Res High Educ (2009) 50:354–367
DOI 10.1007/s11162-009-9121-3

Multi-Campus Studies of College Impact: Which Statistical Method is Appropriate?

Alexander W. Astin · Nida Denson

Received: 18 June 2007 / Published online: 28 January 2009

Abstract In most multi-campus studies of college impact that have been conducted over the past four decades, investigators have relied on ordinary least squares (OLS) regression as the analytic method of choice. Recently, however, some investigators have advocated the use of Hierarchical Linear Modeling (HLM), a method specifically designed for analyses that involve both individual (student) and aggregate (institutional) level measures. Cross-validation analyses using a national database show that the two methods yield an equally good "fit" with empirical data. Existing OLS software has the advantage of enabling one to perform path analytical causal modeling; HLM has the advantage of yielding a more conservative estimate of the significance of institution-level effects.

Keywords Ordinary least squares · Hierarchical linear modeling · Methodology · College effects · Stepwise regression

Studies of student development that involve longitudinal data from several campuses have one obvious advantage over single-campus studies: they enable the researcher to obtain empirical estimates of (1) the comparative effects of different *institutions* and (2) the effects of institutional *characteristics* on students. These characteristics typically include measures such as size and selectivity, but—depending on the data available to the researcher—they can also include more subtle attributes such as peer group measures and characteristics of the faculty.

A. W. Astin (✉)
Higher Education Research Institute, University of California, Los Angeles,
3005 Moore Hall, Mailbox 951521, 90095-1521 Los Angeles, CA, USA
e-mail: aastin@gseis.ucla.edu

N. Denson
Learning and Teaching at UNSW, University of New South Wales, Level 4,
Mathews Building, UNSW, Sydney, NSW 2052, Australia
e-mail: n.denson@unsw.edu.au

From the time when these multi-campus studies first began to appear in the early 1960s, researchers have tended to rely primarily on the multivariate procedure known as ordinary least squares (OLS) regression to assess the effect of institutional-level variables on student development. More recently, however, some investigators have recommended using an alternative multivariate procedure known as Hierarchical Linear Modeling (HLM; also known as multilevel modeling) (e.g., Bryk and Raudenbush 1992; Burstein 1980; Ethington 1997; Raudenbush and Bryk 1986, 2002). HLM was designed specifically for analyses of data that involve both individual (e.g., student) and aggregate (e.g., institutional or school) level measures. The purpose of this paper is to compare OLS regression and HLM in multi-campus studies of students.

OLS Regression

As the number of multi-campus studies continued to grow during the 1960s and ensuing decades, the application of OLS regression became increasingly sophisticated and complex. These advances were made possible primarily because of three features of the SPSS Regression software package: (1) the option to "block" variables in order of their presumed temporal sequencing; (2) the "forward" entry option which permits the investigator to add variables to the regression equation one at a time based on their relative predictive power; and (3) the "Beta in" feature in the SPSS output, which shows, for each variable that has not yet entered the equation, what its standardized regression coefficient (Beta) *would be* if it were entered at the next step. By fully combining these features in the same analysis, the investigator is able to conduct a form of path analysis which—like traditional path analysis—shows all the direct and indirect paths to the dependent variable, but which also shows—*un*like traditional path analysis—*how* each indirect path has been mediated by the action of specific intervening variables (see Astin 1991; Astin and Dey 1996).

This form of analysis is particularly useful in multi-campus longitudinal studies, where the independent variables can be temporally ordered in logical sequence as follows: entering freshman ("input") measures, college characteristics such as size or selectivity ("environment I"), and college experiences such as courses taken ("environment II"). Even finer ordering can be used if desired (e.g., among input characteristics, demographics such as gender and race can be blocked before high school achievements). If, for example, selectivity is found to have an indirect effect on some dependent ("outcome") variable, it becomes possible to determine *which* particular college experiences mediate the effect, and by how much.

HLM

HLM was originally developed in response to methodological problems that arose in analyzing hierarchical or multilevel data (Bryk and Raudenbush 1992; de Leeuw and Kreft 1986; Goldstein 1986; Raudenbush and Bryk 2002). For example, in higher education students are nested within institutions. These hierarchical differences in units of analysis (i.e., student versus institution) raise certain issues in attempting to estimate the effects of institutions on students. Most commonly, the student is chosen as the unit of analysis, although some investigators have chosen the institution as the unit of analysis; but both of these approaches raise certain methodological issues (Burstein 1980; Ethington 1997; Seltzer 1995). On the one hand, disaggregating institutional-level variables to the student

level violates the assumption of independence that is a basic assumption of OLS regression. This results in misestimated standard errors for parameters of interest that are far too small, which in turn increase the chances of making erroneous conclusions. On the other hand, aggregating student-level variables to the institution level introduces the aggregation bias problem where aggregate relationships tend to be much stronger (and sometimes quite different) than those at the student level. In addition, analyzing the aggregate relationships *only* eliminates the possibility of disentangling the effects of the student-level and institutional-level variables on the outcome.

