DOCUMENT RESUME

ED 464 942 TM 033 873

AUTHOR Schulz, E. Matthew; Betebenner, Damian; Ahn, Meeyeon
TITLE A Comparison of Hierarchical and Nonhierarchical Logistic Regression for Estimating Cutoff Scores in Course Placement.
PUB DATE 2002-04-00
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Algebra; *College Students; Comparative Analysis; *Cutting Scores; Estimation (Mathematics); Higher Education; *Regression (Statistics); *Sample Size
IDENTIFIERS *Hierarchical Analysis

ABSTRACT

This study was performed to determine whether hierarchical logistic regression models could reduce the sample size requirements of ordinary (nonhierarchical) logistic regression models. Data from courses with varying class size were randomly partitioned into two halves per course. Grades of students in college algebra courses were obtained from 40 colleges. The largest sample size group, Group 4, contained 11 colleges with half counts ranging from 171 to 563 (average 307). Nonhierarchical and hierarchical analyses were performed on each half. Compared to their nonhierarchical counterparts, hierarchically estimated cutoff scores from different halves were closer together in value and predicted course outcomes in the other half more accurately. These differences were most pronounced with small samples. It is concluded that the sample size requirements could be substantially reduced if hierarchical logistic regression were used to estimate cutoff scores. (Contains 2 tables, 7 figures, and 28 references.) (SLD)

A Comparison of Hierarchical and Nonhierarchical Logistic Regression for Estimating Cutoff Scores in Course Placement

E. Matthew Schulz
ACT, Inc.

Damian Betebenner
University of Colorado

Meeyeon Ahn
University of Iowa

2255 North Dubuque Road
P. O. Box 168
Iowa City, IA 52243-0168
(319) 337-1468

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 2-6, 2002.

A Comparison of Hierarchical and Nonhierarchical Logistic Regression for Estimating Cutoff Scores in Course Placement

Abstract

This study was performed to determine whether hierarchical logistic regression models could reduce the sample size requirements of ordinary (nonhierarchical) logistic regression models. Data from courses with varying class size were randomly partitioned into two halves per course. Nonhierarchical and hierarchical analyses were performed on each half. Compared to their nonhierarchical counterparts, hierarchically estimated cutoff scores from different halves were closer together in value and predicted course outcomes in the other half more accurately. These differences were most pronounced with small samples. We conclude that the sample size requirements could be substantially reduced if hierarchical logistic regression were used to estimate cutoff scores.

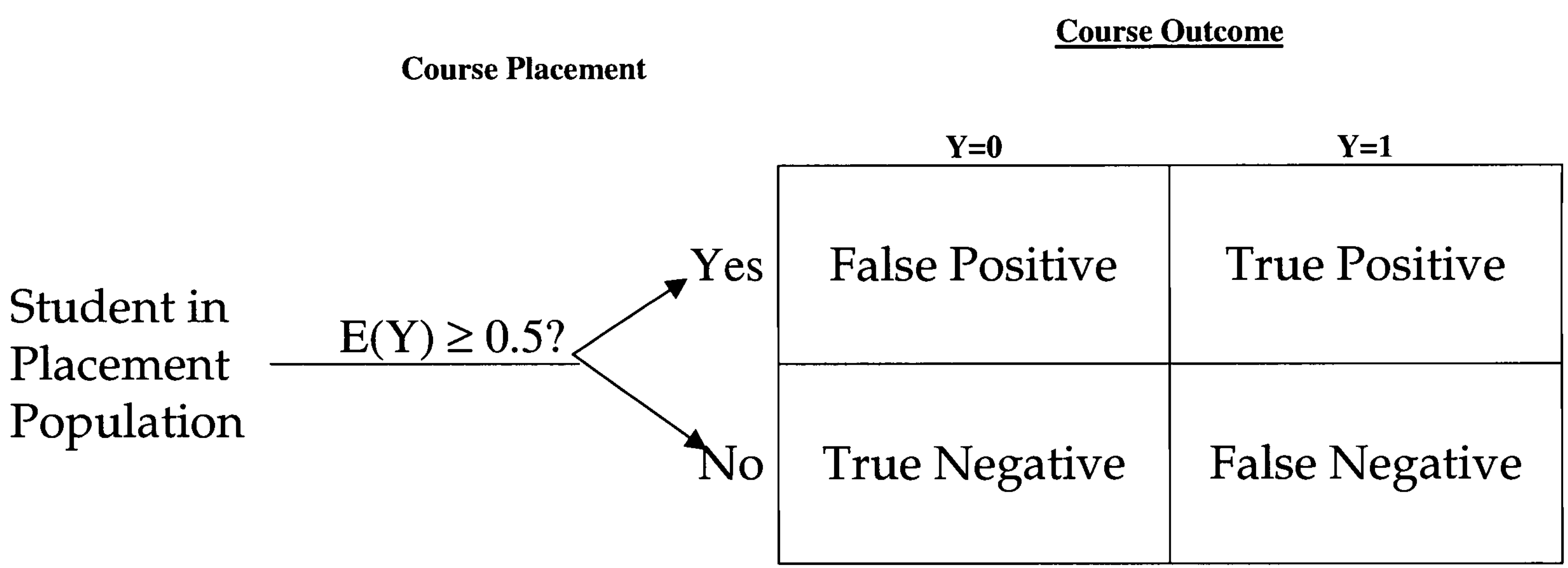
Key Words: Hierarchical Logistic Regression, Accuracy Rates, Cutoff scores, Course Placement, Cross Validity

