856	0.01718971	0.02604293

Note: Tabled results are for the n=10 sample size. OLS: ordinary least squares; LAD: least absolute deviations; WIN10: 10% Winsorized regression; WIN20: 20% Winsorized regression; TLS: trimmed least squares; MON: monotonic regression; Theil: median of pairwise slopes; Wtd. Theil: weighted median of pairwise slopes.

The classical nonparametric estimators for population slope included in this study are monotonic regression, the Theil median based estimator, and the modified (weighted) Theil estimator. Since our design employs X-values such that each x_i value equals its index number (i.e. $x_i = i$, for all i), all the previously described methods for weighting pairwise slopes are equivalent and hence are simply referred to as the weighted Theil slope estimator. The nonparametric Y-intercept estimators described previously and investigated by Dietz (1987) were also investigated in the present study.

Summary measures for each estimator were obtained for the set of 1000 replications in each of the 15 cells in the study. Summary measures of minima and maxima, mean, and median were collected. To measure the quality of parameter estimation, estimator variance, bias, mean square error (MSE), and relative mean square error (RMSE) were computed for the estimators under each condition. MSE can be a useful measure of the quality of parameter estimation (Stone, 1996), and is computed as $\text{MSE} = \text{Var}(\theta') + \text{bias}(\theta')^2$, in which θ' is an estimate of the population parameter θ.

Relative mean square error has also been used as a measure of the quality of parameter estimation (e.g., Yale & Forsythe, 1976). We computed RMSE as $(\text{MSE}_{\text{OLS}} - \text{MSE}_{\theta})/\text{MSE}_{\text{OLS}}$. We believe this formulation is useful for comparing estimator performance within a given condition, and is interpreted as a proportionate (or percent) change from baseline, using the OLS estimator MSE within a given data condition as a baseline value. Positive values of RMSE refer to the proportional reduction in the MSE of a given estimator with respect to OLS estimation. Hence, RMSE is interpreted as a relative measure of performance above and beyond that of the OLS estimator.

Results

Effects of sample size

Across sample sizes, estimator variances (and, to some lesser degree estimator bias) decreased with increasing sample size. For example, the variances for the OLS slope estimator under the uncontaminated unit normal distribution are 0.011, 0.00043, and 0.000098 for sample sizes n = 10, 30, and 50 respectively. This pattern of decreasing variance and bias holds for all estimators under all error distributions. The patterns seen in the variances are also exhibited in the estimator MSE values. Because the results for the n = 30 sample size are intermediate to those for the n = 10 and n = 50 sample sizes, they are not reported here.

Slope estimator performance

Tables 1 and 2 present summary results for the estimation of population slope under the unit normal, contaminated normal, and nonnormal error distributions for sample sizes n = 10 and n = 50, respectively. For the OLS slope estimator, note the increase in MSE as the degree of contamination in the data increases. OLS slope estimator MSE values for the lognormal and t-5df error distributions also show increases as compared to the unit normal error distribution.

Under most conditions, the results for monotonic regression in Tables 1 and 2 show small variances for this slope estimator accompanied by large (in absolute value) bias values. For example, in Table 1, the variance for monotonic regression under the uncontaminated unit normal condition is 0.00096 as compared to the variance for the OLS slope estimator of 0.01115. While monotonic regression yields reduced variances, bias values for this slope estimator can be quite large. Bias values in Table 1 for monotonic regression are often several orders of magnitude higher than the corresponding bias values for the other slope estimators. Note that bias values for monotonic regression are not only large in absolute magnitude, but also negative. These negative bias values indicate the monotonic regression slope estimator consistently under estimated the population slope value of $\beta = 1.0$.

Under ideal conditions (unit normal error distribution, no contamination), MSE values in Tables 1 and 2 indicate inflation in MSE for all robust and nonparametric estimators (with the exception of monotonic regression) as compared to OLS. MSE for these slope estimators are larger than for OLS for this condition and thus corresponding RMSE values are negative. LAD and TLS slope estimators exhibit the largest inflation in MSE as compared to OLS with corresponding reductions in relative estimator performance of approximately 64% for the LAD estimator and 47% (n = 10) and 37% (n = 50) for the TLS estimator.

For the 10% data contamination condition, all robust and nonparametric slope estimators (with the exception of monotonic regression) show strong performance gains with 75-84% decreases in MSE as compared to OLS under this moderate level of data contamination. Comparing estimator performance across the two sample sizes, one sees that performance gains are generally lower for the n = 10 sample size with the exception of the TLS slope estimator. The TLS slope estimator yields an 83.1% reduction in MSE under the n = 10 sample size and a 74.5% reduction in MSE under the larger sample size condition.

Under the 30% contamination condition, the LAD slope estimator shows superior performance for both the small and large sample sizes. RMSE values in the two tables indicate reductions in MSE of 80.3% and 88.8% for the n = 10 and n = 50 sample sizes respectively. In this extreme contamination

Table 2. Summary Measures for Estimating Population Slope (β = 1.0).

Estimation	**Error Distribution: N(0,1) - 0% contamination**			
Method	Variance	Bias	MSE	RMSE
OLS:	0.00009810	0.00027520	0.00009818	0
LAD:	0.00016128	0.00031704	0.00016138	-0.64382254
WIN10:	0.00010321	0.00022429	0.00010326	-0.05175852
WIN20:	0.00010363	0.00022180	0.00010367	-0.05600182
TLS:	0.00013426	0.00001423	0.00013426	-0.36749649
MON:	0.00000043	-0.00214771	0.00000504	0.94866569
Theil:	0.00010445	0.00024160	0.00010451	-0.06451524
Wtd. Theil:	0.00010419	0.00015729	0.00010421	-0.06148768
	Error Distribution: N(0,1) - 10% contamination			
OLS:	0.00088268	0.00146208	0.00088482	0
LAD:	0.00017786	0.00016588	0.00017789	0.79895156
WIN10:	0.00015842	0.00036745	0.00015855	0.82081015
WIN20:	0.00019381	0.00049379	0.00019406	0.78068165
TLS:	0.00022498	0.00053138	0.00022527	0.74541118
MON:	0.00010839	-0.01713325	0.00040194	0.54574057
Theil:	0.00014566	0.00026384	0.00014573	0.83529971
Wtd. Theil:	0.00014591	0.00017933	0.00014594	0.83506178
	Error Distribution: N(0,1) - 30% contamination			
OLS:	0.00255733	0.00110911	0.00255856	0
LAD:	0.00028630	0.00049167	0.00028655	0.88800458
WIN10:	0.00086501	0.00036565	0.00086514	0.66186374
WIN20:	0.00084013	0.00011337	0.00084015	0.67163290
TLS:	0.00047307	0.00037883	0.00047321	0.81504716
MON:	0.00032382	-0.04740485	0.00257104	-0.00487937
Theil:	0.00034353	0.00008385	0.00034354	0.86573035
Wtd. Theil:	0.00034969	0.00000300	0.00034969	0.86332687
	Error Distribution: Lognormal			
OLS:	0.00041453	-0.00022877	0.00041458	0
LAD:	0.00015974	0.00023784	0.00015979	0.61456832
WIN10:	0.00016734	-0.00003671	0.00016734	0.59635229
WIN20:	0.00016585	-0.00005373	0.00016586	0.59994409
TLS:	0.00025751	0.00016274	0.00025753	0.37881494
MON:	0.00008034	-0.00906487	0.00016251	0.60800790
Theil:	0.00006711	-0.00000821	0.00006711	0.83811642
Wtd. Theil:	0.00006952	-0.00015675	0.00006954	0.83225400
	Error Distribution: *t*-5*df*			
OLS:	0.00015653	0.00035306	0.00015665	0
LAD:	0.00016443	0.00007171	0.00016443	-0.04966430
WIN10:	0.00013112	0.00026626	0.00013119	0.16251913
WIN20:	0.00013395	0.00026166	0.00013402	0.14449768
TLS:	0.00015108	-0.00001635	0.00015108	0.03555771
MON:	0.00000383	-0.00367558	0.00001734	0.88929977
Theil:	0.00013153	0.00029653	0.00013162	0.15980954
Wtd. Theil:	0.00013104	0.00015973	0.00013106	0.16336140

Note: Tabled results are for the n=50 sample size. OLS: ordinary least squares; LAD: least absolute deviations; WIN10: 10% Winsorized regression; WIN20: 20% Winsorized regression; TLS: trimmed least squares; MON: monotonic regression; Theil: median of pairwise slopes; Wtd. Theil: weighted median of pairwise slopes.

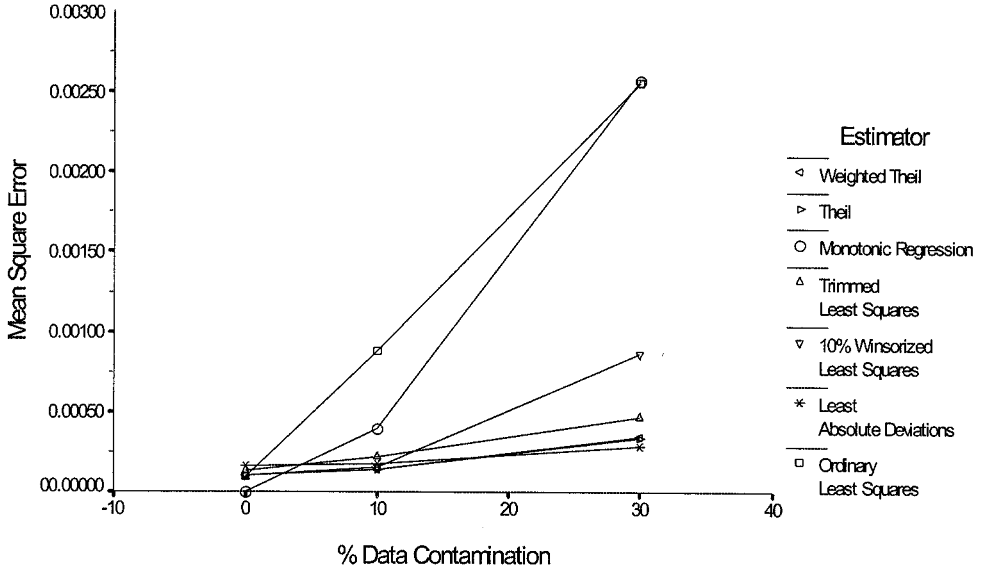

Figure 1. Mean square error in estimation of population slope under varying levels of data contamination. Results charted are for the *n* = 50 sample size.

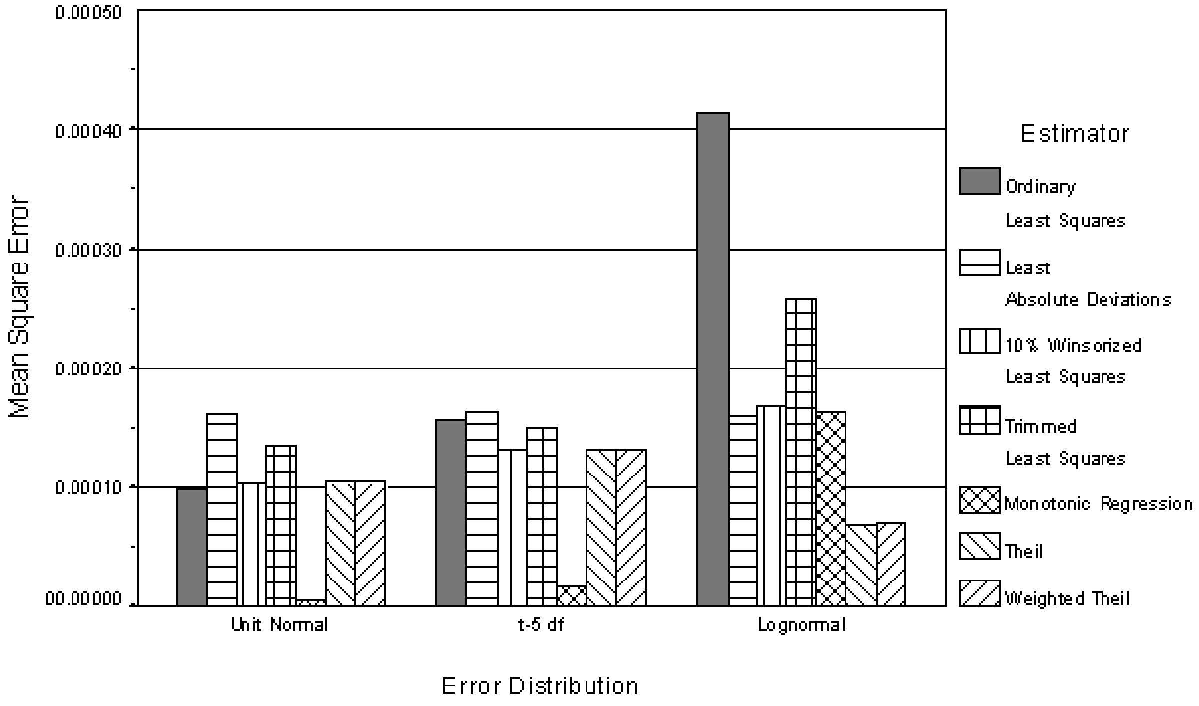

Figure 2. Mean square error in estimation of population slope for normal and nonnormal error distributions. Results charted are for the *n* = 50 sample size.

Table 3. Summary Measures for Estimating Population *Y*-Intercept (α = 2.0)

Estimation	**Error Distribution: N(0,1) - 0% contamination**			
Method	Variance	Bias	MSE	RMSE
OLS:	0.46623625	-0.04029911	0.46786027	0
LAD:	0.69815192	-0.03000075	0.69905197	-0.49414688
WIN10:	0.49958621	-0.04058599	0.50123343	-0.07133146
WIN20:	0.53987082	-0.04371208	0.54178157	-0.15799866
TLS:	0.63496493	-0.04029981	0.63658900	-0.36063915
MON:	0.02906177	-1.74140000	3.06153573	-5.54369673
med (a_{ij}):	0.60101301	-0.04286027	0.60285001	-0.28852576
Conover:	0.75428972	-0.05273348	0.75707054	-0.61815522
Median-1:	0.55398222	-0.03755454	0.55539257	-0.18709068
Median-2:	0.51962863	-0.04548496	0.52169752	-0.11507120
Wtd. Mdn-1:	0.54440377	0.01445567	0.54461273	-0.16404996
Wtd. Mdn-2:	0.50696225	0.00284416	0.50697034	-0.08359348
	Error Distribution: N(0,1) - 10% contamination			
OLS:	4.26531434	-0.10433764	4.27620068	0
LAD:	0.97412322	-0.01615455	0.97438419	0.77213787
WIN10:	0.82937738	-0.02296355	0.82990470	0.80592475
WIN20:	1.03965787	-0.03238871	1.04070690	0.75662814
TLS:	0.69793756	-0.01368506	0.69812484	0.83674180
MON:	0.61935638	-1.15086667	1.94385047	0.54542581
med (a_{ij}):	0.79369545	-0.02981975	0.79458466	0.81418443
Conover:	1.21925681	-0.06955337	1.22409448	0.71374251
Median-1:	0.76404907	-0.02949380	0.76491895	0.82112183
Median-2:	0.77021114	-0.03544536	0.77146751	0.81959043
Wtd. Mdn-1:	0.75465027	0.03493991	0.75587107	0.82323770
Wtd. Mdn-2:	0.75285108	0.03346361	0.75397089	0.82368206

condition, the Theil and weighted Theil estimators also show strong slope estimator performance. For the *n* = 50 sample size, slope estimator MSE values for the uncontaminated and contaminated data conditions are plotted in Figure 1.

For the lognormal error distribution, the nonparametric Theil and weighted Theil methods exhibit the strongest performance in both the small and large sample sizes. For the *n* = 10 sample size, Table 1 reports relative reductions in MSE of 71-72% for these nonparametric estimators. For the large sample size, RMSE values in Table 2 show even higher performance gains with relative reductions in MSE of 83-84%. Close to one another, but running a distant second, are the robust LAD and Winsorized least squares estimators with relative reductions in MSE of about 51% for the small sample size and 60% for the large sample size. Under the *t*-5df error distribution, the Winsorized least squares estimators and the nonparametric Theil and weighted median estimators yield only small reductions in MSE relative to the OLS MSE under this condition. Table 2 shows reductions in MSE of about 16% for these estimators under the large sample size while for the small sample size, RMSE values in Table 1 show reductions in MSE of only 2-3%. Figure 2 displays the estimator MSE results from the unit normal, lognormal, and *t*-5df error distributions for the *n* = 50 sample size. Note that the MSE values for the N(0,1) condition in Figure 2 represent the same summary measures as the 0% contaminated data in Figure 1.

