

The Effects Of Different Methods Of Handling Missing Data
On Institutional-Level Models Of Student Persistence

Tracy Cerrillo

Mary Hansen

Michael Harwell

University of Pittsburgh

April 20, 2000

Correspondence concerning this paper should be sent to Michael Harwell, 5C01 WWPH Building, University of Pittsburgh,
PGH, PA 15260 [email: harwell+@pitt.edu]

Abstract

The problem of college student persistence has produced a flood of research devoted to understanding why students drop out and how it can be remedied. Studies of student persistence have increasingly relied on sophisticated statistical techniques, such as structural equation modeling, to attempt to understand this phenomenon. These studies often rely on data supplied by colleges and universities, which almost always contain missing values. How missing values are handled can have a substantial effect on the conclusions drawn from these studies, a fact that may not be widely appreciated by researchers modeling student persistence. This study investigated the effects of four methods of handling missing data on the conclusions drawn from statistical models of student persistence. Information about similarities and discrepancies among the methods in their conclusions should be useful to persistence researchers struggling with missing data, and to institutional policymakers charged with developing programs to support students.

The problem of persistence (attrition, dropout) among college students has produced a flood of research devoted to understanding why some students persist and others drop out (Tinto, 1993). Considering that approximately 40% of college students drop out this emphasis seems well placed (Horn & Carroll, 1996). Although Tinto (1987, 1993) and others have shown that dropping out of college A does not preclude a student from subsequently enrolling in college B and graduating, or from returning to college A and graduating at a later time, the consequences of dropping out have important implications for students and institutions. As a result, studies of student persistence are now routinely done at many post-secondary institutions (Shelton, 1995; Tinto, 1993).

With relatively few exceptions, studies of student persistence increasingly rely on sophisticated statistical techniques, particularly structural equation modeling or SEM (Braxton, Duster, & Pascarella, 1988). Ultimately, these studies share a common goal of wanting to be able to identify the academic, social, personality, and background factors that predict which students are most prone to dropping out. Accurate prediction of these factors would allow institutions to modify admissions criteria in an attempt to take into account the likelihood of persistence, as well as assist institutions in generating academic and social support structures for students in ways that increase the likelihood of persistence.

Unfortunately, progress in understanding this phenomenon has been hampered by contradictory empirical findings of the efficacy of various theoretical models of persistence. Some of the variation in findings is likely attributable to variation in the quality of the methodological work. The results of performing regression analyses (e.g., Kim & Curry, 1977; Little, 1992) and SEM in the presence of missing data (Brown, 1994; Enders & Bandalos, 1999; Little, 1992; Muthen, Kaplan, & Hollis, 1987) suggests that how researchers handle missing data can have important effects on the conclusions of the data analyses. Yet the potential effects of missing data in persistence research or of the effects of statistical techniques for handling missing data have not received much attention. Meanwhile, even a cursory glance through the persistence literature suggests that missing data are a frequent problem. For example, an institution that agrees to participate in a national study of persistence and subsequently supplies information on variables such as the demographic composition of students at that institution, but fails (for some reason) to turn over financial aid information, would produce missing data. How the missing financial data are handled may have significant implications for the conclusions drawn from a study of student persistence.

Statement of the Problem

This study investigated the effects of four methods of handling missing data on the conclusions drawn from two statistical models of student persistence. This issue has apparently not been addressed in the persistence literature. Information about the similarities and discrepancies among the four methods in their conclusions should be useful to

persistence researchers struggling with how to handle missing data.

Conceptual Framework

Why some students stay in college and others leave is a complicated question that has been tackled by a number of theorists (e.g., Astin, 1984; Bean, 1980, 1986; Cabrera, Castaneda, Nora, & Hengstler, 1992; Spady, 1971; Tinto, 1987, 1993). In general, theoretical models of student persistence have focused on the preparedness of students entering college (more or less prepared) and the institutional support for students (Nagda, Gregerman, Jonides, von Hippel, & Lerner, 1998; Tinto, 1993). Among the two most studied and accepted theoretical models are those attributed to Tinto and Bean. Both of these models have largely been directed toward traditional college students (younger, full time, residential), although nontraditional students have not been ignored. Similarly, both have focused on traditional four-year institutions, as opposed to two-year institutions.

In general, Tinto's model mixes Durkheim's theory of egotistic suicide and cost-benefit analyses to explain student persistence at the institutional level, emphasizing the interactions between students and the institution which over time lead to different kinds of dropout decisions (e.g., voluntary dropout, involuntary dropout, temporary dropout). A key to Tinto's model is the interplay between the academic and social systems of the institution, although his revised models (1987, 1993) also take the previous group memberships of students (e.g., family, community) into account. Students who do not find a place within a college's social or academic system are much more likely to dropout, a process which is exacerbated if a student finds other non-academic activities attractive (e.g., working). Tinto's model has been received empirical support from several studies (e.g., Braxton, et al. 1988; Grosset, 1990; Pascarella & Chapman, 1983; Pascarella & Terenzini, 1988). Bean's (1986) model of persistence is similar to Tinto's in that it incorporates existing knowledge of student withdrawal patterns and characteristics, but it differs from Tinto's in that it draws heavily on theories of work turnover and organizational behavior.

Missing Data

In the plethora of student persistence theorists, contradictory empirical findings, and data analysis techniques one constant emerges: Empirical studies of student persistence relying on data from colleges and universities must deal with missing (incomplete) data. This occurs for many reasons, such as poor record keeping, an unwillingness to provide institution-wide information, and because much of the information is sensitive, e.g. average SAT or ACT scores.

Incomplete data can affect the conclusions of studies of persistence in several ways, the most important of which is

that inferences are based on biased parameter estimates. Another is that incomplete data typically lower the statistical power of hypothesis tests and the precision of estimation because the sample size for the incomplete data is less than it would be if the data were complete (no missing values). This occurs because most statistical procedures require complete data for the analyses to proceed.

Data can be missing in several ways. For example, suppose that data for Y_1, Y_2, \dots, Y_p variables for n cases are to be collected, where Y represents a variable and n is the sample size. Suppose also that 20% of the data are reported to be missing. In describing missing data it is the proportion of missing data that is referred to, not the percentage of cases lost because of missing values. For example, it could be that data for all of the variables are collected (i.e. show complete data) except one which has 20% of its values missing, that a subset of the Y_p shows missing data, etc. It is also important to identify the pattern of missing data. Certain missing data patterns, such as monotone patterns (Little & Rubin, 1987) may allow particularly efficient estimation, but we assume the missing values follow a general missing data pattern.

Underlying Theory

We briefly describe some of the theory underlying missing data methodologies, and refer readers seeking more detailed descriptions to Rubin (1976), Little and Rubin (1987), Rubin (1987), and Schafer (1997).

Suppose that complete-data based on n ($i = 1, 2, \dots, n$) cases and p variables is arranged in array \mathbf{Y} :

$$\mathbf{Y} = \begin{array}{cccc} \text{Cases} & & \text{Variables} & \\ & 1 & 2 & \dots & p \\ 1 & & & & \\ 2 & & & & \\ 3 & & & & \\ \cdot & & & & \\ \cdot & & & & \\ n & & & & \end{array}$$

Assuming that rows can be modeled as independent and identically-distributed draws from a multivariate-normal probability distribution, the probability function of the complete data can be written

$$P(\mathbf{Y}|\theta) = \prod_{i=1}^n f(y_i|\theta) \quad (1)$$

where $f(y_i|\theta)$ represents the probability function for the i th row and θ represents a vector of unknown parameters for the complete data. Once data have been collected we may follow standard practice by finding the likelihood function and using this information to obtain maximum likelihood estimates.

