DOCUMENT RESUME

ED 445 022 TM 031 627

AUTHOR	Witta, E. Lea
TITLE	Four Methods of Handling Missing Data in Predicting Educational Achievement.
PUB DATE	2000-04-00
NOTE	34p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Estimation (Mathematics); *High School Students; High Schools; *Prediction; *Research Methodology
IDENTIFIERS	*EM Algorithm; *Missing Data; National Education Longitudinal Study 1988

ABSTRACT

Four methods of handling missing data were applied to missing values for variables selected from the National Education Longitudinal Study of 1988. Variables used were those selected by K. Singh and M. Ozturk (1999) for a study concerning high school students' academic achievement and work. Samples selected consisted of 100 cases, 300 cases, and 500 cases. The proportion of incomplete cases was manipulated to represent 30%, 50%, and 70% for each sample. In addition, composite variables were created and tested. Results indicate the expectation maximization (EM) algorithm and regression procedures provide accurate estimates under all conditions. Listwise and pairwise deletion were effective with small proportions of missing data and when composites were created. (Contains 1 figure, 8 tables, and 19 references.) (Author/SLD)

Running Head: missing data - predicting achievement

Four methods of handling missing data in predicting educational achievement

E. Lea Witta

University of Central Florida

Department of Educational Foundations

Paper presented at the annual conference of the American Educational Research Association, New Orleans, April 24-28, 2000.

For further information contact: Lea Witta lwitta@mail.ucf..edu

Abstract

Four methods of handling missing data were applied to missing values for variables selected from the National Educational Longitudinal Study of 1988. Variables used were those selected by Singh and Ozturk (1999) for a study concerning high school students' academic achievement and work. Samples selected consisted of 100 cases, 300 cases, and 500 cases. The proportion of incomplete cases was manipulated to represent 30%, 50%, and 70% for each sample. In addition, composite variables were created and tested. Results indicate the EM algorithm and regression procedures provide accurate estimates under all conditions. Listwise and pairwise deletion were effective with small proportions of missing data and when composites were created.

Four methods of handling missing data in predicting educational achievement

When data is analyzed in survey research, often there are missing values. If the mechanism causing the missing values is known, the solution to this problem may be incorporated in the study. Many times, however, the mechanism causing the missing values is not known. Ignoring this problem may lead to analysis of data that is of dubious value.

In addition, different methods of handling missing values may produce different results. When Jackson (1968) entered data on all the available variables in a discriminant analysis, the significance of the regression coefficients of individual variables, as well as the interpretation of the importance of these variables, changed with the missing value method used. Witta and Kaiser (1991) also reported that the regression coefficients and total variance accounted for by the variables changed depending on the method used to handle missing values. After re-analyzing three studies of private/public school achievement, Ward and Clark III (1991) concluded that the method used to handle missing data influenced the outcome of these studies.

In using the National Educational Longitudinal Study database to investigate the effects of part-time work on school outcomes Singh and Ozturk (1999) eliminated more than half of the selected cases by listwise deletion of the incomplete data. In addition, composite variables were created to help explain the school outcomes.

Statement of the Problem

The purpose of the current study was tri-fold: (a) to investigate the effectiveness of four methods of handling missing data using the 26 variables in the Singh and Ozturk (1999) study, (b) to compare the effectiveness of the missing data methods after creating composite variables, and

(c) to compare the effectiveness of each missing data treatment using composite variables to the same treatment when using the individual predictor variables. Effectiveness was defined as the probability of accurately predicting achievement on standardized tests. Effectiveness of the missing data methods was assessed by manipulating the proportion of cases containing missing values, the sample size, and the number of variables. The missing data handling methods studied were listwise deletion, pairwise deletion, regression and expectation maximization. Sample sizes investigated were 100, 300, and 500. The proportion of incomplete cases in each sample was 30%, 50%, and 70%.

Methods Studied

Listwise Deletion

Listwise deletion is probably the most frequently used method of handling missing data and is available as a default option in several statistical software programs. This method discards cases with a missing value on any variable and thus is very wasteful of data. Listwise deletion, however, has been shown to be more effective with low average intercorrelation, less than four variables and a small proportion of missing values (Chan, et.al., 1976; Haitovsky, 1968; Timm, 1970). The assumption of missing completely at random is crucial to the use of this method. It is more likely, however, to find the complete sample different in important ways from the incomplete sample (Little & Rubin, 1987). Problems for a researcher using this method include a reduction in power and an increase in standard error due to reduced sample size and the elimination of sub-populations.