Y-Intercept estimator performance

Tables 3 and 4 present summary results for the estimation of population *Y*-intercept under the unit normal, contaminated normal, and nonnormal error distributions for the small and large sample sizes, respectively. Similar to the slope estimator, notice (for both the large and small sample sizes) the OLS *Y*-intercept estimator yields increases in MSE as the contamination in the data increases. Increased MSE values (as compared to the unit normal error distribution) for OLS are also reported for the nonnormal error distributions. For the small sample size, Table 3 reports the largest MSE for the OLS *Y*-intercept under the 30% data contamination condition with a value of 12.17. Unlike the small sample size, inspection of MSE values for the OLS *Y*-intercept in Table 4 reveals the largest MSE value falls under the lognormal error distribution with a reported value of 3.10.

Table 3 (continued). Summary Measures for Estimating Population *Y*-Intercept (α = 2.0)

Estimation	Error Distribution: Lognormal			
Method	Variance	Bias	MSE	RMSE
OLS:	1.97547147	1.61204928	4.57417434	0
LAD:	1.00661028	1.17225486	2.38079173	0.47951443
WIN10:	1.11938148	1.46240345	3.25800534	0.28773914
WIN20:	1.08117248	1.31667353	2.81480166	0.38463175
TLS:	1.38701960	1.39707222	3.33883039	0.27006928
MON:	0.40113847	-1.39933333	2.35927225	0.48421899
med (a_{ij}):	0.76859984	1.07407962	1.92224688	0.57976091
Conover:	1.17645633	1.53385117	3.52915574	0.22846060
Median-1:	0.69420592	1.14652836	2.00873319	0.56085338
Median-2:	0.74321559	1.31729607	2.47848453	0.45815696
Wtd. Mdn-1:	0.72207252	1.20571787	2.17582810	0.52432331
Wtd. Mdn-2:	0.76095906	1.37362165	2.64779550	0.42114242
	Error Distribution: *t*-5*df*			
OLS:	0.66246365	0.01522071	0.66269532	0
LAD:	0.89415823	0.00021390	0.89415828	-0.34927507
WIN10:	0.63214117	0.01585090	0.63239242	0.04572674
WIN20:	0.64452998	0.01393173	0.64472407	0.02711842
TLS:	0.77244571	0.00765923	0.77250437	-0.16570065
MON:	0.09712767	-1.60820000	2.68343491	-3.04927396
med (a_{ij}):	0.75091053	0.00304771	0.75091982	-0.13312981
Conover:	1.00145486	-0.02732700	1.00220163	-0.51231131
Median-1:	0.71263706	0.00291803	0.71264557	-0.07537438
Median-2:	0.67747509	0.00604977	0.67751169	-0.02235773
Wtd. Mdn-1:	0.67256536	0.06871781	0.67728750	-0.02201944
Wtd. Mdn-2:	0.63839635	0.06975511	0.64326212	0.02932449

Note: Tabled results are for the n=10 sample size. OLS: ordinary least squares; LAD: least absolute deviations; WIN10: 10% Winsorized regression; WIN20: 20% Winsorized regression; TLS: trimmed least squares; MON: monotonic regression; med (a_{ij}): median of pairwise intercepts; Conover: Conover Y-intercept; Median-1: median of (y_i - $\widehat{\beta}$ X_i), Theil slope; Median-2: pairwise average of (y_i - $\widehat{\beta}$ X_i), Theil slope; Wtd. Mdn-1: median of (y_i - $\widehat{\beta}$ X_i), weighted Theil slope; Wtd. Mdn-2: pairwise average of (y_i - $\widehat{\beta}$ X_i), weighted Theil slope.

Results for the monotonic regression *Y*-intercept estimator show extremely poor estimator performance under both the large and small sample sizes. Notice in both Tables 3 and 4, bias values in the *Y*-intercept for this estimator (under all conditions) are large and negative. These negative bias values indicate that the monotonic regression *Y*-intercept estimator consistently underestimates the population value of α = 2.0. For the large sample size, and looking across error distributions, MSE values for the monotonic regression *Y*-intercept estimator are generally larger than the OLS *Y*-intercept estimator under similar conditions. Thus, most RMSE values in Table 4 for monotonic regression are negative, indicative of a loss in estimator performance as compared to OLS. Similar to the monotonic regression *Y*-intercept estimator, the Conover *Y*-intercept (Conover, 1980) did not perform well. For the small sample size, the Conover *Y*-intercept shows reductions in MSE as compared to the OLS MSE baseline, but these reductions are not evidenced in Table 4 for the *n* = 50 sample size. For the large sample size, the Conover *Y*-intercept yields MSE values that are larger than the corresponding OLS MSE values. Thus, RMSE values in Table 4 for the Conover *Y*-intercept are negative.

Under the uncontaminated, unit normal error distribution, all robust and nonparametric *Y*-intercept estimators yield inflation in MSE as compared to OLS. These inflated MSE values are seen for both sample sizes in the two tables. After the monotonic regression and Conover *Y*-intercept estimators, the LAD and TLS estimators exhibit the most substantial loss in estimator performance.

Under the 10% data contamination all nonparametric and robust *Y*-intercept estimators show strong performance relative to OLS. Discounting the monotonic regression and Conover intercepts, all

Table 4. Summary Measures for Estimating Population *Y*-Intercept (α = 2.0)

Estimation	**Error Distribution: N(0,1) - 0% contamination**			
Method	Variance	Bias	MSE	RMSE
OLS:	0.08306252	-0.01121545	0.08318831	0
LAD:	0.13422339	-0.01715318	0.13451762	-0.61702554
WIN10:	0.08685171	-0.01015017	0.08695474	-0.04527592
WIN20:	0.08794651	-0.00868562	0.08802195	-0.05810480
TLS:	0.10876400	-0.00432918	0.10878275	-0.30766868
MON:	0.00027777	-1.94523347	3.78421102	-44.48969748
med (a_{ij}):	0.10683096	-0.01485406	0.10705160	-0.28685873
Conover:	0.43071144	-0.00363965	0.43072469	-4.17770702
Median-1:	0.09617646	-0.01358010	0.09636088	-0.15834637
Median-2:	0.08781186	-0.01081533	0.08792884	-0.05698550
Wtd. Mdn-1:	0.09576219	-0.01185850	0.09590282	-0.15284007
Wtd. Mdn-2:	0.08751654	-0.00873591	0.08759285	-0.05294667
	Error Distribution: N(0,1) - 10% contamination			
OLS:	0.72959431	-0.03815976	0.73105048	0
LAD:	0.14628716	-0.00046989	0.14628738	0.79989428
WIN10:	0.13127624	-0.00755356	0.13133329	0.82034990
WIN20:	0.15145369	-0.01096891	0.15157401	0.79266273
TLS:	0.17187647	-0.01173481	0.17201417	0.76470275
MON:	0.07048059	-1.56310204	2.51376858	-2.43857044
med (a_{ij}):	0.13115023	-0.00677069	0.13119607	0.82053760
Conover:	0.97688948	0.00936086	0.97697710	-0.33640170
Median-1:	0.12444677	-0.00521617	0.12447398	0.82973272
Median-2:	0.12000176	-0.00529341	0.12002979	0.83581191
Wtd. Mdn-1:	0.12399224	-0.00303475	0.12400145	0.83037909
Wtd. Mdn-2:	0.11985011	-0.00315264	0.11986005	0.83604409
	Error Distribution: N(0,1) - 30% contamination			
OLS:	2.22128658	0.00799708	2.22135053	0
LAD:	0.25062945	-0.01285950	0.25079481	0.88709805
WIN10:	1.01203716	0.02938913	1.01290088	0.54401574
WIN20:	0.72494948	0.02057523	0.72537282	0.67345414
TLS:	0.48621344	0.01482624	0.48643326	0.78101914
MON:	0.21056519	-0.79117633	0.83652517	0.62341596
med (a_{ij}):	0.25423256	-0.01322216	0.25440739	0.88547175
Conover:	2.36044148	0.01979859	2.36083347	-0.06279195
Median-1:	0.28495314	-0.00119235	0.28495456	0.87172013
Median-2:	0.30048783	0.00552801	0.30051839	0.86471366
Wtd. Mdn-1:	0.28922996	-0.00037240	0.28923010	0.86979538
Wtd. Mdn-2:	0.30494324	0.00740885	0.30499813	0.86269698

Y-intercept estimators under both sample sizes yield reductions in MSE of 75-83%. The *Y*-intercept nonparametric estimators show slight advantage over the robust estimators. Also, notice the TLS estimator shows weaker performance in the large sample size condition as compared to the *n* = 10 sample size cell for this moderately contaminated data condition.

For the 30% contamination, the LAD Y-intercept estimator and the Y-intercept estimator based on the median a_{ij} values yield the lowest MSE values with the other nonparametric *Y*-intercepts all very close. These results hold for both the small sample size MSE values in Table 3 and for the *n* = 50 sample size presented in Table 4.

Under the lognormal error distribution, all estimators of *Y*-intercept had difficulty in recovering the population value of $\alpha = 2.0$. Note the large bias values for the estimators under this condition, suggesting large discrepancies between the means for the estimators and the population value. The median a_{ij} estimator showed the strongest relative performance under both sample sizes. The nonparametric techniques using the median of the (y_i - $\widehat{\beta}$ X_i) terms (using either the Theil slope or weighted Theil slope) also yield relative strong estimator performance with RMSE values of 0.64 for the *n* = 50 sample size. For the large sample size, the LAD *Y*-intercept estimator was also competitive.

Table 4 (continued). Summary Measures for Estimating Population *Y*-Intercept (α = 2.0)

Estimation	**Error Distribution: Lognormal**			
Method	Variance	Bias	MSE	RMSE
OLS:	0.38500603	1.64740505	3.09894942	0
LAD:	0.14386289	1.01840542	1.18101249	0.61889908
WIN10:	0.16699356	1.38154934	2.07567214	0.33020135
WIN20:	0.15463921	1.28493558	1.80569865	0.41731910
TLS:	0.21240544	1.25722778	1.79302712	0.42140807
MON:	0.05224150	-1.76884571	3.18105666	-0.02649519
med (a_{ij}):	0.09562958	0.90239609	0.90994829	0.70636878
Conover:	0.64892005	1.64056750	3.34038177	-0.07790781
Median-1:	0.07920296	1.01310474	1.10558416	0.64323904
Median-2:	0.08361419	1.22682737	1.58871960	0.48733607
Wtd. Mdn-1:	0.08168383	1.01704798	1.11607041	0.63985523
Wtd. Mdn-2:	0.08588540	1.23072013	1.60055744	0.48351611
	Error Distribution: *t*-5*df*			
OLS:	0.12955685	-0.00977434	0.12965239	0
LAD:	0.13381946	-0.00339975	0.13383102	-0.03222951
WIN10:	0.11127929	-0.00778139	0.11133984	0.14124342
WIN20:	0.11042154	-0.00831868	0.11049074	0.14779248
TLS:	0.12116747	-0.00114135	0.12116877	0.06543357
MON:	0.00249143	-1.90627265	3.63636686	-27.04704888
med (a_{ij}):	0.11836755	-0.00596709	0.11840316	0.08676456
Conover:	0.56742654	-0.02887466	0.56826029	-3.38295273
Median-1:	0.11431813	-0.00951134	0.11440859	0.11757436
Median-2:	0.10786731	-0.00879750	0.10794471	0.16742983
Wtd. Mdn-1:	0.11455410	-0.00583959	0.11458820	0.11618908
Wtd. Mdn-2:	0.10789910	-0.00524933	0.10792666	0.16756907

Note: Tabled results are for the n=50 sample size. OLS: ordinary least squares; LAD: least absolute deviations; WIN10: 10% Winsorized regression; WIN20: 20% Winsorized regression; TLS: trimmed least squares; MON: monotonic regression; med (a_{ij}): median of pairwise intercepts; Conover: Conover Y-intercept; Median-1: median of (y_i - $\widehat{\beta}$ X_i), Theil slope; Median-2: pairwise average of (y_i - $\widehat{\beta}$ X_i), Theil slope; Wtd. Mdn-1: median of (y_i - $\widehat{\beta}$ X_i), weighted Theil slope; Wtd. Mdn-2: pairwise average of (y_i - $\widehat{\beta}$ X_i), weighted Theil slope.

For the *t*-5df error distribution, Tables 3 and 4 report only modest reductions in MSE as compared to the OLS MSE benchmark. Table 3 shows increases in MSE for the LAD estimator as well as for most of the other *Y*-intercept estimators. For the large sample size, the nonparametric pairwise methods demonstrate slightly smaller MSE as compared to OLS, with the Winsorized regression methods exhibiting good performance. The LAD *Y*-intercept estimator exhibited poor performance with a MSE value slightly larger than that of the OLS *Y*-intercept estimator.

Discussion

Findings in the present study have substantive implications for educational researchers and research methodologists. The poor performance of OLS estimation under the contaminated data conditions and nonnormal error distributions serves to reaffirm both the importance of assessing underlying assumptions as part of any regression analysis and the need for alternatives to OLS regression. This study has also replicated past findings that have suggested that when the appropriate assumptions are met, OLS regression is the method of choice. Our results have shown, under all sample sizes and for estimation of both population slope and *Y*-intercept, the OLS estimator yields the lowest MSE under ideal conditions.

Findings in the present study have demonstrated the merits of alternatives to OLS regression under non-ideal conditions. Our results indicate that estimator performance is dependent upon the nature of the error distribution. Figure 1 shows that under mild (10%) data contamination there is no real preference for one alternative slope estimator over another. When the degree of data contamination was increased to 30%, the LAD slope estimator moderately outperformed the other slope estimators by yielding the smallest MSE.

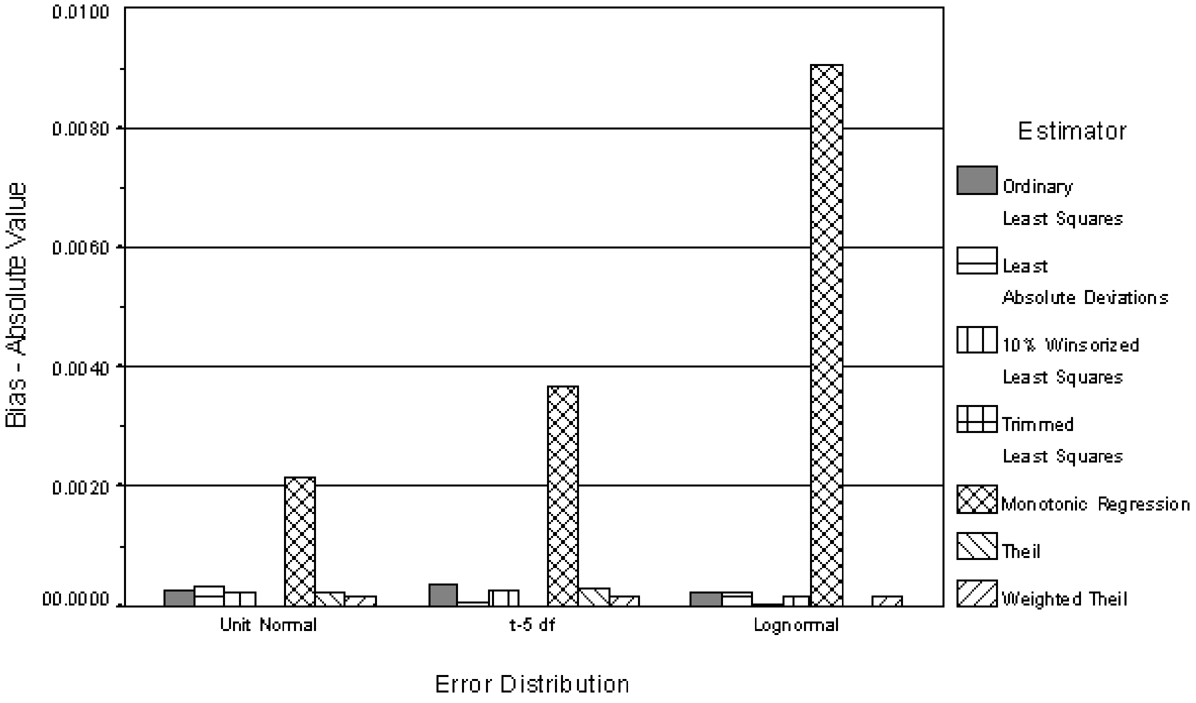

Figure 3. Bias in estimation of population slope for normal and nonnormal error distributions. Results charted are for the *n* = 50 sample size.