The problems of neglecting the hierarchical or nested nature of educational data gathered by using a single-level statistical model have been acknowledged and addressed by a number of researchers (Bryk and Raudenbush 1992; Burstein 1980; Cronbach 1976; Pascarella and Terenzini 1991; Raudenbush and Bryk 1986, 2002). In fact, the investigation of issues surrounding multilevel educational data is nothing new, and has been of interest to educational researchers for over 30 years. For example, as Burstein (1980) pointed out, there are at least eight chapters in the first five volumes of the annual *Review of Research in Education* which address (either directly or indirectly) multilevel issues. Seltzer (1995) credits Burstein's (1980) slopes-as-outcomes approach as the precursor to HLM. With this approach, both estimates of group means *and* within-group slopes become outcomes themselves. For example, in the context of school effects research, those school policies and practices that promote high levels of achievement *and* that result in more equitable distributions of achievement may be identified. Thus, the relationships between student-level predictors and student outcomes are viewed as potentially varying across organizational contexts (e.g., classrooms, schools, institutions).

Raudenbush and Bryk (1986, 1992, 2002) provide an extremely accessible explanation of HLM in education for modeling the structure of multilevel data that is now becoming more commonly applied in higher education research. With this approach, it is no longer necessary to choose between the student or institution as the unit of analysis and to worry about biases inherent in aggregating data. HLM basically resolves the aggregation bias that occurs when using single-level linear models to examine multilevel effects, by decomposing the relationships between variables into separate student-level and institution-level components. Specifically, HLM disentangles the student-level and institution-level effects by partitioning the variance-covariance components into separate within- and between-institution elements (see Raudenbush and Bryk 2002, for a more detailed explanation).[1]

HLM can also provide improved estimation of individual effects by borrowing information from the data as a whole.[2] For example, Braun et al. (1983) examined the use of standardized test scores for selecting minority applicants to graduate business schools. Because most of the applicants were Caucasian, their data had a large influence over the estimated prediction equations. Thus, these equations were not as accurate in estimating prediction equations for the minority applicants. By using HLM, Braun et al. (1983) were able to efficiently use all of the available information to provide each school with separate prediction equations for both Caucasian and minority applicants.[3]

[1] The problem of dependence can be handled in OLS regression by computing for each student a unique institutional-level score (e.g., selectivity) that omits that student's score (i.e., SAT).

[2] One could, of course, handle this problem with OLS regression by performing separate regressions for minority and white students.

[3] Specifically, the estimator for each school was weighted by its precision (see Braun et al., 1983, for a more detailed explanation).

Further, HLM permits the investigator not only to assess separately the main effects of both individual-level and institutional-level measures, but also to explore possible *cross-level* effects between individual measures and institutional units, i.e., does this student characteristic predict the dependent variable differently in different types of institutions? In college effects studies, this would be the equivalent of asking whether an individual level variable such as the student's SAT score predicts an outcome variable such as college GPA differently in different colleges.

In comparing HLM with OLS regression, one of the theoretical advantages of HLM is in the manner in which the statistical significance of aggregate (institutional) effects is tested. The basic problem with using OLS results (when the student is the unit of analysis) for this purpose is that the measures of institutional characteristics are not free to vary *within* institutions, e.g., the institutional "size score" assigned to the students attending any given institution is, by definition, always the same. The error term will therefore be underestimated because the number of degrees of freedom is inflated. In other words, the degrees of freedom in OLS regression are based on the number of *students* when in fact the correct degrees of freedom is presumed to be the number of *institutions*.[4] Under these conditions, then, in using OLS the investigator will tend to commit more Type I errors in assessing the effects of institutional characteristics. While one could guard against such a possibility by decreasing alpha when assessing institutional characteristics, HLM offers a more precise solution because it assesses significance for such variables using a smaller number of degrees of freedom (i.e., based on the number of aggregate units).