Pairwise Deletion

When using pairwise deletion, covariances are computed between all pairs of variables having both observations, eliminating those that have a missing value for one of the two variables (Glasser, 1964). Means and variances are computed on all available observations. The assumption made is that the use of the maximum number of pairs and all the individual observations yield more valid estimates of the relationship between the variables. It is assumed that when two variables are correlated, information on one improves the estimates of the other variable. It is also assumed that the pairs are a random subset of the sample pairs. If these assumptions are true, pairwise deletion produces unbiased estimates of the variable means and variances (Hertel, 1976). When missing data are not missing completely at random, however, the correlation matrix produced by pairwise deletion may not be Gramian (Norusis, 1988).

Marsh (1998) investigated the estimates produced when using pairwise deletion for randomly missing data. From this study, which included five levels of missing data and three sample sizes, Marsh concluded parameter variability was explained, parameter estimates were unbiased, and only one covariance matrix was nonpositive definite.

Regression

Regression as an imputation method has many variations. The variations rely on information from other variables to estimate missing values. As the average intercorrelation and the number of variables from which these methods can obtain information increases, the regression methods, theoretically, perform better. Too many variables, however, can cause problems with over prediction (Kaiser & Tracy, 1988) and too high an average intercorrelation can result in a singular matrix. In these cases, regression does not perform well.

Variations in the regression methods include differences in methods of developing the initial correlation matrix (listwise deletion, pairwise deletion, and mean substitution) and the presence or absence of iteration procedures. Differences in regression methods also include the use of randomly selected residuals for iterations and assumptions of a normal distribution. Theoretically, the more variables considered that provide additional information, the better the estimate. Mundfrom and Whitcomb (1998) investigated the effects of using mean substitution, hot-deck imputation, and regression imputation on classification of cardiac patients. Mean substitution and hot-deck imputation correctly classified patients more frequently than regression imputation.

Expectation Maximization

Dempster, Laird, and Rubin (1977) recommended the use of the EM (expectation maximization) algorithm which imputes estimates simultaneously in an iterative procedure. The alternative is to estimate values and to adjust them one at a time using the Gauss-Seidel method. Both methods converge to the same final estimates, but the speed of convergence differs. The EM algorithm was advocated to hasten convergence. The E step of this algorithm finds the conditional expectation of the missing values. The M step performs maximum likelihood estimation as if there were no missing data. The primary difference between this procedure and the regression procedure is that the values for the missing data are not imputed and then iterated. The missing values are functions based on the conditional expectation (Little & Rubin, 1987). This method of handling missing data represents a fundamental shift in the way of thinking about missing data (Schafer & Olsen, 1998).

Pattern of Missing Values

All of the missing data handling procedures discussed require data missing at random (MAR) or missing completely at random (MCAR). Yet Cohen and Cohen (1983) suggested that in survey research the absence of data on one variable may be related to another variable and may be due to the value of the variable itself. When investigating simultaneously missing values, Witta (1996/97) found concurrently missing values (p<.001) in three of four samples using data from a national database.

Schafer and Olsen (1998), however, argue convincingly that "every missing-data method must make some largely untestable statistical assumptions about the manner in which the missing values were lost" (p551). Consequently, when analyzing real data, researchers typically assume missing at random.

Procedure

All high school seniors who had reported working during their senior year of high school and for whom base-year and first follow-up data were available were included in this study. The initial sample contained the 26 variables used in the Singh and Ozturk study for 4664 subjects. These subjects were split into three populations: those containing one or more missing values but less than 14 and not having any missing values for standardized test scores (n=504), those containing more than 13 missing values (n=19) or missing values on the dependent standardized test variables (n=1038), and those containing no missing values on any variable (n=3103). The 19 subjects having missing values for more than half the variables and the 1038 containing missing values for the standardized test scores were eliminated from further analysis. The remaining two populations (n=3607) were used to create samples for analysis.

Creating Test Samples

A sample containing 500 cases was randomly selected from the non-missing population. This target sample was duplicated twice. A sample of 350 cases was randomly select from the missing population. These cases were used to replace an equal number of randomly selected cases from one of the target samples. This provided a test sample of 500 with 70% of the cases containing missing values. This process was repeated with the second target sample to provide a test sample with 50% (250) of the cases containing missing values. The process was repeated again with the third target sample to provide a test sample with 30% (150) of the cases containing missing values.

This entire procedure was repeated twice to provide test samples with 30%, 50%, and 70% of the cases containing missing values in test samples of 100 and 300 cases. Thus, 9 test samples were created. The missing values of each test sample were treated by each of the four missing data handling methods using SPSS 8.0 and SPSS Missing Data Analysis 7.3.