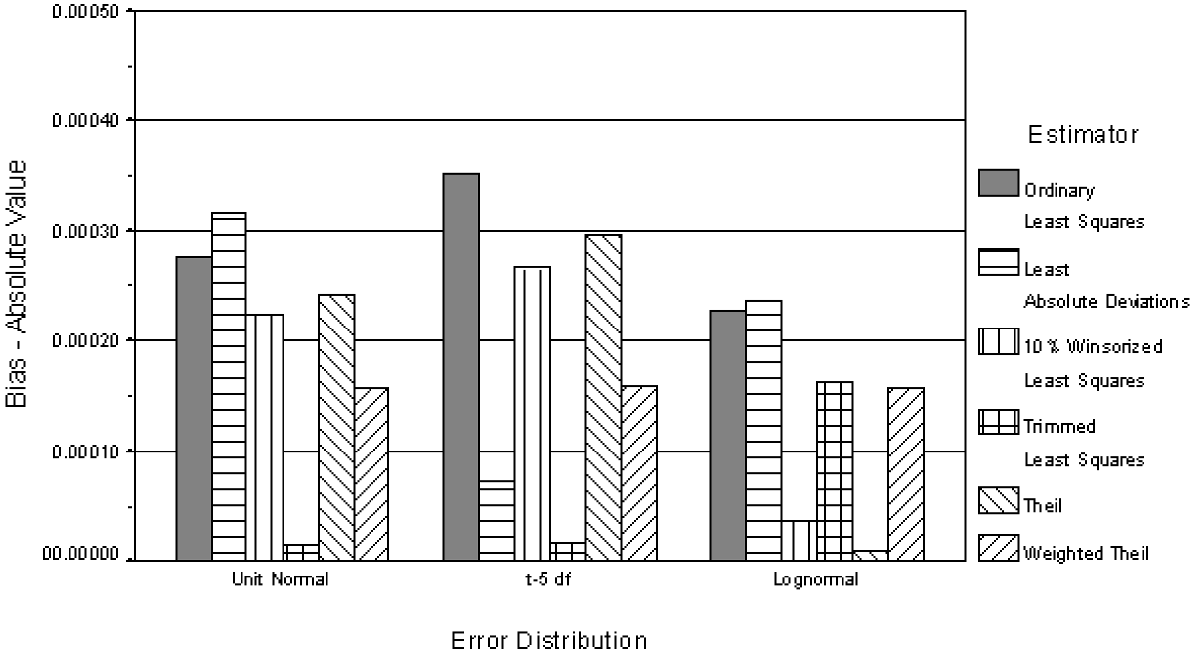

Figure 4. Bias in estimation of population slope for normal and nonnormal error distributions – monotonic regression slope estimator removed. Results charted are for the *n* = 50 sample size.

For the case of nonnormal error distributions, our results demonstrate that the symmetry of the error distribution substantially impacts estimator performance. Figure 2 illustrates that when the error distribution is nonnormal and symmetric (*t*-5df errors) the robust LAD estimator, which demonstrated strong performance under the contaminated normal conditions, is not a desirable choice. Under this condition, the Winsorized least squares and nonparametric methods employing medians of pairwise slopes (Theil and weighted Theil) exhibited superior performance. Figure 2 also demonstrates that when the error distribution is skewed, the nonparametric Theil methods yield very strong performance.

The monotonic regression and the TLS methods investigated in this study were generally not competitive. The poor results obtained for monotonic regression are not entirely unexpected. In their proposal of this alternative method of regression, Iman and Conover (1979) caution the use of this method under situations in which there are outliers in the observed data. They recommend this method only for situations in which observed data exhibits a monotonically increasing or decreasing trend - curvilinear data. Additionally other investigators have found the rank transformation procedure to be problematic (McKean & Vidmar, 1994; Sawilowsky, Blair & Higgins, 1989). Our results have served to substantiate these findings with empirical evidence of the unacceptability of rank transformation in the form of monotonic regression with respect to bias and RMSE. Large bias values in the summary tables reflect monotonic regression's inability to recover the true population values under our data conditions.

The results for monotonic regression in this study also provide valuable insight into the use of MSE as a sole indicator of the quality of parameter estimation. A useful estimator is one in which both bias and variance are minimized. Figure 2 shows monotonic regression as having very low MSE under the N(0,1) and *t*-5*df* error distributions. The small values for monotonic regression in this figure can be misleading with respect to choice of estimator. Table 2 reports bias values for monotonic regression that are approximately 10 times larger than the bias values for the other slope estimators under each condition. We present Figure 3 which charts bias values for the various estimators under the unit normal, *t*-5df, and lognormal error distributions for the n = 50 sample size. When considering bias as a measure of the quality of parameter estimation, this figure readily demonstrates that monotonic regression is not an optimal estimator under the conditions of our study. For clarity of presentation, we also present Figure 4 which shows the same results as in Figure 3, with the monotonic regression estimator removed. With respect to assessing the quality of parameter estimation, our recommendation for methodological researchers is to evaluate MSE with the caveat that bias should also be simultaneously considered.

The TLS estimator was included in the study to address the issue of case deletion, an approach frequently adopted in applied scenarios in which there are outliers in the observed data. For the TLS estimator, data points corresponding to the 10% largest positive and the 10% largest negative residuals from an initial OLS regression were deleted. Under the contaminated data conditions in this study, the case deletion approach to estimation of population slope did not generate unattractive results, although comparison of the TLS slope estimator in Tables 1 and 2 suggests the performance of this estimator is sample size dependent. Under the small sample size, the TLS slope estimator performed well under the 10% data contamination, but not under the 30% contamination condition. For the larger sample size, Table 2 reports weaker performance under the moderate contamination condition (with respect to the other slope estimators) but stronger performance under the more extreme 30% data contamination condition. While the performance of the TLS slope estimator was not unreasonable, for both the 10% and 30% contamination conditions, robust and nonparametric methods (discounting monotonic regression) which utilize all the available data outperformed TLS. Additionally, for the conditions in which the distribution of errors was nonnormal, the TLS slope estimator was not competitive. Figure 4 shows very low bias for this estimator, but the variance for this slope estimator tends to be inflated. Thus the MSE values for TLS shown in Figure 2 tend to be higher than some of the other slope estimators. Our results demonstrate that methods which utilize all available data, but are resistant to outlying values, provide more accurate long run estimates of true population values. This conclusion is consistent with previous research in resistant methods of regression (Birkes & Dodge, 1993; Rousseeuw & Leroy, 1987).

With respect to the estimators investigated in the present study, our results have demonstrated that the nonparametric approaches based on the Theil method are very strong alternatives to OLS regression. This conclusion holds for the small sample size investigated here as well for the large sample size. This study has demonstrated that this approach provides accurate estimates of true population parameters under both outlier contaminated data conditions and under nonnormal error distributions. While these median based nonparametric methods did not outperform the LAD estimator under the heavily contaminated conditions (30% outliers) they were nearly as strong as the LAD regression method under this condition. Under the nonnormal error conditions, no estimator outperformed the Theil methods.

Additionally, under the lognormal error distribution, the Theil based regression methods showed superior performance. The Theil based estimation methods were never the worst, sometimes nearly the best and in some cases the best methods for parameter estimation under the simple linear model.

Median based nonparametric methods for parameter estimation have found little attention in social science research and deserve further consideration by applied researchers. This study has demonstrated that the Theil based regression methods provide strong parameter estimation under a variety of non-ideal conditions. There is also literature available that provides an extension of this method, using a weighted form of the Theil method, to multiple regression (Birkes & Dodges, 1993). Hypothesis testing procedures have been developed for testing both model adequacy and individual regression coefficients (for reviews see Tam, 1996; Birkes & Dodge, 1993). Finally, the modified form of the Theil regression method has been incorporated into at least one of the commonly available applied statistics packages (RANK REGRESSION in Minitab) available for researchers. performs nonparametric regression estimation based on the weighted Theil method.

We recommend the following approach to applications in educational research. First, data analyses should always involve checking for outliers in the observed data and testing the underlying assumptions under OLS estimation. Secondly, researchers may be well advantaged to routinely estimate regression parameters using both OLS and alternative methods when conducting regression based analyses. Should the assumptions of normality and homoscedasticity hold, researchers might adopt and report OLS estimates in their findings. Under applied settings in which the OLS assumptions are not tenable, researchers may turn to estimates of population values using an outlier-resistant method.

The present study only considered estimators under the simple linear regression situation. Further study might compare the performance of nonparametric median based estimators against robust regression estimators under the multiple regression. In addition, future studies might be warranted to compare the nonparametric median based estimators against robust regression methods such as M-regression (Birkes & Dodge, 1993), iteratively reweighted least squares (Holland & Welsch, 1977), or least median squares regression (Rousseeuw & Leroy, 1987). These robust methods are known to be resistant to more extreme forms of data contamination such as leverage points. Finally, additional research investigating power and Type I error rates using nonparametric median based methods would be useful to more fully characterize the behavior of these methods under hypothesis testing paradigms.

Correspondence should be directed to:
Jonathan Nevitt
1228 Benjamin Building
University of Maryland
College Park, MD 20742-1115.
E-mail: jnevitt@wam.umd.edu

References

Aptech Systems. (1996). *GAUSS System and Graphics Manual*. Maple Valley, WA: Aptech Systems, Inc.

Birkes, D., & Dodge, Y. (1993). *Alternative Methods of Regression*. New York, NY: Wiley.

Conover, W. J. 1980. *Practical Nonparametric Statistics* (2nd edition). New York, NY: Wiley.

Dietz, E. J. (1987). A comparison of robust estimators in simple linear regression. *Communication in Statistics-Simulation, 16,* 1209-1227.

Dietz, E. J. (1989). Teaching regression in a nonparametric statistics course. *The American Statistician, 43,* 35-40.

Draper, N., & Smith, H. (1981). *Applied Regression Analysis* (2nd edition). New York, NY: Wiley.

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical Distributions* (2nd edition). New York, NY: Wiley.

Hamilton, L. C. (1992). *Regression with graphics, A second course in applied statistics*. Pacific Grove, CA: Brooks/Cole.

Holland, P. H., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least squares. *Communications Statistics: Theory and Methods*, *6*, 813-827.

Hussain, S. S., & Sprent, P. (1983). Nonparametric regression. *Journal of the Royal Statistical Society, series A*, *146*, 182 - 191.

Iman, R. L., & Conover, W. J. (1979). The use of rank transformation in regression. *Technometrics, 21*(4), 499-509.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals. *Annals of Mathematical Statistics*, *43*, 1449-1458.

McKean, J. W., & Vidmar, T. J. (1994). A comparison of two rank-based methods for the analysis of linear models. *The American Statistician*, *48*(3), 220-229.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research* (2nd edition). Fort Worth, TX: Harcourt Brace Jovanovich.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York, NY: Wiley.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*, *14*(3), 255-267.

Stone, C. J. (1996). *A Course in Probability and Statistics*. Belmont, CA: Duxbury.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, *12,* 85-91.

Tam, H. P. (1996, April). *A review of nonparametric regression techniques.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Yale, C., & Forsythe, A. B. (1976). Winsorized regression. *Technometrics, 18,* 291 - 300.

Analysis Options for Testing Group Differences on Ordered Categorical Variables: An Empirical Investigation of Type I Error Control and Statistical Power

Jeffrey D. Kromrey **Kristine Y. Hogarty**
University of South Florida

Type I error control and statistical power of four methods of testing group differences on an ordered categorical response variable were evaluated in a Monte Carlo study. Data were analyzed using the independent means *t*-test, the chi-square test of homogeneity, the delta statistic, and a cumulative logit model. The number of categories of the response variable, sample size, population distribution shape, and effect size were examined. These experimental conditions were crossed with each other providing a total of 192 conditions. The independent means *t*-test provided the best control of Type I error, but was rarely the most powerful. For the 5-point response scale, the chi-square was most often the most powerful. Results varied for the 7-point response scale. Small power differences (in many instances) among these procedures suggest that researchers' choices should be driven by the interpretations that are appropriate for the research questions being addressed.

Response variables that are measured as ordered categories, such as Likert scale and other rating scale items, present a variety of analysis options for researchers. For example, in testing for the equality of two groups on such a response variable, the data are usually analyzed using either a Pearsonian chi-square test of homogeneity or a test for the equality of population means such as the independent means *t*-test. Implicit in the former analysis is the treatment of the response variable as nominal-level measurement, while the latter analysis implies an assumption of interval-level data. In between these two extremes are analysis options that are infrequently seen in applied educational research, specifically, logistic regression models (Agresti, 1996; Agresti & Finlay, 1997) and ordinal indices of association (Cliff, 1996a). Arguments about the relationship between levels of measurement and appropriate statistical analyses have been ongoing since Stevens' (1951) classic work, and, no doubt, will continue in the future.

Although the present paper is not intended to directly address the logical arguments related to Stevens' levels of measurement issues, the influence of his work is unavoidable. For example, recent arguments on the level-of-measurement/appropriate-statistics issue have been advanced by Davidson and Sharma (1988) and by Velleman and Wilkinson (1993). Rather than examining such analysis issues in terms of "appropriate statistics," the issues surrounding the analysis of ordered categorical data may be productively addressed in terms of Type I error control and statistical power. For example, Cliff (1996a) has argued that ordinal measures of association such as Tau and delta are useful both descriptively and inferentially because of their robustness properties when compared to traditional parametric tests such as the independent means *t*-test. Similarly, Agresti (1989) suggested that researchers may realize power advantages in the use of cumulative logit models rather than Pearsonian chi-square tests when testing hypotheses about ordered categorical data. Unfortunately, neither Cliff nor Agresti presented empirical evidence of the magnitude of power differences or the extent of improvement in robustness when these ordinal-level statistics are used.

It is important to recognize that different statistical null hypotheses are tested with each of these procedures. For example, the independent means *t*-test provides a test of the null hypothesis of equivalence of population means and the chi-square test of homogeneity tests the equivalence of the population proportions at each level of the response variable. In contrast, the G^2 statistic used in testing the cumulative logit model provides a test of the null hypothesis of equal cumulative log odds, while the delta statistic is used to test equivalence of probabilities of scores in each group being larger than scores in the other (the property that Cliff (1993) referred to as "dominance"). However, as Cliff (1993, 1996a) has pointed out, despite the differences in statistical null hypotheses tested, each of these procedures may be used to test the same, conceptual research hypothesis (e.g., "the two groups respond differently on the dependent variable").

The purpose of the present study was to empirically compare the Type I error control and statistical power of four tests of group differences on ordered categorical response data: a parametric test of mean differences (independent means *t*-test), the

Pearsonian chi-square test of homogeneity, the cumulative logit model recommended by Agresti (1989, 1996), and the delta statistic recommended by Cliff (1993, 1996a). Such a comparison was made for a variety of sample sizes and distribution shapes likely to be encountered in educational research. Although previous research has investigated the Type I error control and statistical power of parametric and nonparametric statistics (primarily comparisons of the *t*-test and the Wilcoxon-Mann-Whitney *U* test), such comparisons have typically been conducted using continuous outcome variables (see, for example, Blair & Higgins, 1980, 1985). A notable exception is the recent work of Nanna and Sawilowsky (1998), comparing the *t*-test with the Wilcoxon rank-sum test based on resampling from actual data obtained on ordered categorical variables.

Test Statistics Examined

Four test statistics were examined in this study. These test statistics will be presented in reference to the set of data presented in Table 1. These data, consisting of responses to a 5-point Likert item, were obtained from six members of an experimental group and ten members of a control group. The research question to be addressed is whether the two populations from which the samples were obtained differ in their response to this item.

Table 1. Sample of Two Groups' Responses to a 5-Point Likert Item

Control Group	Experimental Group
1	1
1	2
2	3
2	4
2	4
3	5
3	
3	
4	
5	

Independent Means t-test. The independent means *t*-test is used to test the null hypothesis of equivalent population means (H_O: $\mu_1 = \mu_2$). The test statistic is given by

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{[(n_1 - 1) + (n_2 - 1)]S_{pl}}$$

where $(\bar{X}_1 - \bar{X}_2)$is the difference in sample means, n_1 and n_2 are the sample sizes, S_{pl} is a pooled estimate of the population standard deviation given by

$$S_{pl} = \sqrt{\frac{(SS_1 + SS_2)}{(n_1 + n_2 - 2)}}$$

and SS_1 and SS_2 are the sums of squares computed in each of the samples. The obtained value of this test statistic is compared to the sampling distribution of *t* with degrees of freedom equal to $n_1 + n_2 - 2$.

