Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: http://www.tandfonline.com/loi/hmbr20

A Multilevel Multivariate Analysis of Academic Performances in College Based on NCAA Student-Athletes

John J. McArdle , Thomas S. Paskus & Steven M. Boker

To cite this article: John J. McArdle , Thomas S. Paskus & Steven M. Boker (2013) A Multilevel Multivariate Analysis of Academic Performances in College Based on NCAA Student-Athletes, Multivariate Behavioral Research, 48:1, 57-95, DOI: 10.1080/00273171.2012.715836

To link to this article: https://doi.org/10.1080/00273171.2012.715836

Published online: 29 Mar 2013.

Multivariate Behavioral Research, 48:57–95, 2013

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273171.2012.715836

Routledge
Taylor & Francis Group

A Multilevel Multivariate Analysis of Academic Performances in College Based on NCAA Student-Athletes

John J. McArdle
Department of Psychology, University of Southern California, Los Angeles

Thomas S. Paskus
Research Department, National Collegiate Athletic Association

Steven M. Boker
Department of Psychology, University of Virginia

This is an application of contemporary multilevel regression modeling to the prediction of academic performances of 1st-year college students. At a first level of analysis, the data come from $N > 16{,}000$ students who were college freshman in 1994–1995 and who were also participants in high-level college athletics. At a second level of analysis, the student data were related to the different characteristics of the $C = 267$ colleges in Division I of the NCAA. The analyses presented here initially focus on the prediction of freshman GPA from a variety of high school academic variables. The models used are standard multilevel regression models, but we examine nonlinear prediction within these multilevel models, and additional outcome variables are considered. The multilevel results show that (a) high school grades are the best available predictors of freshman college grades, (b) the ACT and SAT test scores are the next best predictors available, (c) the number of high school core units taken does not add to this prediction but does predict credits attained, (d) college graduation rate has a second-level effect of a small negative

Correspondence concerning this article should be addressed to John J. McArdle, Department of Psychology, University of Southern California, Los Angeles, SGM 501, 3620 South McClintock Avenue, Los Angeles, CA 90089-1061. E-mail: jmcardle@usc.edu

outcome on the average grades, and (e) nonlinear models indicate stronger effects for students at higher levels of the academic variables. These results show that standard multilevel models are practically useful for standard validation studies. Some difficulties were found with more advanced uses and interpretations of these techniques, and these problems lead to suggestions for further research.

Over the past few decades, there have been many research studies on the predictive validity of high school academic performance on college academic performances (for prior reviews, see Astin, 1971; Crouse & Trusheim, 1988; Pascarella & Terenzini, 1991; Willingham, Lewis, Morgan, & Ramist, 1990). This is practically important multivariate research because many persons, many schools, many educational systems, and many societies have used these validation studies to create and justify policies for entrance into higher education. Although the important policy issues are far reaching and well beyond the scope of the current study, these prior validation studies have proven to be an essential first step in providing a rational basis for policy debates going at the current time. The current statistical study, even though it was largely completed more than 10 years ago, extends these prior validation studies by using a few contemporary statistical models using national transcript data with the hope to better understand the 1st-year academic performance of college students who also participate in high-level athletic competition.

In most prior research studies, multiple linear regression analyses have been used for prediction purposes, and these results are widely known, discussed, and criticized (Aitkin & Longford, 1986; Beatty, Greenwood, & Linn, 1999; Crouse & Trusheim, 1988; Willingham et al., 1990). In many studies of college grades the natural first set of predictors is based on high school grades, and the second predictor set is based on some nationally standardized measurement, such as the American College Testing Program (ACT; 1995) or the SAT (Scholastic Aptitude Test; see College Entrance Examination Board [CEEB], 1995). Because these two predictors are often correlated, a multiple linear regression model is used to separate the prediction of college grade point average (GPA) into two independent components. The costs and benefits of including these additional test scores into this predictive model equation, especially the consequences for selection of minority groups, were studied in many ways by many researchers (e.g., Beatty et al., 1999; Bowen & Bok, 1998; Crouse & Trusheim, 1988; McArdle, 1998; National Collegiate Athletic Association [NCAA], 1992; Willingham et al., 1990).

One key methodological aspect of these critiques comes from the fact that the high school students and the college students are "nested" within a particular high school or college. Many research studies using data collected from a wide variety of different colleges are often thought to lead to a more repeatable, robust, and generalizable prediction equation (e.g., Willingham et al., 1990). It is not that the high school level is unimportant, but a more thorough examination of the college level is possible here. That is, although it is possible to examine both effects at the same time, we certainly can use the college-level data we have more effectively.

The nesting of "students within schools" was a key reason for the development of the recent class of models termed *multilevel models* (Wilks & Kempthorne, 1955, as reported by Kreft & de Leeuw, 1998; see also Aitkin & Longford, 1986; Bock, 1989; Bryk & Raudenbush, 1992; Cronbach & Webb, 1975; Laird & Ware, 1982; McArdle & Hamagami, 1996; Wong & Mason, 1985, 1991). Previous research has shown how the use of multilevel modeling analyses, using previously developed software (e.g., HLM, MLn, Mplus, VARCL, SAS-MIXED, SAS-NLMIXED, etc.; see Longford, 1990), should be informative in the context of classical regression analyses and should increase the accuracy of the prediction of the student-level academic achievements.

In a previous study we used multilevel models for the prediction of graduation from college for students who were also high-profile athletes (e.g., McArdle & Hamagami, 1994, 1996). In that study we examined $N > 3{,}000$ students in $C > 65$ colleges, and the observed graduation rate of students was predicted from both individual-level high school academic variables (e.g., student core GPA and ACT or SAT scores) as well as college-level characteristics (e.g., college graduation rate and overall student body ACT or SAT scores). The key results included moderate and equal-size effects for both high school GPA and ACT/SAT, and this prediction was improved substantially by the inclusion of the college graduation rate (but not the overall test scores). Other studies using the same data examined the potential selection and measurement biases apparent in these data (McArdle, 1998). In a related methodological presentation we showed how standard multilevel models could be directly merged together with multiple group structural equation concepts (using software such as LISREL or Mx) to accurately display multilevel results as path diagrams (later see Figure 2) and allow for unusual multivariate options (e.g., latent growth models, common factor models; McArdle & Hamagami, 1996). More up-to-date work on the statistical basis of multilevel models is presented by both Hox (2002) and Snijders & Bosker (2012) and it is seems quite consistent with the analyses summarized here.

We expand upon this multicollege strategy in the current research using new data. The data used here include more than 16,000 high school students who entered colleges during the years 1994 and 1995 obtained from more than 260 colleges who currently participate in the highest level of athletic competition of the National Collegiate Athletic Association (i.e., NCAA Division I). We focus on methodological issues expanding the classical prediction models of the academic performances of 1st-year college students. We outline the new multi-level models to be used (i.e., college-level clustering) using a common structural

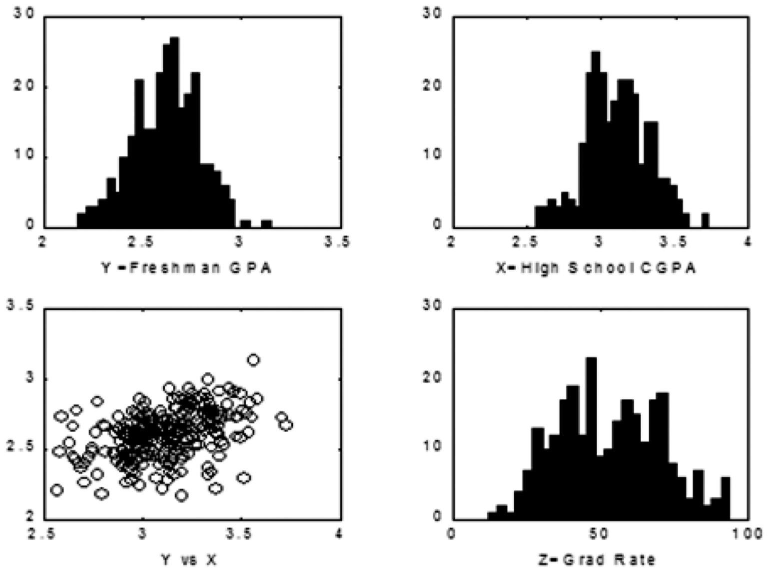


FIGURE 1 Individual and college level data for 1994–1995 NCAA freshman student-athletes.

equation notation (see Littell, Miliken, Stoup, & Wolfinger, 1996, p. 492; see also Laird & Ware, 1982), and we use these new models to make predictions that take into account the differences between colleges attended. The modeling results are presented in a sequence where we try to build up information about the student-level and college-level variables, apply the equations to several outcomes, and examine potential nonlinear relations. We summarize the results obtained and we make suggestions for future research on this topic.

METHODS

Participants

Starting in the 1993–1994 academic year, all high school students who wished to participate in an NCAA college sports program were evaluated for academic eligibility by the NCAA Initial Eligibility Clearinghouse (IEC). The same basic process continues in place today. The eligibility certification process at that time included (a) the student submission of a "Student Release Form"; (b) the IEC collection of a complete high school transcript for the student as well as national test score information from administrations of the SAT or ACT; and

(c) the IEC evaluation of minimum grades (2.0), minimum test scores (SAT = 820 or ACT = 17 in current 2012 units), and the minimum number of credit units of predefined high school "core courses" (11–13 courses with an approved curriculum in English, Math, Science, Social Studies, etc.). Students who met these minimum academic requirements were allowed to play in any NCAA sport in the 1st year of college. The same basic process is used today although the specifics of the required standards has changed somewhat (see Paskus, 2012; Petr & McArdle, 2012).

Starting in 1994 the Academic Performance Census (APC) survey was sent to NCAA representatives at 285 Division I colleges. Soon afterward, in 1995, this contact person was asked to fill out information on "all freshmen receiving any form of athletically related aid." These survey data on all student-athletes were then merged with high school information previously obtained from the IEC. Additional information was obtained by collating this individual-level information with a set of college characteristics obtained from the College Board (CEEB, 1995) computerized files and from the *NCAA Graduation Rates Disclosure Data* (NCAA, 1998). These publicly available files were then merged with the student-level information to create a multilevel data structure.

To create a sample that was comparable to our previous research (e.g., McArdle & Hamagami, 1994, 1996), we limited our analyses to the study of those student-athletes who attended college as freshmen starting in 1994 and who were either Black (African American) or White (Caucasian Non-Hispanic). Some Division I schools did not offer athletically related aid (e.g., the Ivy League, the Patriot League, and the military academies) so data on these students were not available. Also, 13 colleges reported having fewer than 10 student-athletes enrolled, and these data were not used. In sum, the focal data set used here includes N = 16,348 student-athletes at C = 267 different colleges; 43% reported they were female and 23% reported they were Black (see Figure 1).

Variables Measured

The NCAA longitudinal records include information about each student's (a) high school academic profile (e.g., in HS attended, the NCAA measured HS GPA, classes taken, ACT and/or SAT scores, and a variety of other indices), (b) multiple outcomes in 1st year of college (e.g., GPA, credit hours, athletic eligibility, etc.), (c) numerous demographic features of both the student (e.g., gender, ethnicity, family income, etc.) and (d) the schools they attended (high school and college). These IEC high school data appear to show both grades and test scores for prospective student-athletes are well above the national averages at that time (cf. ACT, 1995; CEEB, 1995). However, because the IEC deals with eligibility decisions, the scores available are the "highest scores" obtained from multiple repeated measurements of the national test scores, the ACT, or the SAT. These scores will necessarily have a different distribution than traditional scores—they will be higher than the "averages" typically used in other validation research, and the available test score data may also be restricted in range due to the use of a test score minimum in the initial eligibility process (i.e., for these persons, test scores of 820 approximately represented a national z score = −1; see McArdle, 1998).

From this multivariate array of data, only a few variables were selected for presentation. Table 1a presents a description of three college academic outcome variables: *Freshman Grade Point Average* (F-GPA with mean of 2.62), the number of *Freshman Credit Hours* taken (F-CREDITS with a mean of 28.5), and the number of *Freshman Quality Points* achieved (F-QUALPTS defined as the product of F-GPA and F-CREDITS, with a mean of 76.3). Table 1b describes three high school variables used in eligibility decisions: the high school GPA in core courses (optimal HS-CGPA with mean of 3.12), the highest ACT and/or SAT score as reported and converted to a common scale (HS-TEST with mean 1015 in SAT units), and the number of high school core course units taken (HS-UNITS with a mean of 16.3). Table 1c describes three selected college characteristics that have been used in prior research: the college's graduation rate in 1995 (COL-GRAT); the designation of a college type as a Public (−1) or Private (1) school (COL-PUPR); and the annual costs of tuition, fees, room, and board for the college (COL-COST) as estimated by the College Board (CEEB, 1995).