Analysis

To answer research question 1, "to investigate the effectiveness of four methods of handling missing data using the 26 variables in the Singh and Ozturk (1999) study", the SPSS missing data analysis 7.3 (Hill, 1997) subroutine was used to estimate values for regression and the EM algorithm. Each individual standardized test was then regressed on the remaining variables (not on other standardized tests) using the data produced by the missing analysis procedure and the pairwise and listwise procedures within the regression subroutine of SPSS 8.0. Predicted values from each regression were recorded. The mean vectors of the predicted values for each missing data method were then contrasted in MANOVA (multivariate analysis of variance).

To answer research question 2, "to compare the effectiveness of the missing data methods after creating composite variables", the mean of the four standardized test scores was used as the dependent variable. Composite predictor variables were created by determining the mean of the questions forming that construct (see Table A-1). When measurement scales differed, questions were converted to z scores prior to determining the mean.

After treatment by a missing data method the standardized test score mean was regressed on each of the test samples. The predicted standardized test score for each test sample was compared to the actual standardized test mean using analysis of variance (ANOVA) with Dunnett's test for comparing all treatments to a control (Howell, 1992) used as a post hoc.

To answer research question 3, "to compare the effectiveness of each missing data treatment using composite variables to the same treatment when using individual variables", the composite mean standardized test score was regressed on the individual questions after treatment by a missing data method. A predicted standardized test score was recorded for each method. The predicted score for each missing data method was contrasted with the actual score and with the score produced by that method using ANOVA with Dunnett's and the Tukey post hoc tests.

Results and Discussion

Initially data was examined to determine the pattern of missing values as depicted in Table A-2. When individual questions were used, data was never missing completely at random. This assumption was only violated in one condition (70% incomplete of 300) when composite variables were used. As expected and as shown in Figure 1, use of composite variables increased the number of complete cases in each condition.

Insert Figure 1 About Here

When the mean vectors of the four standardized tests produced by each missing data method and the actual mean vector were compared, statistically significant differences were detected in three conditions; when 50% of the cases were incomplete with a sample size of 500, and when 70% of the cases were incomplete with sample sizes of 300 and 500. These results are depicted in Table 1.

Insert Table 1 About Here

When 50% of the 500 cases were incomplete, none of the means produced by listwise deletion accurately reproduced the target means (see Table A-3). Under these conditions, pairwise deletion could not accurately replicate the standardized mathematics mean. All other missing data methods adequately reproduced the target means.

When 70% of the cases were incomplete, the standardized test means produced by listwise deletion did not accurately reproduce the target means whenever the sample 300 or 500 cases. Under these condition, pairwise deletion reproduced adequately the target standardized reading test mean and the target standardized history mean, but not mathematics or science when the sample size was 300, but could not accurately reproduced any of the target means when the sample size was 500. The EM algorithm and regression procedures accurately reproduced the target sample means under all conditions. It should also be noted, the difference in missing data method never explained more than 1% of the variance in mean vectors and the actual difference between predicted and actual mean never exceeded 5 points. Thus, in response to research question 1, the EM algorithm and regression missing data procedures were more effective in reproducing mean vectors than were pairwise or listwise deletion. In fact, both the EM and regression procedures produced mean vectors almost identical to the target mean vector. There was a reduction, however, in variability as has been noted by other researchers.

When composite variables were created, there were no statistically significant differences in predicting standardized test score based on missing data method under any conditions as shown in Table 2. Again, method of handling missing data did not explain more than 1% of the variance in standardized test score. Apparently the reduction in proportion of cases was beneficial to the listwise and pairwise deletion methods. Composite standardized test means for each missing data method as well as actual means are included in Table A-4.

Insert Table 2 About Here

When the predicted composite standardized test scores (created by regressing composite test score on individual questions) produced by each missing data method were contrasted with the target composite test score, results were similar to those using the mean vector of each test score. As shown in Table 3, statistically significant differences were detected when the sample size was 500 with 50% incomplete cases, and when the sample size was 300 or 500 with 70% incomplete cases. Whenever these differences were detected, listwise and pairwise deletion were significant contributors (see Table A-5). The actual difference between the predicted and actual test mean was not more than 5 points. In this instance, however, 2% of the variance in test score could be attributed to group.

Insert Table 3 About Here

Conclusion

This study used one sample for each set of conditions. Consequently it is limited in generalizability. In addition, there were a relatively large number of variables (26) with a small sample size (100). Thus, larger samples may produce different results. Considering these limitations, the following conclusions offered.

