Journal of Educational Measurement
Fall 2002, Vol. 39, No. 3, pp. 187–206

Alternative Missing Data Techniques to Grade Point Average: Imputing Unavailable Grades

Niels Smits
Gideon J. Mellenbergh
Harrie C. M. Vorst
University of Amsterdam

In this article, grade point average (GPA) is considered a missing data technique for unavailable grades in school grade records. In Study 1, theoretical and empirical differences between GPA and seven alternative missing grade techniques were considered. These seven techniques are subject mean substitution, corrected subject mean, subject correlation substitution, regression imputation, expectation maximization algorithm imputation and two multiple imputation methods—stochastic regression imputation and data augmentation. The missing grade techniques differ greatly. Data augmentation and stochastic regression imputation appear to be superior as missing grade techniques. In Study 2, the completed grade records (observed and imputed values) were used in two prediction analyses of academic achievement. One analysis was based on unweighed grades, the other on weighed grades. In both analyses, alternative missing grade methods produced better and more consistent predictions. It is concluded that some alternative missing grade methods are superior to GPA.

The use of the grade point average (GPA) as a measure of academic achievement has been discussed extensively in the literature. The GPA features often as a criterion in the prediction of academic performance and in the validation of college or university admission tests like SAT, ACT, or GRE (e.g., Goldberg & Alliger, 1992; Nilsson, 1995; Stricker, 1991). In addition, person characteristics like personality traits or time management skills are also used to predict GPA (e.g., Britton & Tesser, 1991; Wolfe & Johnson, 1995). The GPA has also been used as an independent variable. It has been used as a predictor of success in higher education (e.g., Cariaga-Lo, Enarson, Crandall, Zaccaro, & Richards, 1997; House & Johnson, 1992) and of job performance after graduation (e.g., Bretz, 1989). It has been used both in psychological or educational research and in economic research (e.g., Betts & Morell, 1999).

The appeal of the GPA is that it is well defined, widely understood, and easily obtainable from university records (Young, 1993). However, the use of GPA as a measure of performance does have its flaws. The GPA may differ between students in exact meaning. It is often based on ratings or grades obtained on a set of courses that differ in content from student to student. In practice, the GPA is treated as if it represents the same construct regardless of the student's exact curriculum (e.g., Linn & Hastings, 1984). Variation in its exact meaning over students may decrease its reliability (e.g., Young, 1993). The widespread use of GPA is in part attributable to the fact that it is available for all students, even though variation in the

composition of curriculum may render the averages not entirely comparable. The use of a more subject-specific measure of achievement would give rise to a missing data problem, because many students will not have followed the courses relating to the specific subject.

Another problem in validation studies of pre-admission measures is the possible variation in grading standards within and between departments. Stricker, Rock, Burton, Muraki, and Jirele (1994) and Stricker, Rock, and Burton (1996) state that GPA lacks reliability and validity, because it cannot be used to compare students who take courses with severe grading standards and students who take courses with lenient standards.

Several attempts have been made to adjust grades to render them more comparable (see, e.g., Young, 1990, 1993). Stricker et al. (1994) compared the effectiveness of several statistical methods to adjust GPA criteria for differences in grading standards of individual courses. One method involved the treatment of missing grades (missing because a student had not taken a given course) as a missing data problem. The unavailable grades were viewed as missing at random in the sense that they are predictable from the available grades. Stricker et al. (1994) pointed out that the expectation maximization (EM) algorithm may be used to obtain maximum likelihood estimates of the missing grades. Approaching the problem of unavailable grades as a missing data problem is not new. At least implicitly, this approach has often been taken. In calculating the GPA, the average grade for courses that a student has taken is implicitly substituted for missing grade scores. In this procedure, it is assumed that grades for each course are exchangeable and that the data are missing completely at random. From this point of view, this method of calculating GPA can be considered as a simple missing data technique. A number of alternative techniques exist that may be applied to the missing grade problem. It should be noted that these techniques are not adjustments of GPA in the sense of Stricker et al. (1994), but missing data methods.

In this article, the assumptions of GPA and alternative missing grade techniques are described and their use is demonstrated. This article includes two studies. In Study 1, theoretical and empirical similarities among the missing grade techniques were examined. In Study 2, the performance of these techniques was examined when the imputed values were used as exact grades in the prediction of academic achievement among Dutch university students. We first provide a short description of the Dutch educational system, as it differs from the Anglo-Saxon system.

The Dutch Educational System

In the Netherlands, primary education starts at age four and continues for 8 years to the age of 12. Following primary education, there are four levels of secondary education, of which one level is pre-university education (Dutch abbreviation: VWO) that lasts 6 years. At the end of secondary education, students take final examinations. Prior to the last 2 years of secondary education, students choose their examination subjects. Students in the VWO take exams in at least 7 of 15 subjects. These are Dutch (obligatory), English (obligatory), French, German, history, geography, two mathematics subjects, physics, chemistry, biology, two economics sub-

jects, Latin, and Greek. Students with a VWO certificate have automatic access to university education. They are not required to take additional admission tests.

At university, master's degree curriculum officially takes 4 years. Students are expected to earn 42 credits per annum (representing 42 weeks of work). Lectures are given, but attendance in most courses is not obligatory. Students vary greatly in the number of credits they actually obtain in a year and in the total time they need to graduate. The amount of time they take to graduate does have financial consequences, however, because grants supplied by the Dutch government, to which students are entitled, are performance related. Students receive their grants initially in the form of a loan, which is converted to a nonrepayable grant if they meet certain performance criteria.