Unless the college effects investigator is interested in whether the slopes of the individual-level predictors vary across institutional units (cross-level effects), it is not clear whether HLM offers any other advantages over OLS regression. Some might argue on theoretical grounds that HLM might offer a better overall solution ("fit") to the multivariate data, but this is by no means a given. Over the years, OLS regression has proven to be a very robust method, one that produces a reasonably good fit even when some of its assumptions (e.g., homoscedasticity, linearity) are not satisfied (e.g., Bohrnstedt and Carter 1971; Hanushek and Jackson 1977; Snedecor and Cochran 1967). Moreover, on a purely practical level, the existing HLM 6.0 software package (Raudenbush et al. 2004) does not offer the same options as the SPSS software package (e.g., blocking, forward entry, "Beta in") that make it possible to conduct the kinds of modeling of variables described in the discussion of OLS regression.

Design of the Comparative Analyses

The purpose of the analyses described below was to explore the relative performance of HLM and OLS regression using a large national longitudinal sample of college undergraduates. The principal aim of the analyses was to assess the relative "fit" of the different models derived from the two methods by means of cross-validation. A secondary purpose was to determine the extent to which the two methods might lead to different conclusions regarding the effects of institutional-level variables.

For this purpose, we relied on a longitudinal sample of 8,634 college graduates who entered 229 colleges and universities as freshmen in the fall of 1994 and were followed up ten years later in the summer-fall of 2004. The students were also followed up as they were

[4] It should be pointed out that, in assessing the relationship between an institution-level and a student-level variable (as opposed to two institution-level variables), there are really two different "degrees of freedom": institutional and (the larger) student. Within institutions, all student variation constitutes error variance.

completing college in spring-summer of 1998. Pre-college (input) information was obtained from the 1994 CIRP freshman questionnaire.[5] Information concerning the student's college experiences was obtained from the first follow-up conducted in 1998, and information concerning the dependent variable was obtained from the 2004 post-college follow up. The average number of students per institution was 37.7. Measures of student peer group characteristics were based on the responses of all freshmen to the 1994 entering survey[6] (see Astin et al. 1994 for details), while measures of faculty characteristics were derived from Higher Education Research Institute's (HERI) 1998 national faculty survey (see Sax et al. 1999 for details). For more detail concerning the sampling and weighting, see Astin and Denson (2007).

Variables

The dependent variable selected for this analysis was the student's response in 2004 to a political identification question that had five response options: far left, liberal, middle-of-the-road, conservative, and far right (scored 5, 4, 3, 2, and 1, respectively). The same question had been included in the 1994 and 1998 student surveys as well as in the faculty survey. Since one of the main questions explored in the original study was whether the student's 2004 political identification was affected by the political identification of either the peer group and/or the faculty, we also developed two parallel measures by aggregating the responses to this question of all 1994 freshmen and of all faculty separately for each of the 229 institutions.

To identify entering student and institutional characteristics that contribute to the prediction of the student's 2004 political identification, we first ran a series of correlational analyses using the entire sample of 8,634 students. From these preliminary analyses we identified a total of 24 independent variables, consisting of 20 entering student characteristics and 4 institutional characteristics (including the peer group and faculty characteristics described above).

Significance of Institutional-Level Effects

Our first OLS-HLM comparison involved the coefficients generated by each of the models for institutional-level variables (political orientation of the peer group, faculty political orientation, selectivity, and a dummy variable: Catholic versus non-Catholic): How comparable are they in size? What confidence level does the model assign to the coefficient for each variable? This second question was of particular interest, since it was expected that HLM would generate larger *p*-values than OLS regression.

Cross-Validation Test

For this comparison, the sample was divided randomly into two halves: sample A consisting of 4,317 subjects and sample B consisting of 4,317 subjects. Using only sample A,

[5] Students were surveyed as part of UCLA's Cooperative Institutional Research Program (CIRP) (see Astin et al. 1994). For more details on the sample, see Astin and Denson (2007).

[6] Although the individual students for whom we had longitudinal data represent less than 10 percent of the sample upon which we based our peer group measures, our individual and peer group measures are, strictly speaking, not entirely independent. However, unpublished analyses show that making the peer group and individual-level measures completely independent—i.e., by computing for each student a separate peer group measure that omits that student's score on the variable—has a negligible effect on the results.

we ran an OLS multiple regression analysis with all 24 independent variables using the SPSS "enter" command. This analysis yielded a single equation which included a nonstandardized *b* coefficient for each of the 24 variables in addition to an intercept constant. Using the same sample A, we also ran a parallel HLM analysis. In the HLM analysis, the 20 student-level variables were modeled at Level-1, while the 4 institutional-level variables were modeled at Level-2 with nonrandomly varying slopes. This analysis also yielded a single (combined) equation which included a nonstandardized *b* coefficient for each of the 24 variables as well as an intercept constant. For simplicity, we shall refer to these two equations as the "OLS model" and the "HLM model," respectively.

