6-2002

Modeling Student Outcomes in a General Education Course with Hierarchical Linear Models

Christine E. DeMars
James Madison University, demarsce@jmu.edu

Recommended Citation

DeMars, C., & Erwin, T. D. (2001, June). Applications of item response theory in higher education. Paper presented at the Assessment Conference of the American Association for Higher Education, Denver.

Modeling Student Outcomes in a General Education Course with Hierarchical Linear Models

Christine E. DeMars

James Madison University

DeMars, C. (2002, June). *Modeling student outcomes in a general education course with hierarchical linear models (HLM)*. Paper presented at the annual meeting of the Association for Institutional Research, Toronto.

Abstract

When students are nested within course sections, the assumption of independence of residuals is unlikely to be met, unless the course section is explicitly included in the model. Hierarchical linear modeling (HLM) allows for modeling the course section as a random effect, leading to more accurate standard errors. In this study, students chose one of four themes for a communications course, with multiple sections and instructors within each theme. HLM was used to test for differences by theme in scores on a final exam; the differences were not significant when SAT scores were controlled.

Modeling Student Outcomes in a General Education Course with Hierarchical Linear Models

Traditional analysis of variance and regression procedures have an assumption of independence of residuals. If this assumption is not met, the standard errors of the parameter estimates (and p-values) will be incorrect (Stevens, 1996, Ch. 6, Raudenbush & Bryk, 1988/89), and the parameter estimates themselves may or may not be affected greatly. In institutional research, research on learning outcomes may involve students who were grouped into course sections. Within each section, group interactions and the instructor will influence learning in somewhat different ways. If multiple universities or colleges are studied, the institutions will influence student learning. If these group influences are ignored, their effects will become part of the residual term, thus leading to dependent residuals.

In this study, students were grouped into 67 course sections of a required communications course. Each section followed one of four themes, but all sections addressed the same learning objectives and used the same final exam. The themes were intended to spark student interest and provide continuity across a package of three courses covering communications, writing, and critical thinking for first-year students (students chose one theme for all three courses). The research question was whether student learning, measured by performance on a final exam, differed by theme. This question was important on an institutional level because all students were to have an opportunity to achieve the same general education objectives, regardless of the specific courses they chose. If students from one of the theme areas consistently scored lower on a test covering these objectives, it could be an indication program changes were needed to be sure the curriculum was covered. Scores from students in the same course section were not independent, so hierarchical linear modeling (HLM) was used to model section as a random effect.

HLM is a technique for modeling multilevel data when observations at lower levels are nested within observations at higher levels. In this study, students were nested within course

sections. Multilevel models allow for modeling factors measured at the student level, such as gender, and at the course section level, such as theme, as well as interactions between effects at different levels (Bryk & Raudenbush, 1992, Ch. 1). These models address the violation of the independence assumption by modeling the course section as a random effect, so that course section effects are removed from the residual. When effects are treated as random, not fixed, parameter estimates are generally not obtained for each level of the factor (course section), though they could be. The focus is instead on estimating the variance of the factor.

HLM and similar multilevel models are not the only way to analyze grouped data. One approach to this problem is to conduct the analysis at the group level (Stevens, 1996, Ch. 6); the course sections could be treated as the observations with the mean score as the outcome. The resulting standard errors will be correct, but only group-level factors can be modeled; student-level covariates can not be included, potentially reducing power (Raudenbush & Bryk, 1988/89). For example, the effect of a section with a high or low percentage of females could be included, but there would be no way to estimate the average gender difference within a section. An alternate approach focuses only on the student level, but uses special procedures to obtain the correct standard errors (Thomas & Heck, 2001). With this approach, student-level factors but not group-level factors can be modeled. Multilevel models take into account factors at both levels.

The purpose of the proposed study was to examine differences between themes in student performance. Gender and verbal SAT scores were controlled. Gender and SAT score were studied at the student level, and section theme and mean verbal SAT score were studied at the section level. Including SAT at both levels allowed for separating the effect of taking a course with classmates of high SAT from the effect of SAT score within a course section.