Suppose now that \mathbf{Y} is incomplete. Rubin (1976) characterized incomplete data through ignorable and nonignorable

response mechanisms. Missing data that are ignorable in the sense of being missing at random (MAR) allows the missingness to be ignored and unbiased and efficient maximum likelihood parameter estimates to be generated. In general, MAR means that the reasons values are missing for a variable can be explained by data available for other variables. As an example, suppose that Y_1 is the graduation rate of an institution and Y_2 is a predictor indicating the nature of funding for the institution (1 =public, 2 = private), and suppose that Y_1 has missing values for a sample of institutions. The missing Y_1 values are MAR if the probability of missingness can be predicted by Y_2 , i.e., whether the institution is publicly or privately funded.

The missing Y_1 values are missing completely at random (MCAR) if the probability of missingness cannot be predicted by whether the institution is publicly or privately funded or by the graduation rate of other institutions. If the probability of missingness for Y_1 depends on Y_1 itself, meaning, for example, that the failure of an institution to report its graduation rate is attributed to this value being low, the missing graduation rate would not be missing at random (NMAR).

Following Schafer (1997), let \mathbf{Y} be partitioned such that $\mathbf{Y} = (\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ and let \mathbf{R} be an $n \times p$ matrix of (0,1) variables indicating whether values in \mathbf{Y} are observed or missing. \mathbf{R} is related to \mathbf{Y} via the missingness model $P(\mathbf{R}|\mathbf{Y}, \psi)$, meaning that there are two statistical models $[P(\mathbf{Y}|\theta), P(\mathbf{R}|\mathbf{Y}, \psi)]$ that need to be considered jointly, i.e., the likelihood function needs to incorporate a component reflecting the missingness model. The joint probability distribution of the observed data is then $P(\mathbf{R}, \mathbf{Y}|\theta, \psi)$ which, after integrating across the missing data distribution, has the form

$$\begin{aligned} P(\mathbf{R}, \mathbf{Y}_{\text{obs}}|\theta, \psi) &= \int P(\mathbf{R}, \mathbf{Y}|\theta, \psi) d\mathbf{Y}_{\text{mis}} \\ &= \int P(\mathbf{R}|\mathbf{Y}, \psi) P(\mathbf{Y}|\theta) d\mathbf{Y}_{\text{mis}} \end{aligned} \quad (2)$$

Under ignorability,

$$\begin{aligned} P(\mathbf{R}, \mathbf{Y}_{\text{obs}}|\theta, \psi) &= P(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \psi) \int P(\mathbf{Y}|\theta) d\mathbf{Y}_{\text{mis}} \\ &= P(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \psi) P(\mathbf{Y}_{\text{obs}}|\theta) \end{aligned} \quad (3)$$

i.e., distribution of \mathbf{R} does not depend on \mathbf{Y}_{mis} . Assuming that the parameters in ψ and θ are distinct, meaning that knowledge of ψ does not provide much information about θ and vice versa, the likelihood of the observed data (\mathbf{Y}_{obs}) under MAR can be factored into two pieces, the piece of substantive interest $[P(\mathbf{Y}_{\text{obs}}|\theta)]$ and a piece reflecting the missing data mechanism $[P(\mathbf{R}|\mathbf{Y}_{\text{obs}}, \psi)]$ that can be ignored. The implication is that the observed data \mathbf{Y}_{obs} can be used for likelihood-based inferences about θ . Put another way, under MAR the distributions of \mathbf{Y}_{obs} and \mathbf{Y}_{mis} are the same conditional on a set of predictors.

Although the MAR assumption cannot be empirically tested, several authors have suggested that missing data methods that require the MAR assumption to be satisfied may still be usefully employed when MAR is not completely satisfied (Little & Rubin, 1987; Schafer, 1997). Possible robustness associated with violations of the MAR assumption is important because MCAR is unlikely to be satisfied in many educational settings, particularly those involving observational studies.

Methods For Handling Missing Data

The presence of missing data has generated a literature investigating the effects of different methods of handling missing data in educational research, particularly for SEM (e.g., Brown, 1994; Enders & Bandalos, 1999; Wothke, in press). This literature typically describes the merits of various methods for handling incomplete data. One is to simply use the incomplete data and hope that the results are not so biased and underpowered as to render them useless. For example, suppose that a regression model predicting graduation rates for a sample of colleges and universities is fitted to an incomplete dataset with n cases. Typically, data analysis software will employ listwise deletion (LD) when confronted with missing data, meaning that only those cases with complete data on all variables used in the analysis, say n^* , will be used. Thus, $n^* < n$, perhaps significantly so, depending on the pattern of missing data and the statistical analysis employed. For example, a dataset based on $n = 300$ cases may be described as having 5% of the values missing. However, it's possible that the 5% is distributed over the variables in such a way that, under LD, n is reduced substantially, e.g., $n^* = 125$. Moreover, parameter estimates under LD are only unbiased if the missing data are MCAR.

A second approach to handling missing data is to employ an estimation procedure that will make use of all available data. One such method is pairwise deletion in which all available data for particular computations is used. Here the correlation between Y_1 and Y_2 could be based on a different number of cases than the correlation between Y_1 and Y_3 . It is well known that this may produce a covariance matrix that is not positive-definite and that parameter estimates are only unbiased if the missing data are MCAR. Although there are some contradictory findings on the effects of pairwise deletion, we side with Muthen, et al. (1987) and consider this method to be unacceptable. Both LD and PD are ad hoc methods in the sense that missingness is not taken into account in a formal way.

A third approach, and the recommended one, is to handle missing data in a principled way. Two likelihood-based (principled) approaches to handling missing data assuming MAR are to use whatever data are available in estimating parameters via full information maximum likelihood (FIML), and imputation of missing values via the EM algorithm or data augmentation. FIML (Arbuckle, 1996) produces estimates that are unbiased and are more efficient compared to LD/PD.

To see how FIML estimates parameters in the presence of missing data, suppose that the realized missing data patterns in \mathbf{Y} were grouped into $s = 1, 2, \dots, S$ up to 2^p unique patterns, with $i = 1, 2, \dots, t$ cases per pattern. Under the assumption that $\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that the missing values are MAR, the observed data likelihood can be written

$$\prod_{s=1}^S \prod_{i=1}^t | \boldsymbol{\Sigma}_s^* |^{-1/2} \exp[-1/2(\mathbf{y}_{i*}^* - \boldsymbol{\mu}_{s*}^*)' \boldsymbol{\Sigma}_s^{*-1} (\mathbf{y}_{i*}^* - \boldsymbol{\mu}_{s*}^*)] \quad (4)$$

where \mathbf{y}_{i*}^* is the observed part of row i of \mathbf{Y} , $\boldsymbol{\mu}_{s*}^*$ is the subvector of $\boldsymbol{\mu}$ corresponding to unique pattern s , and $\boldsymbol{\Sigma}_s^{*-1}$ is the square submatrix of covariance for $\boldsymbol{\Sigma}$ corresponding to the data in unique pattern s . Maximizing the likelihood in (4) provides unbiased and efficient estimates in the presence of missing data. The AMOS (Arbuckle, 1995) data analysis program has incorporated this procedure for handling missing data.

Another likelihood-based strategy for handling missing data is to impute missing values, producing complete data that can be analyzed with standard procedures. This method is similar to Buck's (1960) method in which a regression model using available complete data is used to generate predicted values, which in turn are imputed for the missing values. Suppose that $\mathbf{Y} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the missing values are MAR. An EM-based approach here is based on the fact that a normal distribution is a member of the so-called regular exponential family, meaning that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ contain sufficient statistics. Imputation focuses on computing these sufficient statistics and has the following general steps:

- (a) E-step--Calculate the expected values of the complete-data sufficient statistics given the model for the data, treating the current estimates of means/covariances as true values. To calculate the expected values a series of regressions are done for each missing data pattern. Every pattern of missing data has a set of missing values and a set of completely observed values. For a given missing data pattern, compute predicted values using observed data and current parameter estimates of means/covariances. These predicted values are used to calculate expected values of the sufficient statistics.
- (b) M-step—Use the computed sufficient statistics as if they were true values to obtain new estimates of means/covariances.