For the sample of data presented in Table 1, the means for the experimental and control groups are 3.167 and 2.600, respectively, and the pooled variance estimate is 1.802. The obtained value of *t* for these data is -0.817, and the probability associated with this value under the null hypothesis is 0.427. The *t*-test, thus, fails to reject the null hypothesis of equal population means.

Chi-Square Test of Homogeneity. In contrast to the *t*-test which compares sample means, the Pearsonian chi-square test of homogeneity tests the null hypothesis of equivalent population proportions in each response category (H_0: $\pi_{1j} = \pi_{2j}$, for all *j*). For computation of the chi-square statistic, the data may be arranged in a contingency table as illustrated in Table 2. The sample value of this test statistic is given by

$$\chi^2 = \frac{\sum_i \sum_j (O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency in cell *ij* of the contingency table, E_{ij} is the expected frequency in the cell under the null hypothesis of homogeneity, and the summation is over all of the cells in the table. The obtained value of χ^2 is compared to the sampling distribution of χ^2 with degrees of freedom equal to $(n_{rows} - 1)(n_{cols} - 1)$.

Table 2. Contingency Table for the Sample Data

	Response Category				
Group	1	2	3	4	5
1	2	3	3	1	1
2	1	1	1	2	1

For the sample of data presented in Table 1, the obtained value of chi-square is 1.778. In comparison to a four degree of freedom chi-square sampling distribution, this value has a probability of 0.777 under the null hypothesis. Thus, like the *t*-test, the chi-square test fails to reject the null hypothesis of equal population proportions for each level of the response variable.

Delta Statistic. Cliff (1993, 1996a) has proposed the use of the delta statistic for testing null hypotheses about group differences on ordinal level measurements. The population parameter for which such tests are intended is the probability that a randomly selected member of one population has a higher response than a randomly selected member of the second population, minus the reverse probability. That is,

$$\text{delta} = \Pr(x_{i1} > x_{j2}) - \Pr(x_{i1} < x_{j2}) ,$$

where x_{i1} is a member of population one and x_{j2} is a member of population two.

A sample estimate of this parameter can be obtained by enumerating the number of occurrences of a sample one member having a higher response value than a sample two member, and the number of occurrences of the reverse. This gives the sample statistic

$$d = \frac{\#(x_{i1} > x_{j2}) - \#(x_{i1} < x_{j2})}{n_1 \; n_2}$$

This statistic, and inferential methods associated with it, are readily addressed by considering the data in an arrangement called a dominance matrix. This n_1 by n_2 matrix has elements taking the value of 1 if the row response is larger than the column response, -1 if the row response is less than the column response, and 0 if the two responses are identical. The sample value of d is simply the average value of the elements in the dominance matrix. The dominance matrix for the Table 1 data is presented in Table 3. The row and column marginals of this table provide mean values of the elements in the respective rows and columns of the matrix. These marginals are used in the inferential statistics associated with d. The null hypothesis tested in such inferential statistics (representing no relationship between the grouping variable and the response variable) is that delta is equal to zero.

Table 3. Dominance Matrix for the Sample Data

	1	2	3	4	4	5	*di.*
1	0	-1	-1	-1	-1	-1	-0.833
1	0	-1	-1	-1	-1	-1	-0.833
2	+1	0	-1	-1	-1	-1	-0.500
2	+1	0	-1	-1	-1	-1	-0.500
2	+1	0	-1	-1	-1	-1	-0.500
3	+1	+1	0	-1	-1	-1	-0.167
3	+1	+1	0	-1	-1	-1	-0.167
3	+1	+1	0	-1	-1	-1	-0.167
4	+1	+1	+1	0	0	-1	0.333
5	+1	+1	+1	+1	+1	0	0.833
d.j	0.8	0.3	-0.3	-0.7	-0.7	-0.9	-0.250

Cliff (1996b) presented three methods of inference for d. The first method uses an "unbiased" estimate of the variance of d. This estimate is given by

$$S_d^2 = \frac{n_2^2 \sum_i (d_{i.} - d)^2 + n_1^2 \sum_j (d_{.j} - d)^2 + \sum_i \sum_j (d_{ij} - d)^2}{n_1 n_2 (n_1 - 1)(n_2 - 1)}$$

where $d_{i.}$ is the marginal value of row i, $d_{.j}$ is the column marginal of column j, and d_{ij} is the value of element ij in the matrix.

For the sample data in Table 1, the value of d is -0.25 and the value of S_d^2 is 0.098. The square root of this variance is used as the denominator of a z statistic: $z_{\text{unbiased}} = d / S_d$

For the sample data, the value of z is -0.798, yielding a probability under the null hypothesis of 0.425. The unbiased test fails to reject the null hypothesis of delta = 0.

The second method of inference for d uses a “consistent” estimate of the variance:

$$S_{dc}^2 = \frac{(n_2 - 1)S_{di.}^2 + (n_1 - 1)S_{d.j}^2 + S_{dij}^2}{n_1 \; n_2}$$

where $S_{di.}{}^2 = \Sigma(d_{i.} - d)^2/(n_1\text{-}1)$, $S_{d.j}{}^2 = \Sigma(d_{j.} - d)^2/(n_2\text{-}1)$, and $S_{dij}{}^2 = \Sigma\Sigma(d_{ij} - d)^2 / [(n_1 - 1)(n_2 - 1)]$.

As with the "unbiased" estimate of variance, the square root of this "consistent" estimate of the variance of d can be used as the denominator of a z statistic: $z_{\text{consistent}} = d / S_{dc}$.

For the Table 1 data, the value of $S_{dc}{}^2$ is 0.106, yielding a value for $z_{\text{consistent}}$ of -0.768, with a probability under the null hypothesis of 0.443. The conclusion with this sample is the same as that reached with the unbiased test, that is, a failure to reject the null hypothesis of delta = 0.

The final method of inference regarding d uses S_{dc} to construct an asymmetric confidence interval around the sample value of d. When such an interval does not include the value of zero, the null hypothesis of delta = 0 can be rejected. The limits of this asymmetric confidence interval are given by

$$\frac{d - d^3 \pm Z_{\alpha/2} S_{dc}[(1 - d^2)^2 + Z_{\alpha/2}^2 S_{dc}^2]}{1 - d^2 + Z_{\alpha/2}^2 S_{dc}^2}$$

where $Z_{\alpha/2}$ is the normal deviate corresponding to the $(1 - \alpha/2)^{\text{th}}$ percentile of the normal distribution.

For the Table 1 data, the lower limit of the 95% confidence interval is -0.713, and the upper limit is 0.364. Because this interval contains the value of zero, the null hypothesis is not rejected at the .05 level.

Cliff (1996a) has pointed out that the well-known Mann-Whitney-Wilcoxon statistic can also provide a test of delta = 0 (because d and U are related by $d = 2U/[n_1 n_2 - 1]$). However, the rank test is not recommended by Cliff because it is actually testing for the equivalence of the two groups’ distributions rather than focusing on the parameter delta.

Cumulative Logit Models. Logistic regression is a technique used to construct models of the probabilities of values of categorical variables. In its simple, binary form, a model relating the probability of response 1 as a function of an explanatory or predictor variable X, can be thought of as:

$$\pi = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

where π is the probability of response 1, exp is the exponential function or the antilog function of the natural logarithms, and α and β are regression parameter estimates. This equation describes an S-shaped curve called the logistic regression model. However, the relationship between π and X is often expressed as logits, yielding the linear logit model: $\text{logit}(\pi) = \log[\pi/(1 - \pi)] = \alpha + \beta X$.

A relatively minor modification of this linear logit model can be used with ordinal response variables having more than two levels (Agresti, 1990; McCullough & Nelder, 1989). With a response variable having J ordered categories, the probability associated with any category j can be denoted π_j, where $\Sigma\pi_j = 1$. The cumulative logit model is formed from logits of cumulative probabilities. For example, the probability of a response less than or equal to an arbitrary category j is given by

$$\text{logit}[\Pr(Y \le j)] = \log[(\pi_1 + ... + \pi_j)/(\pi_{j+1} + ... + \pi_J)]$$
$$= \alpha_j + \beta X$$

This model treats the response as binary by forming the cumulative probability over the first j categories, and the remaining (J - j) categories. This model has J - 1 values of α_j, one for each of the adjacent category differences. The parameter of primary interest in this model is β, which describes the relationship of the X variable to the cumulative probabilities of response. When β is equal to zero, the variable X is not related to the response variable.

Two methods for testing the null hypothesis that $\beta = 0$ are available. The first method uses the standard error of the sample estimate of β to form a z test (or an equivalent chi-square test), called the Wald test. The standard error of β is obtained from the inverse of the information matrix, the matrix of second partial derivatives of the log likelihood function. For the Table 1 data, the sample estimate of β is -0.834, with a standard error of 0.936. The value of the Wald z test is -0.891, with a probability of 0.373 under the null hypothesis. As with the other tests examined thus far, the Wald test fails to reject the null hypothesis of $\beta = 0$.

The second method of testing the null hypothesis that $\beta = 0$ is with a likelihood ratio test. This test is based on the likelihood ratio statistic:

$$G^2 = 2\Sigma_j O_j \log(O_j/E_j),$$

where O_j and E_j are the observed and expected counts, respectively, and log is the natural logarithm.

The likelihood ratio test of $\beta = 0$ is obtained as the difference in the values of G^2 for the model that includes X, and the model that does not (i.e., a model with intercepts only). This difference in G^2 values is distributed as a chi-square with a single degree of freedom. For the sample data, the value of G^2 for the model that includes X is 49.808, while that for the intercept only model is 50.586. The difference in these G^2 values is 0.737, which has a probability of 0.391 under the null hypothesis. Thus, consistent with the other tests conducted on these data, the likelihood ratio tests does not reject the null hypothesis that $\beta = 0$.

Method

This research was a Monte Carlo study designed to provide an empirical comparison of the Type I error control and statistical power of the four methods of testing group differences on an ordered categorical response variable. Two of these tests are frequently used with ordered categorical data: the independent means t-test and Pearsonian chi-square test of homogeneity. The other two methods, the cumulative logit model and the delta statistic, have been recommended for the analysis of ordinal level data because of increased power (relative to the chi-square test) or increased robustness (relative to tests of mean differences). Although four methods for testing group differences were examined in this study, a total of seven statistical tests were compared (i.e., three tests associated with the d statistic and two tests associated with the logistic regression method).

All of the conditions simulated provided tests of differences between two groups on an ordered categorical dependent variable. Four factors were investigated in the Monte Carlo study: number of categories of the response variable, sample size, population distribution shape, and effect size. The number of categories of the response variable was examined at two levels (5-category and 7-category responses). Six sample sizes were examined (equal sizes of 10:10, 30:30, and 100:100; and unequal sizes of 10:30, 10:100, and 30:100). Four population distribution shapes were investigated (a uniform response distribution, a moderately skewed distribution, a highly skewed distribution, and a unimodal symmetric distribution). Finally, small, medium and large population effect sizes (Cohen, 1988) were examined as well as a null condition. These experimental conditions were crossed with each other providing a total of 192 conditions examined.

Table 4. Type I Error Rate Estimates for 5 Point Response Scale at nominal α = .05

Marginal	Sample	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Square	t-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1	10, 10	0.033	0.055	0.081	0.073	0.043	0.057	0.068
	10, 30	0.042	0.048	0.071	0.067	0.049	0.052	0.057
	10,100	0.045	0.054	0.084	0.082	0.067	0.055	0.059
	30, 30	0.046	0.050	0.057	0.056	0.046	0.053	0.054
	30,100	0.051	0.050	0.056	0.055	0.049	0.051	0.052
	100,100	0.055	0.051	0.054	0.054	0.050	0.052	0.053
6:1:1:1:1	10, 10	0.015	0.050	0.073	0.066	0.046	0.039	0.067
	10, 30	0.046	0.049	0.078	0.076	0.060	0.038	0.061
	10,100	0.053	0.048	0.086	0.085	0.075	0.031	0.059
	30, 30	0.043	0.054	0.060	0.059	0.051	0.053	0.057
	30,100	0.048	0.050	0.056	0.056	0.050	0.047	0.051
	100,100	0.050	0.046	0.048	0.048	0.046	0.046	0.047
16:1:1:1:1	10, 10	0.004	0.033	0.084	0.038	0.035	0.004	0.083
	10, 30	0.038	0.036	0.115	0.114	0.108	0.026	0.070
	10,100	0.083	0.046	0.132	0.132	0.126	0.036	0.101
	30, 30	0.018	0.048	0.054	0.053	0.050	0.034	0.055
	30,100	0.050	0.046	0.062	0.062	0.059	0.035	0.053
	100,100	0.049	0.048	0.051	0.051	0.049	0.047	0.052
1:2:4:2:1	10, 10	0.030	0.049	0.077	0.071	0.042	0.047	0.063
	10, 30	0.048	0.049	0.074	0.071	0.055	0.052	0.057
	10,100	0.049	0.051	0.079	0.078	0.063	0.053	0.054
	30, 30	0.046	0.050	0.058	0.056	0.048	0.050	0.053
	30,100	0.047	0.050	0.057	0.056	0.051	0.051	0.052
	100,100	0.053	0.050	0.053	0.052	0.050	0.051	0.052

Programming for the Monte Carlo Study. The program for the Monte Carlo study was written in SAS/IML version 6.12. The data were generated using uniform random numbers on the zero to one interval (the SAS RANUNI function). A separate seed value was used for each execution of the simulation and the accuracy of the program code was verified using benchmark data sets. To simulate samples, a separate series of random numbers was generated for each of the two groups. The observations were then assigned to values of the ordered categorical response variable based upon the value of the random number.

For example, with a 5-point response scale with equal marginals and an effect size of zero, two series of random numbers were generated. Observations with random numbers between zero and .20 were assigned to the first category of the response variable, those with random numbers between .20 and .40 were assigned to the second category, etc. This procedure yields tables in which the expected proportion in each cell is equal, providing a uniform response across the five categories and the two groups.

The marginal skewness of the response variable was controlled by assigning larger or smaller ranges of the uniform random numbers to each of the ordered categories. For example, to simulate a 60:10:10:10:10 marginal distribution, 60% of the observations were assigned to the first value of the response variable, and 10% to each of the other values. Four marginal distributions were examined in this study. The equal marginal condition provided equal proportions at each level of the response variable. A slightly skewed marginal distribution was produced by generating data in which 60% of the observations were in the first category of the response variable, and the remaining 40% were evenly dispersed over the other values. Similarly, a more highly skewed marginal was produced by generating data in which 80% of the observations were at the first value and the remaining 20% were evenly distributed over the remaining values. Finally, a unimodal symmetric distribution was generated with the mode at the middle of scale and descending proportions of observation for scale values towards the scale endpoints.

Non-null effects were generated by assigning observations to response categories in proportions that differed from the products of the row and column marginal proportions. By varying the extent of discrepancy between the products of the marginals and the actual proportions of observations, effect sizes corresponding to *w* values of 0.10, 0.30, and 0.50 (Cohen, 1988) were produced.