Study 1: Similarities and Differences Between Missing Grade Techniques

Methods for dealing with missing data usually involve implicit or explicit assumptions about the process that causes missing data. If missingness depends on neither the values of the observed variables nor those of unobserved variables, the missing data are called "missing completely at random" (MCAR). If missingness does not depend on the values of unobserved variables, but possibly does depend on the values of the observed variables, the data are called "missing at random" (MAR) (e.g., Little & Rubin, 1987). If missingness depends on the missing values of unobserved variables, then the data are "not missing at random" (NMAR). For example, let us suppose that in a group of students, grades on course A are available for all students, but that the grades on course B are missing in certain cases. If the probability that grades on course B are available is equal for all students, regardless of their grades on course A or course B, then the missing grades are said to be MCAR. If the probability that grades on course B are available is related to grades on course A, but not to grades on course B (e.g., if a high grade on course A is a prerequisite for attending course B), the grades are MAR. If the probability of availability is related to the grades of course B itself (e.g., when a student decides not to attend to course B because he or she has a low level on the ability that is associated with course B), then the data are NMAR. (For a more extensive description of missing data mechanisms, see, e.g., Little & Rubin, 1987; or Little & Schenker, 1995.)

Rubin (1976) has discussed the problem of statistical inference in the presence of missing data. In likelihood or Bayesian based inference, it is assumed that the data are MAR. In classical (common) statistical inference, missingness may be ignored only if the data are MCAR.

In this article, several methods of handling missing grades are considered. The methods described ignore the missing data process. It is likely that the students' choice of subjects is prone to selection and therefore that the missing grades are NMAR. Students may choose subjects or courses they are relatively good at, which may result in higher grades. This implies that some missing data techniques may violate assumptions when applied to the missing grade problem.

Missing data techniques can be divided into two categories: first, methods that estimate parameters, and second, methods that replace missing data by derived data (e.g., imputations). In the present article we are concerned only with techniques in

the latter category for two reasons. In Study 1 various methods were compared at the individual grade level. In Study 2 cross-validation was performed, which requires complete (observed and imputed) grades.

Method

Data

The VWO school records of six cohorts of psychology freshmen at the University of Amsterdam were used in the present studies. The total number of records is 2,080. Sex of the students is known in the last three cohorts (40% of the students). Of the students in this group, 64% are female and 36% are male. It is known that the ratio of women to men was also about 2 to 1 in the first three cohorts. Every student made a choice of at least 7 out of 15 subjects at VWO, which means that grades on eight or fewer subjects are missing. The grades are expressed in integer values on a scale from 1 to 10. To pass the VWO exam a grade of 6 or higher on nearly all subjects is required; the school records contained no grades lower than 4.

In this study, five school subjects were selected. The subjects Dutch and English were selected as grades because these subjects are available for all students. Three other subjects selected were biology, French, and history. In Table 1, the missing grade pattern as observed in the present sample is reported. It should be noted that this pattern is not monotone (e.g., Little & Rubin, 1987), that is, there is no accumulation of missingness as one may encounter in panel data with definite dropout. As can be seen in the last row of Table 1, only one third of the sample took examinations in biology.

Regardless of the exact technique, all imputed values were rounded to the nearest integer so that they are expressed in the same manner as the observed grades.

Description of Eight Missing Grade Techniques

Grade point average. The first missing grade technique is the traditional method GPA. This technique implicitly imputes a personal mean: every missing course grade of a given student is replaced by the mean of the student's available grades. Contrary to the common use of GPA, missing grades were explicitly replaced by GPA. In calculating GPA in this manner, the grades are assumed to be exchangeable. In addition each available grade is weighed in the same manner (differential weighing of grades is not applied). A consequence of the procedure is that the imputed grades of a given student are identical. GPA is a special case of regression imputation.

Subject mean substitution. Unconditional mean imputation, which is a naive missing data technique, involves the replacement of missing values by the sample mean of the variable. It can yield satisfactory estimates of unconditional means and totals, but it does not yield consistent estimates of other parameters, even if the data are MCAR. This method results in the underestimation of variances and a distortion of the associations between variables (Little & Schenker, 1995). The missing grades version of unconditional mean imputation involves the imputation based on the average grade. This technique is called "subject mean substitution" (SMS). The student's missing grade on a given subject is replaced by the average of the

TABLE 1
Missing grades pattern for 2,080 students

	Subjects				
Number of students	Dutch	English	Biology	French	History
140					
611			X		
374			X	X	
224				X	
197				X	X
128					X
238			X		X
168			X	X	X
Percentage Missing	0	0	67	46	35

Note. The X's indicate missing values. Blanks indicate observed subjects.

available grades on the subject. Although it is obvious that this method is far from optimal, it is included in this study for reasons of comparison.

Corrected subject mean. Another mean imputation method is corrected subject mean (CSM) substitution. This method replaces missing grades by an average subject grade, which is corrected for the student's ability. The subject mean is multiplied by a factor reflecting the ratio between the student's grades on available subjects and the average available grades on these subjects. It is given by

$$CSM_{vi} = \left[\frac{\sum_{i \in obs(v)} x_{vi}}{\sum_{i \in obs(v)} \bar{x}_{.i}} \right] \bar{x}_{.m}, \tag{1}$$

where x_{vi} is the grade of student v on subject i, $\bar{x}_{.i}$ is the mean grade of the observed values on subject i, $\bar{x}_{.m}$ is the mean grade based on the available values, and *obs(v)* is the set of student v's available grades. CSM can be regarded as a compromise between GPA and SMS, because it uses both student and school subject information. The numerator in Equation 1 is equal to GPA times the number of available grades of student v; SMS is identical to $\bar{x}_{.m}$. CSM takes into account the mean grade of the unavailable subject, so that the subjects are no longer considered exchangeable, but may differ in difficulty. CSM is equivalent to the corrected item mean score substitution developed by Huisman and Molenaar (2001).

Subject correlation substitution. Subject correlation substitution (SCS) replaces the students' unavailable grades on a given subject by the available grade for the school subject that has the highest correlation with the subject in question. This method is similar to the item correlation substitution method proposed by Huisman (1999).