Through an iterative process final estimates of the means and covariances are obtained that are unbiased and efficient. Available software includes SPSS Missing Value Analysis (SPSS Inc., 1999), BMD PAM (BMDP, 199), and NORM (Schafer, 1999).

Data augmentation (DA) is a simulation-based approach that is similar in many ways to an EM approach. The result of DA is a predictive distribution of missing values that is used to impute missing values. DA also follows a 2-step process:

- (a) I-step—Simulates a random draw of the complete-data sufficient statistics. Given a current estimate of $\boldsymbol{\theta}^r$, select a value of the missing data from the conditional predictive distribution of $\mathbf{Y}_{\text{mis}}, \mathbf{Y}_{\text{mis}}^{(r+1)} \sim P(\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}}, \boldsymbol{\theta}^r)$.

(b) P-step—Conditioning on $\mathbf{Y}_{\text{mis}}^{(r+1)}$, draw a new value of θ from its complete-data posterior, $\theta^{(r+1)} \sim P(\theta | \mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(r+1)})$. Through an iterative process two distributions are obtained, $P(\theta | \mathbf{Y}_{\text{obs}})$ and $P(\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}})$. For a suitably large r , $\theta^{(r+1)}$ can be considered as a random draw from $P(\theta | \mathbf{Y}_{\text{obs}})$ and $\mathbf{Y}_{\text{mis}}^{(r)}$ as a random draw from $P(\mathbf{Y}_{\text{mis}} | \mathbf{Y}_{\text{obs}})$. This process leads to complete-data that are then analyzed using standard methods to produce parameter estimates that are unbiased and efficient.

Among the imputation procedures, the missing data literature recommends the use of multiple imputation. Consider a dataset in which 20% of the values are missing. Using a data imputation procedure, the missing values would be replaced with imputed values, say y_i' . The shortcoming of this procedure is that y_i' is not a true value and contains uncertainty. This shortcoming is remedied through multiple imputation (MI) in which DA is used to impute missing values for a dataset m times. For example, suppose the proportion of missing data was 20%. In MI, DA is used to impute the missing values m times. Since DA is a simulation-based technique the m sets of imputed values would be expected to vary. In this way, a case (subject) with a missing value on Y_p would have m imputed values on Y_p , allowing the uncertainty associated with imputation to be taken into account by estimating the variance among the imputed values. Unless the proportion of missing data is quite large, setting $m = 5$ is likely to be adequate in most settings (Rubin, 1987). Following Rubin (1987), the m estimated parameters (e.g., a slope) can be combined in a way that takes uncertainty associated with the imputed values into account. The computer program NORM (Schafer, 1999) can be used to impute missing values under DA.

Before continuing, we note that there should be consistency between the predictors selected because they are believed to capture the reasons for missingness under MAR (the so-called imputation model) and subsequent statistical analyses (Schafer, 1997). Suppose that Y_1 has missing data and Y_2 and Y_3 are completely observed. Specification of a data imputation model requires selecting all of the variables believed to be predictive of the missingness on Y_1 , not only those that might be used in subsequent statistical analyses for the complete-data. According to Schafer (1997), there is typically little harm in using a data imputation model that contains variables that may not be the focus of subsequent complete data analyses (but that are believed to be predictive of missingness). On the other hand, performing subsequent statistical analyses on complete-data using variables that were not part of the data imputation model may have untoward consequences, such as biased estimates. In general, it is best to use comprehensive data imputation models.

An Example Of Incomplete Data

Data containing information for the 1993-94 school year for a sample of 1,302 U.S. colleges and universities were

made available by US News & World Report. These colleges and universities varied in a number of ways, from graduation rates to in-state tuition to percent of faculty with Ph.Ds. Table 1 lists all of the available variables. For example, an analysis of the graduation rates (GRADRATES) (in percentages) for the sample show that 7.5% of the colleges and universities did not report their graduation rate and were missing. Table 2 summarizes the amount of missing data for several of the U.S. News & World Report variables (AVESAT is the average of SAT Verbal and SAT Math, SELECT is the ratio of number admitted to number applied).

Provided that the missing data can be treated as MAR, either an EM or DA approach can be employed to produce complete data and lead to unbiased and efficient estimates. For example, if the missing GRADRATE values are MAR then the reasons for their missingness can be predicted by available observed data. It seems reasonable to expect that the probability of missingness for GRADRATE could be predicted by observed data for variables such as whether a school is publicly or privately funded, SAT average scores, and tuition. If this argument is unconvincing then there is evidence that treating missing data as MAR can be beneficial (e.g., reduce bias) even when this assumption is unlikely to be met perfectly (Muthen, et al., 1987; Schafer, 1997).

Table 1. Variables in the US News & Word Report Dataset

Public/private indicator (public=1, private=2)*+
Average Math SAT score*+
Average Verbal SAT score*+
Average ACT score+
First quartile - Math SAT
Third quartile - Math SAT
First quartile - Verbal SAT
Third quartile - Verbal SAT
First quartile - ACT
Third quartile - ACT
Number of applications received
Number of applicants accepted
Number of new students enrolled
Pct. new students from top 10% of H.S. class*+
Pct. new students from top 25% of H.S. class
Number of fulltime undergraduates*+
Number of parttime undergraduates
In-state tuition*+
Out-of-state tuition*+
Room and board costs*+
Room costs
Board costs
Additional fees*+
Estimated book costs*+
Estimated personal spending

Pct. of faculty with Ph.D.'s*+
 Pct. of faculty with terminal degree
 Student/faculty ratio*+
 Pct.alumni who donate
 Instructional expenditure per student*+
 Graduation rate

*Variable used in the regression imputation model

+Variable used in the SEM imputation model

Procedure

We investigated four ways of handling incomplete data:

1. Listwise Deletion (LD). As described above, cases showing even one missing value are deleted altogether. The advantage of this procedure is that deleting cases with missing values produces complete data, allowing standard methods to be employed and if the missing data are MCAR then parameter estimates are unbiased. The disadvantage of this procedure is reduced statistical power for hypothesis testing and lowered precision because of the reduced sample size. As noted above, the percent of cases discarded under this option can be substantial. For example, Table 2 indicates that the

Table 2. Missing Data For Selected US News & World Report Variables

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes ^a	
				Count	Percent	Low	High
SCHLTYPE	1302	1.6390	.4805	0	.0	0	0
SATVERB	777	461.2239	58.2984	525	40.3	5	24
INSTATE	1272	7897.2744	5348.1626	30	2.3	0	2
OUTSTATE	1282	9276.9056	4170.7709	20	1.5	0	6
RMBOARD	1226	4162.1069	1179.2831	76	5.8	0	11
ADDFEES	1028	392.0126	469.3792	274	21.0	0	55
BOOKCOST	1254	549.9729	167.3554	48	3.7	56	70
PERSPEND	1121	1389.2917	714.2479	181	13.9	0	24
PCTPHD	1270	68.6457	17.8256	32	2.5	10	0
PCTTERM	1272	75.2311	17.1082	30	2.3	2	0
STUDFAC	1300	14.8588	5.1864	2	.2	3	19
AVESAT	777	968.0618	123.4949	525	40.3	2	0
SELECT	1289	.7548	.1596	13	1.0	50	0
GRADRATE	1204	60.3904	18.8505	98	7.5	0	0

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

percentage of missing data for GRADRATE was 7.5% and 40.3% for SAT-VERBAL, but the correlation between these two variables under listwise deletion results in a loss of 43.8% of the cases.

2.FIML Estimation. We used the AMOS (Arbuckle, 1995) to estimate parameters in the presence of missing data under the assumption the missing values are MAR.

3. EM-Based Data Imputation. Assuming the missing data are MAR, we estimated the missing values using the SPSS Missing Values program.

4. MI Imputation. We used DA to impute missing values for $m = 5$ using the NORM (Schafer, 1999) program. Because this method takes the uncertainty associated with imputed values into account (under MAR), it is the "gold standard" against which other results can be compared.