Table 5. Type I Error Rate Estimates for 7 Point Response Scale at nominal α = .05

Marginal	Sample	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Square	t-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1:1:1	10, 10	0.020	0.053	0.077	0.068	0.041	0.058	0.066
	10, 30	0.038	0.049	0.070	0.066	0.048	0.055	0.057
	10,100	0.044	0.050	0.084	0.082	0.067	0.054	0.055
	30, 30	0.047	0.052	0.061	0.059	0.048	0.055	0.057
	30,100	0.050	0.051	0.060	0.058	0.052	0.051	0.053
	100,100	0.050	0.048	0.050	0.049	0.047	0.049	0.049
9:1:1:1:1:1:1	10, 10	0.006	0.049	0.072	0.064	0.044	0.037	0.065
	10, 30	0.043	0.048	0.077	0.074	0.061	0.037	0.060
	10,100	0.061	0.044	0.088	0.086	0.076	0.031	0.058
	30, 30	0.032	0.052	0.058	0.057	0.048	0.051	0.055
	30,100	0.048	0.050	0.058	0.058	0.054	0.049	0.054
	100,100	0.045	0.049	0.048	0.048	0.046	0.047	0.048
24:1:1:1:1:1:1	10, 10	0.001	0.028	0.083	0.038	0.034	0.004	0.082
	10, 30	0.037	0.040	0.118	0.117	0.111	0.030	0.069
	10,100	0.096	0.041	0.125	0.125	0.121	0.031	0.093
	30, 30	0.008	0.046	0.056	0.056	0.050	0.035	0.057
	30,100	0.048	0.042	0.063	0.062	0.060	0.035	0.051
	100,100	0.037	0.051	0.052	0.052	0.051	0.048	0.052
1:2:3:8:3:2:1	10, 10	0.018	0.049	0.075	0.070	0.044	0.047	0.063
	10, 30	0.040	0.050	0.073	0.070	0.054	0.053	0.058
	10,100	0.052	0.050	0.078	0.077	0.063	0.054	0.054
	30, 30	0.034	0.054	0.064	0.061	0.051	0.056	0.060
	30,100	0.046	0.051	0.061	0.060	0.053	0.053	0.054
	100,100	0.049	0.053	0.055	0.054	0.051	0.053	0.054

For each of the 192 conditions, 10,000 samples were generated using SAS IML, version 6.12 (SAS, 1992). The use of 10,000 samples provides an adequate level of precision for this study, yielding maximum 95% confidence intervals of ±.0098 around the observed proportion of null hypotheses rejected (Robey & Barcikowski, 1992). For each condition, seven test statistics were computed: (a) the independent means *t*-test, (b) Pearson's chi-square test of homogeneity, (b) Cliff's *Unbiased* test of *d*, (c) Cliff's *Consistent* test of *d*, (d) Cliff's asymmetric confidence interval (CI) for *d*, (e) the Wald test associated with the cumulative logit model, and (f) the likelihood ratio (LR) test associated with the cumulative logit model. Estimates of the Type I error control and the statistical power of each test were conducted at nominal alpha levels of .10, .05, and .01.

Results and Discussion

Before turning to an examination of statistical power, attention must first focus on a comparison of the relative ability of the seven tests to control Type I error. Estimates of Type I error rate were calculated for each of the seven procedures based on 10,000 randomly generated samples for each null condition under examination. Bradley's (1978) liberal criterion of robustness (actual α within $\alpha_{\text{nominal}} \pm 0.5\alpha_{\text{nominal}}$) was used to evaluate the capacity of each of the seven procedures to control Type I error under the various conditions. To save space, results are provided only for nominal alpha equal to .05. Type I error rates and power estimates for alpha level equal to .10 and .01 are available from the first author.

Estimates of Type I Error Control

Five Point Response Scale. The estimates of Type I error rates for the 5-point scales are provided in Table 4. A broad overview of the robustness of all of the seven tests across all conditions at alpha = .05 is presented in a series of box and whisker plots in Figure 1. The two horizontal lines in this figure are Bradley's limits of robustness. Examination of these plots revealed the *t*-test best able to control Type I error, followed closely by the LR, the Wald test, Cliff's confidence interval, and the chi-square test. Considerably less control was exhibited by Cliff's consistent and unbiased tests. The *t*-test stood alone in its ability to maintain the appropriate level across all conditions. The CI, Wald, LR, and Chi-Square were able to maintain alpha within acceptable limits for all but the most skewed conditions coupled with small and unequal sample sizes. Both the unbiased and consistent tests failed to maintain acceptable control in several instances when small and unequal samples were involved.

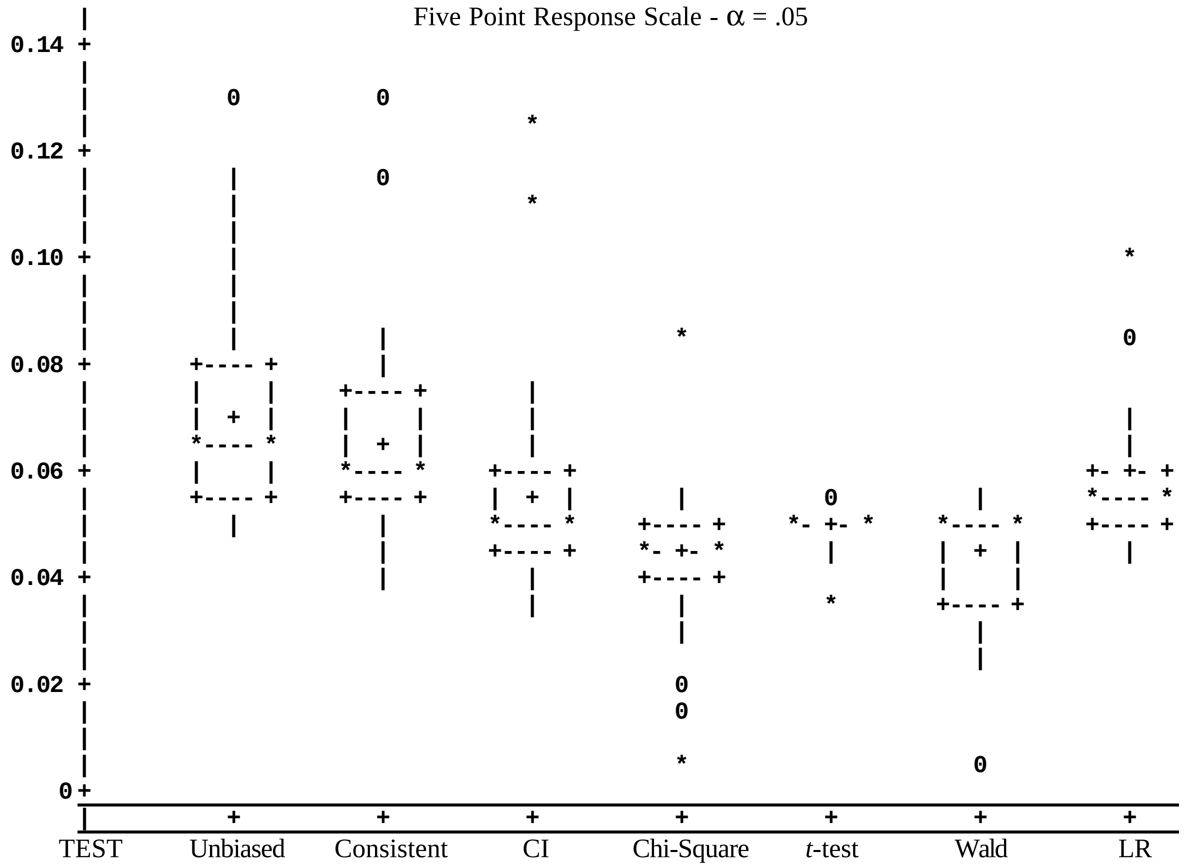

Figure 1. Distribution of Type I Error Rate Estimates for Seven Tests across Experimental Conditions.

Seven Point Response Scale. The estimates of Type I error rates for the 7-point scales are provided in Table 5. The box and whisker plots presented in Figure 2 provide a general overview of the robustness of all seven procedures across all conditions when alpha was set equal to .05. Again, the t-test maintained Type I control across all conditions. Generally, the seven procedures maintained alpha within acceptable limits when large sample sizes were examined. The most skewed condition presented problems for several of the tests, as liberal estimates were observed, on several occasions, for the LR, and Cliff's Confidence Interval, Unbiased, and biased tests. However, there were a few instances in which the Chi-Square and Wald test became conservative. For the unimodal, symmetric distribution, Cliff's Unbiased and consistent tests were liberal only for the unequal sample sizes of 10 and 100, while the Chi-Square test was conservative with the smallest samples.

Estimates of Statistical Power

Five Point Response Scale. Table 6 contains power estimates for the seven procedures. Statistical power estimates are provided only for conditions in which Type I error was controlled. In addition, the Wald test used with the cumulative logit model was not calculable for most samples when the distribution was highly skewed and a non-null condition was simulated (conditions which typically yielded a singular covariance matrix). Estimates of power for only those samples in which it was calculable would be misleading, so these power estimates have also been omitted.

An examination of statistical power at nominal $\alpha = .05$ revealed the chi-square to be superior to all other tests under the equal marginal and slightly skewed marginal conditions. Under the highly skewed marginal conditions, the Chi-Square was the most powerful only under the largest samples examined. For smaller samples, or unequal samples, other tests were more powerful. For example, Cliff's Consistent test and CI produced the highest power under a highly skewed, small sample condition with a large effect size (power = .625 for both). However, it should be noted that in this instance only one other test, the *t*-test, was able to control Type I error.

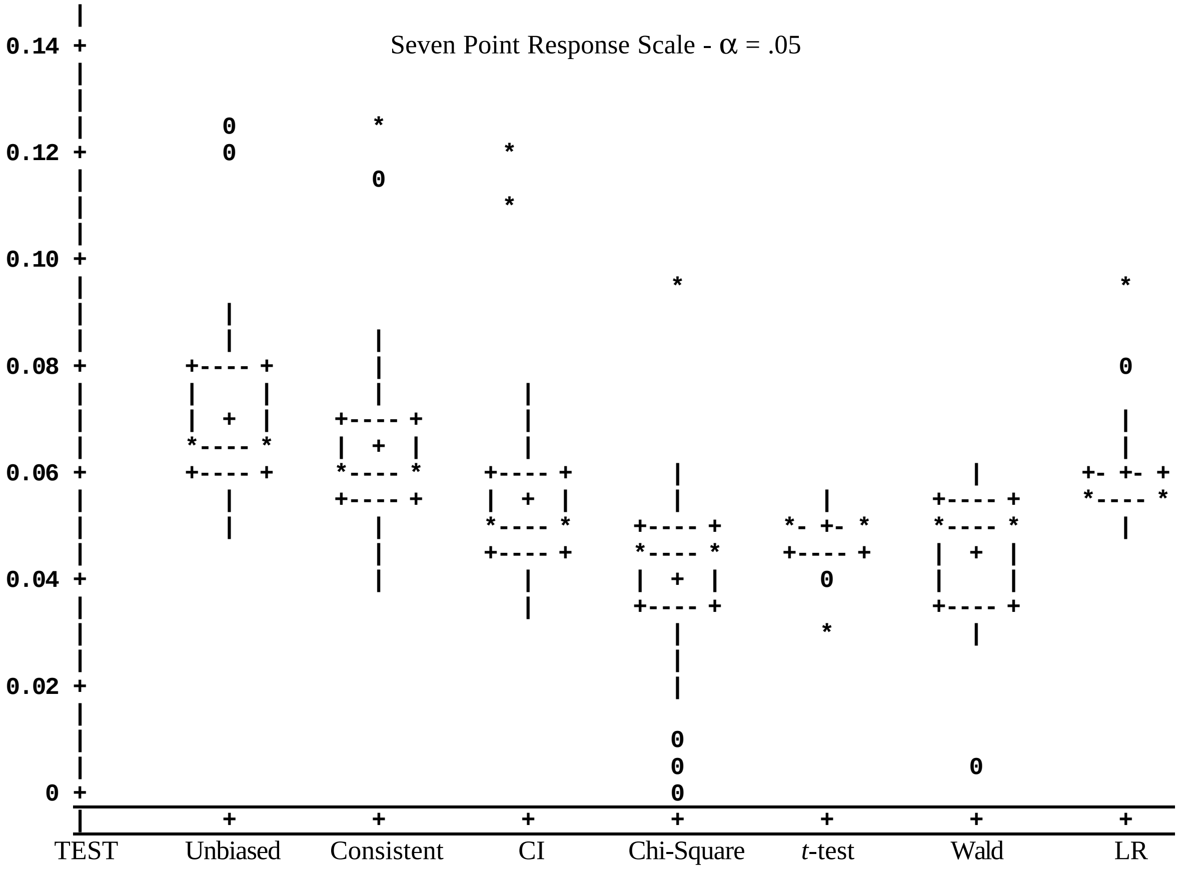

Figure 2. Distribution of Type I Error Rate Estimates for Seven Tests across Experimental Conditions

The *t*-test produced the highest power under highly skewed and unbalanced design conditions, but again, it was one of only two tests that were able to control Type I error under these conditions. For the unimodal, symmetric distribution, the chi-square test was never the most powerful. Rather, for samples drawn from this distribution shape, either the LR test or Cliff's Unbiased or Consistent tests were the most powerful.

Seven Point Response Scale. Table 7 contains power estimates for the seven procedures for nominal alpha level equal to .05. Examination of these results, revealed the Chi-Square to be the most powerful test only under the equal marginal condition, except with small sample sizes. When small sample sizes were examined, Cliff's Consistent test and the LR produced more power than the other tests. Under the slightly skewed and highly skewed marginal distributions, the power produced by several tests was very similar. For example, under the slightly skewed condition with small samples, Cliff's delta tests and the LR produced similar estimates. With larger samples under this condition, it was difficult to choose a superior test from among Cliff's delta tests, the Wald test, or the LR. Similar circumstances surrounded the highly skewed distribution with large sample sizes. For small sample sizes under this condition, the consistent test and CI produced the most power, but many of the tests were unable to control Type I error. For the unimodal, symmetric distribution, the most powerful tests were typically Cliff's Unbiased or Consistent tests. As with the results obtained with the 5-point scales, neither the Chi-Square nor the *t*-test were the most powerful in any sample size condition with this distribution shape.

The differences in the results obtained between the 5-point and 7-point data prompted a further examination of the populations from which samples were generated. Recall that these populations were constructed based on differences between proportions at each scale point to produce desired values of Cohen's *w* (the effect size for differences in population proportions). These populations were examined in terms of the effect size for standardized mean difference (Cohen's *d*) and Cliff's delta. Although the latter is not an effect size, per se, it represents the proportional non-overlap of the two populations from which samples were drawn. These results are presented in Table 8.

Table 6. Statistical Power Estimates for 5 Point Response Scale at nominal α = .05

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	*t*-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1	10, 10	.10	0.040	0.061	-----	0.078	0.048	0.067	0.075
	10, 10	.30	0.115	0.139	-----	0.161	0.104	0.147	0.162
	10, 10	.50	0.345	0.296	-----	0.310	0.226	0.321	0.339
	10, 30	.10	0.061	0.068	0.086	0.082	0.062	0.076	0.079
	10, 30	.30	0.220	0.211	0.206	0.200	0.158	0.242	0.234
	10, 30	.50	0.638	0.482	0.391	0.385	0.307	0.535	0.505
	10,100	.10	0.060	0.074	-----	-----	0.080	0.082	0.083
	10,100	.30	0.310	0.265	-----	-----	0.185	0.298	0.280
	10,100	.50	0.807	0.597	-----	-----	0.337	0.646	0.596
	30, 30	.10	0.080	0.079	0.087	0.084	0.071	0.080	0.083
	30, 30	.30	0.427	0.320	0.330	0.325	0.293	0.336	0.336
	30, 30	.50	0.924	0.710	0.689	0.685	0.651	0.729	0.723
	30,100	.10	0.103	0.108	0.111	0.109	0.099	0.115	0.114
	30,100	.30	0.649	0.482	0.410	0.407	0.383	0.513	0.489
	30,100	.50	0.994	0.888	0.769	0.768	0.743	0.900	0.875
	100,100	.10	0.170	0.142	0.146	0.145	0.140	0.145	0.146
	100,100	.30	0.950	0.777	0.768	0.766	0.759	0.779	0.778
	100,100	.50	1.000	0.996	0.994	0.994	0.993	0.996	0.995
6:1:1:1:1	10, 10	.10	-----	0.058	0.078	0.071	0.048	0.039	0.071
	10, 10	.30	-----	0.111	0.115	0.109	0.080	0.071	0.112
	10, 10	.50	-----	0.225	0.192	0.186	0.143	0.138	0.191
	10, 30	.10	0.045	0.050	-----	-----	0.080	0.031	0.065
	10, 30	.30	0.147	0.113	-----	-----	0.150	0.056	0.119
	10, 30	.50	0.551	0.261	-----	-----	0.284	0.127	0.237
	10,100	.10	0.063	0.047	-----	-----	0.100	0.022	0.063
	10,100	.30	0.243	0.117	-----	-----	0.201	0.036	0.126
	10,100	.50	0.770	0.297	-----	-----	0.370	0.111	0.268
	30, 30	.10	0.058	0.068	0.070	0.069	0.058	0.062	0.068
	30, 30	.30	0.394	0.218	0.192	0.189	0.171	0.185	0.194
	30, 30	.50	0.956	0.508	0.430	0.426	0.400	0.435	0.443
	30,100	.10	0.082	0.072	0.094	0.093	0.087	0.063	0.074
	30,100	.30	0.602	0.290	0.298	0.297	0.286	0.223	0.256
	30,100	.50	0.998	0.662	0.622	0.620	0.605	0.546	0.589
	100,100	.10	0.165	0.109	0.097	0.096	0.093	0.095	0.097
	100,100	.30	0.958	0.572	0.491	0.489	0.481	0.495	0.496
	100,100	.50	1.000	0.952	0.898	0.897	0.893	0.907	0.906

Note that, for the null condition, the populations are identical regardless of how population “differences” are represented. Further, when differences are represented in terms of Cohen's *w*, the 5-point and 7-point populations have identical effect sizes. However, when differences are represented by Cohen's *d*, the effect sizes differ across the two sets, and the difference is not consistent across the distribution shapes. For example, with the "small effect" populations under the slight skew condition, Cohen's *d* was 0.10 for the 5-point data and 0.17 for the 7-point data. A similar difference was evident for the high skew. However, for the unimodal, symmetric distributions, the Cohen's *d* values were nearly identical (0.17 vs. 0.19). Similar differences were noted across the remaining non-null conditions examined. Such discrepancies were also evident when the population differences were measured as Cliff's delta. These observed deviations across effect sizes reflect variations in the magnitude of population differences that result from the design variables of distribution shape and number of scale points. These design variables produced differential effects when inequalities were measured as discrepancies in population standardized mean difference or proportion of non-overlap.