TABLE 1
A Brief Description of the Available Data on NCAA Freshman College Student-Athletes ($N \sim$ 16,500 Student-Athletes in 1994–1995, and N(c) $\sim$ 280 Division I Colleges)

Label	*Description*	*Scaling*	*Mean*	*(Stan. Dev.)*
1a: College Outcome Variables (Source = NCAA Academic Performance Survey of College Freshman)				
F-GPA	Freshman Grade Point Average	0–4	2.62	0.64
F-CREDIT	Freshman Course Credits	3–50	28.5	5.9
F-QUALITY	Freshman Quality Points	0–200	76.3	27.4
1b: High School Academic Variables (Source = NCAA Initial Eligibility Clearinghouse Data)				
HS-CGPA	High School Core GPA	1.3–5.0	3.12	0.59
HS-ASAT	High School ACT or SAT	490–1,520	1,015	143
HS-UNITS	High School Carnegie Units	4.0–25.5	16.3	2.1
1c: College Characteristics (Source = NCAA Graduation Rates Disclosure Survey and College Board Annual Data)				
COL-GRAT	College Graduation Rate	12–94	55.9%	Grad = 17.5%
COL-PUNP	Public versus Private	−1 or 1	73.7%	Public = 26.3%
COL-COST	Ann Tuition, Fees, Room & Board	$4.5–33.5 K	$13,900	$6,800
1d: Student Level Demographic Variables (Source = NCAA Initial Eligibility Clearinghouse Data)				
DEM_Sex	Male versus Female Student	43.3% Female	56.7% Male	
DEM_Ethnic	Black versus White Non-Hispanic	22.7% Black	77.3% White NH	
DEM-Sport	Male Revenue vs Non Revenue	25.6% MalRev	74.4% Other	

Data Description

Several features of the college-level data are illustrated in the plots of Figure 1. The first plot (labeled 1a) is a histogram of a well-known outcome variable-freshman grade point average (*F-GPA*)—but here we plot averages of the freshman class within each college. Although there is no doubt that GPA represents limited information (range 0–4), we do find some spread of scores, with the great

TABLE 2
Descriptive Statistics for 1994–1995 NCAA Academic Census Data

2a: Summary Statistics for All Student-Athletes (N ~ 16,500)

	Freshman			*High School*			*College*		
	F-GPA	*F-Credits*	*F-QualPts*	*CGPA*	*TEST*	*UNITS*	*GRAT*	*PUPR*	*COST*
Means	2.62	28.3	75.76	3.12	1,015	16.26	0.559	−0.471	13,911
Stan.Devs.	0.641	6.27	29.33	0.576	142.7	2.11	0.175	0.882	6,831
Minimum	0	0	0	1.32	490	4	0.12	−1	4,464
Maximum	4	50	200	4	1,520	25.5	0.939	+1	33,386
Correlations									
F-GPA	1								
F-Credits	0.429	1							
F-QualPts	0.884	0.764	1						
HS-CGPA	0.551	0.300	0.533	1					
HS-TEST	0.433	0.237	0.428	0.581	1				
HS-UNITS	0.220	0.163	0.212	0.364	0.372	1			
COL-GRAT	0.051	0.127	0.090	0.179	0.311	0.166	1		
COL-PUPR	0.057	0.076	0.081	0.084	0.171	0.101	0.334	1	
COL-COST	0.062	0.095	0.090	0.120	0.238	0.131	0.523	0.847	1

2b: Summary Statistics for All Colleges (N(c) ~ 268) Weighted by Sample Size

	Freshman			*High School*			*College*		
	F-GPA	*F-Credits*	*F-QualPts*	*CGPA*	*TEST*	*UNITS*	*GRAT*	*PUPR*	*COST*
Means	2.62	28.5	75.45	3.1	1,010	16.21	0.531	−0.392	13,911
Stan.Devs.	0.126	5.5	9.6	0.214	59.7	0.70	0.175	0.89	6,831
Minimum	2.18	21.2	56.4	2.31	857	13.98	0.135	−1	4,464
Maximum	3.14	36.9	104.8	3.66	1,231.5	18.23	0.935	+1	33,386
Correlations									
F-GPA	1								
F-Credits	0.257	1							
F-QualPts	0.755	0.813	1						
HS-CGPA	0.416	0.207	0.367	1					
HS-TEST	0.299	0.246	0.324	0.768	1				
HS-UNITS	0.094	0.274	0.237	0.303	0.672	1			
COL-GRAT	0.189	0.370	0.320	0.567	0.769	0.557	1		
COL-PUPR	0.229	0.262	0.295	0.262	0.441	0.358	0.333	1	
COL-COST	0.243	0.374	0.325	0.355	0.591	0.442	0.523	0.874	1

Note. The mean sample size per college is 61.2 Student-Athletes, with a range of 10 to 202; please see text for description of all acronyms.

majority between 2.0 (i.e., a "C" average) and a 3.0 (i.e., a "B" average). Figure 1b is a histogram of high school core GPA averages for the freshmen at each college, and here we find more spread at the upper levels. Figure 1c illustrates the positive relationship (r_{gg} = .26) between the college grades (Y) and the high school grades (X). For simplicity, these scores were averaged within each college, so there are only $C = 267$ scores plotted in Figure 1c. Figure 1d is a histogram of the graduation rates within each college, and a wide spread of scores is evident here (between 12% and 95%). Other relationships among these variables are described in later analyses.

More detailed statistical information on the variables used here is presented in Table 2. In Table 2a we present means, standard deviations, and correlations for nine variables using the student-level data (i.e., $N > 16{,}000$ students over all colleges). The three outcome variables have interpretable means (*F-GPA* m_g = 2.62, *F-CREDITS* m_c = 28.3, *F-QUALPTS* m_q = 75.8), standard deviations (s_g = 0.64, s_c = 6.3, s_q = 29.3), and moderate to high positive correlations (r_{gc} = .43, r_{gq} = .88, r_{cq} = .76). Because the third outcome variable is a direct product of the first two, some of this information is redundant. In Table 2b the second-level between-college means are weighted by the within-college sample size (with a mean of 61 students per college, ranging from 10 to 202). Here the same three outcome variables have a similar set of means (m_g = 2.62, m_c = 28.5, m_q = 75.5), a smaller set of deviations (s_g = 0.16, s_c = 5.5, s_q = 9.6), and a slightly different set of moderate to high positive correlations (r_{gc} = .26, r_{gq} = .76, r_{cq} = .81).

Multiple Linear Regression Models

The basic methodological issues of our analyses can be summarized in a series of equations. First, we used a multiple linear regression equation typically written for individual n as

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + e_n, \tag{1}$$

where Y is any set of outcomes (e.g., *F-GPA, F-CREDITS, F-QUALPTS*), the X variables are any pairs of predictors (e.g., *HS-CGPA, HS-TEST, HS-UNITS*), and we estimate an intercept (β_0) and regression coefficients (β_1 and β_2) and unobserved residual error scores (e). Typically we assume these errors are independently sampled, are independent of any other predictor scores, and normally distributed with mean zero and constant variance (symbolized as ϕ_e^2 here). Under these standard regression assumptions we can estimate the linear and additive effects of high school characteristics on the prediction of freshman outcomes.

It is well known that one of the threats to the validity and stability of regression equations comes from the use of samples of participants aggregated over different units or subgroups (i.e., colleges). A common approach for dealing with these issues is to create variables that characterize the different groups or persons and include these new variables in the prediction equation. For example, we might include variables representing different colleges in z-score form (e.g., either *COL-GRAT, COL-COST, COL-PUPR*) to adjust the predicted score for the mean differences in college graduation rates across colleges. Using similar multiple regression strategy, we can examine models where we include a product term as a third variable

$$Y_n = \beta_0 + \beta_1 X_n + \beta_2 Z_n + \beta_3 (X_n Z_n) + e_n \tag{2}$$

so the additional regression coefficient (β_3) reflects the differences among the slopes of the X (high school) relationship over groups (colleges) with different scores on z. The aforementioned equations can be interpreted in the traditional regression fashion (e.g., the partial product represents the interaction; see Cohen & Cohen, 1983) as long as the model variables and relationships meet the traditional assumptions of regression analysis (and see Aiken & West, 1991). This common approach for dealing with group differences in a regression model with interaction is depicted in symbolic form in the path diagram of Figure 2a (using the notation detailed in McArdle & Prescott, 1992; see also McArdle & Hamagami, 1996).

But there is a formal difference between the two sets of predictor variables used here. In Equations (1) and (2) each individual ($n = 1$ to N) has the possibility of a unique score on the X variables but the Z variable is a group score so the same score is assigned to each individual within the group (e.g., college). In Equation (2) we have assumed a subscript for each individual student even though the college-level variable should be understood to have been measured only on a specific group ($C < N$). This implies the standard errors of the coefficients will be understated and, correspondingly, our confidence in the point estimates (β) may be overstated. This statistical problem becomes more complicated when we include products of variables at different levels of aggregation (e.g., the product term in Equation (2)). There may be different relationships between the variables for subsets of groups, and this may alter the size or sign of the standard regression estimates (see Kish & Frankel, 1974; Lee, Forthofer, & Lorimor, 1989).

Variance Component and Random Coefficients Models

There are many ways to examine this fundamental problem of "nesting," "clustering," or "aggregation." A regression coefficient that ignores the group-level variation can be written as

$$\beta_{total} = \beta_{within\ group}(1 - \eta^2) + \beta_{between\ group}\eta^2, \text{ where}$$
$$\eta^2 = \phi^2_{between\ group}/(\phi^2_{between\ group} + \phi^2_{within\ group}), \tag{3}$$

with regression coefficients (β), variance terms (ϕ^2), and the well-known *intraclass correlation* (η^2; e.g., Kreft & de Leeuw, 1998; McArdle & Goldsmith, 1990). In this equation the total regression coefficient (β_{total}) is decomposed into a within-group regression ($\beta_{within\ group}$, e.g., for persons within a college) and a between-group regression ($\beta_{between\ group}$, e.g., between the mean scores of colleges) both weighted by the intraclass correlation (η^2). In this expression, the total regression coefficient is equivalent to the within-group regression coefficient when the intraclass coefficient is zero. In other cases the total regression coefficients (i.e., Equation (1)) will be a function of both (a) the size of the intraclass correlation and (b) any difference in the regression coefficients for between and within groups.

After identifying the between and within group regression differences there are many possible methods to correct for potential biases. One fundamental way to deal with this problem is to initially assume a separate regression model for each group and rewrite Equation (1) or (2) as

$$Y_n^{(g)} = \beta_0^{(g)} + \beta_1^{(g)} X_n^{(g)} + e_n^{(g)}, \text{ for } g = 1 \text{ to } G, \tag{4}$$

where the superscript "(g)" designates the individual groups (G), in this case we assume the specific college is a specific group, and the regression coefficients for the prediction of Y from X are allowed to vary between one group and another (i.e., over the Z variable). The parameters of an unrestricted model with many groups can be estimated using multiple runs of multiple regression. To estimate structural models with some restrictions on the across-group parameters (e.g., $\beta_1^{(1)} = \beta_1^{(2)} = \beta_1^{(G)}$), we can use standard structural equation modeling software with multiple group solutions (e.g., LISREL, Mx, RAMONA, AMOS, etc.; see McArdle, 2007; McArdle & Hamagami, 1996). The kind and number of invariance restrictions to be imposed is a general theoretical question, and this approach can prove to be quite practical if the number of groups is relatively small (e.g., $G < 10$). This well-known multiple group SEM approach to regression is depicted in the path diagram of Figure 2b.