Methods and Design

We investigated the effects of four methods of handling missing data on institutional models of student persistence using the sample of 1,302 U.S. colleges and universities. Because our cases are colleges and universities, we used each institution's graduation rate (GRADRATE) as our indicator of persistence. These models were simple representations of three types of variables often described in this literature as predictors of student persistence: financial (room and board, additional student fees, in-state tuition, out-of-state tuition, cost of books), quality of the institution (percent of applicants admitted, percent of faculty with Ph.D.'s, student to faculty ratio, percent fulltime students), and student quality/preparedness (SAT Math and Verbal scores). These variables are similar to empirical studies in the student

persistence literature with two differences: (1) Variables are collected for institutions rather than individual students (2) No measures of the social support available for students at an institution, a key factor in many persistence model, were available.

With these data, we generated multiple linear regression and classical path analysis models that were consistent with models reported in the student persistence literature using GRADRATE as the outcome variable. Univariate summary statistics for the US News & World Report variables used in our regression and path analyses appear in Table 3. These results indicate that most of these variables were no worse than modestly skewed, although several were quite kurtic. Table 4 reports correlations among these variables (A correlation of .92 between SATVERBAL and SATMATH prompted the creation of AVESAT).

For each method of handling missing data we examined the effect on model conclusions using estimated (standardized) slopes/path coefficients, and standard errors of the estimated slopes/path coefficients. We also examined indices of model-data fit (e.g., R^2 for multiple regression and the chi-square test for the path analysis). Information about the similarities or discrepancies in these conclusions should be useful to persistence researchers struggling with how to handle missing data.

Initially, we constructed imputation models using variables we believe relate conceptually to missingness. All variables thought to contribute to the missingness should be included in the imputation model, even if those variables are not included in subsequent analyses (Schafer, 1997). The imputation model for the regression analysis consisted of the variables marked in Table 1 plus SELECT (no. admitted/no. applied).

Model 1 Multiple Linear Regression Results

A regression model of the form

$$Y_{ip} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_{ip} \quad (5)$$

was fitted to the GRADRATE data using backward elimination for the same variables used in the imputation

Table 3
Summary Statistics

	Descriptive Statistics						
	N	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Out of state tuition	1302	9134.2650	4293.3808	.606	.068	.181	.136
Room and Board	1302	3918.6321	1505.4841	-.522	.068	1.010	.136
Average additional fees (\$)	1302	307.6214	447.9477	4.514	.068	31.071	.136
Average cost of books	1302	529.3656	195.1340	1.186	.068	12.088	.136
Average expenditure per student	1302	1194.9055	820.5964	1.010	.068	3.762	.136
Percent of faculty with PhD	1302	66.7373	21.3209	-1.215	.068	2.232	.136
Percent full time students	1302	73.2903	21.1133	-1.513	.068	3.415	.136
Student to faculty ratio	1300	14.8588	5.1864	4.186	.068	50.079	.136
No admitted / No applied	1289	.7548	.1596	-1.106	.068	1.611	.136
Graduation Rate	1204	60.3904	18.8505	-.011	.071	-.488	.141
Average Verbal SAT	1302	271.6175	235.1063	-.280	.068	-1.763	.136
AVETSAT	1302	575.5361	488.5747	-.281	.068	-1.755	.136
Valid N (listwise)	1197						

Table 4
Correlations Among Selected Variables

		Correlations														
		Type of school	Average Verbal SAT	In-state tuition	Out of state tuition	Room and Board	Average additional fees (\$)	Average cost of books	Average expenditure per student	Percent of faculty with PhD	Percent full time students	Student to faculty ratio	No admitted / No applied	AVESAT	Graduation Rate	
Type of school	Pearson Correlation	1.000	.196*	.783*	.558*	.457*	.285*	-.018	-.238*	-.149*	-.108	-.367*	.010	-.129*	-.407*	
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.725	.000	.000	
	N	1302	1272	1272	1282	1226	1028	1254	1121	1270	1272	1289	777	1289	1204	
Average Verbal SAT	Pearson Correlation		1.000	.535*	.626*	.410**	.036	-.133*	-.171**	.483*	-.484**	-.305*	-.386**	.976*	.595*	
	Sig. (2-tailed)			.000	.000	.000	.370	.000	.000	.000	.000	.000	.000	.000	.000	
	N		777	777	757	765	738	629	758	699	755	760	776	772	731	
In-state tuition	Pearson Correlation			1.000	.928**	.650**	-.153*	.044	-.304**	.202*	-.245**	-.460**	-.191**	-.496**	.605*	
	Sig. (2-tailed)				.000	.000	.000	.120	.000	.000	.000	.000	.000	.000	.000	
	N			1272	1270	1209	1011	1236	1104	1242	1245	1270	1261	757	1176	
Out of state tuition	Pearson Correlation				1.000	.699**	-.051	.065*	-.269*	.377**	-.414**	-.441**	-.271**	.608*	.624*	
	Sig. (2-tailed)					.000	.102	.000	.000	.000	.000	.000	.000	.000	.000	
	N				1282	1219	1022	1246	1115	1254	1280	1271	765	1187		
Room and Board	Pearson Correlation					1.000	-.122*	-.189*	-.156*	-.342**	-.413**	-.311**	-.293**	.398**	.477**	
	Sig. (2-tailed)						.000	.000	.000	.000	.000	.000	.000	.000	.000	
	N					1226	1219	1226	974	1195	1080	1196	1202	1225	1217	
Average additional fees (\$)	Pearson Correlation						1.000	.069*	.013	.175**	-.170**	.062	.181**	.067	.043	
	Sig. (2-tailed)							.000	.028	.000	.000	.047	.000	.003	.188	
	N						1028	1012	911	1007	1009	1027	1020	629	952	
Average cost of books	Pearson Correlation							1.000	.165*	.038	-.116*	-.056*	-.146**	.145*	.630*	
	Sig. (2-tailed)								.000	.209	.000	.046	.000	.000	.300	
	N							1254	1119	1224	1228	1253	1245	738	1188	
Average expenditure per student	Pearson Correlation								1.000	-.038	-.067*	.081**	.052	-.142*	-.239*	
	Sig. (2-tailed)									.208	.026	.007	.040	.000	.000	
	N								1121	1097	1121	1114	689	1056		
Percent of faculty with PhD	Pearson Correlation									1.000	.881**	-.102*	-.255**	.536**	.279*	
	Sig. (2-tailed)										.000	.000	.000	.000	.000	
	N									1270	1270	1248	1270	1259	735	
Percent full time students	Pearson Correlation										1.000	-.134**	-.252**	.516*	.272*	
	Sig. (2-tailed)											.000	.000	.000	.000	
	N										1272	1272	1263	760	1183	
Student to faculty ratio	Pearson Correlation											1.000	-.154**	-.292**	-.341**	
	Sig. (2-tailed)												.000	.000	.000	
	N												1300	1289	776	
No admitted / No applied	Pearson Correlation												1.000	-.416**	-.276**	
	Sig. (2-tailed)													.000	.000	
	N													1289	772	
AVESAT	Pearson Correlation													1.000	.590*	
	Sig. (2-tailed)														.000	
	N														777	
Graduation Rate	Pearson Correlation														1.000	
	Sig. (2-tailed)														.000	
	N														1204	

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

model. This method was utilized because it is the method of choice for many statisticians (Neter, Kutner, Nachtsheim, & Wasserman, 1996). The backwards elimination approach eliminated eight variables that were included in the imputation

model, and the final regression model appears in Figure 1. Each of the four methods of handling missing data were examined for this regression model.

Fitting the model in equation (5) to GRADRATE data under LD resulted in a loss of 53.3% of the cases, whereas FIML used all available data. For the EM approach we used the SPSS for Windows (1999) software, and for MI we used the computer program NORM (Schafer, 1999) to generate $m = 5$ multiple imputations using data augmentation. There is no agreed upon minimum but efficient estimation is usually achieved with $m = 5$ to 10 (Schafer, 1997). For MI, the model in equation (5) was fitted to each of the five now-complete datasets, and the estimated slopes and standard errors were aggregated using Rubin's (1987) method. Table 5 reports the estimated slopes and standard errors for each missing data method, and Figures 2 and 3 provide graphical representations of these values.