Regression imputation and stochastic regression imputation. Regression imputation (RI) represents a more principled missing data method. This technique replaces missing values with predicted values obtained from a linear regression of the variable with missing values on other observed variables. When the MCAR assumption holds, the averages of the observed and imputed values are consistent estimates of the means. The covariance matrix of the imputed data however does require an adjustment (Little & Rubin, 1987). RI can be interpreted as the imputation based on a conditional mean and hence tends to be less variable than the observed values (Little & Schenker, 1995). In RI, it is assumed that, conditionally on the available data, the missing data are normally distributed.

Stochastic regression imputation (SRI) involves the addition of a random error to the regression prediction obtained using RI. SRI compensates for the underestimation of the variance of variables with missing data that is associated with RI (e.g., Little & Schenker, 1995). However, SRI is not proper in the sense of Rubin (1987) if the imputed values are used as exact values in secondary analyses, because the method underestimates error due to imputation (Beaton & Johnson, 1990). RI and SRI were applied to the missing grade problem by replacing students' missing grades by the predicted value obtained in the regression of the subject, which includes the missing values, on available subjects. As mentioned SRI involves the addition of a random error term to the predicted grade. GPA can be seen as a special case of RI. GPA is a regression with the regression coefficients fixed at one. RI and SRI are two missing data techniques in the Missing Variable Analysis (MVA) module in SPSS 8.0.0 for Windows. In the case of SRI, three optional random error terms can be added to the predicted values obtained using RI. We chose to add a random normal deviate, scaled by the standard error of the estimated grade.

In this study, to account for the improperness of SRI as a single imputation method, SRI was applied as a multiple imputation (MI) technique rather than a single imputation technique. In this case, the missing values were imputed five times. MI will be described in more detail.

EM algorithm imputation. Some missing data methods are based on a likelihood approach (Little & Rubin, 1987). Maximum likelihood (ML) estimation of missing data requires the specification of a model for both the distribution of the data and the missing data mechanism (Little & Schenker, 1995). Frequently ML inference is based on the assumption that the data are MAR or, in other words, that the missing data mechanism is ignorable (for ML methods that estimate parameters for nonignorable models, see, e.g., Liou, 1998). The EM algorithm (Dempster, Laird, & Rubin, 1977) is commonly used to obtain ML estimates of the parameters of data with missing values. In the E step of the EM algorithm the expected value of the incomplete data likelihood is calculated, given the observed data and current parameter estimates. In the M step the ML function is maximized, given the expected values obtained in the E step, to provide new parameter estimates (Little, 1992). The EM algorithm when applied to normally distributed data can be viewed as an iterative form of RI. In the E step the best linear predictions of the missing values are calculated, using current estimates of the parameters. In the M step, the mean vector and covariance matrix of the completed data are calculated, incorpo-

rating corrections of the covariance matrix for imputing predicted means (Little & Schenker, 1995).

The EM algorithm is applicable to the missing grade problem. Stricker et al. (1994) used the EM algorithm in order to adjust GPA, assuming the missing grades were MAR. As mentioned, this assumption might not be correct. However, ignorable ML methods reduce missing data bias even when the assumption of MAR is not strictly valid (Little & Rubin, 1990). Contrary to Stricker et al. (1994), in the present article, the EM algorithm was applied as a method to estimate unavailable grades and not as a way to adjust GPA.

The EM algorithm for normally distributed data was used in this study. The assumption of multivariate normality may not apply to school grades. However, the EM algorithm can provide consistent estimates of the underlying distribution under weaker distributional assumptions (Little & Rubin, 1987). Although the EM algorithm is primarily a method to obtain estimates of the means and covariance matrix, here we require actual imputed values. We use the imputed values as provided by SPSS 8.0.0 for Windows. This imputation is referred to as EM imputation (EMI). EMI is the same procedure that Stricker et al. (1994) used in their analysis.

Multiple imputation. Another general model based technique for the imputation of missing data is MI. A single imputed value does not adequately represent the uncertainty about the imputed value. Therefore, in analyses that treat imputed values as observed values, the uncertainty associated with imputing data is not properly accounted for, even if the missingness is modeled correctly and random imputations are created (Little & Schenker, 1995). In MI, a missing value is replaced by a set of $m > 1$ plausible values drawn from a predictive distribution. The variability among the m imputations provides a measure of the uncertainty with which the missing values are derived from the observed ones (Schafer & Olsen, 1998). Standard complete data methods are used to analyze each data set. The m complete data inferences can be combined, using simple rules provided by Rubin (1987), to arrive at a single inference. These rules produce overall estimates and standard errors that reflect the uncertainty due to missing data and imputation. Generally, MI does not require or assume that the missingness is ignorable. Imputation may in fact be created under any kind of model for nonresponse. The resulting inferences will be valid under the postulated model (Little & Rubin, 1990).

One model that can be used to create MIs is SRI. Little and Rubin (1990) call SRI a crude version of MI. In the case of SRI, the imputations in each of the m data files are dependent, because imputed values only differ by a stochastic error term. In this study, SRI was used five times to create five complete grade files.

Schafer (1997) and Schafer and Olsen (1998) made use of another procedure to create MIs called "data augmentation" (DA). This procedure was first developed by Tanner and Wong (1987). DA is an iterative procedure that alternately performs a random imputation of missing data under assumed values of the parameters and draws new parameters from a Bayesian posterior distribution based on the observed and imputed data. The procedure creates a Markov chain that eventually converges in distribution. The distribution of the missing data stabilizes to a predictive distribution, out of which values are drawn to create MIs (Schafer & Olsen, 1998).

For DA, a very small value of m will usually suffice (Schafer, 1997). As is the case with the EM algorithm, inference on the basis of DA is valid only if the missing data mechanism is ignorable.