Method

Participants

Participants were all first-year students who completed the final exam in a required communications course in fall 2000 at a Mid-Atlantic public university, except those who enrolled in honors sections. There were 67 sections; 23 used theme A (Effective Arguments), 28 used theme B (Critical Skills in the World of Business), 9 used theme C (Critical Skills and Historical Inquiry), and 7 used theme D (Media Literacy and Communication). Complete data was available for 1304 students.

Instrument

The assessment instrument was a 100-item multiple choice test with a coefficient alpha reliability of .78. Interpersonal communications, group interactions, and public speaking were covered. Some items included video or audio clips; computers equipped with headphones, and separated by carrels, were used to minimize distractions. The assessment was administered in a computer lab during each section's regularly scheduled exam time, with the instructors present. Multiple sections took the test at the same time, but no other students were allowed in the computer lab during testing.

Results - Multilevel Model

About 81% of the variance in scores was due to differences among students within sections, and about 19% of the variance was due to differences in sections. Some of this 19% could potentially be explained by theme. When no control variables were in the model, theme did account for about 14% of the between-section variance, a statistically significant amount ($p = .011$). In other words, there were significant differences among themes, which might indicate they were differentially effective. However, when SAT verbal scores and gender were included in the model,

theme essentially accounted for none of the variance[1]. The themes did not differ when SAT verbal scores were controlled, so student outcomes did not depend on the section theme.

The model with all of the predictors (theme, gender, verbal SAT) scores, then, was:

Level 1 (within-sections, or student level)

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{1ij}) + \beta_{2j}(X_{2ij}) + r_{ij}, \qquad (1)$$

where

Y_{ij} was the test score for student i in section j,

X_{1ij} was the verbal SAT score for student i in section j (centered around the grand mean),

X_{2ij} = 0 for females, 1 for males, centered around the proportion of males (after centering, the codes were: male = 0.60, female = -0.40), and

r_{ij} was the error term (difference between observed and predicted score).

The first subscript of each X identifies the factor it is associated with (1 for SAT, 2 for gender), and the i,j connect it with student i in section j. The first subscript of each β identifies the X it is associated with (or 0 for the intercept), and the second subscript (j) links it to section j.

Gender was centered such that the course section intercepts were predicted means after adjusting the section's gender mix to reflect the average across sections (because in a section with an average mix of students, the average gender group would equal zero), and similarly for verbal SAT scores. In HLM, the intercepts for individual sections are not estimated, but their variance is. The variance terms discussed in the results are interpreted in the context of the predictors in the model and their centering (see Bryk and Raudenbush, 1992, pp. 25-29, for a discussion of centering issues). Although the intercept is dependent on the choice of centering, the size of the effects would be the same regardless of whether or not the variables were centered.

[1] The between-section variance actually increased when theme was added to the model with verbal scores and gender. While this can not occur with least-squares regression, it is possible with maximum-likelihood regression, and can be

Level 2 (between-sections) (2)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(W_{1j}) + \gamma_{01}(W_{2j}) + \gamma_{01}(W_{3j}) + \gamma_{01}(W_{4j}) + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

$$\beta_{2j} = \gamma_{20},$$

where

γ_{00} was the grand mean (across sections and students),

W_{1j} was the average verbal SAT for section j, centered around the grand mean verbal score,

W_{2j}, W_{3j}, and W_{4j} were dummy codes for the themes, with W_{2j} coded 1 if section j was theme A, W_{3j} coded 1 for theme B, and W_{4j} coded 1 for theme C (each W coded zero otherwise), grand-mean centered,

and the u_{0j} was a random section effect.

The first subscript on the γ link it with a particular β, and the second subscript identifies which section predictor variable the γ is associated with (or 0 for the intercept). Only the section intercepts (average score) were freed to vary; the SAT and gender effects were constrained to be the same in each section. With this relatively small number of sections, stability would be decreased by estimated additional parameters, and the variance in SAT slopes and gender effects were not central to the research question. The centering of the SAT and theme variables meant that the intercept $\gamma 10$ was the grand mean. Because the themes were coded 0/1 before centering, each of the parameters for theme would show the difference between the associated theme coded 1 and theme E, which was always coded 0 (though the focus was on the test of all three theme parameters simultaneously, parallel to the omnibus F-test in ANOVA, not on the individual contrasts). The coefficients are shown in Table 1.