Although statistically significant differences in standardized test scores were detected between the missing data method treatments, the variance accounted for by those differences was never more than 2%. Use of imputation procedures (EM and regression), however, provide more responses and correspondingly higher power. While reduction in variability by the EM and regression procedures is troubling, these methods provide greater power and produced more accurate estimates of mean vectors. Thus, it is recommended that researchers begin to implement these procedures more frequently.

The use of composite variables produced no differences based on missing data method. Because the use of multiple similar variables provides more reliable indicators (although less precision) of a construct, this procedure is also recommended. If researchers do not wish to use procedures such as the EM algorithm or regression, creating composite variables provides an alternative that helps reduce the number of incomplete cases - possibly to an acceptable level.

Finally, when the proportion of incomplete cases was small (30%), there were no statistically significant differences in the performance of the missing data methods. Therefore, if the proportion of incomplete cases is small, any procedure will work. The best solution, however, is no missing data.

Further research is needed to investigate more thoroughly the problems associated with variability reduction with the EM algorithm and regression procedures. In addition, further research is needed using actual data with real patterns of missing values.

References

Chan, L.S., Gilman, J.A., & Dunn, O.J. (1976). Alternative approaches to missing values in discriminant analysis. *Journal of the American Statistical Association, 71*, 842-844.

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

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

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

Hertel, B.R. (1976). Minimizing error variance introduced by missing data in survey analysis. *Sociological Methods & Research, 4*, 459-474.

Hill, M.A. (1997). *SPSS Missing Value Analysis 7.5* [Computer program manual]. Chicago: SPSS Inc.

Jackson, E.C. (1968). Missing values in linear multiple discriminant analysis. *Biometrics, 24*, 835-844.

Joreskog, K.G. & Sorbom, D. (1988). *Lisrel 7 A guide to the program and applications* (2nd ed.). Chicago: SPSS Inc.

Kaiser, J. & Tracy, D.B. (1988). Estimation of missing values by predicted score. *Proceedings of the Section on Survey Research, American Statistical Association 1988*. 631-635.

Little, R.J.A., & Rubin, D.R. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Marsh, H.W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling, 5* (1), p 22-36.

Mundfrom, D.J. & Whitcomb, A. (1998). *Imputing missing values: The effect on the accuracy of classification*. Paper presented at the annual meeting of the American Educational Research Association, San Diego. ED419817.

Norusis, M.J. (1988). *SPSS-X Introductory Statistics Guide: Release 3* [Computer program manual]. (pp 107-108). Chicago: SPSS Inc.

Schafer, J.L. & Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavior Research, 33* (4), p 545-571.

Singh K., & Ozturk, M. (1999). *Part-time work and school-related outcomes for high school seniors: An analysis of NELS:88*. Paper presented at the 1999 Annual Conference of the American Educational Research Association, Montreal, Canada.

Timm, N.H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika, 35*, 417-437.

Ward, Jr., T.J. & Clark III, H.T. (1991). A reexamination of public-versus private-school achievement: the case for missing data. *Journal of Educational Research, 84*, 153-163.

Witta, E.L. (1996/97). Randomness of missing values in survey data. *Louisiana Education Research Journal, XXII* (2), p 73-86.

Witta L. & Kaiser, J. (1991, November). *Four methods of handling missing data with GSS-84*. Paper presented at the meeting of the Mid-South Educational Research Association, Lexington, KY