A Comparison of Hierarchical and Nonhierarchical Logistic Regression for Estimating Cutoff Scores in Course Placement

Course placement is an important area of decision-making at postsecondary institutions. By the end of the 1980s a large majority of U.S. colleges and universities had remedial placement programs (McNabb, 1990). In 1989, 68% of all postsecondary institutions provided remedial instruction in mathematics, and 65% provided remedial instruction in writing (*Education Week*, 1994). By 1993-1994, 90% of all 4-year colleges and 93% of all 2-year colleges offered remedial instruction and tutoring. Moreover, 30% of all first-year students took at least one remedial course, and 90% of all institutions with remedial placement programs used placement tests to identify those needing help (*Education Week*, 1994).

Course placement services at ACT are based upon the decision model illustrated in Figure 1. The variable, Y, is a dichotomous variable describing a student's performance in the course as successful (Y=1) or unsuccessful (Y=0). Y might, for example, be defined as completing the course with a B or higher grade. As a criterion for placement into the course, a 0.5 probability of success maximizes the course placement accuracy rate—the sum of true positive and true negative placement outcomes (Petersen, 1976; Sawyer, 1996) among students in the placement population. The application of this framework to course placement is explained in interpretive guides (ACT, 1995, 1994) and in research literature (Petersen, 1976; Petersen & Novick, 1976; Sawyer, 1996).

Figure 1: Decision diagram for course placement.



A commonly used procedure for predicting a student's probability of success in a given course, *j*, given a score, *x*, on a content-valid placement test is the logistic regression model:

$$\pi(x, j) = \mathrm{E}(Y \mid x, j) = \frac{\exp(\beta_{0j} + \beta_{1j}x)}{1 + \exp(\beta_{0j} + \beta_{j1}x)}. \tag{1}$$

The test score that maximizes the placement accuracy rate in the course, hereafter called the optimal cutoff score, is a simple function of the logistic regression coefficients:

$$K_j = -\frac{\beta_{0j}}{\beta_{1j}} \tag{2}$$

The optimal cutoff score, K_j, corresponds to a .5 probability of success: $\pi(K_j,i)$ = .5. In practice, when estimates of the logistic regression coefficients are substituted in (2), the estimate, $\hat{K}_j$, is truncated or rounded to the nearest integer. For a continuous predictor variable, *X*, the accuracy rate corresponding to K_j is:

$$A(K_j) = \int_{x<K_j} (1 - \pi(x, j)) f_j(x)dx + \int_{x \geq K_j} \pi(x, j) f_j(x)dx, \tag{3}$$

where $f_{xj}(x)$ is the density function of *X* in the placement population for course *j*.