---

interpreted as zero additional explained variance.

Table 1

Multilevel model

Within-Section (Student Level) Effects	Coefficient	standard error	p-value
Verbal SAT (β_{1j})	0.053	0.004	<.001
Gender (β_{2j})	-2.982	0.417	<.001
Between-Section Effects			
Intercept (γ_{00})	72.787	0.390	<.001
Section Average SAT (γ_{01})	0.024	0.018	.187
theme A dummy code (γ_{02})	-0.527	0.799	.509
theme B dummy code (γ_{03})	-0.879	0.892	.324
theme C dummy code (γ_{04})	-1.983	1.145	.083

The effect of SAT scores can be summarized as follows: with theme controlled, for every 100 point increase in the class mean SAT verbal, the class mean final exam score increased by 2.4 points (on a 100 point scale). Within a class section, for every 100 point increase in a student's SAT verbal, the student's final exam score increased above the class average by 5.3 points (0.64 standard deviations), controlling for gender. Gender was also a significant predictor of final exam scores, though controlling gender did not affect the differences among themes. When SAT scores were controlled, men scored an average of 3.0 points (0.36 standard deviations) lower than women in the same section.

Results - Course Section as the Unit of Analysis

Because the research question was primarily at the course section level, analyses were also conducted using solely the course section data and ordinary least-squares regression. For this analysis, all of the variance in the model is among sections. When SAT scores were not controlled, course theme accounted for about 17% of the variance among sections (R^2=.168, p=.008), roughly

similar to the 14% seen at this level in the multilevel model. When the section mean SAT score was controlled, course theme accounted for only about 3.3% (change in R^2=.033, p = .380). The primary difference between the models was in the estimated regression slope for verbal SAT. For each 100-point increase in the average SAT within the section, the average exam score in the section was predicted to increase by about 7.1 points, roughly the sum of the section-level and student-level effects in the multilevel model.

The full model for the course-section analysis was:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_4 X_{4j} + e_j$$

where

Y_j was the mean test score for section j,

X_{1j} was the mean SAT verbal score for section j (centered around the grand mean),

X_{2j}, X_{3j}, and X_{4j} were dummy codes for the themes, coded as for the multilevel model,

and e_j was the residual for section j.

The coefficients are shown in Table 2.

Table 2

Course Section Model

	Coefficient	standard error	p-value
Intercept (β_0)	72.537	0.392	<.001
Section Average SAT (β_1)	0.071	0.018	<.001
theme A dummy code (β_2)	-0.555	1.400	.693
theme B dummy code (β_3)	-1.819	1.493	.228
theme C dummy code (β_4)	-2.044	1.620	.212

Results - Student as the Unit of Analysis

The data could also be analyzed at a student level, ignoring the fact that the students are nested within course sections, though the standard errors would be incorrect. When the data were analyzed this way, theme explained 3.5% of the variance ($p<.0001$). When verbal SAT was controlled, theme explained only 1.1% of the variance, which was statistically significant ($F_{3,1299}=6.05$, $p=.0004$) given the (incorrectly) large degrees of freedom. In other words, because the analysis was conducted as if there were many more independent pieces of information than there were, the error term was artificially small, and, from the results of the statistical tests, the conclusion would have been that scores from the themes differed even after controlling verbal SAT scores. This is somewhat mitigated because many educators would consider 1% of the variance not enough to be meaningful, even though it was statistically significant.

The coefficient for verbal SAT was 0.055; for every 100 point increase in verbal score, the test score was predicted to increase by 5.5 points (0.66 standard deviations). This was very similar to the SAT effect in the student level within the multilevel model. Male students were predicted to score 3.0 points lower than female students, identical to the difference in the multilevel model. Essentially, the results (effects and their standard errors) for the variables measured at the student level were similar to the results from the multilevel model[2], but the results from the theme variable, measured at the section level but treated as if it were a student-level variable, had incorrect standard errors (the significance level for theme was too small--this can also be seen by noting that the standard errors for the dummy variables were too small).