Table 6 (continued). Statistical Power Estimates for 5 Point Response Scale at nominal α = .05

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	*t*-test	Unbiased	Consistent	CI	Wald	LR
16:1:1:1:1	10, 10	.10	-----	0.032	-----	0.038	0.033	-----	-----
	10, 10	.30	-----	0.068	-----	0.058	0.056	-----	-----
	10, 10	.50	-----	0.542	-----	0.625	0.625	-----	-----
	10, 30	.10	0.036	0.025	-----	----	-----	0.015	0.076
	10, 30	.30	0.089	0.035	-----	----	-----	-----	0.116
	10, 30	.50	0.054	0.670	-----	----	-----	-----	0.981
	10,100	.10	-----	0.021	-----	----	-----	0.018	-----
	10,100	.30	-----	0.013	-----	----	-----	-----	-----
	10,100	.50	-----	0.874	-----	----	-----	-----	-----
	30, 30	.10	-----	0.060	0.062	0.062	0.057	0.044	0.063
	30, 30	.30	-----	0.197	0.146	0.145	0.138	0.105	0.147
	30, 30	.50	-----	0.998	1.000	1.000	1.000	-----	1.000
	30,100	.10	0.073	0.056	0.093	0.093	0.090	0.035	0.065
	30,100	.30	0.584	0.212	0.257	0.256	0.251	0.098	0.180
	30,100	.50	0.998	1.000	1.000	1.000	1.000	-----	1.000
	100,10	.10	0.154	0.099	0.083	0.083	0.081	0.078	0.083
	100,10	.30	0.972	0.495	0.340	0.340	0.335	0.331	0.343
	100,10	.50	1.000	1.000	1.000	1.000	1.000	-----	1.000
1:2:4:2:1	10, 10	.10	0.035	0.070	-----	0.095	0.062	0.069	0.089
	10, 10	.30	0.105	0.222	-----	0.283	0.210	0.227	0.274
	10, 10	.50	0.319	0.536	-----	0.625	0.533	0.481	0.628
	10, 30	.10	0.063	0.078	0.115	0.111	0.088	0.086	0.094
	10, 30	.30	0.216	0.299	0.379	0.372	0.322	0.334	0.352
	10, 30	.50	0.654	0.725	0.787	0.783	0.733	0.778	0.799
	10,100	.10	0.064	0.079	-----	-----	0.100	0.086	0.089
	10,100	.30	0.305	0.378	-----	-----	0.411	0.411	0.424
	10,100	.50	0.821	0.821	-----	-----	0.797	0.871	0.872
	30, 30	.10	0.075	0.103	0.123	0.120	0.104	0.111	0.116
	30, 30	.30	0.417	0.540	0.616	0.611	0.581	0.603	0.612
	30, 30	.50	0.934	0.948	0.974	0.974	0.968	0.974	0.975
	30,100	.10	0.103	0.139	0.162	0.161	0.149	0.150	0.153
	30,100	.30	0.650	0.729	0.782	0.780	0.764	0.785	0.787
	30,100	.50	0.995	0.994	0.996	0.996	0.995	0.998	0.998
	100,100	.10	0.166	0.235	0.266	0.265	0.257	0.262	0.264
	100,100	.30	0.954	0.967	0.982	0.982	0.981	0.981	0.981
	100,100	.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Estimates are based on 10,000 samples of each condition. Power estimates are provided only for conditions in which Type I error was controlled.

Conclusions

The results of this research need to be interpreted in the light of the limitations of the study. First, only analyses based on two independent groups were conducted. Although all of the statistical procedures investigated can be extended to multiple group applications, the resulting Type I error rates and power estimates will not necessarily be comparable to those obtained here. Secondly, a limited number of distribution shapes were examined in this study. Extensions to other shapes, such as bimodal distributions, are important areas to explore because distribution shape was seen to influence both Type I error control and the relative power of these tests. Finally, in the consideration of statistical power, the nature of the differences between groups can assume several forms. Although ordered categorical data preclude the consideration of simple shifts in location (because of the boundedness of the response scale), types of non-null effects other than those modeled here need to be investigated.

In light of these limitations, the superiority of the *t*-test and the cumulative logit model in their control of Type I error is evident in these data. Problems with the control of Type I error rates were frequently encountered in conditions with skewed marginal distributions and with unbalanced or small samples. Specific limitations in Type I error control were

Table 7. Statistical Power Estimates for 7 Point Response Scale at nominal α = .05

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	t-test	Unbiased	Consistent	CI	Wald	LR
1:1:1:1:1	10, 10	.10	-----	0.053	-----	0.071	0.041	0.059	0.067
	10, 10	.30	-----	0.060	-----	0.079	0.046	0.067	0.077
	10, 10	.50	-----	0.081	-----	0.106	0.064	0.093	0.104
	10, 30	.10	0.054	0.055	0.076	0.072	0.052	0.059	0.062
	10, 30	.30	0.181	0.073	0.093	0.087	0.066	0.083	0.084
	10, 30	.50	0.548	0.103	0.126	0.120	0.090	0.119	0.121
	10,100	.10	0.064	0.053	-----	-----	0.063	0.061	0.059
	10,100	.30	0.265	0.078	-----	-----	0.081	0.097	0.090
	10,100	.50	0.757	0.127	-----	-----	0.111	0.156	0.145
	30, 30	.10	0.067	0.052	0.059	0.056	0.047	0.053	0.055
	30, 30	.30	0.347	0.083	0.095	0.092	0.077	0.088	0.091
	30, 30	.50	0.868	0.143	0.160	0.155	0.136	0.151	0.155
	30,100	.10	0.084	0.056	0.064	0.063	0.055	0.058	0.058
	30,100	.30	0.578	0.111	0.114	0.112	0.102	0.122	0.119
	30,100	.50	0.988	0.212	0.205	0.203	0.187	0.230	0.223
	100,100	.10	0.137	0.065	0.068	0.067	0.064	0.066	0.067
	100,100	.30	0.923	0.165	0.171	0.170	0.163	0.168	0.169
	100,100	.50	1.000	0.378	0.383	0.381	0.372	0.383	0.384
9:1:1:1:1	10, 10	.10	-----	0.064	0.092	0.084	0.059	0.048	0.084
	10, 10	.30	-----	0.207	0.293	0.278	0.218	0.172	0.278
	10, 10	.50	-----	0.521	0.665	0.646	0.568	0.394	0.654
	10, 30	.10	0.017	0.065	-----	0.132	0.116	0.048	0.100
	10, 30	.30	0.028	0.260	-----	0.424	0.385	0.241	0.382
	10,100	.10	0.023	0.060	-----	-----	----	0.030	0.106
	10,100	.30	0.022	0.296	-----	-----	----	0.295	0.441
	10,100	.50	0.350	0.761	-----	-----	----	0.670	0.879
	30, 30	.10	0.048	0.094	0.117	0.114	0.102	0.107	0.114
	30, 30	.30	0.317	0.521	0.631	0.626	0.603	0.615	0.628
	30, 30	.50	0.862	0.939	0.981	0.981	0.978	0.972	0.982
	30,100	.10	0.036	0.120	0.181	0.180	0.170	0.141	0.158
	30,100	.30	0.406	0.680	0.806	0.805	0.791	0.797	0.807
	30,100	.50	0.971	0.990	0.998	0.998	0.997	0.992	0.999
	100,100	.10	0.139	0.229	0.283	0.283	0.276	0.279	0.282
	100,100	.30	0.916	0.958	0.986	0.986	0.985	0.986	0.986
	100,100	.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000

observed for the tests of delta suggested by Cliff (1993, 1996a). Of special interest is that in many conditions, Cliff's Confidence Interval approach to inferences regarding delta were superior to the two *z* test approaches examined. The asymmetric approach to the confidence interval estimation appeared to improve the control of Type I errors in several of the conditions examined in this study. However, for researchers working with small samples or unequal sample sizes, the *t*-test or cumulative logit model appear to be the tests of choice to maintain Type I error control.

Finally, in terms of statistical power, although the independent means *t*-test provided the best control of Type I error rates across the conditions examined, this test was rarely the most powerful, and, consequently, should not be the first choice in most applications. For the 5-point response scales, the chi-square test of homogeneity was clearly the most powerful test for those conditions in which it maintained Type I error control. In contrast, for the 7-point scales, the Chi-Square test was only the most powerful when the marginal distribution was symmetric. For the skewed marginal distributions, the cumulative logit models or the tests of delta tended to be the most powerful. However, the variation in power across these scales should not be interpreted as a simple function of the number of scale points. Rather, such variations represents changes in the magnitude of the population differences in terms of standardized mean difference or proportion of non-overlap of the populations.

Table 7 (continued). Statistical Power Estimates for 7 Point Response Scale at nominal α = .05.

Marginal	Sample	Effect	Chi-		Cliff's *d* Tests			Cumulative Logit	
Distribution	Size	Size	Square	*t*-test	Unbiased	Consistent	CI	Wald	LR
24:1:1:1:1	10, 10	.10	-----	0.046	-----	0.063	0.057	-----	-----
	10, 10	.30	-----	0.173	-----	0.232	0.217	-----	-----
	10, 10	.50	-----	0.506	-----	0.622	0.622	-----	-----
	10, 30	.10	0.011	0.024	-----	-----	-----	0.007	0.126
	10, 30	.30	0.001	0.145	-----	-----	-----	-----	0.441
	10, 30	.50	0.004	0.620	-----	-----	-----	-----	0.982
	10,100	.10	-----	0.011	-----	-----	-----	0.005	-----
	10,100	.30	-----	0.107	-----	-----	-----	-----	-----
	10,100	.50	-----	0.818	-----	-----	-----	-----	-----
	30, 30	.10	-----	0.105	0.129	0.128	0.119	0.090	0.131
	30, 30	.30	-----	0.575	0.694	0.692	0.677	0.535	0.699
	30, 30	.50	-----	0.999	1.000	1.000	1.000	-----	1.000
	30,100	.10	0.017	0.100	0.224	0.223	0.220	0.094	0.167
	30,100	.30	0.129	0.707	0.894	0.894	0.890	0.708	0.859
	30,100	.50	0.960	1.000	1.000	1.000	1.000	-----	1.000
	100,10	.10	0.120	0.233	0.288	0.288	0.284	0.278	0.288
	100,10	.30	0.920	0.972	0.992	0.992	0.992	0.991	0.992
	100,10	.50	1.000	1.000	1.000	1.000	1.000	-----	1.000
1:2:3:8:3:2:1	10, 10	.10	-----	0.072	-----	0.098	0.066	0.072	0.089
	10, 10	.30	-----	0.245	-----	0.307	0.232	0.244	0.293
	10, 10	.50	-----	0.611	-----	0.680	0.596	0.539	0.667
	10, 30	.10	0.050	0.082	0.116	0.111	0.089	0.091	0.096
	10, 30	.30	0.208	0.369	0.410	0.402	0.345	0.401	0.408
	10, 30	.50	0.632	0.810	0.799	0.792	0.741	0.821	0.837
	10,100	.10	0.082	0.087	-----	-----	0.098	0.095	0.094
	10,100	.30	0.352	0.432	-----	-----	0.405	0.472	0.464
	10,100	.50	0.882	0.891	-----	-----	0.802	0.910	0.902
	30, 30	.10	0.054	0.114	0.131	0.126	0.111	0.118	0.123
	30, 30	.30	0.310	0.602	0.653	0.645	0.612	0.634	0.643
	30, 30	.50	0.876	0.978	0.984	0.983	0.979	0.981	0.983
	30,100	.10	0.088	0.139	0.156	0.154	0.144	0.150	0.152
	30,100	.30	0.624	0.803	0.806	0.804	0.790	0.831	0.829
	30,100	.50	0.994	0.998	0.997	0.997	0.997	0.999	0.998
	100,100	.10	0.136	0.259	0.282	0.280	0.272	0.276	0.279
	100,100	.30	0.919	0.982	0.987	0.987	0.987	0.987	0.987
	100,100	.50	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Estimates are based on 10,000 samples of each condition. Power estimates are provided only for conditions in which Type I error was controlled.

Further, the power differences among these procedures were small suggesting that researchers' choices may be based on the types of interpretations that are appropriate for the research questions being addressed. For interpretations based on simple dominance, the d statistics and their inferential tests would be the most appropriate. In contrast, a more rigorous modeling of response probabilities is provided by the cumulative logit models.

In summary, ordered categorical data, such as those investigated in this study, are frequently encountered in educational research. Unfortunately, the analysis strategies most frequently employed with these types of data are not necessarily the best strategies to use. This research has provided information about the operating characteristics (Type I error control and statistical power) of the commonly used tests employed with ordered categorical data, and has provided evidence of the advantages (in some data conditions) associated with two recently recommended options for testing hypotheses. Although additional research is certainly needed to further explore the performance of these tests and their limitations, this initial examination suggests that for many data conditions, the choice of an appropriate test statistic is vitally important to the validity of research inferences.

Table 8. Indices of Differences in the Simulated Populations

Population		Index of Group Difference					
Group	Marginal	Effect size *W*		Effect Size *d*		Cliff's *d*	
Differences	Distribution	5-point	7-point	5-point	7-point	5-point	7-point
Null Model	Uniform	0.00	0.00	0.00	0.00	0.00	0.00
	Slight Skew	0.00	0.00	0.00	0.00	0.00	0.00
	High Skew	0.00	0.00	0.00	0.00	0.00	0.00
	Unimodal Sym	0.00	0.00	0.00	0.00	0.00	0.00
Small Effect	Uniform	0.10	0.10	0.13	0.05	0.07	0.03
	Slight Skew	0.10	0.10	0.10	0.17	0.05	0.10
	High Skew	0.10	0.10	0.09	0.18	0.03	0.08
	Unimodal Sym	0.10	0.10	0.17	0.19	0.10	0.11
Medium Effect	Uniform	0.30	0.30	0.39	0.14	0.21	0.08
	Slight Skew	0.30	0.30	0.31	0.53	0.14	0.29
	High Skew	0.30	0.30	0.28	0.57	0.09	0.24
	Unimodal Sym	0.30	0.30	0.54	0.58	0.31	0.32
Large Effect	Uniform	0.50	0.50	0.67	0.23	0.36	0.13
	Slight Skew	0.50	0.50	0.52	0.95	0.23	0.49
	High Skew	0.50	0.50	1.40	1.36	0.40	0.40
	Unimodal Sym	0.50	0.50	1.00	1.05	0.51	0.54

Correspondence should be directed to
Jeffrey D. Kromrey
Educational Measurement & Research,
University of South Florida
4202 East Fowler Avenue, Tampa, FL 33620

References

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, *105*, 290-301.

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

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

Blair, R. C. & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of the student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, *5*, 309-335.

Blair, R. C. & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, *97*, 119-128.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*, 494-509.

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, *31*, 331-350.

Cliff, N. (1996b). *Ordinal methods for behavioral data analysis*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Davidson, M. L. & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin*, *104*, 137-144.