Often, however, the number of groups is relatively large and the desired structural restrictions are of a relatively simple form. In these common cases, we can use an approach commonly referred to as a "variance components" or "random coefficients" model. Here we can rewrite the simple regression of Equation (4) so the regression coefficients for the prediction of grades are

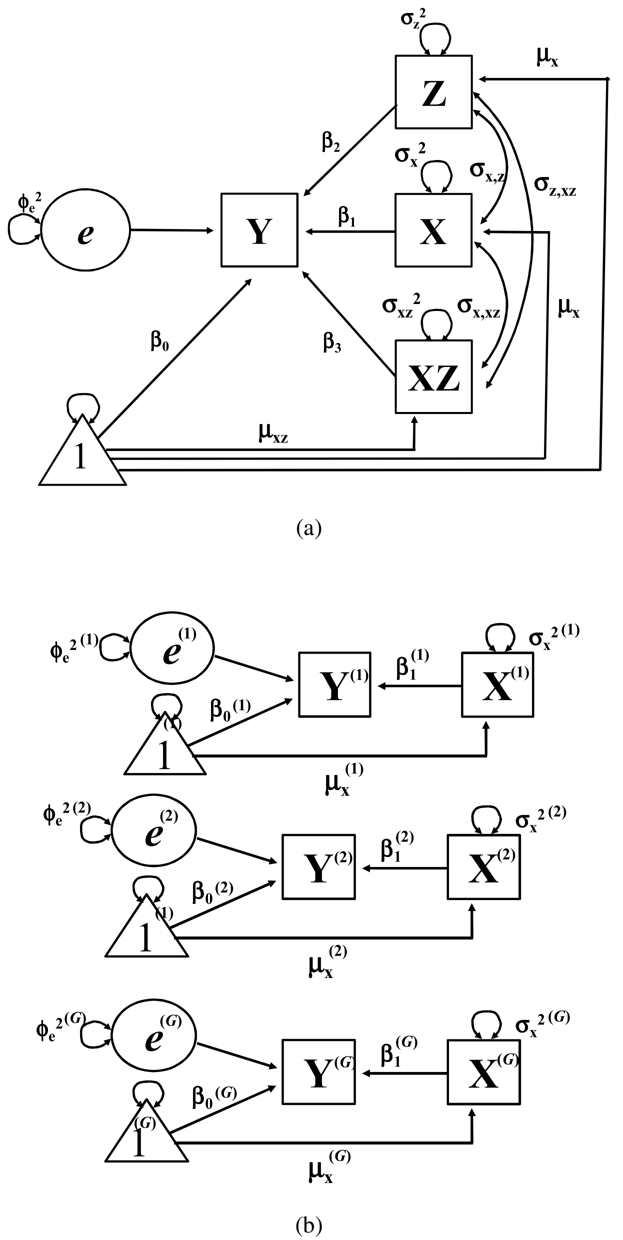


FIGURE 2 Alternative path diagrams of multilevel models (from McArdle & Hamagami, 1996). (a) Path diagram of standard regression model with interaction. (b) Groups separated in the MG-SEM regression. (c) Variance components in MG-SEM for one group. (d) Multilevel constraints in MG-SEM for one group. (*continued*)

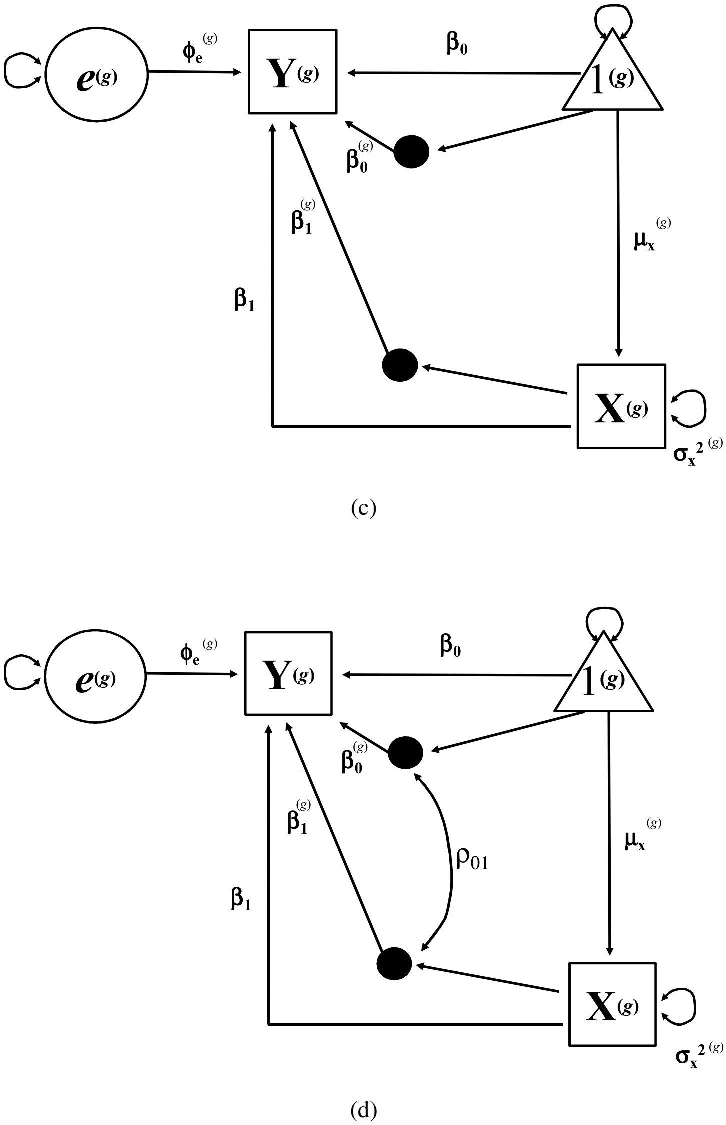


FIGURE 2 (*Continued*).

assumed to vary between one college and another by stating

$$
\begin{aligned}
Y_n^{(g)} &= \beta_0^{(g)} + \beta_1^{(g)} X_n^{(g)} + e_n, \text{ with} \\
\beta_0^{(g)} &= \beta_0 + d_0^{(g)} \text{ and} \\
\beta_1^{(g)} &= \beta_1 + d_1^{(g)},
\end{aligned} \tag{5}
$$

where the group variation is parameterized by a set of *fixed* means (β_0 and β_1) and *random* deviations or *disturbances* (d_0 and d_1) at a second level of aggregation. In standard *variance components* models we estimate all fixed terms as well as the variance of these second-level variance terms (ϕ_0^2 and ϕ_1^2). In the standard *random coefficients* model we additionally estimate the covariance of the second-level terms (ϕ_{01}). These multilevel equations can be accurately described by the multiple group structural model where the standard regression model is presented for each separate group but the model allows some sources of variation. A summary path diagram of the kind presented in Figure 2c can be used to generate statistical expectations (e.g., means, variances, and covariances) for any one group from the parameters estimated (for details, see Bryk & Raudenbush, 1992; McArdle & Hamagami, 1996).

The extent of potential bias due to nesting can be seen by rewriting Equation (5) as

$$
\begin{aligned}
Y_n^{(g)} &= (\beta_0 + d_0^{(g)}) + (\beta_1 + d_1^{(g)}) X_n^{(g)} + e_n, \\
&= [\beta_0 + \beta_1 X_n^{(g)} + e_n] + d_0^{(g)} + d_1^{(g)} X_n^{(g)},
\end{aligned} \tag{6}
$$

where, by substitution, the first (bracketed) part of this equation is seen as being identical to a standard regression model (e.g., Equation (1)]) but the remainder now includes two additional unobserved deviation terms (d_0 and $d_1 X$). If these latent second-level deviations are all zero, or if they are independent of all other variables, they can be subsumed into the first-level unobserved error term (e_n) and the coefficients of Equation (1) will be unbiased. However, to the degree these new terms reflect a lack of independence within the original residuals, the standard regression assumptions will fail, and a new set of estimates will be necessary. Furthermore, this inclusion of intraclass clustering does not seem to create any further biases even when clustering is not present, so we can routinely use this form of a nested regression model (Equation (5)) in all analyses.

Multilevel Linear Models

In analyses presented here additional group-specific information (e.g., college graduation rate) is used from a variety of different colleges. In order to examine these interactions in the context of the intraclass effects, we use a regression equation referred to as a multilevel (or hierarchical model) written in the form of

$$Y_n^{(g)} = \beta_0^{(g)} + \beta_1^{(g)} X_n^{(g)} + e_n, \text{ with}$$
$$\beta_0^{(g)} = \gamma_{00} + \gamma_{01} Z^{(g)} + d_0^{(g)} \text{ and} \tag{7}$$
$$\beta_1^{(g)} = \gamma_{10} + \gamma_{11} Z^{(g)} + d_1^{(g)},$$

where the individual groups have regression coefficients ($\beta_j^{(g)}$) that are allowed to vary between one school and another, and this variation is related to another set of fixed coefficients (γ_{jk}) and disturbances (d_j) for variable Z at the second level of aggregation (e.g., *COL-GRAT, COL-COST, COL-PUPR*). The intercepts in these equations are interpreted as means only when we force the Z variables to have zero means (and we could rewrite $\gamma_{00} = \beta_0$). In this form (i.e., when all Z variables have zero means) it is easier to see the potential bias by rewriting Equation (7) as

$$\begin{aligned} Y_n^{(g)} &= (\gamma_{00} + \gamma_{01} Z^{(g)} + d_0^{(g)}) + (\gamma_{10} + \gamma_{11} Z^{(g)} + d_1^{(g)}) X_n^{(g)} + e_n, \\ &= \beta_0 + \gamma_{01} Z^{(g)} + d_0^{(g)} + \beta_1 X_n^{(g)} + \gamma_{11} (Z^{(g)} X_n^{(g)}) + d_1^{(g)} X_n^{(g)} + e_n, \\ &= [\beta_0 + \beta_1 X_n^{(g)} + \gamma_{01} Z^{(g)} + \gamma_{11} (Z^{(g)} X_n^{(g)}) + e_n] + [d_0^{(g)} + d_1^{(g)} X_n^{(g)}], \end{aligned} \tag{8}$$

where we have reorganized (by brackets) the first five terms to distinguish the standard interaction model components (Equation (2)) from the two new unobserved variance components (d_j). Although the first five coefficients could be estimated using product variables in a standard regression model of Equation (2), ignoring the last two components can lead to biased estimates. The extent of this bias is typically a matter of empirical observation, is related to the intraclass correlation (Equation (3)), and can be directly estimated using available multilevel analysis software. In practice, there seems little reason not to try. A path diagram of this algebraic expression of the more complete multilevel model is presented for a single group (of many) in Figure 2d.

Multivariate Multilevel Models

The models fitted in the next section include different academic outcomes and more predictors at each level. One system of equations we examine has a form written as

$$
\begin{aligned}
Y_n^{(g)} &= \beta_0^{(g)} + \beta_1^{(g)} X_{1n}^{(g)} + \beta_2^{(g)} X_{2n}^{(g)} + \beta_3^{(g)} X_{3n}^{(g)} + e_n, \text{ with} \\
\beta_0^{(g)} &= \gamma_{00} + \gamma_{01} Z_1^{(g)} + \gamma_{02} Z_2^{(g)} + d_0^{(g)}, \\
\beta_1^{(g)} &= \gamma_{10} + \gamma_{11} Z_1^{(g)} + \gamma_{12} Z_2^{(g)} + d_1^{(g)}, \\
\beta_2^{(g)} &= \gamma_{20} + \gamma_{21} Z_1^{(g)} + \gamma_{22} Z_2^{(g)} + d_2^{(g)},
\end{aligned} \tag{9a}
$$

which, in matrix terms, can be expressed more simply as

$$\mathbf{Y}^{(g)} = (\boldsymbol{\gamma}\mathbf{Z}^{(g)} + \boldsymbol{d}^{(g)}) + \boldsymbol{e}, \tag{9b}$$

where $\mathbf{Y}^{(g)}$ is the $(N \times 1)$ data score matrix of all people for all groups; $\boldsymbol{e}$ is the $(N \times 1)$ error or residual score matrix of all people for all groups; $\boldsymbol{\gamma} = [\gamma_{00}\gamma_{01}\gamma_{02}, \gamma_{10}\gamma_{11}\gamma_{12}, \gamma_{20}\gamma_{21}\gamma_{22}, \gamma_{30}\gamma_{31}\gamma_{32}]$, the (4×3) second-level coefficient matrix; $\mathbf{Z}^{(g)} = [1^{(g)}, Z_1^{(g)}, Z_2^{(g)}]$, the $(3 \times N)$ second-level score matrix; and $\boldsymbol{d} = [d_0, d_1, d_2]$, the $(N \times 3)$ second-level disturbance or residual matrix. In this case we assume there are three measured variables at the first level and two measured variables at the second level. In this more complex system, the general multilevel interpretations are the same as described earlier: all first-level regression coefficients $(\beta_i^{(g)})$ can be interpreted to vary between groups, and these differences may be decomposed into direct effects by the second-level regression coefficients $(\gamma_{j\mathbf{k}})$ in matrix $\boldsymbol{\gamma}$.