DA for multivariate normal data with missing values is similar to the EM algorithm. The deterministic E and M steps are replaced by stochastic I and P steps, respectively. The I step of DA involves the independent simulation of random normal vectors for each row of the data matrix, with means and covariances equal to those estimated in the previous iteration. The P step simulates a mean vector and a covariance matrix conditioned on the observed data and estimated values of the missing data in the current I step (Schafer, 1997).

The Bayesian nature of DA appears in the specification of a prior distribution of the parameters of the missing data, which depends on the imputer's belief about these parameters. In practice, Bayesian inference is more sensitive to the choice of the data model than to the choice of the prior distribution. DA allows one to specify noninformative prior distributions, in case one has no information about the parameters of the missing data. Usually, the use of a noninformative prior works well (Schafer & Olsen, 1998).

For DA to work it is assumed the data are at least MAR. As mentioned before, school grades may be NMAR. Yet, DA tends to perform better than ad hoc procedures such as listwise deletion or mean imputation, even when the assumption of MAR is unrealistic (Schafer & Olsen, 1998).

DA was carried out as described by Schafer (1997) and Schafer and Olsen (1998) using an S-plus library, called NORM, that was written by Schafer (1998) (this library is downloadable at http://www.stat.psu.edu/~jls/misoftwa.html). NORM was developed for the multiple imputation of multivariate continuous data under normality. Like RI, SRI, and EMI, DA is based on the assumption that the data are normally distributed. Commonly, school grades may at best be approximately normally distributed. However, DA tends to be quite robust against violations of the imputation model (Schafer & Olsen, 1998).

As starting values for the incomplete data parameters, ML estimates produced by the EM algorithm in NORM were used. The default noninformative prior was used. The incomplete grade records were imputed five times and complete data analysis was performed on each of the five complete data sets. Inferences were based on the results obtained from all five analyses (for an extensive description of MI inference, see, e.g., Rubin, 1987, chap. 3). Convergence of the estimates of the means was assessed for each of the three subjects by using time series and autocorrelation function plots, as described by Schafer (1997, chap. 4 & 5).

Evaluation of the Eight Missing Grade Techniques

Differences and similarities between the techniques described above were evaluated by several formal principles, as presented by Little (1988). According to Little's first principle, imputations should be based on the predictive distribution of the missing values, given observed values of each case. The second principle states that in each case all observed variables should be taken into account in obtaining imputations. Third, imputations should be based on contextual knowledge about the imputed variables. The fourth principle states that excessive extrapolation beyond

the range of the data should be avoided. Fifth, imputations should be drawn from the predictive distribution (mentioned in Principle 1) to preserve the distribution of the variables in the completed data set. According to the sixth principle, a method should be provided to calculate sampling errors of estimates that takes into account the fact that values have been imputed. In the present article, Principles 1, 2, 5, and 6 were used to classify the missing grade methods. Principle 3 is not applied because "contextual" knowledge about the imputation of missing grades is not available. Principle 4 does not apply because none of the imputed values falls outside the range of 1 to 10. The second principle states that in each case all observed variables should be taken into account. Here the other observed variables are simply the observed, non-missing, grades. Table 2 indicates whether each method satisfies the four principles. There clearly are large differences between the eight methods. SMS is the only method that does not satisfy any of the principles; it merely imputes one value for all missing grades on a subject. SCS only satisfies the principle of using a predictive distribution; this predictive distribution is based on one observed grade only. GPA, CSM, RI, and EMI each satisfy two of the four principles. SRI and DA satisfy all four principles. SRI satisfies the principle of accounting for uncertainty due to missingness because it was applied five times. However, SRI was not designed as a MI method. If it is applied only once, it does not meet this principle. On the basis of the four principles, DA and SRI look most promising, followed by GPA, CSM, RI, and EMI. Given the present criteria, SMS and SCS seem to be inappropriate methods.

Analysis

First, methods were compared with respect to the imputed grade values. To this end, the means and standard deviations of the imputed values were calculated. To investigate the agreement in ranking produced by the various methods, imputed values were correlated. Correlations were also computed in the whole data files, including both observed and imputed values. Next, sex and cohort effects on the differences between imputed values were investigated. The differences between imputed grades were assessed by calculating the imputed values variance of each grade, for all students. The variances among methods were computed using the means of the five imputed values obtained using DA and SRI.

Results

Convergence of the EM and DA Algorithms

The EM algorithm converged within the specified maximum number of 100 iterations in SPSS. Little's MCAR test, which tests the hypothesis that data are MCAR, was significant (χ^2 (23, $N = 2080$) = 113.09, $p < .00$), indicating that missing grades are not MCAR. In carrying out DA, the EM algorithm implemented in NORM took less than 35 iterations to estimate the mean vector and covariance matrix of the five subjects. It appeared that 35 cycles of DA sufficed to converge in distribution. For an extra margin of safety, it was decided to carry out 150 cycles of DA between imputations. DA was run for a total of 750 cycles, producing an imputation at every 150th cycle for a total of $m = 5$ imputations. Both the time

TABLE 2
Classification of the eight missing grade techniques on Little's (1988) four principles.

Missing data technique	Little's principles			
	Predictive distribution	All grades	Drawn from distribution	Accounting for uncertainty
GPA	+	+	−	−
SMS	−	−	−	−
CSM	+	+	−	−
SCS	+	−	−	−
RI	+	+	−	−
SRI	+	+	+	+[a]
EMI	+	+	−	−
DA	+	+	+	+

Note. A plus indicates that a technique satisfies the principle.