McCullough, P. & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

Nanna, M. J. & Sawilowsky, S. S. (1998). Analysis of Likert data in disability and medical rehabilitation research. *Psychological Methods*, *3*, 55-67.

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.

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

Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *American Statistician*, *47*, 65-72.

Think Different.
Comments on Alternative Regression Procedures

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

I chose Apple Computers' slogan, not because I happen to be a Macintosh user, but because the issues raised in these three articles should lead us to "Think Different" as statisticians, data analysts, and researchers. One key issue underlying these articles is the ultimate question, "What are the data trying to tell us?" Several statistics texts have used the signal-to-noise analogy for analyzing data. Therefore, if we are simply trying to detect a signal amongst random, ambient noise then it does not seem as problematic to transform the data or to perform alternative procedures that potentially test different statistical hypotheses. If exact parameter estimation is of interest, however, data transformations may lead to interpretive difficulties.

Nevitt and Tam (*pp*. 54-69) approach this issue from the parameter estimation perspective of: What should be done in order to detect an *accurate* signal if the data are not "well behaved" or do not conform to the statistical assumptions of the regression model? These authors examine three general approaches for estimating parameters when data are not well behaved (i.e., nonnormal): (a) treat outliers differently (i.e., Trimming, Winsorizing), (b) transform the data (i.e., Monotonic Regression), or (c) compute parameter estimates in a different manner (i.e., LAD, Theil estimators).

The authors make an important distinction between robust and nonparametric estimators. Robust methods were developed for situations in which *symmetric* error distributions have heavy tails due to outliers in the observed data. Thus, the normality assumption is simply relaxed. Robust estimators are therefore resistant to violation of assumptions while testing the same null hypothesis as the normal theory methods (Draper & Smith, 1981). By contrast, nonparametric and distribution-free methods may involve (a) transforming data to ranks or other metrics or (b) computing the parameter estimate in an entirely different way. Therefore, the normality assumption may not apply whatsoever. In these cases, the statistical hypothesis tested, although conceptually similar, may be quite different than the hypothesis evaluated by a normal theory counterpart. Because of this difference, the performance of nonparametric methods relative to OLS methods is often hard to assess except under conditions where many parameters (i.e., skew and kurtosis) are held to normal theory assumptions which of course favors OLS procedures.

Recall the question posed in the foreword (*p*. 2), "How do these techniques integrate with what is already known about statistics?" There are extremely interesting relationships between OLS and nonparametric estimators of slope. By using the geometric definition of a regression slope and taking the $n(n - 1)$ pairwise slopes,

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \text{ where } x_i \neq x_j,$$

the Theil estimator of slope is the median of all the b_{ij} slopes. Interestingly, when all values of X are distinct, Sprent (1993) demonstrates that significance testing of the Theil median slope is based on Kendall's (1970) tau statistic which is related to Cliff's (1994) ordinal multiple regression (Long, *pp*. 45-53). This can be inferred from the fact that $n(n - 1)$ pairwise values are used in both procedures. Other relationships can be shown by making an aggregate of these slopes such that

$$\hat{\beta} = \frac{\sum_{i<j} w_{ij} b_{ij}}{\sum_{i<j} w_{ij}} \quad ,$$

OLS regression defines the weight as $w_{ij} = (x_j - x_i)^2$. Other nonparametric approaches defines $w_{ij} = |x_j - x_i|$ (Birkes & Dodge, 1993) which reduces to a sign function for the X variable over the sum of the absolute deviations of X (Huynh, 1978). For Kendall's tau, the weights would be defined in terms of the absolute value of both the Y and X deviations (i.e., $w_{ij} = |x_j - x_i|/|y_j - y_i|$) which would then reduce to a sign function for both Y and X. Kendall's tau is the simple (i.e., unweighted) average sign of the $n(n - 1)$ pairwise slopes.

In terms of Nevitt and Tam's methodology, it is questionable whether the sequential series of X is realistic. First of all, the sequential series of X is a uniform distribution. This implies that Nevitt and Tam are examining fixed-effects models because the underlying assumption of a random-effects model is that Y and X are sampled from a bivariate normal distribution (Hays, 1994). Although fixed-effects models are applied most commonly, even when random-effects are of interest (Clark, 1973), the use of a uniform distribution as the basis for the parameter model seems realistic only if the population relationship among ranks is of interest. Secondly, in the population, the uniform distribution of X yields a uniform distribution for Y through the linear transformation described in the methods section (*p*.

57). When a normal, random error component is added to Y then the conditional distribution of Y is normal which again is adequate for a fixed-effects model. In this case however, the overall distribution of Y is neither uniform nor normal which violates the bivariate normality assumption of a simple linear regression random-effects model. Thus in regression applications where random-effects models are of interest, the fact that the data for Y are nonnormal could stem from either (a) the structural component (i.e., population distributions of X and Y are nonnormal), (b) the error component being nonnormal, or (c) from both (a) and (b).

Thus, from this fixed-effects perspective, Nevitt and Tam's methodological approach assumes that "bad" (i.e., nonnormal) data originates from the error distribution of a regression model. Therefore, they rightfully suggest that "data analyses should always involve checking for outliers in the observed data and testing the underlying assumptions under OLS estimation" (*p*. 68). The idea that outliers and heteroscedasticity may stem from nonnormal error distributions is certainly interesting and leads to the question: How does one know if the error distribution is normal when the data are nonnormal? The possibilities are that: (a) the variable itself is nonnormal; (b) there are outliers present in a symmetric error distributions; or (c) the error distribution is skewed or nonnormal. Thus, there is an important distinction between: (1) a normal distribution with outliers that create skewness and (2) a skewed distribution such as reaction time. Nevitt and Tam's results show that as expected (Draper & Smith, 1981), robust estimators perform better under condition (1) but nonparametric methods perform better under condition (2). Nevitt and Tam report the surprisingly good overall performance of the Theil estimator. Furthermore, the Theil estimates were accurate especially with nonnormal error distributions. As expected with contaminated normal error distributions, the robust procedures (i.e., LAD, Trimming, Winsorizing) performed well. As a personal bias, however, I am not fond of Trimming because this form of discarding data creates a situation where the data are systematically missing which is know to lead to biased estimates.

One astounding and important result, mainly because of the common application of rank transformations, was the poor performance of Monotonic Regression. This finding should be viewed in a certain light, however. The authors note that their results substantiate the unacceptability of rank transformation in the form of Monotonic Regression with respect to Bias and root mean square error (RMSE). Namely, large Bias values reflect the inability of Monotonic Regression to "recover the true population values" (see *p*. 67). The fact that Nevitt and Tam used sequential values of X would seem to have benefited Monotonic (rank) Regression because the transformation was linear for X. With the addition of a random error distribution, however, the rank transformation for Y was not linear in most cases. Thus, Monotonic Regression did not perform well in general. Yet, procedures that transform the original data should not be expected to perform as well. How would rank values transform back to the original metric of Y if sequential X values were not used? Furthermore, one must consider that Monotonic Regression tests a different null hypothesis; it tests OLS hypotheses in the metric of ranks.

As with the Brockmeier et al. article (*pp*. 20-39), if the purpose of a study is simply to establish a relationship (i.e., just detecting the signal) then finding non-zero correlations (or standardized regression slopes) is the major issue rather than exact parameter estimation. Perhaps the rank transform procedure (Monotonic Regression) would not perform so poorly in these circumstances (e.g., a simulation study where Type I error and Power rates, instead of estimation bias, would be reported). Yet, if exact parameter estimation is of interest then the precision of both α and β parameter estimates is important. The RMSE and Bias reported by Nevitt and Tam are both valuable indicators because procedures such as Monotonic Regression can maintain stable Type I error rates and demonstrate superior power (e.g., Harwell & Serlin, 1989) yet provide consistently bad parameter estimates. Thus, it would appear that rank (as well as other non-linear) transformations are not appropriate when exact parameter estimates are to be "recovered." By contrast, if researchers are merely attempting to establish a relationship, then they could consider the signal-to-noise analogy where transformations (and other alternative approaches) do not seem so disabling.

To elaborate, in experimental designs and other group comparison research, ANOVA models that test for mean differences are employed. In contrast to single-sample statistics where a relevant population parameter must be known *a priori*, the fact that there is a comparison group makes the signal more detectable. One may think of this in terms of perceptual research which has demonstrated that judging the length or orientation of an object is much easier when there are perceptual cues that allow for comparisons (e.g., Witkin & Goodenough, 1981). Using the same analogy, violations of assumptions and other data problems can be viewed as the factors that create perceptual (statistical) distortions and illusions, and thus, the use of statistics in many behavioral research contexts may be seen as a field-dependent endeavor. Similar to the ANOVA model, a linear regression model is a comparison of means in the sense that as X increases the expected value of Y increases by the slope on average. Again, if one is simply trying to detect a signal, rather than estimating a parameter precisely, then the fact that Y generally increases with increases in X may be good enough. And it does not matter too much how the variables are expressed.

Popular sources such as Tabachnick and Fidell (1996) discuss transforming data when assumptions are violated. This is even more systematized as the "ladder of re-expression" when power and logarithmic transformations are used to transform data (Hoaglin, Mosteller, & Tukey, 1983). Yet most researchers have problems with such nonlinear transformation with the exception of the rank transform concept. That is, unlike taking the square root of a variable to quell an outlier or reduce asymmetry, ranking the data still retains the "meaning" of the data to many researchers (Zimmerman, 1996). Furthermore, there are conveniences because there are many rank-based tests already in existence and ranks have known means and variances. Despite these conveniences, Nevitt and Tam's results are consistent with other research (Zimmerman, 1996, 1998; Zimmerman & Zumbo, 1993) that has demonstrated serious problems in applying rank transformations. Therefore, the reliance on rank transformation may be somewhat superstitious because its historical prevalence and intuitive appeal are more convincing than empirical evidence showing its statistical viability for estimating parameters.

One issue that all researchers have with any re-expression is what do the data "mean" after transformation. For the sake of symmetry in a variable, a researchers may be left with the question: What does the square root of achievement scores mean? Cliff (1996) argues that researchers usually do not want their conclusions to be confined to the current, somewhat arbitrary version of the variables. Moreover the current measurements are often assumed to be manifest versions of latent variables that are not linearly related to them. Therefore, a poignant question for researchers to ask would be: What did my scores mean in the first place? From this perspective the central question of data analysis can be posed as: What question should I be asking? That is, the null hypotheses associated with OLS regression may not be what is really of interest. Cliff (1993, 1996) contends that most of the answers behavioral researchers want to get from their data are ordinal ones. Furthermore, most of the observed variables have only ordinal justification, at least as measures of the theoretical constructs they are used to represent. Therefore, because the questions asked are ordinal and the data are ordinal, ordinal methods are suggested.

Based on this perspective, **Long** (*pp*. 45-53) explicates another less common re-expression, the transformation of data into what Cliff (1993) has termed the "dominance" metric. Many of us have been familiarized with this concept through Kendall's (1970) measure of concordance. Not only does this notion lead to testing statistical hypotheses that are different from their OLS counterparts, the procedures require us to "Think Different" because the hypotheses are different conceptually. From the Pearsonian perspective, relationships are an issue of the average value of Y conditional on X. From the dominance perspective, however, relationships are expressed as the proportional alignment of Y with X. In terms of group comparisons where the OLS solution involves an ANOVA model, ordinal methods address what proportion of scores in group one are larger than the scores in group two. In terms of a linear regression, they assess what proportion of Y scores become larger as X increases.

Marascuilo and McSweeney (1977, *pp*. 439-440) discuss Kendall's tau as a measure of concordance and as a measure of correlation. However, Kendall's tau as a measure of correlation is not interpreted in the Pearsonian sense but as a measure of "array." That is, it is an index of the amount of agreement between two sets of ranks. When teaching the Pearson product-moment correlation, I prefer demonstrating the z-score formula and discussing the Pearson r as an averaged leverage (i.e., product-moment) value. Similarly, the notion of the dominance metric is appealing because it allows a perspective of what Kendall's tau (as well as ordinal multiple regression and Cliff's d statistic) actually measures. Thus, from the dominance matrix, it can be seen that Kendall's tau measures the proportional agreement between the dominance scores on two variables. Therefore, Kendall's tau coefficient, as a summary measure, is an average of proportional increase.

As is the case with OLS regression, a second predictor makes the interpretation more complicated but there are analogies in ordinal multiple regression (OMR). However, there are some unresolved issues in OMR which again force us to "Think Different." First of all, OMR does not yield truly partialled values. Similar to Marascuilo and McSweeney's discussion of the relationship of Kendall's tau to Pearson's r, one cannot interpret the coefficients that result from OMR as OLS regression weights. Furthermore, although Kendall (1970) developed a "partial tau," its properties are quite different from those in OLS. For example, suppose there are three variables that have positive intercorrelations and a trivariate normal distribution. Although the first two are statistically independent, conditional on the third ($r_{12.3} = 0$), Kendall's partial tau will not be zero (Cliff, 1996). Thus, Long rightfully warns that the "OMR function is much more ambiguous in its specification of the relationship between the weights and the criterion. In fact, an algebraic formula expressing the criterion in terms of the weighted predictors is not possible" (*p*. 47). This means that there is no final "regression equation" where a line or plane of best fit is described. Predicted values for each subject are not rendered. Long states that it would be possible "if $\hat{d}_{ihy}$ were used in the loss function" (*p*. 47). To elaborate, one can use either equation (4) or (6) and calculate, $\hat{d}_{ihy} = .40(d_{ih1}) + .33(d_{ih2})$, a "prediction equation" for the $n(n - 1)$ pairwise dominance scores. Then based on these $n(n - 1)$ predicted values an "average predicted dominance score" of the form $\hat{d}_{iy} = \sum_h(\hat{d}_{ihy})/(n - 1)$ can be computed for each of the n subjects. It can be shown that both $\hat{d}_{ihy}$ and $\hat{d}_{iy}$ sum to zero as would standardized predicted values from an OLS regression. However, this approach violates the logic of ordinal analysis. Therefore, only a verbal description of the functional relationship between the weights and the criterion is appropriate. Thus, the OMR weights are the constants that when applied to the predictor dominance scores best predict order on the criterion, "best" meaning that Q is optimal (*p*. 47). Thus using equation (5), Q = .5945; however, this is only the optimization of the weights. That is, Q can be viewed as analogous to Multiple R^2, but it is not the "variance accounted for" typically associated with OLS regression. Furthermore, to date there is not an omnibus test for Q analogous to the F-test for the full model R^2. Given the confidence interval approach taken by Long this may be less problematic. Still Q is only a statistic descriptive of the loss function. Thus, in its current state OMR has many statistical and interpretive limitations despite the compelling arguments of Cliff (1996). Possibly the OMR methodology forces us to "Think *too* Different." When applied to group comparison research, however, the dominance metric approach has many foreseeable advantages. In research in which two or more groups are compared, ANOVA models are applied to test differences in means which addresses the question: "Do the groups have different average values?" Cliff (1993) suggests that through using the dominance metric one can answer the question many behavioral researchers *really* want to ask, "Which group has higher scores?" Yet a similar and even more general question is: "Did the groups respond differently?"

Kromrey and Hogarty (*pp*. 70-82) address the differences among these three questions. They present an interesting situation in which two groups are compared on an ordered categorical response, as opposed to analyzing a dependent variable that is truly continuous in nature. This is a common practice in a variety of educational and psychological studies where Likert-type responses are elicited and groups are subsequently compared. Kromrey and Hogarty evaluate the statistical properties of four general procedures (*t*-test, Pearson chi-square test, Cliff's *d*, and Cumulative Logit model). They contend that despite the differences among the statistical null hypotheses tested, each of these procedures may be used to test the same, "conceptual" research hypothesis (*p*. 70). Although these methods may seem to address conceptually similar research questions, statistically they are *not* the same. Thus, a review of the procedures, their null hypotheses, and the questions addressed should be examined carefully.