It is also possible to create several alternative SEMs for the comparison of the prediction equations that use the available information from multiple outcome variables (see Goldstein, 1995; Hox, 2002; McArdle & Hamagami, 1996; McDonald, 1993, 1994; McDonald & Goldstein, 1988; Muthén, 1994). One SEM considered here is a classical linear measurement-structural model in the multilevel form of

$$
\begin{aligned}
Y_{1n}^{(g)} &= \nu_1^{(g)} + \lambda_1^{(g)} F_n^{(g)} + u_{1n}, \\
Y_{2n}^{(g)} &= \nu_2^{(g)} + \lambda_2^{(g)} F_n^{(g)} + u_{2n}, \\
Y_{3n}^{(g)} &= \nu_3^{(g)} + \lambda_3^{(g)} F_n^{(g)} + u_{3n}, \\
Y_{4n}^{(g)} &= \nu_4^{(g)} + \lambda_4^{(g)} F_n^{(g)} + u_{4n},
\end{aligned} \tag{10a}
$$

with

$$
\begin{aligned}
F_n^{(g)} &= \beta_0^{(g)} + \beta_1^{(g)} X_{1n}^{(g)} + \beta_2^{(g)} X_{2n}^{(g)} + e_n \text{ and} \\
\beta_0^{(g)} &= \gamma_{00} + \gamma_{01} Z_1^{(g)} + \gamma_{02} Z_2^{(g)} + d_0^{(g)}, \\
\beta_1^{(g)} &= \gamma_{10} + \gamma_{11} Z_1^{(g)} + \gamma_{12} Z_2^{(g)} + d_1^{(g)}, \\
\beta_2^{(g)} &= \gamma_{20} + \gamma_{21} Z_1^{(g)} + \gamma_{22} Z_2^{(g)} + d_2^{(g)},
\end{aligned}
\tag{10b}
$$

which, in matrix terms, can be expressed more simply as two expressions,

$$
\begin{aligned}
\mathbf{Y}^{(g)} &= \nu^{(g)} + \mathbf{\Lambda}^{(g)} \boldsymbol{F}^{(g)} + \boldsymbol{u}^{(g)}, \text{ with} \\
\boldsymbol{F}^{(g)} &= (\boldsymbol{\gamma} \mathbf{Z}^{(g)} + \boldsymbol{d}^{(g)}) + \boldsymbol{e},
\end{aligned}
\tag{10c}
$$

where $\mathbf{Y}^{(g)}$ is the $(N \times 4)$ data score matrix of all people for all groups; $\boldsymbol{e}$ is the $(N \times 4)$ error or residual score matrix of all people for all groups; $\boldsymbol{F}^{(g)}$ is an $(N \times 1)$ set of unobserved common factor scores for the observations; $\mathbf{\Lambda}$ is a (4×1) matrix of factor loadings; $\boldsymbol{u} = [u_1, u_2, u_3, u_4]$, the $(N \times 4)$ matrix of residual scores; and the multilevel structure is a decomposition of the common factor score with γ, (a 3×3) second-level coefficient matrix, multiplied by $\mathbf{Z}^{(g)}$, a (3×3) second-level score matrix, with disturbances $\boldsymbol{d}$ and factor score residuals $\boldsymbol{e}$. In this way, the measured Y variables are related (within each group) to a single latent variable $\boldsymbol{F}$ by a common factor model with variable intercepts (ν_i), factor loadings (λ_i), and uniqueness (u_i) as well as multilevel parameters (β, γ, ϕ^2) and disturbances ($\boldsymbol{d}$) for the common factor scores. In this kind of a multivariate model there are three different sources of random error (uniqueness, errors, and disturbances) and they are all assumed to be uncorrelated. Appropriate identification constraints follow from common factor theory (McDonald, 1993, 1994). In the analysis used here, the common factor score $\boldsymbol{F}$ is considered as a broad academic achievement, and this factor is assumed to carry the full multilevel structure of all high school $\mathbf{X}$ and second-level $\mathbf{Z}$ effects. We note that this model is similar to but not identical to the popular multilevel factor models where factor loadings (λ_i) are compared for a "between-within" multilevel structure (i.e., as in McDonald, 1993, 1994; Muthén, 1994).

Multilevel Computer Software

The computer software for multilevel regression analyses is widely available (for review, see Goldstein, 1995; Hox, 2002; Hox & Kreft, 1994; Kreft, de Leeuw,

& van der Leeden, 1995). The programs that can deal with these problems now include HLM (Bryk & Raudenbush, 1992); LIMDEP (Greene, 1998); MLn (Goldstein, 1995); MIXREG (Hedecker & Gibbons, 1994); VARCL (Longford, 1990); and the newest versions of PROC MIXED and PROC NLMIXED (Littell et al., 1996; Singer, 1998), LISREL 8.30 (Joreskog, Sorbom, du Toit, & du Toit, 1999), S-Plus (S-Plus, 1999), and Mplus (Muthén & Muthén, 2007). Each of these computer programs provides maximum likelihood estimates (MLE) and standard likelihood-based statistical information on goodness of fit—this can be indexed by the function ($f = -2LL$), or as the difference between two nested models (L^2), or as the ratio of the difference compared with its degrees of freedom ($LRR = L^2/\Delta df$) for a variety of multilevel models. Although there are some practical differences among these computer programs (e.g., MLn allows more than two levels, MIXREG is free, etc.), these computer programs are treated here as interchangeable. Results from the PROC MIXED and PROC NLMIXED algorithms (e.g., Littell et al., 1996; Singer, 1998) are presented here, and any further differences in computer programs are only discussed as needed.

The modeling approach we present here is similar to the approach we have used in other latent variable analyses (e.g., McArdle, 1994). However, there are numerous controversies specific to the statistical features of multilevel models, including (a) the lack of true nested models with Restricted MLE (REML but not MLE), (b) the downward biases of MLE estimates of variance components (but not with REML), (c) the appropriate choice of metric for the predictors, (d) the interpretation of explained variance, and (e) the appropriate strategy for model selection (e.g., Goldstein, 1995; Kreft, de Leeuw, & Aiken, 1995; Pinherio & Bates, 2000). In all models reported here we only present MLE solutions but we did not find any notable differences in REML. The outcome variables were retained in their original metric, but the predictor variables were all fitted in rescaled z-score form to eliminate some potential problems due to scaling, centering, and programming (e.g., Kreft, de Leeuw, & Aiken, 1995). We report the statistical significance or accuracy of each model parameter using a t value constructed from the simple ratio of the model parameter divided by the estimated standard error.

We also examine the fit of a model of interest (vs. a baseline model) by calculating two indices of percent reduction in error (PRE) documented by Snijders & Bosker (1994, 2012). Several explained variance-like indices can be written here as the ratios of the first-level error variance (ϕ_e^2), the intercept variance (ϕ_0^2), and the associated sample sizes ($n^{(g)}$). Additional parameters may also be needed in higher order model (e.g., ϕ_1^2, ϕ_{01}). These percentages are considered unbiased estimates of the "modeled variance" at the first level (PRE_1) and at the second level (PRE_2) for a correctly specified population, so we use these as one indicator of the impact of specific predictor variables. Similarly, we assume any negative values of these ratios, possible in any latent variable model, are a warning sign of an improperly specified model in some way or the other (for examples, see Kreft & de Leeuw, 1998; Snijders & Bosker, 1994).

In recent multilevel statistical theory, it is worthwhile to add that any multivariate model (Equation (10)) can also be reparameterized and fit using the available mixed-model software (Goldstein, 1995). The observed outcome data can be lined up into an augmented single vector $\mathbf{Y}' = [Y_1' : Y_2' : Y_3' : \ldots Y_M']$ and, by making appropriate provisions for the expectations of the model for each variable (using dummy variables and nested levels), the standard single outcome optimization can treat this multiple variable model as a single equation. This theory is a bit cumbersome to implement, but it makes it possible to construct a simultaneous test of proportionality of multiple ($M \geq 2$) variables, estimating the proportions (λ_j, λ_k, etc.) as factor loadings and obtaining an overall fit to test the hypothesis of a general common factor (e.g., Equation (10); see McArdle & Woodcock, 1997). In practice, however, these kinds of multiple outcome variable models can lead to new complications and new computer software or modeling techniques may be required.

RESULTS

Several related multilevel models have been fit to the data described in Table 2, and only a selected set of models is presented here. This presentation of results begins with some initial models for freshman grades (Table 3), includes additional predictor variables (Table 4), moves to multiple outcome and college-level variables (Table 5), examines potential nonlinearity in these relationships (Table 6), and concludes with some comments on our modeling strategy.

Preliminary "Two-Stage" Results

To begin we first calculated a simple regression model with HS-CGPA as a predictor of F-GPA in the total sample (Table 1) and also using the sample-size weighted means of the colleges (Table 2). These simple analyses yielded model parameters that were similar for the intercept ($\beta_0 = 2.616$ vs. 2.623) but different for slopes ($\beta_1 = .379$ vs. .191) and for the explained variance ($R^2 = .350$ vs. .175). This total regression exhibits relatively high prediction accuracy for these validation models (cf. Crouse & Trusheim, 1988; Willingham et al., 1990). However, these slope differences make it complex to judge the impact of the omission of the intraclass college differences, and this initially suggests that a more complete multilevel analysis may be useful.

Some expected results from the general multilevel approach are first described using a simple "two-stage" regression approach based on a model with *HS-CGPA*

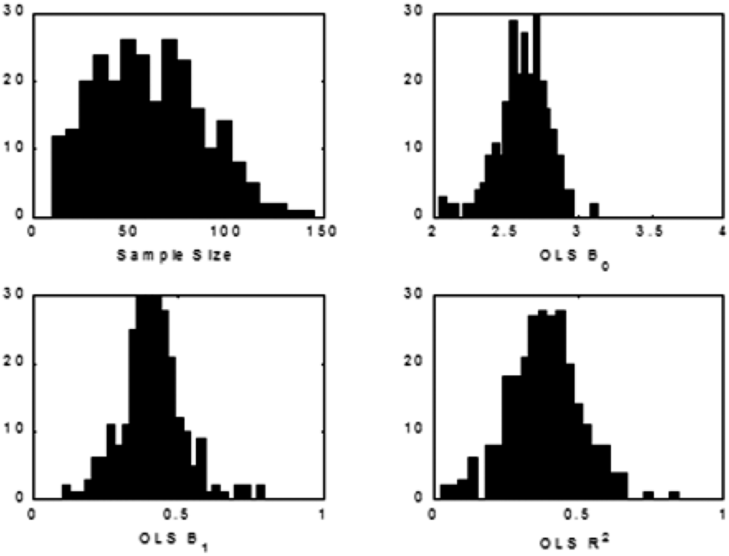


FIGURE 3 College level parameters from $C = 267$ linear regression models of freshman GPA.

as a predictor of *F-GPA* (as in Equation (4)). In a first stage we fitted a separate regression equation to each of the $C = 267$ colleges and examined the data for outliers. In a second stage we extracted the parameters from these models and examined their distribution characteristics. Although these standard Ordinary Least Squares (OLS) models have limited statistical features (e.g., they are not "Bayes" corrected estimates; see Braun, 1989), they are fairly easy to calculate and useful in initial assessments of the models and data.

Figure 3a is a histogram showing the samples sizes used within each college (ranging from 10 to 202). The other plots are displays of the frequency distributions of parameters obtained from fitting separate OLS regressions of HS-CGPA to data within college. In Figure 3b we plot the OLS intercept terms ($\beta_0^{(c)}$) obtained, in Figure 3c we plot the OLS slope terms ($\beta_1^{(c)}$), and in Figure 3d we plot the OLS explained variances ($R^{2(c)}$). It appears that there is notable variation in each distribution, there is at least one outlier in all cases, but each distribution is also symmetric and nearly normally distributed (Q-Q plots may be more informative about the latter concern).

Figure 4 is a display of the predicted value equation for each separate college using these estimates with all predictions in the original units. In these models the regression was fitted so that it would be centered at the average score of the entire distribution (i.e., at HS-CGPA = 3.12) and the resulting intercept is the

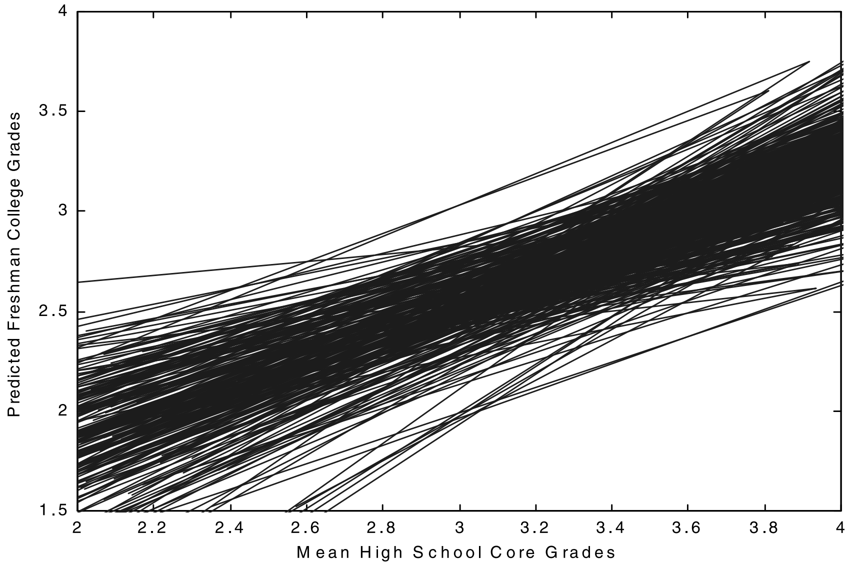


FIGURE 4 Separate college OLS regression functions using a simple linear model.

predicted score for an average student (i.e., $\beta_0 = 2.62$). (Each regression line in this plot is displayed for a two-standard-deviation range around the college mean score.) As this picture shows, this collection of lines has a notable difference in the intercepts across colleges but remarkably similar slopes of increasing high school grades predicting increasing college grades. This intercept variation can be studied in more detail in the multilevel models to follow, but the slope variation is expected to be small in these analyses.