The full model for the student analysis was:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + e_i$$

[2] In many situations where student is used as the level of analysis, the standard errors will be larger than those from the multilevel model, though the coefficients should be similar. In this study, the standard errors were about the same because the gender and verbal skills effects were not allowed to randomly vary across course sections.

where

Y_i was the test score for student i,

X_{1i} was the SAT verbal score for student i (centered around the mean),

X_{2i} was gender for student i (centered around the mean),

X_{3i}, X_{4i}, and X_{5i} were dummy codes for the themes, coded as for the multilevel model,

and e_i was the residual for student i.

The coefficients are shown in Table 3.

Table 3

Student Model

	Coefficient	standard error	p-value
Intercept (β_0)	72.797	0.204	<.0001
SAT (β_1)	0.056	0.003	<.0001
Gender	-2.977	0.428	<.0001
theme A dummy code (β_2)	-0.824	0.693	.235
theme B dummy code (β_3)	-1.612	0.689	.020
theme C dummy code (β_4)	-2.168	0.818	.008

Replication Study

The multilevel study was repeated using data from fall 2001. This time, less of the variance, 7% compared to 19% in fall 2000, was due to differences in sections. However, as in 2000, when no control variables were in the model, theme accounted for a statistically significant ($p = .003$) amount of the between-section variance (about 20%). In contrast to the 2000 findings, though, when SAT verbal scores and gender were included in the model, the themes were still statistically different ($p = .034$). The section-average SAT had almost no effect on the mean score for the

section, so controlling it may little difference. Within sections, SAT and gender had similar effects as in the previous year (for every 100 point increase in verbal SAT, the student's score increased by 5.7 points, and males scored 3.5 points lower than females with the same SAT).

Summary and Conclusions

Selected results from the three analyses of the main study are shown in Table 4. In brief, the results with section as the unit of analysis were similar to the multilevel results, but the verbal skills effect could not be partitioned into within- and between-sections effects, and the gender effect could not be modeled. The results with students as the unit of analysis had similar gender and verbal skills effects as the student-level in the multilevel analysis, but the course section average SAT effect lost was lost, and the standard error was artificially low for theme, leading to different conclusions about the effect of course theme on performance (if the significance test was the main piece of information used in drawing conclusions).

In higher education, students are often nested within courses or sections or dorms or other groups. Research conducted by instructional faculty may involve only one of these groups, but research conducted by institutional researchers often involves a larger sample across a number of groups representing some facet of the university. Multilevel data requires the use of techniques which take into account the structure of the data. Multilevel models, such as HLM, allow for designs with factors at more than one level. In this study, HLM was used to test for differences in communications scores by course theme, after controlling both section-level and student-level variables. The themes did not vary in learning outcomes when SAT verbal scores were controlled. This type of analysis illustrates how institutional researchers can use multilevel techniques in comparing program effectiveness, a task which is increasingly required, particularly in state-supported universities where accountability is emphasized.

Table 4

Summary of Results

Percent of variance explained by theme

	before controlling SAT and gender		after controlling SAT and gender	
	% of total variance	% of between-section var.	% of total variance	% of between-section var.
multilevel analysis	2.6	14	0	0
Section as unit of analysis	--	17	--	3.3
Student as unit of analysis	3.5	--	1.1	--

Slope of Verbal SAT score (controlling theme and gender)

	Within-Sections		Between-Sections	
	coefficient	standard error	coefficient	standard error
multilevel analysis	0.053	0.004	0.024	0.018
Section as unit of analysis	--	--	0.071	0.018
Student as unit of analysis	0.056	0.003	--	--

References

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

Raudenbush, S. W., & Bryk, A. S. (1988/89). Methodological advances in analyzing the effects of schools and classrooms on student learning. Review of Research in Education, 15, 423-475.

Stevens, J. (1996). Applied multivariate statistics for the social sciences, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. Research in Higher Education, 42, 517-540.