One estimate of the maximum accuracy rate for course *j*, assuming *X* = 1,2,…, 36 (e.g., the ACT Mathematics test) is:

$$\hat{A}(\hat{K}_j \mid \mathbf{z}_j) = \frac{1}{N_j}\left[\sum_{x=1}^{\hat{K}_j - 1}(1 - \hat{\pi}(x, j))n_{xj} + \sum_{x=\hat{K}_j}^{36}\hat{\pi}(x, j)n_{xj}\right], \tag{4}$$

where $\mathbf{z}_j$ represents a sample of students in the placement population for the course, N_j is the sample size, $\hat{K}_j$ is an estimate obtained by (2), $\hat{\pi}(x, j)$ is obtained by (1), and n_{xj} is the number of students the sample with $X=x$. If the students have taken the course and received grades, an alternative estimate of the accuracy rate is

$$\hat{A}(\hat{K}_j \mid \mathbf{z}_j) = \frac{1}{N_j}\left[\sum_{x=1}^{\hat{K}_j - 1}(n_{xj} - y_{xj}) + \sum_{x=\hat{K}_j}^{36} y_{xj}\right], \tag{5}$$

where y_{xj} is the number of successful students at $X=x$ in this sample.

In an optimal placement system, the sample enabling estimation according to (5) is not representative of the entire placement population because low-scoring and high scoring students are placed in different courses. Censorship in the estimation sample could, in principle, affect the accuracy of estimated cutoff scores and accuracy rates (Schiel and Noble, 1993; Schiel and King, 1999; Schiel and Harmstron, 2000).

Despite the possible effects of censorship, estimates obtained via (5) may be used to compare alternative cutoff scores resulting from different methods of estimating the logistic regression coefficients in (1). The optimal cutoff score does not depend on the distribution of X (Sawyer, 1996; Petersen, 1976). Of two potential cutoff scores, the one closest to K_j is expected to have the higher value of (5) regardless of how X is distributed in the sample.

The problem

Following a study by Houston (1993) and other findings concerning the reliability and validity of statistics from course placement analyses (Schulz, 1993; Crouse, 1993), ACT initially required samples of fifty or more students per course in order to perform a course placement analysis. Unfortunately, this sample size requirement denies analyses for courses with small enrollments. To achieve the necessary sample size for a given course title, such as algebra, a college may pool students from different sections, instructors, or even years. But pooling takes time and resources, delays the research, and could decrease the value of the analysis if data is pooled too broadly across instructors or campuses. In order to provide the service to more schools on a more timely basis, the sample size requirement has been lowered to forty students per course. But even with this concession, many schools are still excluded or inconvenienced.

Purpose of Study

The purpose of this study was to determine whether hierarchical logistic regression can yield sufficiently stable and valid estimates of cutoff scores with sample sizes less than fifty. ACT currently uses nonhierarchical estimation in its course placement service. The nonhierarchical model consists of (1).

Nonhierarchical estimates are obtained through the likelihood function of the logistic regression parameters. For parameters in (1), this function is:

$$l(\boldsymbol{\beta}_j \mid \mathbf{z}_j) = \prod_x \pi(y_{xj} \mid n_{xj}) , \tag{6}$$

where $\boldsymbol{\beta}_j = [\beta_{0j}, \beta_{1j}]^t$, $\mathbf{z}_j = [(n_{1j}, y_{1j}), (n_{2j}, y_{2j}) \ldots (n_{36j}, y_{36j})]$, and

$$\pi(y_{xj} \mid n_{xj}) = \binom{n_{xj}}{y_{xj}} \pi(x, j)^{y_{xj}} (1 - \pi(x, j))^{n_{xj} - y_{xj}} \ . \tag{7}$$

This function is unstable with small sample sizes.

Hierarchical models are discussed in Bryke and Raudenbush (1992) and Gelman, Carlin, Stern, and Rubin (1995). The hierarchical model consists of the model given in (1), plus a model of how the logistic regression parameters are distributed across courses. In the hierarchical model for course placement, the $\boldsymbol{\beta}_j$, j=1,2,...,J are assumed to have a multivariate normal distribution:

$$\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right), \qquad \text{j=1,2,\ldots, J.} \tag{8}$$

This distribution is called the hyperdistribution, or Level 2, of the hierarchical model. Equation (1) comprises Level 1. The parameters in (2) are called Level 2 coefficients, or hyperparameters. If information that could account for differences between the regression parameters (and cutoff scores) of courses is unavailable or is not used, it is reasonable to treat the courses as exchangeable units (Gelman, et al, 1995, pp 123-124). Exchangeability means the course parameters can be modeled as independently and identically distributed.