As can be seen, results for the AVESAT, OUTSTATE, RMBOARD, and STUDFAC predictors did not differ much by missing data method. However, results for the remaining two predictors (SELECT, SCHLTYPE) varied noticeably by missing data method. As noted earlier, the MI method is the most principled method for handling missing data, although it is also the most computationally tedious and inaccessible. Interestingly, the LD method produces slopes closest to the aggregated values of the MI procedure, but for standard errors the EM method results are most similar to MI. FIML produced standard errors for some slopes that were noticeably smaller than the other methods, while the LD method produces the largest standard errors for several of the predictors. We also compared the various methods by considering the adjusted R^2 values. For the current analyses, the adjusted R^2 for the EM method was the highest at .501, followed by the R^2 's generated by the five MI runs, which ranged from .457 to .476. The R^2 generated by the LD method was .414, while that generated by the FIML method was only .293. Once again, EM was closest to MI.

A particularly appealing feature of the MI approach is that information about the extent to which missing data affect parameter estimation can be obtained. Rubin (1987) described the rate of missing information as a way for estimating how strongly the quantity being estimated is influenced by missing data:

Figure 1 Final Regression Model Fitted To Graduation Data

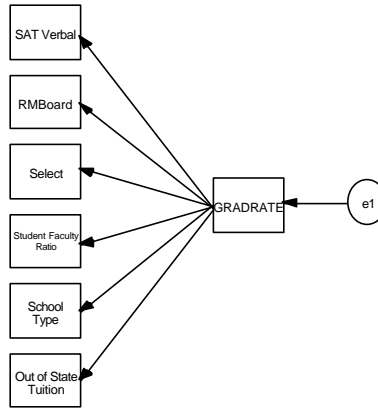


Table 5. Estimates and Standard Errors of Slopes for the Final Regression Model

	LD		EM		FIML		MI	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
(Constant)	-14.386	8.030	-31.446	5.885	27.564	3.571		
Average Verbal SAT	0.119	0.013	0.149	0.010	0.005	0.003	.119	.012
Out of state tuition	0.001	0.000	0.001	0.000	0.003	0.000	-.001	.000
Room and Board	0.002	0.001	0.001	0.000	0.002	0.001	-.001	.000
No admitted / No applied	-4.170	4.120	0.355	2.659	2.067	0.619	-4.745	2.783
Student to faculty ratio	-0.040	0.118	0.026	0.079	-0.268	0.128	-.012	.116
Type of school	6.027	1.506	6.325	0.945	-1.602	1.528	7.029	1.346

$$\%MIS. INF. = \frac{1}{1 - r} = [r + 2/(df + 3)] / (r + 1) \tag{6}$$

where $\frac{1}{1 - r}$ quantifies how much more precise the estimate might have been with no missing data, and

$r = [(1 + m^{-1})B] / \bar{U}$ is the relative increase in variance due to nonresponse (B = between-imputation variance, \bar{U} = average sampling variance of a statistic). The expression $(1 + m^{-1})^{-1}$ estimates the efficiency of estimation. The MI results for the final regression model are reported in Table 6 and the %MIS. INF. values indicate that most of the estimated regression

Figure 2 Plot of Estimated Slopes for the Final Regression Model

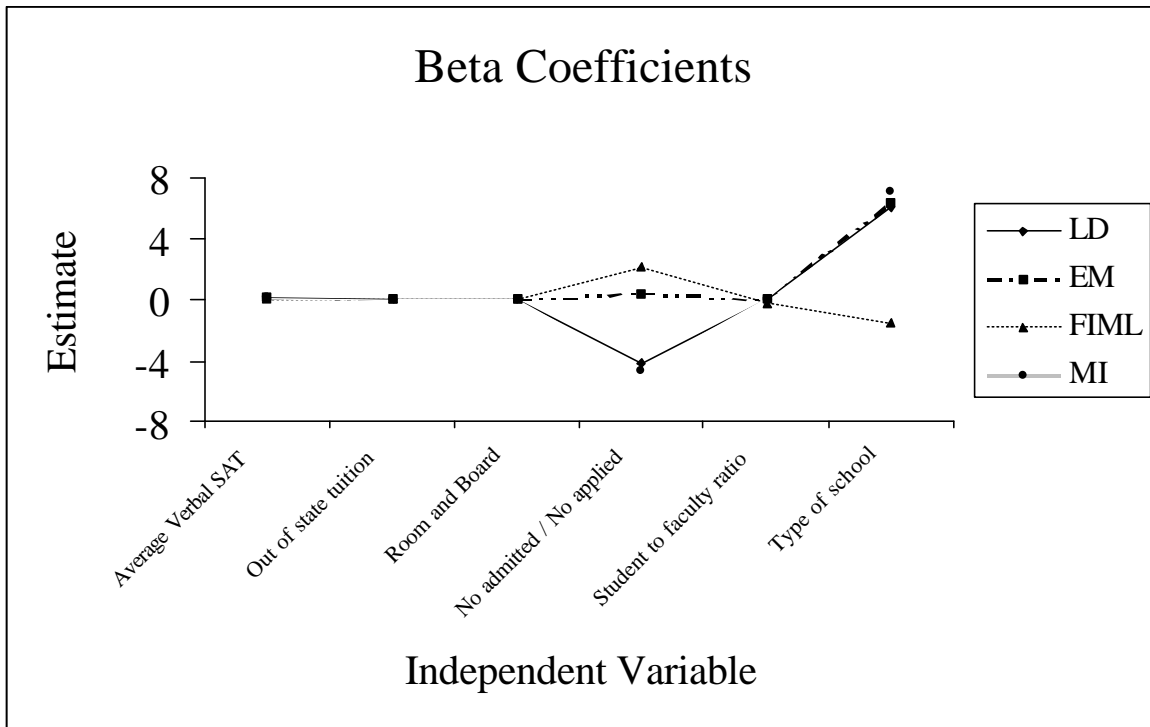


Figure 3 Plot of Estimated Standard Errors of Slopes for the Final Regression Model