Tables & Figures

Figure 1

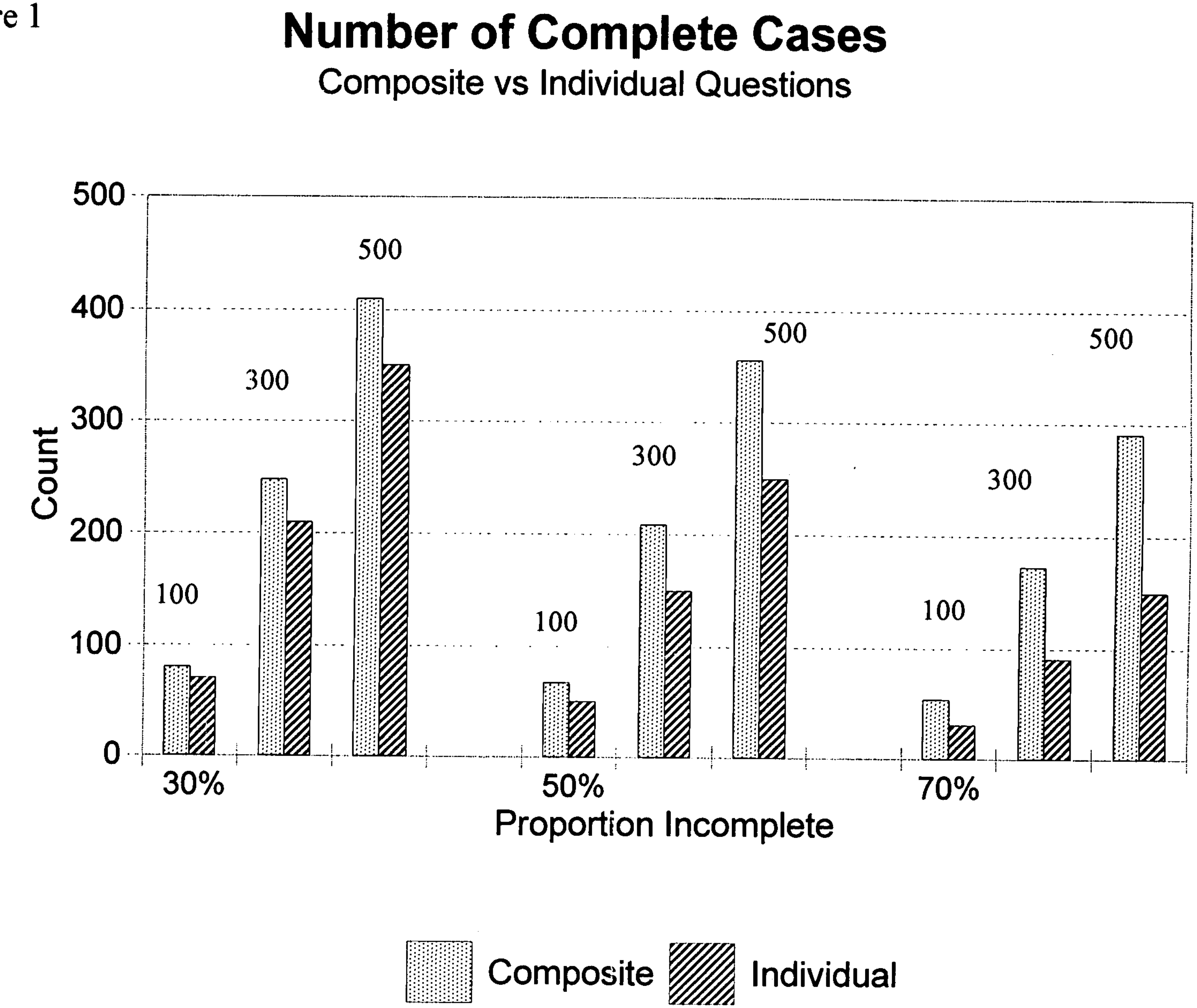
Number of Complete Cases
Composite vs Individual Questions
Count
500
400
300
200
100
0
100
300
500
100
300
500
100
300
500
30%
50%
70%
Proportion Incomplete
Composite
Individual

Table 1

Tests of Statistical Significance Using Individual Questions

n	Wilks' λ	F	df[a]	df[b]	Eta^2
30%					
100	0.985	0.418	16	1320.4	<.01
300	0.992	0.632	16	4008.9	<.01
500	0.995	0.639	16	6697.3	<.01
50%					
100	0.963	0.929	16	1198.2	0.01
300	0.984	1.236	16	3639.2	<.01
500	0.981	**2.38****	16	6086.3	0.01
70%					
100	0.976	0.545	16	1073.0	0.01
300	0.965	**2.40****	16	3275.7	0.01
500	0.976	**2.70****	16	5472.2	0.01

Note. [a] hypothesis. [b] error. *p<.05. **p<.01.

Table 2

Tests of Statistical Significance Using Composite Questions

n	MS (between)	MS (within)	F	df (between)	df (within)	Eta Squared
30%						
100	6.18	57.46	0.107	4	455.0	<.01
300	20.60	35.54	0.58	4	1391.0	<.01
500	30.62	38.02	0.805	4	2152.0	<.01
50%						
100	17.85	51.00	0.35	4	429.0	<.01
300	3.28	38.50	0.085	4	1313.0	<.01
500	9.42	37.53	0.251	4	2205.0	<.01
70%						
100	31.96	47.29	0.676	4	415.0	0.01
300	2.15	42.66	0.05	4	1241.0	<.01
500	1.85	41.26	0.045	4	2077.0	<.01

Note. *p<.05. **p<.01.