Again, the most general question is, "Did the groups respond differently?" It is most likely to be addressed with the Pearson chi-square test for contingency tables which for two groups has the following null hypothesis:

$$\mathrm{H}_{\mathrm{O}(\pi)}: \pi_{1k} = \pi_{2k}, \text{ for all } k \text{ categories.}$$

The question of "Which group has higher scores?" is often thought of terms of the *t*-test. However, this issue is actually more in line with Cliff's *d* statistic. It tests the null hypothesis that the probability that a randomly selected member (i) of one population has a higher response than a randomly selected member (j) of the second population is equal to the reverse probability. That is, the probability that the scores from one group are higher minus the probability that a second group's scores are higher is equal to zero:

$$\mathrm{H}_{\mathrm{O}(\delta)}: \delta = \Pr(y_{i1} > y_{j2}) - \Pr(y_{i1} < y_{j2}) = 0.$$

These population probabilities are measured by the frequencies in the samples. It should be noted that the *d* statistic is equivalent to Kendall's tau performed with a dummy code representing the group distinctions, and thus, Cliff's *d* can extend into multiple group and factorial designs (Cliff, 1996).

The most specific of the three research questions is, "Do the groups have different average values?" It is addressed by the independent samples *t*-test with the following null hypothesis:

$$\mathrm{H}_{\mathrm{O}(\mu)}: \mu_1 - \mu_2 = 0.$$

A fourth approach investigated by Kromrey and Hogarty is a Cumulative Logit model suggested by Agresti (1989). In the current situation, this method treats the categorical response as an ordinal variable and the grouping variable as dichotomous. The impetus for the Cumulative Logit model is that the Pearson chi-square test was designed for variables that have unordered categories. Therefore, it detects any type of deviation from the null hypothesis $\mathrm{H}_{\mathrm{O}(\pi)}$. If the variable is ordinal, however, the categorical data may be represented with fewer degrees-of-freedom which for a fixed noncentrality structure increases the statistical power of a test. Thus, the Cumulative Logit model detects only monotonic deviations but these are the ones of most importance with ordinal variables (Agresti, 1989, *p*. 298).

To explicate this approach, Beasley and Schumacker (1995) demonstrated a method for orthogonally partitioning a contingency table using ANOVA contrast codes. One thing not pointed out by Beasley and Schumacker is that in the situation presented by Kromrey and Hogarty, the contingency table can be partitioned in order to test for mean differences (i.e., $H_{O(\mu)}$). Suppose a linear polynomial contrast (i.e., [-2 -1 0 1 2] for 5 categories, [-3 -2 -1 0 1 2 3] for 7 categories) is applied to the categorical variable. If this contrast variable is weighted by the frequencies and then correlated with a dummy code representing the two groups, the result is identical to the *t*-test. Therefore, with 5 ordered categories and two groups, the null hypothesis for the linear polynomial contrast (ψ) of population proportions in a contingency table,

$$H_{O(\psi)}: -2\pi_{11} -1\ \pi_{12} +0\ \pi_{13} +1\ \pi_{14} +2\ \pi_{15}$$
$$+2\pi_{21} +1\ \pi_{22} +0\ \pi_{23} -1\ \pi_{24} -2\ \pi_{25} = 0\ ,$$

is equivalent to evaluating differences in population means, $H_{O(\mu)}$: μ_1 - μ_2 = 0. The Cumulative Logit model uses a similar approach (Agresti, 1989, *p*. 294); however, a Logit model instead of a Pearsonian model is used. Thus, the null hypothesis associated with the cumulative Logit model ($H_{O(\beta)}$: $\beta = 0$, see *p*. 73 or Agresti, 1989 for details), although not identical to $H_{O(\mu)}$, is extremely similar in concept. Differences among these statistical hypotheses will be discussed later.

When evaluating the performance of these four procedures, one must consider that any test of statistical significance has assumptions. The assumption that each of the observations are *independent* of each other applies to all these procedures. Importantly, the independent *t*-test has the additional assumptions that the two groups are sampled from identical (i.e., *homogeneous variances*), *normal* (i.e., skew and kurtosis of zero) populations. The assumption of homogeneous variances translates into the notion that the group effect is "additive." As a point of distinction, the Cumulative Logit model can be interpreted as the multiplicative effect of the grouping variable on the cumulative odds. Because these odds for cumulative probabilities are expressed in logits, however, this multiplicative effect can be interpreted as "additive." The Cumulative Logit model investigated by Kromrey and Hogarty implies a uniform association of cumulative odds ratios and is referred to as the "Proportional Odds" model (Agresti & Finlay, 1997, *p*. 601). Therefore, this Cumulative Logit Model assumes that the group effect is the same for each cumulative probability, an assumption analogous to an additive model (Agresti, 1989, *p*. 293). Furthermore, under the conditions imposed by Kromrey and Hogarty, the Cumulative Logit model performed similarly to the *t*-test in terms of Type I error (e.g., Table 4, *p*. 74) and power rates (e.g., Table 6, *p*. 78). Unlike the *t*-test, however, the independence of the variables corresponds to the distribution of the response variable being identical, not necessarily *identical and normal*. Therefore, similar to the Pearson chi-square and Cliff's *d*, which are relatively "distribution-free," the Cumulative Logit model makes no assumption about the shape of the response variable. Moreover, like the Pearson chi-square and Cliff's *d*, it can be sensitive to differences in variance and shape even when population means are identical. Thus, the Cumulative Logit model tests a statistical hypothesis that is different from $H_{O(\mu)}$, a topic explicated later.

Kromrey and Hogarty's results confirmed that the Pearson chi-square test should not be used with small sample sizes which accentuates the need for an alternative procedure such as the Cumulative Logit model that reduces the hypothesis degrees-of-freedom in a contingency table analysis. That aside, it is interesting that most tests were generally acceptable for testing the null hypothesis of identical population distributions, but the *t*-test gave the most consistent Type I error rate (see Fig. 1, *p*. 76). The Type I error results (e.g., Table 4, p. 74) also showed that even when the conditional distribution of the dependent variable was highly skewed, the *t*-test was generally robust to violations of the normality assumption thus confirming the seminal work of Norton (1952, cited in Lindquist, 1956) and Boneau (1960). It should be noted, however, that under the conditions simulated the conditional distributions for *Y* were identical in that the population values for variance, skew, and kurtosis, as well as the population means, were the same for both groups. Therefore, the null hypotheses for all procedures were true. Thus, because of the robustness of the *t*-test to violations of the normality assumption, the three research questions are considered the same if the groups have identical distributions in terms of variance, skew, and kurtosis. However, one must consider that outside of violating the normality assumption, the Type I error simulation conditions favored the parametric *t*-test (i.e., identical conditional distributions). Although the Type I error results are valid, they are limited in the sense that there are many situations in which some of the null hypotheses are false while others are true. Moreover, it is difficult to reconcile one procedure being more "robust" when they have different assumptions. That is, a test cannot be robust to a condition for which it makes no assumption (Huber, 1991).

The Power results were even more difficult to interpret because in the conditions simulated, all three null hypotheses were false but to different extents (see Table 8, *p*. 82). Furthermore, because some tests are sensitive to different parameters, a researcher may confirm the "conceptual" research hypothesis for a variety of reasons. For example, the Pearson chi-square test was powerful because it can detect a variety of differences (i.e., mean, variance, skew, kurtosis). By contrast, the *t*-test detects very specific differences. It is designed to detect differences in means but can be sensitive to differences in variance. Thus, as compared to evaluating the robustness of these tests, it is even more problematic to discern which is most "powerful" when the null hypotheses tested are different. To elaborate, the chi-square null hypothesis is the most general. Consequently, if $H_{O(\pi)}$ is true, then $H_{O(\mu)}$, $H_{O(\delta)}$, and $H_{O(\beta)}$ are also true. However, a true $H_{O(\mu)}$ does not imply that $H_{O(\pi)}$ is true. Likewise, a true $H_{O(\delta)}$ does not imply that $H_{O(\delta)}$ is true. This is also the case for $H_{O(\beta)}$.

Imagine the following tables show the population probabilities (π_k) for each of the K = 5 ordered categories in each group. Situation One is identical to the moderately skewed distribution condition simulated by Kromrey and Hogarty for assessing Type I error rates.

Situation One

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	60	10	10	10	10
Group 2	60	10	10	10	10

In this case all four null hypotheses are true. By contrast, imagine the following scenario.

Situation Two

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	44	4	4	4	44
Group 2	10	20	40	20	10

In this case $\mu_1 = \mu_2 = 3$, and thus, $H_{O(\mu)}$ is true. $H_{O(\delta)}$ and $H_{O(\beta)}$ are also true. However, $H_{O(\pi)}$ is false again demonstrating that the Pearson chi-square test of $H_{O(\pi)}$ is sensitive to parameters other the mean differences. It should also be noted that the homoscedasticity assumption of the *t*-test is violated in that $\sigma^2_1 = 3.6$ and $\sigma^2_2 = 1.2$. Therefore, the *t*-test may not maintain a Type I error rate near the nominal alpha in this case (i.e., it can be sensitive to differences in variance).

In the following scenarios the difference between the *t*-test, Cliff's *d*, and the Cumulative Logit Model can be further demonstrated.

Situation Three

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	6	34	30	14	16
Group 2	16	14	30	34	6

In this case the two distributions have identical values for the population mean ($\mu_1 = \mu_2 = 3$), variance ($\sigma^2_1 = \sigma^2_2 = 1.36$), and kurtosis ($\gamma^4_1 = \gamma^4_2 = -0.80$). Therefore, the null hypothesis for the *t*-test is true. The population skews are different ($\gamma^3_1 = -0.39$, $\gamma^3_2 = 0.39$). Furthermore, $H_{O(\pi)}$, $H_{O(\delta)}$, and $H_{O(\beta)}$ are false. Importantly, Cliff's *d* and the Cumulative Logit Proportional Odds model can be sensitive to differences in skew even when population means and variances are identical. However, if the means are the same and both distributions are symmetric, not necessarily identical (e.g., Situation Two) both $H_{O(\delta)}$ and $H_{O(\beta)}$ are true (see Vargha & Delaney, 1998, for a discussion of what they call Stochastic Homogeneity).

To further accentuate how Cliff's ordinal method and Agresti's Cumulative Logit model forces us to "Think Different," all four population moments are different ($\mu_1 = 3.0$, $\mu_2 = 3.1$; $\sigma^2_1 = 0.96$, $\sigma^2_2 = 3.60$; $\gamma^3_1 = -0.34$, $\gamma^3_2 = 0$; $\gamma^4_1 = -0.55$, $\gamma^4_2 = -1.98$) in Situation Four, but $H_{O(\delta)}$ and $H_{O(\beta)}$ are true.

Situation Four

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	4	36	32	22	6
Group 2	44	4	4	4	44

Also, imagine a situation where Cliff's *d* would be equal to 1.0. That is, every subject in Group 1 (n_1 = 100) has a higher score than every subject in Group 2 (n_2 = 100). Furthermore, suppose that 50 people in Group 1 responded to category 5 and the other 50 endorsed category 4.

Situation Five

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	0	0	0	50	50
Group 2	0	50	50	0	0

In terms of maintaining a Cliff's *d* of 1.0, it does not matter what pattern of 3, 2, or 1 categories is endorsed by Group 2.

Situation Six

Probabilities	π_1	π_2	π_3	π_4	π_5
Group 1	0	0	0	50	50
Group 2	50	0	0	0	0

That is, regardless of whether Group 2 responds to categories 3 and 2 only (Situation Five) or all of them endorse category 1 (Situation Six), Cliff's *d* would still be equal to 1.0. Thus, in this scenario, the *d* statistic can be contrasted with the *t*-test in the sense that Cliff's *d* considers rank position and dominance rather than average magnitude. Although the Cumulative Logit Proportional Odds model is somewhat sensitive to these differences in magnitude (Agresti, 1989), in general it seems more similar to Cliff's *d* than to the *t*-test, at least statistically.

It would be interesting to see how the Cumulative Logit model performs empirically under the various conditions elaborated, especially with between group differences in variance and skew (e.g., Situations Two through Four). Agresti (1989, *p.*

294) indicates that the independence of X and Y (i.e., $H_{O(\beta)}$ is true) corresponds to the distribution of the ordered categorical response (Y) being the same for each level of X (the grouping variable). In Situation Two, however, the β parameter is zero although the distributions of the ordered categorical responses are not identical for both groups which presents a violation of the Proportional Odds model. Therefore, it would also be interesting to determine whether the Cumulative Logit model performs more similar to Cliff's *d*, the Pearson chi-square, or to the *t*-test under such conditions.

Because of these statistical issues, the conceptual differences, and other previously elaborated arguments, Cliff (1996) contends that δ (and *Q* for OMR) are *NOT* just surrogates for OLS solutions; they are parameters worth estimating in their own right. From this perspective, Cliff's *d*, Kendall's tau, and OMR make parametric use of "nonparametric" statistics. For such statistical procedures, Bradley (1968) suggested the term "distribution-free," while Cliff prefers the term "ordinal methods."

Bradley (1968) and Zimmerman (1996) have pointed out that much of the confusion concerning the use of nonparametric methods lies in the treatment of nonparametric tests as "different" in most textbooks when actually many nonparametric tests are often algebraic reduction of OLS parametric tests performed on ranks (or signs or a dominance matrix). Under the basic assumptions of parametric tests, ranks have known means and variances. This allows the parametric formula to simplify which in turn makes it seem different. However, many of the problems associated with the original data can be inherited by the ranks (Zimmerman, 1996, 1998). Therefore, they may not be as "robust" as commonly believed. It is also true that the ordinal methods (i.e., Cliff's *d*, OMR) are OLS solutions for the dominance matrix (see Long, *p*. 46). Yet, the dominance matrix is a transformation that partially changes the *meaning* of the score. Therefore, the associated hypotheses are different both statistically and conceptually. Given its statistical similarity to Cliff's *d* , the same may also be said for the Cumulative Logit model.

The differences among the statistical hypotheses of parametric and alternative procedures has been seen as a drawback to employing "nonparametric" methods. Yet, Cliff (1996) argues that the hypotheses tested by alternative methods are often more in line with what behavioral researchers want to know from their data as compared to a null hypothesis of equal means. The point is that mean differences may not always be of interest (Olejnik, 1987). For example, in a randomized experiment if differences in variances occur then, an ANOVA model (*t*-test) may be inappropriate because a non-additive effect is suggested. That is, differences in variance indicate that the treatment did something to change the variability and thus a test of means may not be entirely appropriate. Furthermore, heterogeneous variances may also indicate some non-additive, interaction effect that has not been examined. This emphasizes the importance of data screening, data exploration, descriptive statistics, and graphical display in order to evaluate "What the data are trying to tell us." Moreover, instead of employing parametric statistical tests ritualistically, perhaps researchers should "Think Different" and perform alternative procedures. Again, the conclusions may be similar conceptually. Yet, there is the distinct possibility that the results from an alternative procedure may force investigators to "Think Different" about their research questions.

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

References

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, *105*, 290-301.

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

Beasley, T. M., & Schumacker, R. E. (1995). Multiple regression approach to analyzing contingency tables: Post-hoc and planned comparison procedures. *Journal of Experimental Education*, *64*, 79-93.

Birkes, D., & Dodge, Y. (1993). *Alternative Methods of Regression*. New York: Wiley.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, *57*, 49-64.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

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

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, *31*, 331-350.

Draper, N., & Smith, H. (1981). *Applied Regression Analysis* (2nd edition). New York, NY: Wiley.

Harwell, M. R., & Serlin, R. C. (1989). A nonparametric test statistic for the general linear model. *Journal of Educational Statistics*, *14*, 351-371.

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Huber, P. (1991). *Robust Statistics*. New York: Wiley.

Huynh, H. (1978). A comparison of four approaches to robust regression. *Psychological Bulletin*, *92*, 505-512.

Kendall, M. G. (1970). *Rank Correlation Methods* (4th ed.). London: Charles Griffin.

Lindquist, E. F. (1956). *Design and analysis of experiments in psychology and education* (2nd ed.). Boston: Houghton-Mifflin.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks-Cole.

Sprent, P. (1993). *Applied nonparametric statistical methods*. London: Chapman & Hall.

Norton, D. W. (1952). *An empirical investigation of some effects of non-normality and heterogeneity on the F-distribution*. Unpublished doctoral dissertation, State University of Iowa.

Olejnik, S. (April, 1987). Teacher education effects: Looking beyond the means. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational & Behavioral Statistics*, *23*, 170-192.

Witkin, H. A., & Goodenough, D. R. (1981). *Cognitive styles: Essence and origins*. New York: International Universities Press.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education*, *64*, 351-362.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, *67*, 55-68.

Zimmerman, D. W., & Zumbo, B. (1993). Relative power of the Wilcoxon test, the Friedman test, and the repeated-measures ANOVA on ranks. *Journal of Experimental Education*, *62*, 75-86.