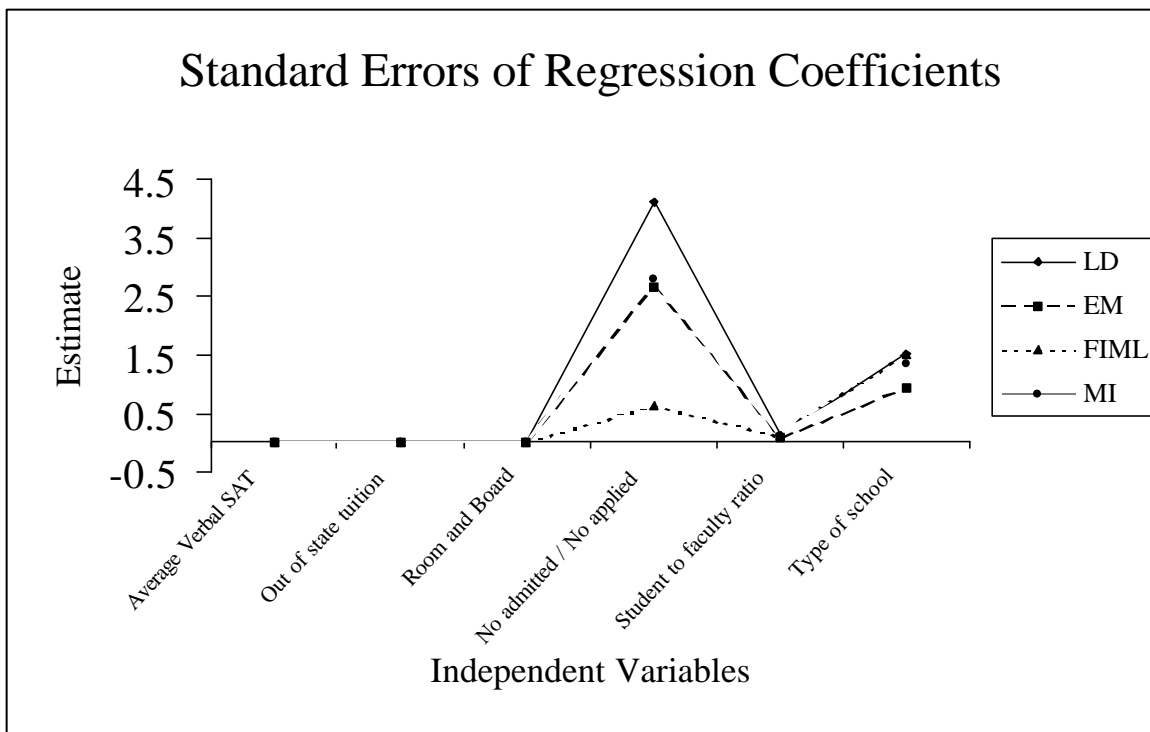


Table 6. Multiple Imputation Regression Results

<u>QUANTITY</u>	<u>ESTIMATE</u>	<u>STD.ERR.</u>	<u>T-RATIO</u>	<u>DF</u>	<u>P-VALUE</u>
SATVERBAL	0.118600	0.117881E-01	10.06	22	0.0000
OUTSTATE	-.961600E-03	0.143832E-03	-6.69	14	0.0000
RMBOARD	-.117000E-02	0.223652E-03	-5.23	6	0.0020
SELECT	-4.74540	2.78339	-1.70	2098	0.0884
STUDFAC	-.117560E-01	0.116276	-0.10	17	0.9207
SCHLTYPE	7.02900	1.34639	5.22	20	0.0000

CONFIDENCE LEVEL FOR INTERVAL ESTIMATES (%): 95.00

<u>QUANTITY</u>	<u>LOW ENDPT.</u>	<u>HIGH ENDPT.</u>	<u>%MIS.INF.</u>
SATVERBAL	0.941529E-01	0.143047	46.2
OUTSTATE	-.127009E-02	-.653112E-03	57.0
RMBOARD	-.171726E-02	-.622744E-03	84.3
SELECT	-10.2039	0.713083	4.5
STUDFAC	-.257078	0.233566	53.2
SCHLTYPE	4.22048	9.83752	49.0

slopes were highly affected by missing data.

The variation in efficiency in Table 6 is also reflected in the reduction in degrees of freedom as a function of between-imputation variance for some estimates. For example, the degrees of freedom for SELECT (df = 2,098) is substantially greater than that associated with RMBOARD (df = 6) and OUTSTATE (df = 14). This occurs because the degrees of freedom are computed using B and \bar{U} , and B was much larger for RMBOARD and OUTSTATE. Thus, the aggregated slopes for RMBOARD and OUTSTATE are not as precisely estimated as that for SELECT.

Model 2 Path Analysis Model

Next, we fitted the path model in Figure 4 to the graduation data for each method of handling missing data. The three latent variables (institutional) Quality, Ability (student preparedness), and Cost correspond to factors often studied in the persistence literature. GRADRATE served as a single indicator of PERSISTENCE, as did AVESAT for ABILITY, meaning that these manifest variables were assumed to be perfectly reliable measures. Table 7 presents the estimates and standard errors of the path coefficients that were found using each of the methods for handling missing data for the SEM model. We did not attempt to find the best-fitting model since it is likely that this would have produced different final models that would have made comparisons among the missing data methods difficult. Nor would it have been clear whether

different final models were attributable to some intrinsic properties of the missing data methods or the vagaries of finding such models.

For the LD procedure, minimization was unsuccessful, which means that the results generated by this method are not interpretable. This is apparent in the large estimates and standard errors for four of the paths. The estimates generated by the EM and FIML methods are relatively similar to each other, although some inconsistencies are present. In addition, although a few of the path coefficients found using the MI method were similar to those obtained using EM and FIML, the majority were dissimilar.

The FIML, EM, and MI methods tended to produce similar standard errors for the path coefficients, although for paths relating to the costs variable MI tends to produce slightly larger values. Figures 5 and 6 present graphical representations of the path coefficients and standard errors for the SEM model for each of the methods of handling missing data. It should be noted that the estimates produced by the LD method are not presented in the graphs because minimization was unsuccessful using this method. The value of the standard error representing the path from costs to out of state tuition is not shown in Figure 6 because its value falls out of the range of the vertical scale. Thus, it is important to note that the magnitude of several of the standard errors is greater than the graph may indicate at first glance.

The EM method produced estimates similar to those generated by FIML, and the standard errors were consistently smaller. The MI method produced path coefficients that were dissimilar to those generated by EM and FIML, and for certain paths, produced larger standard errors. For example, the MI path coefficient for $COST \rightarrow OUTSTATE$ was quite large (295.02) compared to that produced under EM and FIML. For example, the path coefficient among the latent variables $QUALITY \rightarrow PERSISTENCE$ was 4.47 for MI and -3.48 for FIML. Similar conclusions hold for the standard errors. Since the fit of the model was not adequate using any of the methods, it is possible that different conclusions may have been drawn for a model that fit the data. Still, these results provide graphic evidence of the different conclusions that can be reached depending on which missing data method is employed.

Figure 4
Path Model Fitted To Graduation Data

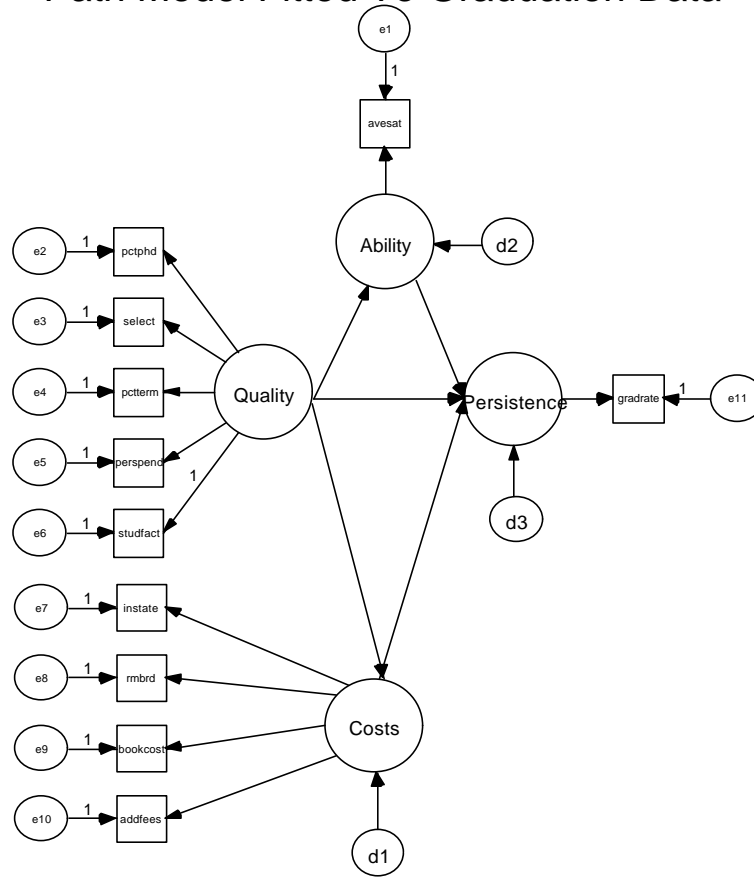


Table 7. Estimates, Standard Errors and t Statistics for Path Coefficients Using the SEM Model