The prediction equations from the two models were then applied blindly to the second ("cross-validation") sample B in order to compute an expected 2004 political orientation score ($\hat{Y}$) for each sample B subject using the OLS model and a parallel expected score using the HLM model. The sample B subjects were then rank-ordered from highest to lowest separately on each of the two expected scores. Finally, we correlated each of the two expected scores with the student's actual 2004 political orientation score and compared the resulting simple *r*s.

However, an even more meaningful test of the "fit" of each model on the sample B data would be through *classification*: how accurately does the model actually classify subjects into relevant political groupings? For this purpose, we examined the sample A data to determine what percentage of the students in 2004 were either "liberal" or "far left" ("liberals") and what percentage were either "conservative" or "far right" ("conservatives"). These percentages turned out to be 33.0 and 33.9, respectively. Using the sample B ranking based on the OLS formula, we next identified the highest-scoring 33.0 percent ("expected liberals") and lowest-scoring 33.9 percent ("expected conservatives") of the sample B subjects. Similarly, a parallel set of expected liberals and expected conservatives was also identified using sample B rankings based on the HLM model. Finally, the percentages of correctly classified sample B subjects ("hits") derived from each of the two models were compared.

Results

Table 1 compares the coefficients and *p*-values derived from the two models using the total sample of 8,634 subjects. Note that the coefficients produced by the two models are very similar, although there are minor differences. More to the point is that the correlations between expected and actual values in the cross-validation sample (sample B) were identical: .612 (OLS) and .612 (HLM). While the *p*-values (confidence levels) produced by the two models were very similar for each of the 20 entering student characteristics, the *p*-values for the four institution-level variables were, as expected, somewhat larger in the case of the HLM model. While such differences would obviously tend to create more Type I errors when OLS rather than HLM is used, in this particular case the decisions concerning significance would be the same for both models: we would conclude that two variables—peer political orientation and selectivity—have significant ($p < .01$) positive effects on the individual student's political orientation, and that the other two institutional-level variables—Catholic college and faculty political orientation—show no significant effects.[7]

[7] It should be pointed out, however, that the *p* value for Catholic colleges was of near-borderline significance in the OLS model, but not in the HLM model.

Table 1 Comparison of multivariate results for OLS regression and HLM (N = 8,634)

	OLS		HLM	
	b	*p*	*b*	*p*
Entering 1994 freshman characteristics				
Gender: Female	.057	.005	.058	.005
Socioeconomic status	.001	.575	.001	.507
SAT composite	.000	.000	.000	.000
Race: Black/African-American	.134	.012	.139	.014
Race: Asian/Asian American	.061	.196	.070	.142
Race: Other	.102	.135	.093	.170
Discussed politics	.079	.000	.079	.000
Influence political structure	.000	.983	−.002	.899
Influence social values	.003	.789	.003	.843
Plan: Be elected to student office	−.033	.005	−.032	.008
Plan: Take part in students protests	.059	.000	.059	.000
View: Married women should stay at home (reverse coded)	.031	.003	.032	.003
View: Employers can test for drugs (reverse coded)	.046	.000	.046	.000
View: Racial discrimination no longer a problem (reverse coded)	.037	.004	.036	.005
View: De-emphasize college sports	.055	.000	.055	.000
Economic Liberalism	.027	.000	.027	.000
Social Liberalism	.046	.000	.047	.000
Liberal views on crime and punishment	.040	.000	.039	.000
Political orientation (pretest)	.246	.000	.245	.000
Attended religious services	−.040	.014	−.042	.009
College-Level Characteristics				
Catholic college	.049	.052	.053	.123
Selectivity	.000	.001	.000	.009
Faculty political orientation	.054	.283	.052	.408
Peer political orientation	.262	.000	.252	.001
Intercept	−1.673	.000	−1.609	.000

Note: Dependent variable is the subject's political orientation in 2004 (5 = far left, 4 = liberal, 3 = middle-of-the-road, 2 = conservative, 5 = far right)