Initial Multilevel Models of Freshman GPA

The numerical results of six initial multilevel models fitted to the data for freshman grades (F-GPA) are presented in Table 3. The first model (M_1) is a baseline model with three parameters: an intercept mean ($\beta_0 = 2.62$), an intercept variance ($\phi_0^2 = .020$), and an error variance ($\phi_e^2 = .391$). This model was fitted with MLE and yields a baseline likelihood (of $-2LL = 31401$) to be used in further model comparisons. The resulting explained variance of the raw scores here depends only on the intracollege variance, so here the $\eta^2 = .049$. Although this intraclass correlation is relatively small, it is not thought of as zero (e.g., the Wald-Ratio = 8), so we expect some reduction in bias by taking college differences in average freshman grades into account in additional equations.

TABLE 3
A Sequence of Multilevel Models for the Prediction of Freshman Grades in College

Parameters (Symbol)	M_1 *Baseline*	M_2 *Var. Comp*	M_3 *Random*	M_4 *Multilevel*	M_5 *Quadratic*	M_6 *Segments*
Fixed Terms at Level 1						
Intercept (γ_{00})	2.62 (>99)	2.62 (>99)	2.62 (>99)	2.62 (>99)	2.57 (>99)	2.53 (>99)
HS_CGPA1 (γ_{10})		0.399 (71)	0.399 (71)	0.400 (71)	0.404 (73)	0.288 (28)
HS_CGPA2 (γ_{20})					0.057 (14)	0.514 (59)
Fixed Terms at Level 2						
COL-GRAT > Intercept (γ_{01})				−0.034 (3)	−0.022 (2)	−0.021 (2)
COL-GRAT > HS_CGPA1 (γ_{11})				−0.003 (1)	−0.020 (4)	−0.006 (1)
COL_GRAT > HS_CGPA2 (γ_{21})					−0.010 (2)	−0.032 (4)
Random Terms at Level 1						
Variance{Error} (ϕ_e^2)	0.391 (90)	0.242 (90)	0.242 (89)	0.242 (89)	0.238 (89)	0.239 (88)
Random Terms at Level 2						
Variance{Intercept} (ϕ_0^2)	0.020 (8)	0.022 (9)	0.023 (9)	0.022 (9)	0.024 (8)	0.025 (7)
Variance{HS_CGPA1} (ϕ_1^2)		0.004 (5)	0.004 (5)	0.004 (5)	0.003 (4)	0.007 (3)
Variance{HS_CGPA2} (ϕ_2^2)					0.001 (2)	0.001 (1)
Covar{Inter,CGPA1} (ϕ_{01})			−0.004 (4)	−0.004 (4)	−0.002 (2)	0.002 (1)
Covar{Inter,CGPA2} (ϕ_{02})					−0.002 (2)	−0.004 (2)
Covar{CGPA1,CGPA2} (ϕ_{12})					∼0 (0)	0.002 (2)
Goodness-of-Fit						
Explained{F_GPA}(R^2)	.049	.412	.412	.412	.421	.418
Log Likelihood (−2LL)	31,401	23,858	23,841	23,841	23,626	23,613
Diff. −2LL/Diff. dfs (LRR)	0/0	7,543/2	7,560/4	7,560/6	7,775/12	7,788/12

Note. (1) All models were fitted here using SAS PROC MIXED with $N = 16{,}348$, and $C = 267$; (2) MLE parameter estimates listed with approximate $t = p/se(p)$ in parentheses; (3) Comparable results found using other programs (e.g., MIXREG, VARCL, LIMDEP, and LISREL 8.13). (4) For F_GPA Observed Mean = 2.617 with Variance = .411. All predictor variables scaled to Z-score form. (5) In Polynomial Model 5, CGPA1 = CGPA and CGPA2 = CGPA1*CGPA1; (6) In Segmented Model 5, CGPA1 = CGPA IFF CGPA < 0 and CGPA2 = CGPA IFF CGPA > 0.

The second model (M_2) is a variance components model (see Equation (5)) where we add the high school core grades (*HS-CGPA*) as a predictor of the freshman college grades (*F-GPA*). This model yields a similar intercept ($\beta_0 = 2.62$), a strong positive slope ($\beta_1 = .399$) reflecting a shift in freshman grades for each standard deviation of high school core grades (or .58 in raw score *HS-CGPA* terms), a slightly larger variance component for the intercept ($\phi_0^2 = .022$), and a small but nonzero variance component for the college-level slopes ($\phi_1^2 = .004$). The third model (M_3) adds a covariance ($\phi_{01} = -.004$) between the two second-order residual components, and the result remains same.

The fourth model (M_4) is a multilevel model (e.g., Equation (8)) with high school grades at the first-level prediction and college graduation rate at the second-level prediction, and the resulting model can be written in full multilevel form as

$$\begin{aligned} \text{F-GPA}_n^{(c)} &= \beta_0^{(c)} + \beta_1^{(c)}\text{HS-CGPA}_n^{(c)} + \underline{.492}e_n \\ \beta_0^{(c)} &= \underline{2.62} + \underline{-.034}\ \text{COL-GRAT}^{(c)} + \underline{.148}d_0^{(c)} \text{ and} \\ \beta_1^{(c)} &= \underline{.400} + -.003\ \text{COL-GRAT}^{(c)} + \underline{.063}d_1^{(c)}, \end{aligned} \tag{11}$$

with the accurate coefficients underlined (i.e., those that are significantly different from zero). In this expression the residual terms are rewritten in a standardized form (e) so the variance components can all be expressed as standard deviations (e.g., $\phi_e^2 = .242$, so $\phi_e = .492$) and included directly in the systems of equations. The results yield a similar intercept ($\beta_0 = 2.62$), a strong positive slope for high school grades ($\beta_1 = .400$), a small but negative effect for the second-level regression of the intercept on the college graduation ($\gamma_{01} = -.034$), and a nonsignificant second-level regression of the slope on the college graduation ($\gamma_{11} = -.003$). Compared with the baseline model, the first-level error variance has become smaller ($\phi_e^2 = .242$), resulting in an increased model variance ($PRE_1 > .36$), with no second-level variance gain ($PRE_2 < -.02$). This initial multilevel model is depicted in a path diagram in Figure 5 using the techniques described earlier (from McArdle & Hamagami, 1996).

In the fifth model (M_5) we examine the possibility of nonlinear relationships by adding a squared term (HS-CGPA*HS-CGPA) and a second coefficient (β_2) to the first-level prediction (as in Equation (2)). This model permits a quadratic polynomial for high school grades with variation in both components. The inclusion of these terms leads to a notable increase in the goodness of fit ($L^2 = 215$ on $\Delta df = 6$) and only a small increase in the model variance ($PRE_1 > .36$ but $PRE_2 < -.05$). The second-level prediction of these components from the college graduation rate is similar to the previous linear model (M_4). The fixed parameters of this quadratic model are all nonzero (i.e., $t > 2$) and the predicted average function is concave upward with a minimum at approximately HS-

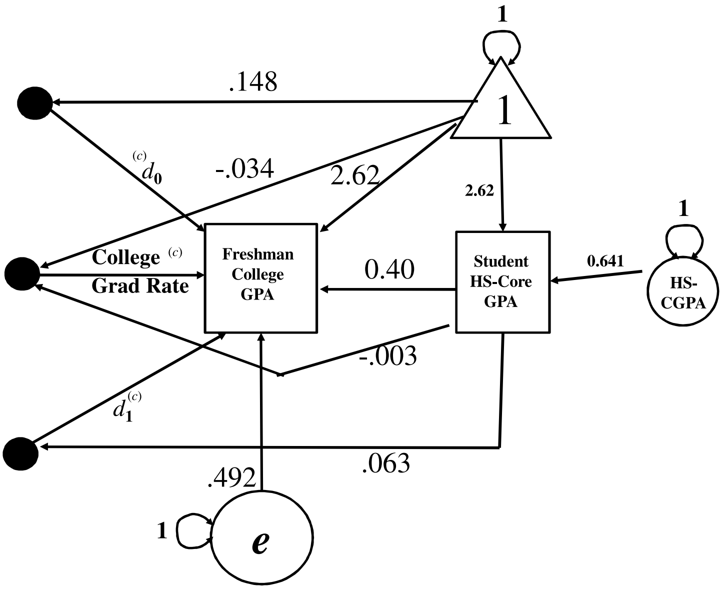
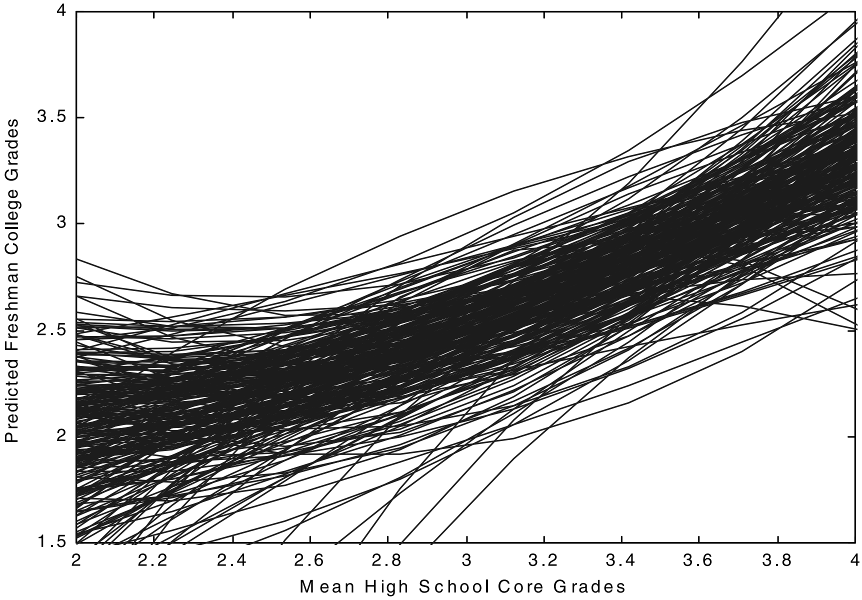


FIGURE 5 Initial multilevel results (Eq. [11]) as a path diagram.

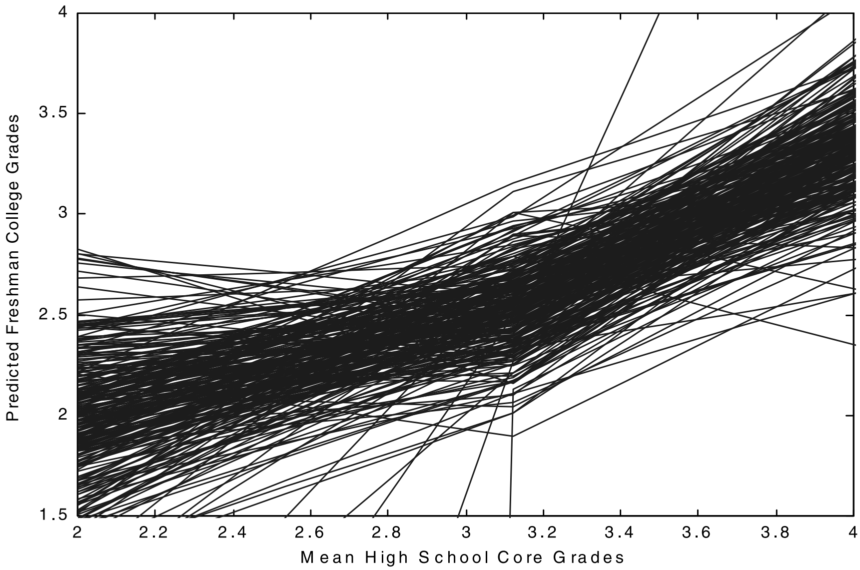
CGPA = 2.85. For descriptive purpose, two-stage OLS estimates of the quadratic model (M_5) are presented for separate colleges in Figure 6a. Although there may be some outliers in this plot, the general tendency shows prediction models with stronger effects of high school grades on freshman college grades after the minimum point.