Paths			LD		EM		FIML		MI	
			Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Ability	<--	Quality	-88.87	17.19	-88.79	15.36	-155.78	33.71	-88.56	15.61
	-									
Costs	<--	Quality	-0.05	14.29	-14.12	8.20	-14.99	8.97	-9.14	9.63
	-									
Persistence	<--	Costs	94.44	27720.1	0.34	0.19	0.57	0.32	0.71	1.45
	-									
Persistence	<--	Ability	0.06	0.01	0.08	0.00	0.00	0.00	.07	0.00
	-									
Persistence	<--	Quality	3.43	1.31	5.03	1.09	-3.48	1.23	4.47	1.06
	-									
Percent of faculty with PhD.	<--	Quality	-15.94	3.01	-19.19	3.26	-18.43	3.49	-19.60	3.43
	-									
Average Total SAT	<--	Ability	1.00	-----	1.00	-----	1.00	-----	1.00	-----
	-									
Student to Faculty Ratio	<--	Quality	1.00	-----	1.00	-----	1.00	-----	1.00	-----
	-									
Average Expenditure per Student	<--	Quality	13.71	31.67	48.62	23.68	-83.61	33.29	47.77	27.15
	-									
Percent Full Time Students	<--	Quality	-13.97	2.64	-18.70	3.18	-22.38	4.27	-19.16	3.35
	-									
No admitted / No applied	<--	Quality	0.06	0.01	0.06	0.01	-0.14	0.04	.06	0.01
	-									
Average Additional Fees	<--	Costs	1.00	-----	1.00	-----	1.00	-----	1.00	-----
	-									
Average Cost of Books	<--	Costs	194.17	56990.0	0.92	0.55	1.69	0.99	1.61	3.40
	-									
Room and Board	<--	Costs	9060.34	2660000	40.91	22.66	44.52	25.22	74.96	152.07
	-									
Out of State Tuition	<--	Costs	38088.2	1120000	155.45	86.12	152.83	86.56	295.02	602.36
	-									
Graduation Rate	<--	Persist	1.00	-----	1.00	-----	1.00	-----	1.00	-----
	-									

*Minimization was unsuccessful

Figure 5. Path Coefficients for the SEM Model

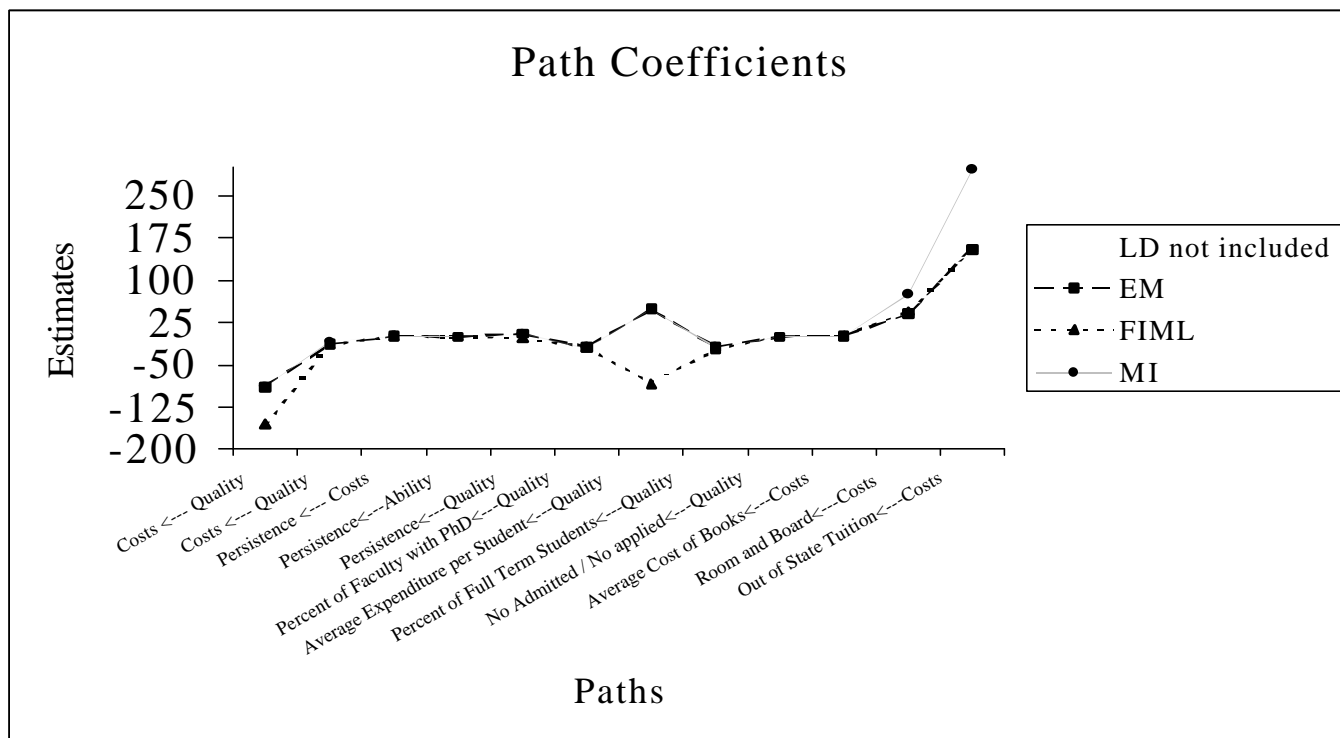


Figure 6 Standard Errors of Path Coefficients for the SEM Model

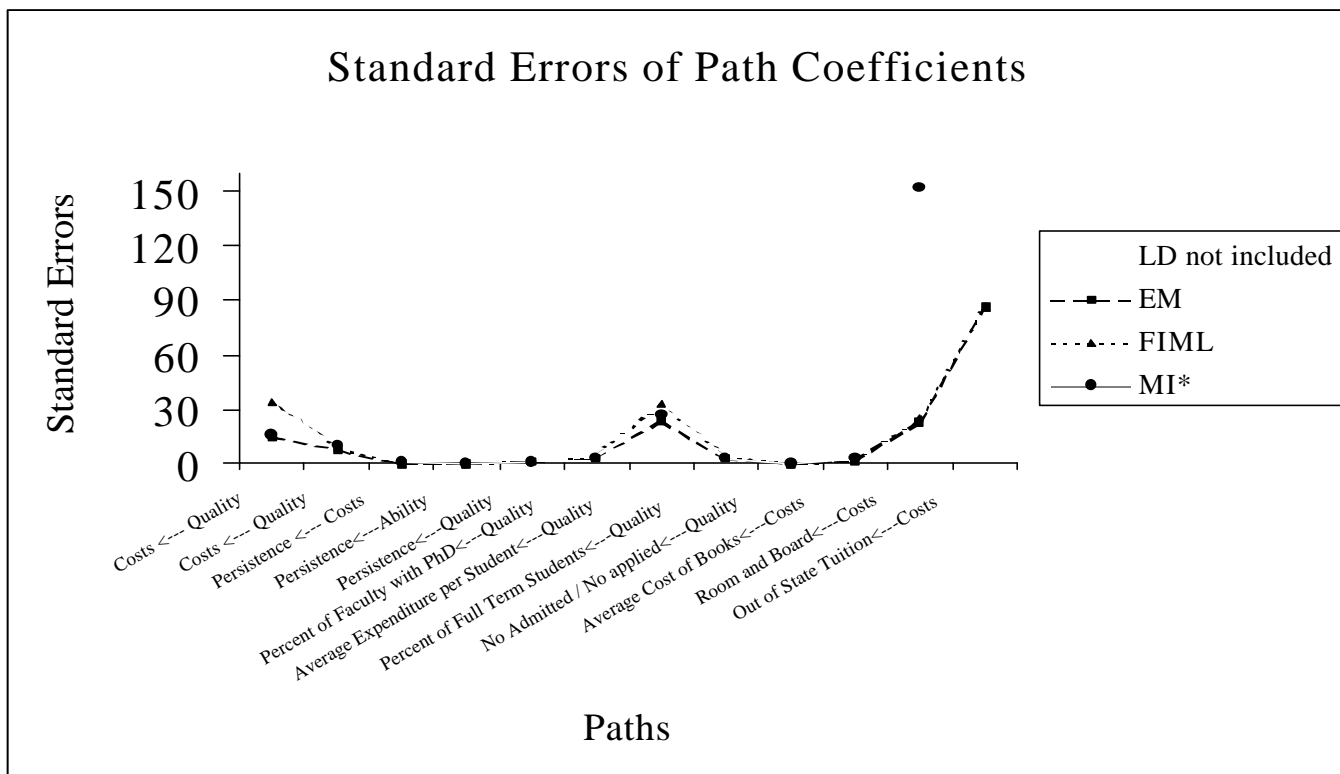


Table 8 reports the MI results for the SEM, and shows some variation in the df for testing various estimated path coefficients. The %MIS. INF. values indicate that path coefficients were generally estimated with better precision than was the case in the regression model.