[a] SRI is not a MI method by definition. When it is applied as a single imputation method, then the fourth principle (accounting for uncertainty) is not satisfied.

series and autocorrelation function plots indicated that the estimates of the means of the three subjects biology, French, and history converged rapidly.[1]

Empirical Comparison of Missing Grade Techniques

Large differences were found between methods in the distribution of the imputed values in the completed files (see Table 3). The imputed values obtained using the two methods based on student information only, that is, GPA and SCS, are characterized by large variances relative to the variances of the observed values. In the case of the subject with the lowest mean observed grade, biology, means of the imputed values obtained using GPA and SCS were higher than the observed values. In other words, on average students who did not take examinations in biology ended up with a higher grade than students who did take examinations in this subject. This is a consequence of imputing biology grades on the basis of available grades on other subjects (GPA imputes an average of available grades), which on average are higher than the biology grades. The means of the imputed grades obtained by CSM and means of the observed values are close in value. The variances of the imputed values however are lower. SMS produced an average imputed biology grade that was lower than the average of the observed biology grade. The average French and history grades obtained by SMS are larger than the corresponding means of the available grades. This is a consequence of the rounding that was employed; the average of 6.32 for biology was rounded to 6, and the average of 6.80 for French and the average of 6.86 for history were both rounded to 7. The standard deviations of SMS are zero, which is a natural result of the imputation of a constant. Both RI and EMI produced mean imputed history grades close to the mean of the observed history grades. The mean imputed French and biology grades are somewhat lower than the corresponding means of the observed grades. The standard deviations of EMI and RI imputations are quite small com-

TABLE 3
Means and standard deviations per subject for the observed and imputed grades

	Subjects									
	Dutch		English		Biology		French		History	
Grades	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Observed Value	6.79	0.75	7.15	0.95	6.32	0.77	6.81	0.97	6.86	0.77
Imputed value										
GPA	—	—	—	—	7.01	0.69	6.89	0.66	6.93	0.70
SMS	—	—	—	—	6.00	0.00	7.00	0.00	7.00	0.00
CSM	—	—	—	—	6.34	0.62	6.81	0.67	6.83	0.72
SCS	—	—	—	—	6.92	0.98	7.06	0.98	7.12	0.97
RI	—	—	—	—	6.19	0.40	6.76	0.58	6.90	0.37
SRI	—	—	—	—	6.33	0.82	6.75	1.05	6.86	0.83
EMI	—	—	—	—	6.21	0.41	6.71	0.59	6.89	0.36
DA	—	—	—	—	6.31	0.53	6.70	0.75	6.87	0.54
Number of grades										
Observed	2080		2080		689		1117		1349	
Imputed	0		0		1391		963		731	

Note. No means and *SD*s are reported for imputed values for Dutch and English, because all grades were available.

pared to the standard deviations of the observed values. The relatively small standard deviations of EMI imputations came as a surprise (see Comments, Study 1). The means for the SRI and DA imputations are very close to the means of the observed values for all subjects. However, DA produced smaller standard deviations than the standard deviations of the observed values. SRI produced standard deviations approximately equal to the standard deviations of the observed values.

In Table 4, the correlations between the imputed values produced by the different methods are reported (the correlation of SMS with other methods is not given; the correlation cannot be computed because the imputed values have a variance of zero). The methods clearly differ greatly in the ranking of the students they produce. Values obtained by SRI and DA had consistently lower correlations with the values obtained by other methods. The correlation between EMI and RI was very high for each of the three grades. The correlations between the imputations were highest for the subject French and lowest for biology.

The correlations among the history grades and biology grades that were imputed by means of SRI and RI were low. This is surprising because the imputed values only differed by a stochastic error term. The correlation between the imputed French grades obtained using SRI and RI, however, was higher. The difference in correlations is due to greater variance of the imputed French grades compared to the variance of the imputed biology and history grades. Because of the low variance of RI imputed history and biology grades, the addition of stochastic error (that had approximately equal variance for history, French, and biology) had a relatively greater effect, making grades imputed by RI and SRI more variable.

TABLE 4
Between-method correlations of imputed values per school subject

Method	GPA	CSM	SCS	RI	SRI	EMI	DA
			Biology				
GPA	1.00						
CSM	.75	1.00					
SCS	.69	.68	1.00				
RI	.74	.70	.57	1.00			
SRI	.37	.39	.33	.35	1.00		
EMI	.75	.72	.54	.95	.34	1.00	
DA	.40	.42	.30	.35	.21	.35	1.00
			French				
GPA	1.00						
CSM	.92	1.00					
SCS	.69	.70	1.00				
RI	.70	.71	.85	1.00			
SRI	.57	.58	.67	.63	1.00		
EMI	.73	.74	.86	.93	.62	1.00	
DA	.62	.63	.73	.67	.54	.67	1.00
			History				
GPA	1.00						
CSM	.91	1.00					
SCS	.72	.75	1.00				
RI	.57	.55	.54	1.00			
SRI	.45	.48	.44	.31	1.00		
EMI	.56	.55	.53	.98	.29	1.00	
DA	.47	.49	.46	.37	.21	.37	1.00

Note. SMS was left out of this table; no correlations can be computed because for this method imputed values do not vary among each other.

The correlations between all the grades, that is, both the imputed and observed data were higher because observed grades are identical from one missing data technique to the next. As the general pattern of correlations is similar to that based on the imputed values, these results are not reported.

To check whether differences between the imputed values were related to person characteristics, the variance among the imputed values per grade and for each student was calculated. First, this variance was related to the number of unavailable grades. To assess the effect of the number of missing grades associated with two of the three subjects (values 0, 1, or 2) on the variance of imputed grades for the remaining subject, ANOVAs were preformed. Eta squared was used as a measure of effect size. The etas squared equaled .012, .005, and .008 for biology, French, and history, respectively. Because these effect sizes are very small (see, Cohen, 1977), it appears that there is no relationship between the number of missing grades and variance of the imputed grades. Second, to assess the effect of cohorts on the variances, an ANOVA with cohort as factor was performed for each subject. As the etas squared equaled .003, .008 and .025 (for biology, French and history, respec-

tively), we concluded that a cohort effect was absent. To check whether sex was related to the differences between imputations, the point biserial correlation between the variance of the imputed values and sex was computed. Information about sex was available only in the last three cohorts (40% of the students) and unavailable in the other cohorts. Therefore, these data were missing by design, which means the data are at least MAR (e.g., Schafer, 1997). In view of this, correlations were computed using the EM algorithm. The correlations equaled .08, −.02, and .04 for biology, French, and history, respectively. Again these results suggest that differences between sex and differences in variance among the imputed values are unrelated.