We turn now to consider the results of the cross-validation test. (Note that "validation" in this case refers to the accuracy with which the freshman formulae can classify students according to their political preferences ten years later, *not* to the "validity" of the weights assigned to each independent variable.) The simple r^2 between the students' expected and actual political orientations in sample B is .375, which represents only 1% shrinkage from the original multiple R^2 of .385 obtained from sample A. Table 2 compares the rates of correct classifications using the OLS and HLM models. When it comes to overall accuracy in identifying both liberals and conservatives, the OLS model is actually slightly better—but not significantly so—than the HLM model: 65.6 percent versus 65.4 percent, respectively. This suggests that both models achieve nearly 100% improvement over chance accuracy: one would expect about 1/3 correct "hits" by chance, compared to nearly 2/3 correct hits using either the OLS or HLM formula. HLM is slightly better at identifying

Table 2 Relative accuracy of OLS and HLM in classifying subjects in the cross-validation sample ($N = 3{,}178$)

2004 Political orientation[a]	*N*	% of total	Percent correct Classification using	
			OLS	HLM
Conservatives (including far right)	1,079	33.9	65.4	65.6
Liberals (including far left)	1,048	33.0	65.7	65.1
Total conservatives & liberals	2,127	66.9	65.6	65.4

[a] Subjects were predicted to be "liberals" if their estimated score was ranked in the top 33.0% (>3.204 for OLS and >3.188 for HLM) and as "conservatives" if their estimated score was ranked in the bottom 33.9% (<2.751 for OLS and <2.739 for HLM)

conservatives (65.6% versus 65.4% accuracy), while OLS is slightly better at identifying liberals (65.7% versus 65.1% accuracy). These differences are, of course, trivial. In short, these findings show clearly that, when it comes to overall "fit" with the multivariate data, there is no difference between OLS and HLM: *both methods yielded essentially the same results.*

What else can be learned in this particular example by using the "modeling" capabilities of the SPSS Regression software package? To explore this question, we re-ran the OLS regression analysis with the total sample using two sets of possible "mediating" variables: college experiences (which were added in a fourth block following the college characteristics variables) and post-college experiences (which were added in a fifth and final block). The college experiences included major field of study, co-curricular activities, and selected curricular variables (e.g., women's studies courses, ethnic studies courses, etc.). The post-college experiences included post-graduate degree attainment, marriage, and several factorially-derived "lifestyle" measures (e.g., community involvement, religious involvement, etc.). Blocks were entered according to the presumed temporal order of occurrence. Within each block, variables were entered *one at a time* according to the size of their contribution to reducing the residual sum of squares in the dependent variable, to the point where no remaining variable was capable of producing a significant ($p < .01$) additional reduction. This approach not only tends to produce the most parsimonious solution within any block (i.e., the fewest number of variables entered), but also permits the investigator to track step-by-step changes in the Beta coefficients associated with the entry of each individual variable.

While these analyses yielded a number of interesting findings (see Astin and Denson 2007), for the purposes of this paper we shall focus on possible mediators of the institutional-level effects. In the case of the political orientation of the peer group, there were two mediating variables that produced a reduction in the partial Beta of at least .02: attending religious services during college (change in partial Beta from .12 to .10), and the post-college religious involvement factor (change from .08 to .06). (This latter factor was a composite that combined attendance (both frequency and hours/week) at religious services/meetings with donating money to a religious organization). Both of these mediating variables have negative coefficients, suggesting that they are associated with a conservative political identification.

The findings suggest some interesting possibilities. Since it appears that the student's political identification six years after completing college has been shaped in part by the college peer group—students whose peer group leans in a conservative direction tend to

become more conservative, while those whose peers lean in a liberal direction tend to become more liberal—it would appear that part of this effect occurs because of how the peer group shapes religious activity. That is, colleges with conservative student bodies appear to encourage religious activity, which in turn reinforces a conservative political identification, whereas colleges with liberal student bodies tend to reinforce a liberal identification in part because they discourage religious engagement. Why the political orientation of the peer group affects religious engagement in this fashion would appear to be an interesting topic for future research.

Since the partial Beta for the attendance at religious services during college shrinks to nonsignificance (from $-.10$ to $-.02$) when post-college religious engagement enters the equation, it appears that the effect of religious engagement during college is entirely indirect. That is, attendance at religious services during college affects the student's post-college political identification *only* because it affects religious engagement after college. However, since peer political orientation retains a significant partial Beta through the final step in the analysis, its effect on post-college political identification appears to be both direct and indirect.