Table 3

Tests of Statistical Significance Using Individual Questions with Composite Dependent

n	MS (between)	MS (within)	F	df (between)	df (within)	Eta Squared
30%						
100	44.54	66.82	0.62	4	435	<.01
300	92.27	41.28	2.33	4	1315	<.01
500	89.92	40.4	2.23	4	2195	<.01
50%						
100	119.04	59.35	2.01	4	395	0.02
300	101.8	44.97	2.26	4	1194	<.01
500	353.2	43.69	**8.08****	4	1995	0.02
70%						
100	98.05	69.04	1.42	4	354	0.02
300	325.27	52.33	**6.22****	4	1075	0.02
500	511.73	52.11	**9.82****	4	1794	0.02

Note. *p<.05. **p<.01.

Appendix

Table A-1

Composite Variable Questions

Composite Variable	Questions
Parttime Work[a]	F1S85 HOW MANY HRS DOES R USUALLY WORK A WEEK F2S88 CURRENT JOB, # HRS WORKED DURING SCHL YR
Attendance 10[a]	F1S10A HOW MANY TIMES WAS R LATE FOR SCHOOL F1S10B HOW MANY TIMES DID R CUT/SKIP CLASSES F1S13 HOW MANY DAYS WAS R ABSENT FROM SCHOOL
Attendance 12	F2S9A HOW MANY TIMES WAS R LATE FOR SCHOOL F2S9B HOW MANY TIMES DID R CUT/SKIP CLASSES F2S9C HOW MANY TIMES DID R MISS SCHOOL
Participation 10	F1S40A OFTEN GO TO CLASS WITHOUT PENCIL/PAPER F1S40B OFTEN GO TO CLASS WITHOUT BOOKS F1S40C OFTEN GO TO CLASS WITHOUT HOMEWORK DONE
Participation 12	F2S24A GO TO CLASS WITHOUT PENCIL/PAPER F2S24B GO TO CLASS WITHOUT BOOKS F2S24C GO TO CLASS WITHOUT HOMEWORK DONE
Homework 10	F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL
Homework 12	F2S25F1 TOTAL TIME SPENT ON HMWRK IN SCHOOL F2S25F2 TOTAL TIME SPENT ON HMWRK OUT SCHL
Grades 12	F2RHENG2 AVERAGE GRADE IN ENGLISH (HS+B) F2RHMAG2 AVERAGE GRADE IN MATHEMATICS (HS+B) F2RHSCG2 AVERAGE GRADE IN SCIENCE (HS+B) F2RHSOG2 AVERAGE GRADE IN SOCIAL STUDIES (HS+B)
Standardized Tests	F22XHSTD HISTORY/CIT/GEOG STANDARDIZED SCORE F22XMSTD MATHEMATICS STANDARDIZED SCORE F22XRSTD READING STANDARDIZED SCORE F22XSSTD SCIENCE STANDARDIZED SCORE

Note. [a] z-score

Table A-2

Data Patterns for the Samples Used

		Homework Time 10			Homework Time 12			Grades12		Other			Totals		Missing Completel at Random	
n	Condition	Out side	Both	In Schl	Out side	Both	In Schl	Grade 12	Engl 12	Abs ence	Other[a]	Partic 10	Com plete	Incom plete	χ^2	df
30%																
100	C[b]					2		15			3		80	50	47.68	34
	I[c]			2	2		2	13			11		70	30	**343.7***	289
300	C[b]		4			4		36			8		248	52	62.35	58
	I[c]	4	4	9		4	7	33		5	24		210	90	**621****	510
500	C[b]							75			16		409	241	55.68	46
	I[c]			11			11	72		8	48		350	150	**671****	584
50%																
100	C[b]		2			2		24			5		67	324	**62.52***	46
	I[c]	4		5	2	2	4	22			11		50	291	343.48	335
300	C[b]		5			5		75			6		209	241	55.44	46
	I[c]	5	4	11	5	5	11	72	4	8	25		150	150	**654***	584
500	C[b]		6			7		116			16		355	145	48.74	40
	I[c]	11	6	28	8	7	12	107		13	58		250	250	**922***	841
70%																
100	C[b]		2			3		35			3	4	53	117	60.48	47
	I[c]	4		6	2	3	2	33	2		18		30	70	484.6	422
300	C[b]		8			8		104			7		173	127	**85.64****	46
	I[c]	4	7	27	9	8	6	98	5	11	35		90	210	**824.9****	733
500	C[b]		9			13		165			12	10	291	209	72.6	75
	I[c]	12	6	37	16	12	13	153		15	86		150	350	**1248****	1097

Note. [a]=Patterns with <1% Missing Values are included in other. [b]=Composite. [c]=Individual Questions. * $p<.05$. ** $p<.01$.