Hierarchical estimates are generally more stable than nonhierarchical estimates. Hierarchical estimates are derived from the likelihood function and the *prior* density function corresponding to the hyperdistribution in (8). The posterior density of $\boldsymbol{\beta}_j$ conditional on the data, $\mathbf{z}_j$, for course j is proportional to product of the likelihood function and the prior density. An expression of this proportionality is:

$$p(\boldsymbol{\beta}_j | \mathbf{z}_j) \propto l(\boldsymbol{\beta}_j | \mathbf{z}_j)\, p(\boldsymbol{\beta}_j). \tag{9}$$

Any nonzero values of σ_0^2 and σ_1^2 in (8) cause estimates of β_{0j} and β_{1j} to be regressed towards their respective prior means, μ_0 and μ_1, and thus to be more stable across random samples of data from the same course. Using course placement data, Houston and Woodruff, (1997) showed that empirical Bayesian estimates of $\boldsymbol{\beta}_j$ in a hierarchical model were more stable than their nonhierarchical counterparts and that the stability effect was stronger as sample size decreased.

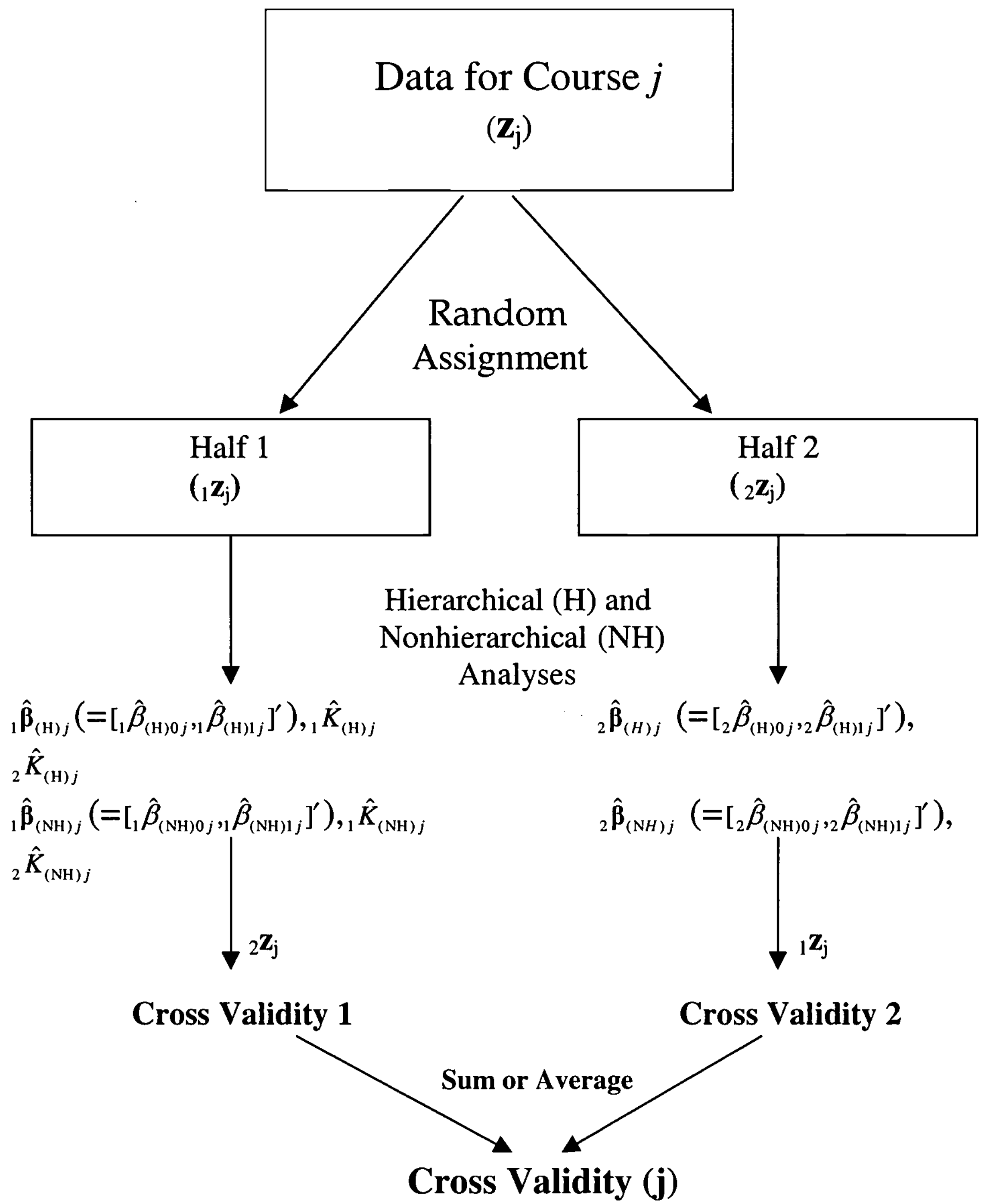
<u>Research Strategy</u>

Ultimately, however, sample size requirements depend on the cross-validity of the estimates (e.g., Algina & Keselman, 2000). Generally speaking, cross-validity refers to how well estimates obtained from one sample can predict the dependent variable in the population from which the sample was drawn. Cross-validity has been extensively studied in the context of ordinary least squares multiple regression (Raju, Bilgic, Edwards, & Fleer, 1997; Algina & Keselman, 2000). Hosmer and Lemeshow (2000) describe procedures for assessing the fit of logistic regression models via external validation. The rationale for external validation is the same as for cross-validation. The fitted model always performs in an optimistic manner on the developmental sample. It is important to assess how the model will perform in predicting outcomes for future subjects.

The approach to cross validation in this study is based on an empirical procedure called double cross-validation (Mosier, 1951). Figure 2 illustrates the procedure as applied in this study. The data within each of *J* courses is randomly assigned to halves. Hierarchical and nonhierarchical estimates are obtained from each half of the data separately for each course. Estimates from half 1 of a given course are then used to predict the half 2 data of the same course and vice versa. Results are pooled across courses as described in the methods section.