One more of the many other uses to which the SPSS Regression step-by-step options can be utilized merits brief mention here. By tracking the Beta changes from step to step, it becomes possible to "decompose" the simple correlation between the dependent variable and *any* independent variable. Take, for example, socioeconomic status (SES), which has a small but highly significant ($p < .001$) positive correlation ($r = .09$). with post-college political identification. The individual step-by-step results show that the reasons why high-SES students lean slightly in a liberal direction six years after completing college are, in descending order of importance, that they tend to (a) obtain high scores on college admissions tests; (b) embrace liberal views on social issues when they start college; and (c) enter college with a slightly more liberal than conservative political identification. Once these three variables are taken into account, SES no longer shows a significant relationship with post-college political identification. One could, if desired, do a similar analysis with race, gender, or any other independent variable that is significantly correlated with post-college political orientation.

We would like to conclude by raising a few points of caution about such uses of multivariate analyses, whether it be HLM, OLS or some other method. When independent variables within a block are correlated with each other (which is almost always the case), the fact that one variable enters and another does not should not necessarily be taken as proof either (a) that the entered one is "important" and the non-entered one is not or (b) that the entered one rather than the non-entered one is causally related to the dependent variable. In such a case it is essential first to examine the "Beta in" for both variables at the step prior to the entry of the first variable. If the two coefficients are similar in magnitude, and if the entry of the first variable causes the "Beta in" for the second variable to become non-significant, then there is really no basis for assuming that one variable is more important than the other since the difference in their "Beta in's" could well be a chance difference that might well reverse itself in an independent sample.

On the matter of making causal inferences (which would include the possible mediating role of a variable), there are two additional considerations. First, is the variable in question clearly antecedent to the dependent variable in its temporal order of occurrence? If not—that is, if the dependent variable could arguably be antecedent to, or coincident with, the independent variable—then this ambiguity should be clearly noted and the causal conclusions appropriately tempered or qualified. Second, there is the problem of the "omitted variable": Is (are) there unmeasured variable(s) that could plausibly account for the

observed "effect" of the variable in question? Could the observed variable be serving as a proxy for the omitted variable(s)?

Finally, it should be stressed that these inferential issues are inherent in any multivariate study, no matter which theoretical position the investigator assumes and no matter which multivariate statistical method is used.

Issues with Stepwise Analysis

Since this paper has discussed some of the "modeling" benefits of multivariate analyses that use a stepwise method for adding independent variables to the model, we should also acknowledge that some investigators (e.g., Thompson 1995) believe that the stepwise approach has serious limitations. Let us briefly comment on the three criticisms of stepwise procedures put forward by Thompson (1995).

The first criticism has to do with the fact that most computer programs assume that the number of degrees of freedom in the numerator of the F-ratio used for testing the statistical significance of the regression equation should be determined by the number of variables in the regression equation. However, since forward stepwise analysis examines *all* non-entered independent variables before entering each variable, the entered variables do not represent a random sample of variables, thereby increasing the likelihood of Type I errors. Thompson recommends inflating the degrees of freedom used in the numerator to equal the total number of entered *and* nonentered independent variables and reducing the degrees of freedom used in the denominator by the same number.

We believe that this "correction" errs in the other direction (Type II errors) because it deflates the F-ratio excessively by treating the entire sum of squares associated with non-entered variables as "error" (i.e., leaving it in the denominator). A better approach would be to fit the dependent variable to the entire set of independent variables and test the resulting equation for statistical significance. If this equation proves to be significant, then it is reasonable to assume that smaller equations composed of subsets of the "best" variables would also be significant.

While there may not be a precise solution to this problem, given the nonrandom nature of the variable entry process, perhaps the best protection against Type I errors is to use a stringent p value for entry of independent variables. With stepwise analyses, the thing to keep in mind is that *every* entered variable must satisfy that p value before it is entered. Non-stepwise approaches, on the other hand, enter all variables at once, a procedure which entails certain risks if the sample is relatively small and the number of independent variables relatively large (see below). Of course, as Thompson himself acknowledges, with very large samples, the question of whether the stepwise equation is statistically significant assumes less importance.

Thompson's second criticism is that stepwise regression should not be used to identify "the best subset of n predictors" within a set of independent variables, and we agree. A better approach would be to compute separate equations for each possible subset of n variables and compare the results.

Thompson's third criticism reflects a common misconception regarding stepwise regression that has become a part of the currently accepted folklore in methodological circles: that stepwise regression is flawed because it "capitalizes on sampling error." However, one might as well say that *all* multivariate methods are flawed because they "capitalize on sampling error." Thompson elaborates this argument by pointing out that a

variable might be "incorrectly" selected at a given step because of sampling error. Such an error might, in turn, result in another variable being "incorrectly" selected, and so on.