Table A-3

Standardized Test Means by Proportion Incomplete, Sample Size, and Missing Data Method

Proportion Incomplete	Sample Size	Criteria	Listwise Mean	Listwise N	Listwise SD	EM Mean	EM N	EM SD	Regression Mean	Regression N	Regression SD	Pairwise Mean	Pairwise N	Pairwise SD	Target Mean	Target N	Target SD
30%																	
	100																
		Reading	51.69	70	8.06	50.38	100	8.23	50.38	100	7.97	51.39	70	8.07	50.38	100	10.1
		Math	52.12	70	8.46	50.46	100	9.04	50.46	100	8.83	51.78	70	8.66	50.46	100	10.4
		History	52.71	70	7.16	50.51	100	8.07	50.51	100	7.88	51.62	70	7.68	50.51	100	10.1
		Science	51.29	70	8.08	50.01	100	8.23	50.01	100	8.07	50.69	70	8.08	50.01	100	9.72
	300																
		Reading	52.02	210	5.04	50.81	300	5.75	51.56	210	5.77	50.81	300	5.98	50.81	300	8.93
		Math	52.74	210	6.1	51.17	300	6.46	51.89	210	6.55	51.17	300	6.64	51.17	300	9.17
		History	52.53	210	5.46	51.12	300	6.00	51.82	210	6.1	51.12	300	6.21	51.12	300	9.33
		Science	52.78	210	5.33	51.24	300	5.91	51.99	210	5.82	51.24	300	6.12	51.24	300	9.64
	500																
		Reading	53.06	350	5.35	52.02	500	5.98	52.02	500	5.89	52.6	350	5.65	52.02	500	9.3
		Math	53.05	350	6.21	51.72	500	6.80	51.72	500	6.76	52.35	350	6.54	51.72	500	9.77
		History	53.19	350	4.82	52.31	500	5.26	52.31	500	5.21	52.75	350	4.99	52.31	500	9.44
		Science	53.04	350	5.05	52.02	500	5.60	52.02	500	5.52	52.54	350	5.24	52.02	500	9.64
50%																	
	100																
		Reading	53.1	50	6.99	49.65	100	7.47	49.65	100	7.12	51.5	50	6.5	49.65	100	9.55
		Math	54.04	50	8.3	50.26	100	8.67	50.26	100	8.4	52.63	50	7.71	50.26	100	10.2
		History	52.16	50	7.34	49.84	100	8.10	49.84	100	7.68	51.3	50	7.02	49.84	100	9.93
		Science	52.86	50	7.76	50.18	100	7.23	50.18	100	6.7	51.49	50	6.45	50.18	100	9.91
	300																
		Reading	52.09	150	5.41	50.83	300	6.41	50.8	299	6.26	51.95	150	5.79	50.83	300	8.82
		Math	53.35	150	6.45	50.98	300	7.44	51	299	7.15	52.23	150	6.82	50.98	300	9.6
		History	52.32	150	5.3	51.28	300	5.94	51.28	299	5.68	52.36	150	5.28	51.28	300	9.4
		Science	52.35	150	6.2	50.99	300	6.51	51.03	299	6.18	52.18	150	5.86	50.99	300	10.1
	500																
		Reading	**53.06***	250	5.14	50.83	500	5.99	50.83	500	5.88	52.05	250	5.6	50.83	500	9.44
		Math	**53.16***	250	6.35	50.42	500	7.35	50.42	500	7.15	**51.93***	250	6.86	50.42	500	9.88
		History	**52.84***	250	4.83	50.94	500	5.52	50.94	500	5.25	52.04	250	5.02	50.94	500	9.37
		Science	**53.56***	250	4.99	50.9	500	6.03	50.9	500	5.72	52.2	250	5.54	50.90	500	9.77

Table A-3 (Continued)