Comments

We first comment on the missing grade techniques that were performed using the MVA module in SPSS 8.0.0 for Windows. Generally speaking, in RI, SRI, and EMI, the covariance matrix that is standard output in MVA differs from the covariance matrix that is calculated on the basis of the completed data file (i.e., including observed and imputed grades). In comparing the matrices obtained using EMI and RI, it turned out that there were large differences in MVA estimates, but only small differences between the covariance matrices of the completed grades. EMI and RI yielded nearly identical rankings of the imputed values (see, Table 4), which raises doubt concerning the correctness of the implementation of the MAR assumption for EMI in SPSS.[2]

A second comment concerns the estimation of the unavailable grades on the basis of ML and Bayesian inference, that is, using EMI and DA, respectively. To ensure convergence of estimation, the ratio of courses to students should not be too high. In the present study this ratio was low, the unavailable grades on just five subjects were estimated in a sample of 2,080 students. However, if this ratio is too high, estimation of the unavailable grades may not converge. Stricker et al. (1994) encountered this problem in estimating GPAs for departments instead of single course grades. In many situations, the ratio of courses to students will probably be acceptable so that application of these methods should not pose a convergence problem. However, if it is not possible to use these two methods, other, noniterative methods that estimate grades, such as RI and SRI, may be used.

Study 2: Prediction With the Completed Grades File

In the second study, the imputed and observed grades were used to predict academic achievement.

Method

Data

In the present study, the grades (both observed and imputed) featured as the independent variables and the number of credits that a student earned within the freshman psychology year featured as the dependent variable. The number of credits is considered to be a good measure of study success because it accurately reflects study progress. The full freshman program comprises of 42 credits. The number of credits was assessed at the end of the freshman year. On the basis of

accuracy of administration criteria, students from two cohorts were selected for the present study. The two cohorts comprised a total 446 students, distributed about equally over the cohorts ($n = 213$ and $n = 233$). The cohorts were considered comparable since the governmental and departmental policymaking had not changed between these two cohorts.

Analysis

Two analyses were performed in the prediction study: one in which grades were weighed identically, and another in which grades were weighed differentially.

Prediction analysis with identically weighed grades. In predicting academic achievement using identically weighed grades, each subject contributed equally to the prediction. The correlation between the average grade and the number of credits was calculated within each cohort and separately for each imputation technique.

Prediction analysis with differentially weighed grades. In the second analysis, linear regression was performed within each cohort. In this regression analysis, the grades on the five subjects were used as predictors, and the number of credits obtained served as the criterion. Regression weights were estimated, allowing for varying contributions of each subject in the prediction. The correlation between the predicted values and the criterion was used as a measure of the predictive validity. The GPA imputation method that was applied here should not be confused with the common use of GPA. Every missing grade was explicitly replaced by the personal mean of the student and regression analyses were performed on the completed grade records.

To assess the error rates of the regression analyses, results obtained in one cohort were cross-validated in the other cohort. Cross-validation is the traditional method to estimate error rates and it produces estimates of error rates that are almost unbiased (Efron & Tibshirani, 1995). Cross-validation involves the computation of a predicted score by weighing the predictors in one sample with the regression weights estimated in the other sample.

This study can be seen as an experiment in which the missing data technique is the independent variable, and the correlation between the actual number of credits and the prediction of the number of credits based upon the completed grade records is the dependent variable.

Results

Prediction Analysis With Unweighed Grades

In Table 5 it can be seen that for all methods the correlations between the average grade over all subjects are higher in Cohort 2 than they are in Cohort 1. In other words, prediction based on unweighed grades explained more variance in the criterion in Cohort 2 than it does in Cohort 1. In both cohorts, the correlations associated with SMS, EMI, and RI are higher than those associated with GPA. The correlations associated with SCS and DA are lower than those associated with GPA in both cohorts. These differences are small, however. In Table 5 MI standard errors for the estimated correlations associated with SRI and DA are also reported.[3]

TABLE 5
The correlations between the number of credits and the average grade over all subjects after imputation for each missing data technique

Missing data technique	Amount of freshman year credits Cohort 1[a]	Amount of freshman year credits Cohort 2[b]
GPA	.313	.326
SMS	.343	.363
CSM	.310	.352
SCS	.296	.319
RI	.332	.359
SRI[c]	.326(.074)	.330(.070)
EMI	.330	.354
DA[c]	.308(.075)	.309(.080)

[a]$n = 213$. [b]$n = 233$. [c]For SRI and DA, correlation standard errors resulting from within and between imputation variability are reported between brackets.

Prediction Analysis With Weighed Grades

As with prediction based on unweighed grades, multiple correlations and cross-validity coefficients are higher in Cohort 2 than in Cohort 1 (see Table 6). In five of eight methods, the cross-validated multiple correlation is greater than the multiple correlation of the prediction function that was obtained in first cohort. In the second cohort, the multiple correlation associated with GPA was the highest (.442) that was observed in the regression analyses. However, cross-validation of this prediction function in the first cohort also resulted in the largest drop in predictive validity: the correlation fell to .295. The prediction functions associated with the other imputation techniques display a much smaller drop in correlation. The correlation between predicted values and the criterion was consistently large for RI, EMI, SRI, and SMS. The prediction functions associated with CSM and SCS performed well in cross-validation. In the first cohort, the results are not as good. The correlations associated with DA are consistently low. Cross-validation revealed that these correlations remained more or less stable. However, we do note that overall the differences between the methods are small. In Table 6 MI standard errors for the estimated correlations associated with SRI and DA are also reported.[3]