While this argument about the effects of sampling error is literally true, it is no less true of *every* multivariate procedure. To understand why, it is important first to understand that the only basic difference between the stepwise approach and the non-stepwise approach is that in the latter instance the variables are entered *all at once.* Sampling error, in short, *does not depend on how variables are entered.* The sampling error that causes a variable to be "incorrectly" entered in a stepwise analysis will cause that same variable to have an inflated regression coefficient in a non-stepwise analysis. And the same error that causes a subsequent "wrong" variable to be entered will distort that same variable's partial regression coefficient in a non-stepwise analysis. Sampling error, in other words, does not go away merely because we choose to enter our independent variables all at once instead of in a stepwise fashion.

As a matter of fact, under certain circumstances the stepwise approach could well minimize the untoward effects of sampling error in comparison to the non-stepwise approach. Thus, when using relatively small samples and relatively large numbers of independent variables, the non-stepwise approach maximizes the risks of "capitalizing" on sampling error because it tries to fit the dependent variable to *all* independent variables simultaneously. These risks increase exponentially as the number of independent variables increases. By contrast, in the stepwise approach, the significance test that must be satisfied before any variable can be entered provides at least some protection against sampling errors because it minimizes the number of variables entered. For example, if there are 20 independent variables in a data set, stepwise results are initially affected by sampling errors in only 20 coefficients, whereas non-stepwise results are affected by sampling errors in 210 coefficients! (These two numbers gradually converge, of course, as variables are added to the stepwise solution.)

It is true, of course, that with small samples there are risks associated with arguing that particular entered variables are "more important" than particular non-entered variables (criticism # 2, above), but the SPSS stepwise regression routine offers certain protections against such problems, provided that investigators choose to employ these protections (our experience is that most investigators don't, unfortunately). We are speaking here of the "Beta in" statistic. By comparing the "Beta in" coefficients associated with the competing variables at the step *immediately prior* to the step where the first competing variable enters, we can determine whether or not the respective Beta ins are substantially different. If the differences are trivial or nonsignificant, then one cannot legitimately claim that the entered variable is "more important" than the non-entered variable(s).

The "capitalizing on error" argument has led some researchers to conclude that the stepwise approach yields "results that are not replicable" (Thompson 1995, p. 525). That this claim is not valid, at least when it comes to large samples, is aptly illustrated by a recent empirical study of two four-year longitudinal samples of undergraduates from the 1980s and 1990s, nearly a decade apart (Astin et al. 2002). The study looked at 36 different dependent variables. Stepwise regression weights derived from the 1980s sample were applied blindly to the 1990s sample. When compared to the 1980s results, the "cross-validated" (i.e., 1990s) multiple Rs showed almost no shrinkage over the nine-year period! Thus, for the 36 different regression equations computed from the 1980s, the median shrinkage in the 1990s was about 1 percent of the variance, and there were actually 12 regressions that yielded slightly *larger* cross-validated multiple Rs in the 1990s! To us this makes it clear that stepwise analyses can produce solutions that are highly replicable.

Discussion

Do these findings provide any guidelines as to which method is to be preferred in which situation? Clearly, if an investigator is interested in temporal "modeling" of the variables in a path analytic fashion, the current HLM software is relatively cumbersome as variables must be modeled in an incremental fashion.[8] The advantage of using the SPSS software and the "Beta in" feature is that the path coefficients for all the independent variables are routinely produced in the SPSS output.[9] But, if one were to use the SPSS software for this purpose, what about the increased risk of Type I errors when conclusions are drawn with respect to institutional-level variables? At a minimum, we would recommend assigning a smaller *p*-value to determine statistical significance for such variables (our experience so far with analyses involving several different dependent variables suggests that a rough guideline would be to use a *p*-value for institutional variables that is half the size of the *p*-value used for individual variables, e.g., .005 rather than .01). Indeed, Type I errors can have quite serious implications if the results are used to inform policy. But there is a more precise solution that we would recommend here: take all of the independent variables from any OLS regression solution where "significant" institutional-level effects were found and re-run the analysis with HLM, modeling the student-level variables at Level-1 and the institutional-level variables at Level-2 with nonrandomly varying slopes. The *p*-values resulting from such an HLM analysis will presumably yield a more valid test of the statistical significance of institutional-level effects.