In the sixth model (M_6) we approach this problem of nonlinearity in a different way: we fit a nonlinear model as two *linear segments* joined at the average score (or knot point) of *HS-CGPA* = 3.12 (see Cudeck, 1996; Draper & Smith, 1981; Smith, 1979). Although we could choose this knot point in any way (i.e., at the minimum of the parabola), we choose the mean as the knot and code the predictors so that there is a segment *before* the average z score ($\beta_{1(b)}$) and *after* the z average score ($\beta_{1(a)}$). The results can be written as

$$
\begin{aligned}
\text{F-GPA}_n^{(c)} &= \beta_0^{(c)} + \beta_{1(b)}^{(c)}\text{HS-CGPA(b)}_n^{(c)} + \beta_{1(a)}^{(c)}\text{HS-CGPA(a)}_n^{(c)} + \underline{.489}e_n \\
\beta_0^{(c)} &= \underline{2.53} + \underline{-.021}\ \text{COL-GRAT}^{(c)} + \underline{.158}d_0^{(c)} \text{ and} \\
\beta_{1(b)}^{(c)} &= \underline{.288} + -.006\ \text{COL\_GRAT}^{(c)} + \underline{.084}d_{1(b)}^{(c)} \text{ and} \\
\beta_{1(a)}^{(c)} &= \underline{.514} + \underline{-.032}\ \text{COL\_GRAT}^{(c)} + .003d_{1(a)}^{(c)}
\end{aligned}
\quad (12)
$$



(a)



(b)

FIGURE 6 (a) Separate college OLS regression functions from a quadratic polynomial model. (b) Separate college OLS regression functions from a bilinear segments model.

with the subscript "(b)" for "below" and "(a)" for "above" the average score. One potentially new effect is apparent: the prediction at the average score is the intercept ($\beta_0 = 2.53$), but the "below average" segment seems to have a lower slope ($\beta_{1(b)} = .288$) than the "above average" segment ($\beta_{1(a)} = .514$). This implies that the prediction model would be more linear if we stretched the HS-CGPA score so the interval 2.0 to 3.0 was considered equivalent to 3.0 to 3.5 and 3.5 to 4.0. In addition, the impact of college graduation rate is now larger in the above average segment. These slope differences obtain a nontrivial improvement in fit ($L^2 = 227$ on $\Delta df = 6$) over the nested baseline model but this model not as accurate as the quadratic model ($PRE = .36, -.10$). The two-stage OLS estimates of this segmented model (M_6) are presented for separate colleges in Figure 6b. Here, again, it is clear that the prediction model suggests a stronger effect of high school grades after the minimum point. We return to this nonlinear possibility in the final series of multilevel models (in Table 6).

Adding Additional Predictor Variables

Table 4 includes a listing of results for a sequence of random components models where we examine three potentially important high school predictors. The first model (M_3) is repeated from Table 3 and shows the random components prediction of the freshman grades from high school grades (with $\beta_1 = .399$ and $PRE = .36, -.02$). The model M_7 is a random components prediction of freshman grades only from high school SAT or ACT scores (HS-TEST), and the results ($\beta_2 = .323$ and $PRE = .22, -.06$) lead to a slightly lower prediction than in the grades-only model (M_3). The next model (M_8) is a random components prediction of freshman grades only from the number of high school core courses (HS-UNITS), and the results ($\beta_3 = .154$ and $PRE = .06, -.02$) show much lower prediction than the other two univariate models. Among these variables, high school grades seem to be the strongest single predictor of freshman grades, with the test scores next best, and the core units with relatively low prediction.

The last four models include multiple predictors. A random components prediction from the combination of high school grades and tests is given first (M_9). In this prediction the independent grade slope ($\beta_1 = .327$) is largest, but the independent ACT/SAT test-score slope ($\beta_2 = .135$) is also positive and accurate. In contrast to either univariate model (M_3 or M_7) the goodness of fit is increased substantially, the first-level modeled variance increases ($PRE_1 = .38$), but there is a decrease in the second level ($PRE_2 = -.13$). This combination of both grades and test variables yields a more accurate prediction than either variable alone, but the second-level model is probably not yet appropriate.

TABLE 4
Alternative Multilevel Models for the Prediction of Freshman Grades in College

Parameters (*Symbol*)	M_3 *GPA*	M_7 *TEST*	M_8 *UNITS*	M_9 $G + T$	M_{10} $G + U$	M_{11} $T + U$	M_{12} $G + T + U$
Fixed Terms at Level 1							
Intercept (γ_{00})	2.62 (>99)	2.63 (>99)	2.62 (>99)	2.63 (>.99)	2.62 (>99)	2.63 (>99)	2.63 (>99)
HS_CGPA (γ_{01})	0.399 (71)	=0	=0	0.327 (53)	0.396 (67)	=0	0.328 (53)
HS_TEST (γ_{02})	=0	0.323 (54)		0.135 (23)	=0	0.302 (49)	0.137 (24)
HS_UNITS (γ_{03})	=0	=0	0.154 (28)	=0	0.009 (2)	0.069 (14)	−0.004 (1)
Random Terms at Level 1							
Variance{Error} (ϕ_e^2)	0.242 (89)	0.297 (89)	0.367 (89)	0.229 (88)	0.242 (88)	0.292 (88)	0.230 (88)
Random Terms at Level 2							
Variance{Intercept} (ϕ_0^2)	0.023 (9)	0.023 (9)	0.021 (9)	0.026 (9)	0.023 (9)	0.026 (9)	0.026 (9)
Variance{CGPA} (ϕ_1^2)	0.004 (5)	=0	=0	0.004 (4)	0.004 (5)	=0	0.003 (4)
Variance{TEST} (ϕ_2^2)	=0	0.003 (4)	=0	0.002 (3)	=0	0.003 (4)	0.002 (3)
Variance{UNITS} (ϕ_2^2)	=0	=0	0.001 (2)	=0	∼0 (0)	0.001 (1)	∼0 (0)
Covariances (ϕ_1^2)	−0.004 (4)	=0	=0	0.003 (2)	−0.004 (4)	=0	∼0 (0)
(ϕ_2^2)		∼0 (0)	=0	−0.002 (2)	=0	∼0 (0)	∼0 (0)
(ϕ_3^2)			∼0 (0)		∼0 (0)	0.001 (1)	∼0 (0)
(ϕ_{12})				∼0 (0)	=0	=0	∼0 (0)
(ϕ_{13})					∼0 (0)	=0	∼0 (0)
(ϕ_{23})						∼0 (0)	∼0 (0)
Goodness-of-Fit							
Explained{F_GPA} (R^2)	.412	.277	.107	.443	.412	.290	.440
Log Likelihood (−2LL)	23,841	27,098	30,442	23,073	23,828	26,877	23,083
Diff. −2LL/Diff. dfs (LRR)	7,757/3	4,303/3	959/3	8,329/7	7,573/7	4,524/7	8,318/13

Note. (1) All models were fitted here using SAS PROC MIXED (Singer, 1998) with $N = 16{,}348$, and $N(c) = 267$; (2) MLE parameter estimates and approximate $t = p/se(p)$ values in parentheses; (3) =0 indicates parameter fixed at zero; (4) ∼0 indicates parameter estimated at zero; (5) For F_GPA Observed Mean = 2.617 with Variance = .411; (6) All predictor variables scaled to Z-score form.

Model M_{10} combines grades and core units and shows a strong effect of grades and only a minor net effect for the addition of core units. Model M_{11} combines tests and units and shows a strong effect of tests and a slightly positive effect of core units. The final model M_{12} combines all three variables and shows the same basic result as found in grades and tests alone (M_9)—there is no net impact of the number of core units on the prediction of freshman grades. Here the number of core units ($\beta_3 = -.004$) is not accurately different from zero (and it would be interpreted as a case of classic suppression in any case; see Table 1). The estimated model variance for these three predictors ($PRE_1 = .38$) is similar to the estimated model variance of M_9 with just two predictors ($PRE_1 = .38$), probably indicating some model misspecification. Also, several covariance parameters were estimated at zero and this probably reflects a numerical boundary condition. In general, the results of M_{12} generally indicate numerical instability in the estimates from the inclusion of all three individual-level variables with all random components.

To avoid further convergence problems (using SAS PROC MIXED) in these more complex models we restricted our estimation to variance terms only ($\phi_{xy} = 0$). In models not presented in detail here, we added college graduation to this prediction and found positive average slopes for grades ($\beta_1 = .327$) and for tests ($\beta_2 = .140$) and again no independent effect for core units. At the second level we found a small but accurate negative effect of college graduation rate on the average score ($\gamma_{01} = -.061$), but we do not find any additional effect of the college graduation rate on the slopes of grades, tests, or units. The variance components of first-order grades and tests were nonzero, but the core units do not add systematic variance here either. In sum, these new models yield basically the same results as the less complex models fitted earlier (M_9 and M_{12}).

Additional College-Level Characteristics and College Outcomes

The next set of multilevel models are listed in Table 5. These models all include a second-level prediction with three college characteristics: college graduation rate (COL-GRAT), a contrast of the public versus private colleges (COL-PUPR), and the cost of attending the college (COL-COST). These models are fitted separately for three college outcomes: freshman grades (F-GPA), freshman credit hours attained (F-CREDITS), and freshman quality points (F-QUALPTS).

In the first model (M_{13}) the previous prediction model for freshman grades (F-GPA) is expanded to include college-level variables. The results obtained with all three college-level variables yield only a small increase in first-level model variance ($PRE_1 = .38$) and no new independent effects beyond those

TABLE 5
Multilevel Models for the Prediction of Three Freshman Academic College Outcomes

Parameters (Symbol) Outcome	M_{13} *F_GPA*	M_{14} *F_CREDITS*	M_{15} *F_QUALPTS*
Fixed Terms			
Intercept (γ_{00})	2.63 (>99)	28.5 (>99)	76.6 (>99)
HS_CGPA (γ_{10})	0.333 (39)	1.46 (15)	12.8 (34)
HS_TEST (γ_{20})	0.143 (18)	0.607 (6)	6.31 (17)
HS_UNITS (γ_{30})	−0.005 (1)	0.173 (2)	0.16 (1)
COL_GRAT>Intercept (γ_{01})	−0.067 (5)	0.262 (2)	−1.6 (3)
COL_GRAT>HS_CGPA (γ_{11})	−0.005 (1)	−0.245 (3)	−0.53 (2)
COL_GRAT>HS_TEST (γ_{20})	−0.009 (1)	−0.054 (1)	−0.33 (1)
COL_GRAT>HS_UNITS (γ_{30})	0.003 (0)	−0.046 (1)	0.009 (0)
COL_PUNP>Intercept (γ_{02})	0.010 (1)	−0.131 (1)	0.22 (0)
COL_PUNP>HS_CGPA (γ_{12})	0.014 (1)	0.061 (0)	0.50 (1)
COL_PUNP>HS_TEST (γ_{22})	0.007 (1)	0.101 (1)	0.45 (1)
COL_PUNP>HS_UNITS (γ_{32})	0.003 (0)	0.040 (0)	−0.05 (0)
COL_COST>Intercept (γ_{02})	0.004 (1)	0.267 (1)	0.58 (1)
COL_COST>HS_CGPA (γ_{13})	−0.010 (1)	0.044 (1)	0.01 (0)
COL_COST>HS_TEST (γ_{23})	−0.004 (0)	−0.128 (1)	−0.40 (1)
COL_COST>HS_UNITS (γ_{33})	0.004 (0)	0.015 (1)	0.15 (0)
Random Terms			
Variance{Error} (ϕ_e^2)	0.230 (88)	27.4 (88)	443 (88)
Variance{Intercept} (ϕ_0^2)	0.023 (9)	2.93 (9)	49.5 (9)
Variance{HS_CGPA} (ϕ_1^2)	0.003 (4)	0.470 (5)	5.9 (4)
Variance{HS_TEST} (ϕ_2^2)	0.002 (3)	0.433 (5)	5.50 (4)
Variance{HS_UNITS} (ϕ_2^2)	∼0 (0)	0.056 (1)	0.48 (1)
Goodness-of-Fit			
Nesteded{Outcome} (η^2)	.047	.100	.064
Explained{Outcome} (R^2)	.440	.214	.412
Log Likelihood (−2LL)	23,134	101,219	146,593
Diff. −2LL/Diff. dfs (LRR)	8,267/20	2,036/20	7,445/20

Note. (1) All models were fitted here using SAS PROC MIXED with $N = 16{,}348$, and $C = 267$; (2) MLE parameter estimates and approximate $t = p/se(p)$ values in parentheses; (3) ∼0 indicates parameter estimated at zero; (4) For F_GPA Observed Mean = 2.617 with Variance = .411; (5) All predictor variables scaled to Z-scores.

found in previous models. It appears that once the overall college graduation rate is taken into account, the independent effects of college type and college cost do not add much to the prediction of college grades of freshman students.

The next model (M_{14}) uses the same multilevel structure but makes a prediction of a new dependent variable of freshman credit hours (F-CREDITS). This model yields accurate first-level estimates for the intercept ($\beta_0 = 28.5$), for the slope of core grades ($\beta_1 = 1.46$), for the slope of tests ($\beta_2 = .607$), and also for the slope of core units ($\beta_3 = .173$). At the second level, accurate estimates include a positive effect of college graduation on average credits hours ($\gamma_{01} = .262$) and a negative effect of college graduation on the slope of the high school grades on the credit hours ($\gamma_{11} = -.245$). All other effects are either the same as before or not accurate. In contrast to the prior freshman grade predictions, the multilevel model of freshman credits attained shows (a) a higher degree of college-level nesting ($\eta^2 = .100$), (b) a much lower first-level model variance ($PRE_1 = .13$), but (c) a much higher second-level model variance ($PRE_2 = .16$). These larger second-level effects of freshman credit hours are related only to the college graduation rate.