Table 8. MI Results For SEM

QUANTITY	ESTIMATE	STD. ERR.	T-RATIO	DF	P-VALUE
Quality/Ability	-88.5554	15.6141	-5.67	119888	0.0000
Quality/Costs	-9.14260	9.62623	-0.95	126	0.3441
Costs/Persisence	0.710800	1.44870	0.49	565	0.6239
Ability/Persistence	0.668000E-01	0.465188E-0	14.36	58	0.0000
Quality/Persistence	4.47280	1.05871	4.22	5272	0.0000
Quality/PctPhD	-19.6028	3.42932	-5.72	28238	0.0000
Quality/Perspend	47.7720	27.1535	1.76	266	0.0797
Quality/Pctterm	-19.1634	3.34612	-5.73	53306	0.0000
Quality/Select	0.564000E-01	0.112499E-0	5.01	69520	0.0000
Costs/Bookcosts	1.61280	3.40189	0.47	529	0.6356
Costs/RoomBoard	74.9636	152.068	0.49	573	0.6222
Costs/Outstate	295.015	602.363	0.49	548	0.6245

CONFIDENCE LEVEL FOR INTERVAL ESTIMATES (%): 95.00

QUANTITY	LOW ENDPT.	HIGH ENDPT.	%MIS. INF.
Quality/Ability	-119.159	-57.9521	0.6
Quality/Costs	-28.1926	9.90742	19.0
Costs/Persisence	-2.13469	3.55629	8.7
Ability/Persistence	0.574882E-01	0.761118E-01	28.5
Quality/Persistence	2.39729	6.54831	2.8
Quality/PctPhD	-26.3244	-12.8812	1.2
Quality/Perspend	-5.69114	101.235	12.9
Quality/Pctterm	-25.7218	-12.6050	0.9
Quality/Select	0.343502E-01	0.784498E-01	0.8
Costs/Bookcosts	-5.07007	8.29567	9.0
Costs/RoomBoard	-223.716	373.643	8.7
Costs/Outstate	-888.207	1478.24	8.9

Conclusion

The selection of a particular missing data method can have important consequences on the data analysis and subsequent interpretation of findings in institutional models of student persistence. Among the missing data methods, multiple imputation is the most widely recommended. But because of the computational intensity of this method and the need for specialized software it is natural to ask if a simpler, more accessible method can be used. In general, no other method produced results that were uniformly

similar to those of multiple imputation, although the EM and FIML methods often were reasonably similar. Among EM and FIML, the EM method may be preferred because it is easily carried out in a statistical package such as SPSS and utilizes a complete data set that ensures that the statistical power of the analyses does not suffer. However, the EM approach has the added assumption that the imputation model is appropriate for predicting missingness.

Our results provide evidence of two kinds of effects associated with choosing a missing data method. First, for the same statistical model, estimated slopes or path coefficients can vary substantially among the different missing data methods, as can their standard errors and indices of model-data fit. Such differences can have a significant impact on interpretation. For example, the slope for the selectivity predictor for the regression model was positive for two of the missing data methods, indicating that less selective institutions are associated with higher graduation rates, whereas for the remaining two methods the slope for selectivity was negative, meaning that increased selectivity was associated with lower graduation rates. This puts student persistence researchers employing this kind of statistical model in the position of reconciling contradictory findings largely on methodological grounds.

Second, the use of the imputation-based methods in structural equation modeling raises a question about achieving adequate model-data fit with versus without imputed values, and the wisdom of model modification attempts. We did not attempt to modify the structural equation model to improve model-data fit because the various missing data methods may have resulted in different modifications, implying a dependency of adequate model-data fit on the selected missing data method. For example, using the multiple imputations method would involve fitting a model to m complete datasets, followed by m separate model modifications. This raises the possibility of having m different final models, an outcome made more likely with higher proportions of missing data. This could also result in substantial compounding of Type I error rates since model modification in this example would involve testing statistical hypotheses on data in m datasets that are mostly the same, inducing a strong dependency test results.

In sum, these results provide evidence of the differences in findings that can occur in student persistence research as a function of the way that missing data are handled. For now, student persistence

researchers are better off adopting the multiple imputations method, despite its difficulties.

References

- Arbuckle, J.L. (1995) Amos for Windows. Analysis of moment structures. Version 3.6. Chicago: SmallWaters Corp.
- Astin, A.W. (1975). *Preventing students from dropping out*. Jossey-Bass: San Francisco.
- Bean, J.P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12, 155-187.
- Bean, J.P. (1986). Assessing and reducing attrition. In D. Hossler (Ed.), *Managing college enrollments: New directions for higher education*. Jossey-Bass: San Francisco.
- Braxton, J.M., Duster, M., & Pascarella, E.T. (1988). Appraising Tinto's theory of college student departure. In J.C. Smart (Ed.), *Higher education: handbook of theory and research*, Vol 12.
- Brown, R.L. (1994) Efficacy of the indirect approach for estimating structural equation model with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 1, 287-316
- 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, Series B*, 22, 302-303.
- Cabrera, A.F., Castaneda, M.B., Nora, A., & Hengstler, D. (1992). The convergence between two theories of college persistence. *Journal of Higher Education*, 63, 143-164.
- Enders, C.K., & Bandalos, D.L. (1999, April). The relative efficacy of full information maximum likelihood estimation for missing data in structural equation models. Paper presented at the annual meeting of the American Educational Research Association, Montreal.
- Grosset, J. (1990, April). A proposed approach for use of Tinto's model with non-traditional students. Paper presented at the annual meeting of the American Educational Research Association, Boston.
- Horn, L.J., & Carroll, C.D. (1996). Nontraditional undergraduates: Trends in enrollment from 1986 to

- 1992 and persistence and attainment among 1989-90 beginning postsecondary students. Office of Educational Research and Improvement, U.S. Department of Education, Technical report NCES 97-578.
- Kim, J.O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, 6, 215-240
- 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.R. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Muthén, B., Kaplan, D., & Hollis, M. (1987) On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431-462
- Nagda, B.A., Gregerman, S.R., Jonides, J., von Hippel, W., & Lerner, J.S. (1998). Undergraduate student-faculty research partnerships affect student retention. *The Review of Higher Education*, 22, 55-72.
- Neter, Kuter, Nachtsheim, & Wasserman, 1996. *Applied Linear Regression Models* (3rd Ed.). Irwin: Chicago, Il.
- Pascarella, E.T., & Chapman, D. (1983). A multi-institutional path analytical validation of Tinto's model of college withdraw. *American Educational Research Journal*, 20, 87-102.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Chapman & Hall.
- Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Shafer, J. L. (1999). NORM program for multiple imputation with normally distributed data. Available from the web site: <http://www.stat.psu.edu/~jls/>
- Shelton, D. (1995). Portrait of a working model for calculating student retention. University of South Carolina. ERIC Document ED 388353.
- Spady, W.G. (1971). An analysis of flunked-out and readmitted students. *Journal of Educational Measurement*, 8, 171-175.

Tinto, T. (1993). *Leaving college: Rethinking the causes and cures of student attrition*. (2nd ed).

University of Chicago Press: Chicago.

Wothke, W. (1999). Longitudinal and multi-group modeling with missing data. In T.D. Little, K.U.

Schnabel, and J. Baumert [Eds.] *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples*. Mahwah, NJ: Lawrence Erlbaum Associates.