What if the investigator is interested in exploring possible cross-level effects (e.g., how does the institutional-level variable affect the relationship between a student-level variable and the dependent variable)? In the study reported here, with 20 student-level variables and four institutional-level variables, there are 20 × 4 or 80 possible cross-level effects that could be explored. In this case, using the HLM software would be preferable since it was specifically designed for such purposes and has a number of other features such as computing separate OLS regressions for *each* institution. In such situations, one would ideally have relatively large numbers of students per institution—certainly larger than our mean figure of about 40 students. If the investigator wants to use OLS regression and the SPSS software instead, he can compute interaction "terms" involving non-additive *combinations* of individual and institutional variables (e.g., the product of institutional selectivity and the individual student's SAT score). Note that such terms should be entered into the regression only *after* the independent simple effects of the same variables have been controlled (just as in analysis of variance one must first control for main effects before interaction effects can be examined).

It has been our experience that increasing numbers of editorial reviewers for scholarly journals are now routinely recommending that HLM rather than OLS regression be used whenever a study using individual and institutional data is submitted. Such a recommendation would seem reasonable *only* under conditions where the investigator is (a) *not* interested either in temporal modeling of the variables or in identifying the direct and indirect effects of particular independent variables, or (b) wishes to explore possible cross-level effects involving individual and institutional variables. OLS regression can, of

[8] See Krull and MacKinnon (2001) for a detailed explanation of mediation in multilevel modeling and Raudenbush and Bryk (2002) for a detailed explanation of mediation of latent variables in multilevel modeling.

[9] There is, of course, no reason why the HLM software could not be modified to include the same modeling capabilities.

course, accommodate interaction terms comprising nonlinear combinations of individual and institutional variables (although it has to be computed manually). Moreover, since the results of this study show that the multivariate solutions produced by OLS regression "fit" the data every bit as well as the solutions produced by HLM, requiring authors to use HLM rather than OLS regression would not seem to be a reasonable demand. (We should point out, however, that since this is but one empirical example, our findings and conclusions might be somewhat different with a different outcome or a different kind of database). Indeed, if the investigator wishes to "model" the independent variables temporally or to explore their mediating effects by means of path analysis, the SPSS Regression software program would be the preferred method. We hasten to add, however, that in the case where "significant" institutional-level effects are found using OLS regression, it would be prudent either to (a) lower the p value required for significance or—preferably—(b) rerun the analysis using HLM in order to obtain the more conservative p-values for the institutional-level variables.

Acknowledgement The authors wish to thank Michael H. Seltzer for his feedback and invaluable assistance in the writing of this manuscript.

References

Astin, A. W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. New York: Macmillan/Oryx.

Astin, A. W., & Denson, N. (2007). *Long-term effects of college on students' political orientation*. Los Angeles, CA: Higher Education Research Institute, UCLA.

Astin, A. W., & Dey, E. (1996). *Causal analytical modeling via blocked regression analysis (CAMBRA): An introduction with examples*. Los Angeles: Higher Education Research Institute, UCLA.

Astin, A. W., Keup, J. R., & Lindholm, J. A. (2002). A decade of changes in undergraduate education: A national study of system "transformation". *The Review of Higher Education, 25*(2), 141–162.

Astin, A. W., Korn, W. S., Sax, L. J., & Mahoney, K. M. (1994). *The American freshman: National norms for fall 1994*. Los Angeles: Higher Education Research Institute, UCLA.

Bohrnstedt, G. W., & Carter, T. M. (1971). Robustness in regression analysis. In H. L. Costner (Ed.), *Sociological methodology 1971* (pp. 118–146). San Francisco: Jossey-Bass.

Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes estimation of coefficients in the general linear model from data of deficient rank. *Psychometrika, 489*, 171–181.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 8). Washington, DC: American Educational Research Association.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis (Occasional Paper)*. Stanford, CA: Stanford Evaluation Consortium, Stanford University.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11*, 57–85.

Ethington, C. A. (1997). A hierarchical linear modeling approach to studying college effects. *Higher Education: Handbook of Theory and Research, 12*, 165–194.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43–56.

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. New York: Academic Press.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36*, 249–277.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1–17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Sax, L. J., Astin, A. W., Korn, W. S., & Gilmartin, S. K. (1999). *The American college teacher: National norms for the 1998–1999 HERI faculty survey*. Los Angeles: Higher Education Research Institute, UCLA.

Seltzer, M. H. (1995). Furthering our understanding of the effects of educational programs via a slopes-as-outcomes framework. *Educational Evaluation and Policy Analysis, 17*, 295–304.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames, IA: The Iowa State University Press.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines proposal. *Educational and Psychological Measurement, 55*(4), 525–534.