The next model (M_{15}) is the first model presented here for freshman quality points (F-QUALPTS). This variable is a product of the freshman grades and credit hours attained (i.e., grades weighted by credits), and these results are most similar to the previous model for freshman grades. This model yields accurate first-level estimates for the intercept ($\beta_0 = 76.6$), for the slope of grades ($\beta_1 = 12.8$), for the slope of tests ($\beta_2 = 6.31$), but no significant slope for core units. At the second level, accurate estimates include a small negative effect of college graduation on average quality points ($\gamma_{01} = -1.6$) and a small negative effect of college graduation on the slope of the high school grades on the credit hours ($\gamma_{11} = -.53$). All other effects are either the same as before or not accurate. As in the prior freshman grade predictions, the model of quality points attained (a) a comparable degree of college-level nesting ($\eta^2 = .064$), (b) similar first-level model variance ($PRE_1 = .35$), and (c) a small but positive second-level model variance ($PRE_2 = .05$).

Segmented Multilevel Models

In another set of models, we now explore the possibility of nonlinearity in the prior multilevel prediction models. Table 6 is a list of results for each of the three freshman academic outcomes: grades, credits, and quality points. In each of the three models we have used two first-level variables (high school grades and tests) and one second-level variable (college graduation rate). The distinguishing feature of these models is the inclusion of two "linear segments around the average" for each predictor variable. These segments were formed for each of the z scores using the same simple method described earlier (see model M_6 in Table 3). These models are estimated as a single multilevel prediction equation with connected segments and with variance components to permit a closer look at some segments of predictability in these models.

The first model (M_{16}) is a multilevel prediction of freshman grades (F-GPA) using the three segmented variables, and the results can be written as

$$\text{F-GPA}_n^{(c)} = \beta_0^{(c)} + \beta_{1(b)}^{(c)}\text{HS-CGPA(b)}_n^{(c)} + \beta_{1(a)}^{(c)}\text{HS-CGPA(a)}_n^{(c)}$$

$$+ \beta_{2(b)}^{(c)}\text{HS-TEST(b)}_n^{(c)} + \beta_{1(a)}^{(c)}\text{HS-TEST(a)}_n^{(c)} + \underline{.476}e_n$$

$$\beta_0^{(c)} = \underline{2.57} + -.010\text{COL-GRAT(b)} + \underline{-.108}\text{COL-GRAT(a)}^{(c)} + \underline{.138}d_0^{(c)}$$

$$\beta_{1(b)}^{(c)} = \underline{.290} + .036\text{COL-GRAT(b)} + \underline{-.070}\text{COL-GRAT(a)}^{(c)} + \underline{.089}d_{1(b)}^{(c)}$$

$$\beta_{1(a)}^{(c)} = \underline{.434} + .000\text{COL-GRAT(b)} + \underline{-.058}\text{COL-GRAT(a)}^{(c)} + .000d_{1(a)}^{(c)}$$

$$\beta_{2(b)}^{(c)} = \underline{.076} + -.028\text{COL-GRAT(b)} + -.005\text{COL-GRAT(a)}^{(c)} + \underline{.084}d_{2(b)}^{(c)}$$

$$\beta_{2(a)}^{(c)} = \underline{.178} + .001\text{COL-GRAT(b)} + -.019\text{COL-GRAT(a)}^{(c)} + .000d_{2(a)}^{(c)} \tag{13}$$

with the accurate effects underlined. This model yields accurate first-level estimates for the intercept ($\beta_0 = 2.57$), some differences in grades for the "below average" slope ($\beta_{1(b)} = .290$) and the "above average" slope ($\beta_{1(a)} = .434$), and similar differences in tests for the "below average" slope ($\beta_{2(b)} = .076$) and the "above average" slope ($\beta_{2(a)} = .178$). At the second level, accurate estimates are found only for the "above average" graduation rate colleges. These include a small negative effect of college graduation on average grades ($\gamma_{01(b)} = -.108$) and small negative effects of college graduation on the slope of the high school grades ($\gamma_{11(b)} = -.070$) and tests ($\gamma_{12(b)} = -.058$). This final *F-GPA* model yields the highest first-level model variance ($PRE_1 = .40$) and a smaller but positive second-level model variance ($PRE_2 = .14$).

The last two models (M_{17} and M_{18}) include the same structure of predictors for freshman credit hours and for quality points. The model for credit hours shows a different pattern of results, and the "below versus above" average GPA result does not seem to apply. Here the "below average" slope ($\beta_{1(b)} = 2.01$) is greater than the "above average" slope ($\beta_{1(a)} = 1.36$). However, as before, the test score parameters yielded a zero "below average" slope ($\beta_{2(b)} = -0.2$) with a positive "above average" slope ($\beta_{2(b)} = 1.4$). This final F-CREDITS model yields a nontrivial first-level model variance ($PRE_1 = .15$) and the largest second-level model variance ($PRE_2 = .28$) of any model presented here.

The first-level model for freshman quality points follows the pattern of the freshman grades model described earlier, but the second-level model seems to enhance the credit hour effects. Strong negative effects are found at the "above average" slopes for the college graduation on the average quality points ($\gamma_{02} = -3.62$) and on the high school grades slope ($\gamma_{11(b)} = -3.37$). This final F-QUALPTS model yields positive model variance (*PRE* = .37, .18).

This general pattern of stronger effects at the "above average" levels seems notable among all variables. The goodness of fit of these models is slightly improved over other models not presented here, including simple single segment models (e.g., for F-GPA, $L^2 = 184$ on $\Delta df = 8$). Furthermore, the positive first-level and second-level model variances suggest the models are more properly specified than before. These kinds of nonlinear results are provocative and seem worthy of further study.

Additional Multivariate Multilevel Models

Various attempts were made to fit more elaborate multivariate models where the interrelationships among the outcomes included some common features (i.e., common factors as in Equation (10)). By comparing the fit of these combined outcome models to the separate outcome models we hoped to obtain a formal test of the fit of a single "academic factor" hypothesis (see McArdle, 1998). Unfortunately, most of our multiple outcome models failed to converge for one reason or another using the standard software.

To minimize this problem we examined a bivariate representation of this common factor model of Equation (10) where the two outcomes were assumed to be proportional (λ) to one another except for constants representing the differences in origin (ν_1) and scale (ϕ_u^2). We fit these bivariate models with restrictions of proportionality of the parameters of the model for grades (M_{13}) and credits (M_{14}). Numerical boundary conditions were quickly located and convergence was not achieved. The results did yield multilevel equations with proportional estimates ($\lambda = 3.5$) but the parameters were unstable and yielded extremely poor fit ($L^2 >$ twice the sum of the prior fits).

These generally poor numerical results we obtained could be because the multiple outcome models fitted were too complex for the available programs. However, we also fit a standard SEM latent variable path model (as in McArdle, 1994, 1998) to the between-group (i.e., Table 2) and within-group matrices (i.e., Table 2a minus Table 2b), and a single common factor model also fit poorly for these data. This SEM is not the model of interest (Equation (10)), but this analysis points out that the poor numerical results could also be because a simple common factor with multilevel structure (as in Equation (10)) was a seriously misspecified model for these data. For one or more of these reasons, the potentially useful multivariate hypotheses about broad academic achievements have not been fully examined here.

TABLE 6
Multilevel Segmented Prediction Models for High School Academic Variables

Parameter (Symbol)	M_{16} *F_GPA*	M_{17} *F_CREDITS*	M_{18} *F_QUALITY*
Fixed Effects			
Intercept (γ_{00})	2.57*	28.3*	72.6*
HS_CGPA1 ($\gamma_{1(b)}$)	0.290*	2.01*	11.02*
HS_CGPA2 ($\gamma_{1(a)}$)	0.434*	1.36*	17.37*
HS_TEST1 ($\gamma_{2(b)}$)	0.076*	−0.18	1.49
HS_TEST2 ($\gamma_{2(a)}$)	0.178*	1.4*	9.98*
COL_GRAD1>Intercept (γ_{01})	−0.010	0.40	0.36
COL_GRAD1>HS_CGPA1 ($\gamma_{11(b)}$)	0.036	0.15	1.02
COL_GRAD1>HS_CGPA2 ($\gamma_{12(b)}$)	~0	−0.15	0.26
COL_GRAD1>HS_TEST1 ($\gamma_{21(b)}$)	−0.028	−0.43	−1.69
COL_GRAD1>HS_TEST2 ($\gamma_{22(b)}$)	0.001	0.25	0.78
COL_GRAD2>Intercept (γ_{02})	−0.108*	−0.19	−3.62*
COL_GRAD2>HS_CGPA1 ($\gamma_{11(a)}$)	−0.070	−1.03*	−3.37*
COL_GRAD2>HS_CGPA2 ($\gamma_{12(a)}$)	−0.058*	0.02	−1.75
COL_GRAD2>HS_TEST1 ($\gamma_{21(a)}$)	−0.005	−0.48	−1.37
COL_GRAD2>HS_TEST2 ($\gamma_{22(a)}$)	−0.019	−0.28	−1.54
Random Effects			
Variance{Error} (ϕ_e^2)	0.227*	27.1*	434*
Variance{Intercept} (ϕ_0^2)	0.019*	2.46*	42.3*
Variance{HS_CGPA1} (ϕ_{11}^2)	0.008*	1.17*	6.96*
Variance{HS_CGPA2} (ϕ_{12}^2)	~0	0.32	10.59*
Variance(HS_TEST1) (ϕ_{11}^2)	0.007*	1.28*	7.55*
Variance(HS_TEST2) (ϕ_{22}^2)	~0	0.43*	9.19*
Goodness-of-Fit			
Explained{Outcome} (R^2)	.448	.223	.424
Log Likelihood (−2LL)	22,891	101,134	146,309
Diff. −2LL/Diff. dfs (LRR)	8,510/19	2,121/19	7,729/19

Note. (1) Same data as listed in Tables 3 and 4; (2) Asterisk indicates a t-value > 2.58; (3) Predictor variable are all segments; (4) Each predictor variable is separated into two segments: $X1 = XiffX < 0$ and $X2 = XiffX > 0$.

DISCUSSION

This study was an application of contemporary multilevel regression modeling to the prediction of 1st-year academic performances in college. The data used here were collected as the first cohorts of a 5-year longitudinal study of student-athletes. Although this study was initiated in 1993–1994, this data collection data marked the first step toward a national census of student-athlete academic performance that is now used for numerous purposes including penalizing colleges when their teams substantially underperform academically. The primary goal of the study was to gather information about the academic successes and failures of student-athletes in all NCAA sports activities. A secondary goal was to provide an improved empirical base for examining the potential impacts of NCAA legislation related to academic issues, including initial eligibility, mandatory degree progress benchmarks, and team-level academic expectations. It is possible that the results of these multilevel models can add to the literature and be of some practical use in future decisions, and we summarize a few of these issues now.

Substantive Findings

Our multilevel analyses initially focused on the prediction of freshman GPA from a variety of high school academic variables. The models used were mainly standard multilevel regression models, but we also examined nonlinear prediction within multilevel models. The basic pattern of results obtained shows the high school grades are the best available predictors of freshman college grades, but the ACT and SAT test scores are the next best available. The number of high school core units taken does not add to this freshman grade prediction, but it was useful in understanding differences in freshman credits taken, and this included college-level differences. One key second-level impact found here was that the college graduation rate, which may serve as a proxy for the student body academic profile and institutional educational resources, has a small but significant negative relationship to the average freshman grades. This result requires further elaboration. We also found some interesting nonlinear effects of higher prediction for students at higher levels of the academic variables.

Some aspects of these results have further substantive interpretation. First, the well-known and widely cited results of freshman grades predicted by high school grades and SAT or ACT test scores were not altered very much by the inclusion of a multilevel structure to the model even though this was quite likely (e.g., Wainer & Brown, 2004). The clustering of student-athletes within colleges was apparent, but this was relatively small ($\eta^2 = 5\%$), and this led to only minor alterations in the traditional regression estimates of fixed effects. Of course, this form of clustering might be different for college students who do not play college sports. Nevertheless, under the restrictive assumption of a homogeneous residual variance across colleges, this representation affords correct estimates of the confidence boundaries (i.e., standard errors).