Comments

In Study 2, regression analyses were performed with the completed grades (imputed and observed) as predictors. In the case of RI, SRI, EMI, and DA, the missing grades were estimated on the basis of observed grades. Generally, in applying these methods, information about the criterion can be included in the calculation of imputations. The use of the criterion to estimate missing data among the predictors has different consequences for these four missing data techniques. In the case of RI, regression on the completed predictors (imputed and observed) results in unbiased estimates only if imputations are made conditionally on the predictors alone (see Little, 1992). To ensure unbiased estimation of regression parameters, the EM algorithm has to be applied to the predictors and the criterion in

TABLE 6
The multiple correlations and cross-validities for the prediction of the number of credits

	Number of credits Cohort 1[a]		Number of credits Cohort 2[b]	
Missing Data Technique	Multiple correlation	Cross Validity of Prediction rule 2	Multiple Correlation	Cross Validity of Prediction rule 1
GPA	.330	.295	.442	.390
SMS	.360	.318	.398	.347
CSM	.331	.295	.385	.352
SCS	.335	.320	.369	.355
RI	.351	.326	.404	.376
SRI[c]	.348(.074)	.326(.077)	.368(.070)	.338(.068)
EMI	.350	.320	.399	.362
DA[c]	.331(.073)	.313(.073)	.352(.078)	.330(.077)

[a]$n = 213$. [b]$n = 233$. [c]For SRI and DA, correlation standard errors resulting from within and between imputation variability are reported between brackets.

calculating means and the covariance matrix (Little, 1992). In this study, the inputed grades, which resulted from the last E step of the EM algorithm, were used in subsequent analyses. It is unclear how the estimates of the regression parameters were affected by the exclusion of a criterion in the production of EMI imputations. MI methods like SRI and DA should involve conditioning on both the predictors and the criterion to ensure unbiased estimates in the regression of the criterion on the predictors. If conditioning is limited to the predictors, the strength of the relationship will be underestimated, that is, regression coefficients will be downwardly biased (Little, 1988, 1992; Schafer, 1997; Schafer & Olsen, 1998).

In Study 2, the imputation model did not include the criterion, which may have resulted in underestimation of regression coefficients in the case of DA. (For an extensive discussion of prediction analysis with missing values in the predictors, see e.g., Little, 1992). In this study, imputations were not derived by conditioning on the criterion (number of credits) for the simple reason that accurate credit records were unavailable in some cohorts. One solution to this problem would be to carry out estimation within each cohort. However, a consequence of this would be that the imputation of missing grades would be based on smaller group of students and thus would be more susceptible to sampling fluctuation. Another, more general reason for not conditioning on the criterion when imputing grades is that the criterion may change from one prediction study to another (the number of credits obtained in the first 2 years of study, or the number of years needed to graduate, etc.). One would then require new imputations for the same grade records file in every new study.

A final comment concerns techniques that estimate means and covariance matrices in the presence of missing data, such as the EM algorithm. Regression analysis on the basis of a mean vector and a covariance matrix is more complicated to carry out because statistical packages no longer seem to have the option to perform regression analysis on the basis of the mean vector and covariance matrix. Generally packages require raw data.

Discussion

In Study 1, the comparison of the eight missing grade methods demonstrated large theoretical differences between methods. The techniques were classified on four imputation principles presented by Little (1988). SRI and DA emerged as the most promising, followed by RI, EMI, GPA, and CSM. SMS and SCS appear to be the least suitable as missing grades techniques.

Little's MCAR test showed that the present data are not MCAR, which means that the data are MAR or NMAR. This result suggests that an important assumption concerning the process causing the missingness is violated in the case of at least six of the eight missing grade techniques (GPA, SMS, CSM, SCS, RI, and SRI assume MCAR). DA assumes that the data are MAR. It is not clear whether the assumption of MAR holds when EMI is applied (see Study 1 Comments). As no test is available to establish that the data are MAR, MAR remains an untested assumption. It is likely that the data are in fact NMAR. It appears that the violation of assumptions relating to the missing data mechanism is less serious for DA (and perhaps for EMI) than for the other methods.

Large differences in the values imputed by the various methods were found, and consequently the means and standard deviations of the imputed values differed as well. The correlations between imputations obtained using the various methods were rather low. One exception is the correlation between the imputations based on EMI and RI. This correlation exceeds .90 for biology, history, and French. Imputations based on SRI and DA display the lowest correlations with imputations obtained using the other methods. This is a consequence of the fact that both methods involve drawing imputations from a predictive distribution. The amount of variance among imputed values was found to be unrelated to the number of missing grades, sex, or cohort.

In Study 2, in the regression analysis using unweighed grades as predictors, prediction based on grades that included imputations by SMS, EMI, and RI outperformed prediction based on grades that included imputations based on GPA. This was found to be the case in both cohorts. In the regression analysis with weighed grades, the largest multiple correlation was obtained with imputations based on GPA. However, in cross-validation, this correlation showed the largest drop. Other techniques showed more consistent results. The largest correlations were obtained using RI, EMI, SRI, and SMS. However, overall the differences between the methods were small.

A number of missing data techniques provided better predictions than those obtained using GPA. This may be the result of the fact that these methods utilized additional information (GPA is limited to information provided by the student's available grades). RI, EMI, and SRI utilize information relating to the relationship between subjects. It is remarkable that the SMS imputation of an observed subject mean produced better predictions than the GPA imputation, which is based on a personal mean. As an imputation technique SMS does not satisfy any of the four imputation principles. The available mean grade on a given subject appears to provide more information than the student's observed grades on other subjects. This is remarkable because the imputation based on a mean value does not differentiate between students who have missing grades.