Means by Proportion Incomplete, Sample Size, and Missing Data Method

Proportion Incomplete	Sample Size		Listwise			EM			Regression			Pairwise			Target		
			Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD
70%																	
	100																
		Reading	51.49	30	9.37	48.78	100	8.40	48.8	99	8.25	51.3	30	8.08	48.78	100	10.2
		Math	51.07	30	9.6	48.26	100	9.21	48.26	99	8.44	52.04	30	7.86	48.26	100	10.3
		History	51.86	30	8.27	49.07	100	8.12	49.14	99	7.81	51.82	30	7.29	49.07	100	9.93
		Science	51.15	30	9.06	49.04	100	8.33	49.04	99	7.55	52.08	30	8.21	49.04	100	10.4
	300																
		Reading	**53.01***	90	5.81	49.85	300	6.68	49.85	300	6.2	51.8	90	6.04	49.85	300	9.7
		Math	**53.86***	90	6.37	50.05	300	7.93	50.05	300	7.53	**52.81***	90	7.12	50.05	300	10.1
		History	**53.66***	90	5.92	50.39	300	6.80	50.39	300	6.48	52.6	90	6.22	50.39	300	9.73
		Science	**53.73***	90	6.19	50.15	300	6.53	50.15	300	6.35	**52.40***	90	5.69	50.15	300	10.1
	500																
		Reading	**53.13**	150	6.41	49.82	500	6.92	49.8	499	6.61	**52.28**	150	6.48	49.82	500	9.79
		Math	**52.89**	150	6.94	49.39	500	7.57	49.38	499	7.24	**52.04**	150	7.29	49.39	500	9.8
		History	**52.46**	150	6.09	49.71	500	6.47	49.7	499	5.91	**51.96**	150	6.04	49.71	500	9.81
		Science	**52.93**	150	6.04	49.55	500	6.53	49.55	499	6.09	**51.87**	150	6.23	49.55	500	9.89

Table A-4

Composite Standardized Test Means by Proportion Incomplete, Sample Size, and Missing Data Method

Proportion Incomplete	Sample Size	Listwise			EM			Regression			Pairwise			Target		
		Mean	n	SD	Mean	n	SD	Mean	n	SD	Mean	n	SD	Mean	n	SD
30%																
	100	50.86	80	7.37	50.78	80	7.30	50.34	100	6.97	50.34	100	6.74	50.34	100	9.18
	300	51.73	248	5.24	51.32	248	5.43	51.09	300	5.14	51.09	300	4.97	51.09	300	8.20
	500	52.17	409	5.51	51.32	409	5.43	52.02	500	5.17	52.02	500	5.04	52.02	500	8.51
50%																
	100	51.12	67	6.59	50.13	67	7.19	49.98	100	6.24	49.98	100	6.21	49.98	100	8.94
	300	50.88	209	5.91	51.23	209	5.92	51.02	300	5.19	51.02	300	4.76	51.02	300	8.42
	500	51.12	355	5.61	50.93	355	5.53	50.77	500	5.09	50.77	500	4.71	50.77	500	8.59
70%																
	100	48.24	53	7.56	50.13	67	7.19	48.79	100	5.59	48.79	100	4.04	48.79	100	9.28
	300	49.9	173	6.19	50.2	173	6.07	50.11	300	5.29	50.11	300	5.09	50.11	300	8.94
	500	49.6	291	6.22	49.79	291	5.96	49.62	500	5.19	49.62	500	4.75	49.62	500	8.89

Table A-5

Composite Standardized Test Means by Proportion Incomplete, Sample Size, and Missing Data Method with Individual Questions as Predictors

		Listwise			EM			Regression			Pairwise			Target		
Proportion Incomplete	Sample Size	Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD
30%																
	100	51.95	70	7.53	50.34	100	8.08	50.34	100	7.87	51.36	70	7.81	50.34	100	9.18
	300	52.52	210	5.25	51.09	300	6.05	51.09	300	5.87	51.81	210	5.88	51.09	300	8.20
	500	53.09	350	5.16	52.02	500	5.76	52.02	500	5.71	52.56	350	5.46	52.02	500	8.51
50%																
	100	53.04	50	7.17	49.98	100	7.60	49.98	100	7.20	51.73	50	6.64	49.98	100	8.94
	300	52.53	150	5.41	51.02	300	6.35	51.03	299	6.12	52.18	150	5.74	51.02	300	8.42
	500	**53.15****	250	5.09	50.77	500	6.10	50.77	500	5.90	**52.06****	250	5.66	50.77	500	8.59
70%																
	100	51.39	30	8.48	48.79	100	8.10	48.81	99	7.64	51.78	30	7.45	48.79	100	9.28
	300	**53.57****	90	5.67	50.11	300	6.76	50.11	300	6.47	**52.41***	90	6.09	50.11	300	8.94
	500	**52.85****	150	6.15	49.62	500	6.69	49.60	499	6.34	**52.03****	150	6.36	49.62	500	8.89

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Four Methods of handling missing data... achievement	
Author(s): E. Lea Witta	
Corporate Source: University of Central Florida	Publication Date: April 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	Level 2A	Level 2B
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: E Lea Witta	Printed Name/Position/Title: E. Lea Witta, Associate Prof	
Organization/Address: University of Central Florida PO Box 161250 Orlando, FL 32816-1250	Telephone: 407-823-3220	Fax: 407-823-5144
	E-mail Address: lwitta@mail.ucf.edu	Date: April 24, 2000

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706
Telephone: 301-552-4200
Toll Free: 800-799-3742
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)