One interesting result was the finding that college graduation rate had a small but significant negative relationship as a second-level predictor (see model M_4, Table 3, $\gamma_{01} = -.034$). This small value implies that, if all else could be held constant, the higher the graduation rate of the school, the lower the 1st-year grades. Of course, because we find that more highly selective schools seem to have higher graduation rates (e.g., McArdle, 1998; McArdle & Hamagami, 1994), this result seems counterintuitive. However, this new result can be interpreted to mean that more selective colleges have more difficult 1st-year grading practices, and major programs may be altered after college matriculation. These college-level parameters can also be interpreted relative to the change in individual z scores—that is, if we estimated the net effect of the differences in high school grades and college graduation rates (i.e., z scores for HS-CGPA minus COL-GRAT), the coefficient would yield a positive difference ($.400 - .034 = .366$). Thus, for a student whose high school grades are "above the average," going to a college where the other students are "at the average" yields predicted freshman grades higher than for those students "at the average." Conversely, this also means that students with grades "below the average" will have lower predicted scores at the specific colleges (cf. Bowen & Bok, 1998).

We generally found some differences between the models for freshman grades and freshman credit hours, and these results may be interpreted in several different ways. The first-level results show a positive effect of all high school academic variables on the credit hours. The new result here is an independent effect of course units, but the fact that the number of high school core courses is positively related to number of college credit hours attained is not so surprising. The new second-level results may be interpreted for colleges with different graduation rates: in those colleges with higher graduation rates the students attain more credits hours as freshmen, but these increased credit hours are highest for students with lower high school grades. Other patterns of college and students can be described in a similar fashion.

In analyses not presented here, we have used multilevel models to examine a variety of other aspects of these data. Multilevel logit (and probit) models were examined for binary variables, including freshman persistence and academic eligibility, and these models are far less predictive and efficient than the ones described here (cf. Hedecker & Mermelstein, 1999; McArdle & Hamagami, 1994; Pascaralla & Terenzini, 1991). Other ethnic and sport group differences have been found before, and these will no doubt arise in these models (e.g., McArdle, 1998; Sawyer, 1986). We have also examined new models where the high school location (state and area) was used as the nested effect, and the intraclass correlations are much larger ($\eta^2 > .15$) than those reported for college groupings ($\eta^2 < .05$) here. This last result raises the possibility that 1st-year college students are more efficiently grouped by the high school context they were a part of (for many years) than the new college context (of 1 year). Combinations of these groupings and contexts may be of further importance because the most important context for any one person—of one ethnic background in one high school going to one college participating in one sport—may be quite different from another person going to the same college.

Continuing Challenges

The multilevel models can be routinely used for these and other kinds of validation studies. In fact, these multilevel multivariate models may eventually be useful as a standard against which we judge all other results. However, multilevel analyses raise more general research challenges.

A key challenge raised here is the general question of exactly "what context is most important" (see Brooks-Gunn, Duncan, & Aber, 1998; Diez-Roux, 1998; Hox & Kreft, 1994), and further methodological and substantive work on this question is needed. The multilevel models used here afford one useful statistical tool for unconfounding the complex issues of "college selectivity" and "college choice" (e.g., Beatty et al., 1999; Manski & Wise, 1983). But, as we have found, parametric interpretations are only one way to deal with the model-based interpretations of comparisons within a manifest group (i.e., the "Frog Pond effect"; e.g., Aitkin & Longford, 1986; Burstein et al., 1989; Cronbach & Webb, 1975; Kreft, 1993; Kreft et al., 1995). The most appropriate clustering of groups of persons may even be considered another kind of latent factor to be found by empirical study (e.g., Muthén, 1994).

Prior research has shown how it is possible to create several alternative SEMs for the comparison of multiple variable equations (see Goldstein, 1995; McArdle & Hamagami, 1996; McDonald & Goldstein, 1988; B. Muthén, 1994). Our efforts to fit a common factor model here proved impractical given the limited outcomes we studied. In practice, this common factor model may be more appropriate at other levels of observation (i.e., separate grades, multiple tests or interest areas, etc.). But this kind of multilevel SEM is important on a theoretical basis. That is, this model is restrictive in the sense that only the common factor scores, and not the specific factor scores, can carry the multilevel structural information. Even using only a bivariate model, this can lead to a formal rejection of the fit of the "same factor" hypothesis. In a general sense, this kind of multivariate multilevel model can help provide a more rigorous test of the validity of any construct in question and, hence, can be a useful scientific tool (as in McArdle & Goldsmith, 1990; McArdle & Prescott, 1992; McArdle & Woodcock, 1997; Nesselroade & McArdle, 1996).

In this sense, we can utilize the lessons learned from several decades of structural equation modeling and add any useful feature of any modeling concepts to improve our research tools. There have been several recent attempts at expanding the possibilities for multivariate multilevel analyses (e.g., see Goldstein & McDonald, 1988; Kaplan & Elliot, 1997; McArdle & Hamagami, 1996; McDonald, 1993, 1994; cf. Hox, 2002; B. Muthén, 1994; Thum, 1997). The inclusion of more explicit measurement and structural models for the outcomes and inputs is a natural extension of multilevel concepts. The current multilevel programs are already fairly complicated but they are easy to use and understand.

Nevertheless, the challenge of finding out exactly how multilevel models can improve multivariate model building may take the time, energy, and patience of the next generation of research and researchers.

ACKNOWLEDGMENTS

This research was supported by the National Collegiate Athletic Association (NCAA). Conclusions drawn from or recommendations based on the data provided by the NCAA are those of the authors based on analyses/evaluations of the authors and may not represent the views of the officers, staff, or membership of the NCAA. We thank Todd Petr (Managing Director of Research) and Maria DeJulio (Research Associate) of the NCAA Research Staff for their diligent collection of all research information used here and for their continuing support of this research. We thank the members of the NCAA Research Committee for their helpful suggestions and continued support of our research. We also thank Karen S. McCollam (UVa), Carol A. Prescott (USC), and two unnamed reviewers (MBR) for their helpful suggestions on this manuscript. Finally, we thank the MBR editor, Joseph Rodgers, for his excellent editorial guidance and general support of this research.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Aitkin, M. A., & Longford, N. T. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, A, 149,* 1–43.

American College Testing Program (ACT). (1995). *ACT assessment student information.* Iowa City, IA: Author.

Astin, A. (1971). *Predicting academic performance in college*. New York, NY: Free Press.

Beatty, A., Greenwood, M. R. C., & Linn, R. L. (1999). *Myths and tradeoffs: The role of tests in undergraduate admissions.* Washington, DC: National Research Council Press.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data*. New York, NY: Academic Press.

Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long term consequences of considering race in college and university admissions.* Princeton, NJ: Princeton University Press.

Braun, H. I. (1989). Empirical Bayes methods: A tool for exploratory analysis. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 19–56). New York, NY: Academic Press.

Brooks-Gunn, J. Duncan, G. J., & Aber, J. L. (1998). *Neighborhood Poverty, Volume II: Policy Implications in Studying Neighborhoods*. New York: Russell-Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Press.

Burstein, L., Kim, K.-S., & Delandshere, G. (1989). Multilevel investigations of systematically varying slopes: Issues, alternatives, and consequences. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 235–282). New York, NY: Academic Press.

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

College Entrance Examination Board (1995). *College Bound Seniors National Report: 1995 Profile of SAT and Achievement Test Takers.* New York, NY: College Entrance Examination Board.

Cronbach, L. J., & Webb, N. (1975). Between class and within class effects in a reported aptitude × treatment interaction: A reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 67,* 717–724.

Crouse, J., & Trusheim, D. (1988). *The case against the SAT.* Chicago, IL: University of Chicago Press.

Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behavioral Research, 31,* 371–403.

Diez-Roux, A. V. (1998). Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health, 88*(2), 216–222.

Draper, N., & Smith, H. (1981). *Applied regression analysis, Second edition*. New York, NY: Wiley.

Goldstein, H. (1995). *Multilevel statistical models*. New York, NY: Halsted Press.

Goldstein, H., & McDonald, R. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455–467.

Greene, W. H. (1998). *LIMDEP: Version 7.0 user's manual* (Rev. ed.). Plainview, NY: Econometric Software, Inc.

Hedecker, D., & Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics, 50,* 933–944.

Hedecker, D., & Mermelstein, R. (1999). A multilevel thresholds of change model for the analysis of stages of change data. *Multivariate Behavioral Research, 33,* 427–455.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J., & Kreft, I. G. G. (1994). Multilevel analysis methods. *Sociological Methods & Research, 22*(3), 283–299.

Joreskog, K. G., Sorbom, D., du Toit, S. H. C., & du Toit, M. (1999). *LISREL 8: New statistical features*. Chicago, IL: Scientific Software International.

Kaplan, D., & Elliot, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling, 4*(1), 1–24.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, B, 36,* 1–37.

Kreft, I. G. G. (1993). Using multilevel analysis to assess school effectiveness: A study of Dutch secondary schools. *Sociology of Education, 66,* 104–129.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–22.

Kreft, I. G. G., de Leeuw, J., and van der Leeden, R. (1995). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *American Statistician, 48,* 324–335.

Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrika, 38,* 963–974.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data*. Newbury Park, CA: Sage.

Littell, R. C., Miliken, G. A., Stoup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Longford, N. (1990). *VARCL: Software for variance components analysis of data with nested random effects (Maximum Likelihood).* Princeton, NJ: Educational Testing Service.

Manski, C. F., & Wise, D. A. (1983). *College choice in America*. Cambridge, MA: Harvard University Press.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McArdle, J. J. (1998). Contemporary statistical models for examining test bias. In J. J. McArdle & R. W. Woodcock (Eds.), *Human abilities in theory and practice* (pp. 157–195). Mahwah, NJ: Erlbaum.

McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. MacCallum, (Eds.), *Factor analysis at 100 years* (pp. 99–130). Mahwah, NJ: Erlbaum.

McArdle, J. J., & Goldsmith, H. H. (1990). Alternative common factor models for multivariate biometric analyses. *Behavior Genetics, 20* (5), 569–608.

McArdle, J. J., & Hamagami, F. (1994). Logit and multilevel logit modeling studies of college graduation for 1984-85 freshman student-athletes. *The Journal of the American Statistical Association, 89*(427), 1107–1123.

McArdle, J. J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 89–124). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Prescott, C. A. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18*(3), 87–115.

McArdle, J. J., & Woodcock, J. R. (1997). Expanding test-rest designs to include developmental time-lag components. *Psychological Methods*, *2*(4), 403–435.

McDonald, R. P. (1993). A general model for two level data with responses missing at random. *Psychometrika, 58,* 575–585.

McDonald, R. P. (1994). The bilevel Reticular Action Model for path analysis with latent variables. *Sociological Methods Research, 22*(3), 399–413.

McDonald, R. P., & Goldstein, H. (1988). Balanced versus unbalanced designs for linear structural relations in two level data. *British Journal of Mathematical and Statistical Psychology, 42,* 215–232.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376–398.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus: The comprehensive modeling program for applied researchers user's guide*. Los Angeles, CA: Author.

National Collegiate Athletic Association (NCAA, 1992). *A statistical comparison of the college graduation of freshman student-athletes before and after proposition 48.* NCAA Research Report #92-02. Overland Park, KA: NCAA Research publications.

Nesselroade, J. J., & McArdle, J. J. (1996). On the mismatching of levels of abstraction in mathematical statistical model fitting. In M. D. Franzen & Reese, H. W. (Eds.), *Life Span Developmental Psychology: Biological and Neuropsychological Mechansims* (pp. 23–49). Mahwah, NJ: Erlbaum.

Pascarella, E. T., & Terenzini, P. T. (1991). How college affects students. San Francisco, CA: Jossey-Bass.

Paskus, T. S. (2012). A summary and commentary on the quantitative results of current NCAA academic reforms. *Journal of Intercollegiate Sport, 5,* 41–53.

Petr, T. A., & McArdle, J. J. (2012). Academic research and reform: A history of the empirical basis for NCAA academic policy. *Journal of Intercollegiate Sport, 5,* 27–40.

Pinherio, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Sawyer, R. (1986). Using demographic subgroup and dummy variable equations to predict college freshman grade point averages. *Journal of Educational Measurement, 23*(2), 131–145.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*(4), 323–355.

Smith, P. L. (1979). Splines as a useful and convenient tool. *The American Statistician, 32*(2), 57–63.

Snijders, T. A. B., & Bosker, R. J. (1994). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Los Angeles, CA: Sage.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.

S-Plus. (1999). S-Plus 2000. Seattle, WA: MathSoft, Inc.

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics, 22*(1), 77–108.

Wainer, H., & Brown, L. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data (with L. Brown). *The American Statistician, 58,* 117–123.

Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (1990). Predicting college grades: An analysis of institutional trends over two decades. New York, NY: The College Board.

Wong, G. Y., & Mason, W. M. (1985). The linear logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80*(391), 513–524.

Wong, G. Y., & Mason, W. M. (1991). Contextually specific effects and other generalizations of the hierarchical linear model for comparative analysis. *Journal of the American Statistical Association, 86,* 487–503.