It is surprising that results obtained with DA are disappointing in both prediction analyses. DA, at least in theory, is superior in that it uses a predictive distribution to draw imputations, it utilizes all available grades, and it takes into account the uncertainty due to missing data. In the comments to Study 2, it is noted that in carrying out DA, the criterion was not included in the imputation model. This may have resulted in lower correlations between predicted values and the criterion in the regression analyses. It is noted, however, that the differences between the techniques are rather small. No single method emerged as really superior.

From a practical point of view, the replacement of missing grades by imputed values is very convenient and will generally result in better prediction. As we saw above, even the method of GPA can be used to improve predictions if unavailable grades are explicitly replaced and regression analysis is performed (however, cross-validation showed that these predictions turned out to be somewhat unstable). A completed data file (in which missing data are replaced by imputations) allows one to carry out analyses in which a subset of subjects can be selected as predictors. One's interest may be limited to a certain selection of subjects, for example, science subjects. Imputation based on all available subjects is likely to increase the accuracy of the imputations within the subset of subjects. The completed grades file also enables one to carry out analyses in which the grades feature not as predictors, but as dependents. A single subject or a combination of subjects may feature as the criteria in such analyses.

In summary, we have found large differences between the various methods both in their theoretical underpinnings and in the actual imputed values that these methods produce. Of the methods considered, DA and SRI appear to be superior. In the two prediction studies reported above, only small differences between the methods were observed. However, some methods, notably RI and SRI, were found to produce more consistent predictions than other methods. One clear message that does emerge from the present results is that GPA should not be considered as a default missing grade technique. On both theoretical and empirical grounds, other techniques should be preferred.

Notes

We thank Conor V. Dolan and the reviewers for their valuable comments.

[1]For MI methods DA and SRI, results of the MI inferences for the means of biology, French, and history, such as estimated fractions of missing information and standard errors of the estimates, are available from the first author.

[2]A full description of the problems concerning the application of RI, SRI, and EMI in the MVA module in SPSS 8.0.0 is available from the first author.

[3]Other results of these MI inferences, such as estimated fractions of missing information and between- and within-imputation variance, and a discussion of these MI inferences are available from the first author.

References

Beaton, A. E., & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15*, 9–38.

Betts, J. R., & Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of Human Resources, 34*, 268–293.

Bretz, R. D. (1989). College grade point average as a predictor of adult success: A meta analytic review and some additional evidence. *Public Personnel Management, 18*, 11–22.

Britton, B. K., & Tesser, A. (1991). Effects of time management practices on college grades. *Journal of Educational Psychology, 83*, 405–410.

Cariaga-Lo, L. D., Enarson, C. E., Crandall, S. J., Zaccaro, D. J., & Richards, B. F. (1997). Cognitive and noncognitive predictors of academic difficulty and attrition. *Academic Medicine, 72*, S69–S71.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Efron, B., & Tibshirani, R. (1995). *Cross validation and the bootstrap: Simulating the error rate of a prediction rule* (Tech. Rep. No. 176). Stanford, CA: Stanford University, Division of Biostatistics.

Goldberg, E. L., & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validation generalization approach. *Educational and Psychological Measurement, 52*, 1019–1027.

House, J. D., & Johnson, J. J. (1992). Predictive validity of graduate record examination scores and undergraduate grades for length of time to completion of degree. *Psychological Reports, 71*, 1019–1022.

Huisman, J.M.E. (1999). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden, NL: DSWO Press.

Huisman, J.M.E., & Molenaar, I. W. (2001). Imputation of missing scale data with item response models. In A. Boomsma, M. van Duin, & T. Snijders (Eds.), *Essays on item response theory* (pp. 221–244). New York: Springer Verlag.

Linn, R. L., & Hastings, C. N. (1984). A meta analysis of the validity of predictors of performance in law school. *Journal of Educational Measurement, 21*, 245–259.

Liou, M. (1998). Establishing score comparability in heterogeneous populations. *Statistica Sinica, 8*, 669–690.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics, 6*, 287–301.

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

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

Little, R.J.A., & Rubin, D. B. (1990). The analysis of social science data with missing values. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 374–409). Newbury Park, CA: Sage Publications.

Little, R.J.A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (p. 39–75). New York: Plenum Press.

Nilsson, J. E. (1995). The GRE and the GMAT: A comparison of their correlations to GGPA. *Educational and Psychological Measurement, 55*, 637–640.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.

Schafer, J. L. (1998). *NORM: Multiple imputation of incomplete multivariate data under a normal model* [Software for S-PLUS 4.0 for Windows, available from http://www.stat.psu.edu/~jls/misoftwa.html].

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

Stricker, L. J. (1991). Current validity of 1975 and 1985 SATs: Implications for validity trends since the mid 1970s. *Journal of Educational Measurement, 28*, 93–98.

Stricker, L. J., Rock, D. A., & Burton, N. W. (1996). Using the SAT and high school record in academic guidance. *Educational and Psychological Measurement, 56*, 626–641.

Stricker, L. J., Rock, D. A., Burton, N. W., Muraki, E., & Jirele, J. J. (1994). Adjusting college grade point average criteria for variances in grading standards: A comparison of methods. *Journal of Applied Psychology, 79*, 178–183.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*, 528–550.

Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement, 55*, 177–185.

Young, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement, 27*, 175–186.

Young, J. W. (1993). Grade adjustment methods. *Review of Educational Research, 63*, 151–165.

Authors

NIELS SMITS is a doctoral student, Scholar Project: Schooling, Labor Market, and Economic Development, Faculty of Economics, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam; nielss@fee.uva.nl. His specializations include educational research methodology and missing data analysis.

GIDEON J. MELLENBERGH is a professor, University of Amsterdam, Department of Psychological Methods, Roetersstraat 15, 1018 WB Amsterdam. His specializations include psychological methods and psychometrics.

HARRIE C. M. VORST is a data manager and assistant professor, University of Amsterdam, Department of Psychological Methods, Roetersstraat 15, 1018 WB Amsterdam. His specializations include test construction methods.