There seems to be no generally preferred index for measuring how well a given set of estimates in logistic regression predict new data in a cross-validity study. Hosmer and Lemeshow (2000) discuss many indices for measuring model fit via external validation. Different indices are recommended for different purposes. Most, however, incorporate log likelihoods or some form of accuracy rate.

Figure 2: Double cross-validation design.



In this study, we will use cross-validated log likelihoods and accuracy rates. Half 2 of the data from course j will be used to compute the log likelihood of logistic regression weights estimated from half 1 and to compute the accuracy rate of the corresponding cutoff score. Likewise, half 1 data will be used to evaluate half 2 estimates. More detail on these procedures is given in the methods section.

Expectations

One expected trend in this study is that the cross-validity of estimates from either model (hierarchical and nonhierarchical) decreases with decreasing sample size. This trend is seen in ordinary least squares regression when other factors, such as the number of predictors and the population validity coefficient are held constant (Algina & Keselman, 2000). The trend is due to the effect of sample size on estimation error. With increasing estimation error, $\hat{K}_j$ should be farther from K_j on average, lowering the accuracy rate. This trend should be apparent for both hierarchically-estimated and nonhierarchically-estimated cutoff scores.

The relative performance of hierarchical and nonhierarchical estimates depends on the relative magnitude of systematic and unsystematic sources of error. Nonhierarchical estimates are asymptotically unbiased, but become relatively unstable as sample size decreases. Hierarchical estimates are more stable, but become more regressed to their Level 2 means as sample size decreases. The tradeoff between these types of error is likely to depend on the specific conditions of a study, including the values supplied for the hyperparameters in the hierarchical model. If the values are realistic, it is possible that generalizations can be made across studies and even types of models. Using empirical Bayes procedures, Houston and Sawyer (1988) compared hierarchical and nonhierarchical models for the linear regression of numerically-coded course grades on multiple predictors. They found that hierarchical estimates from samples of twenty students had a level of cross-validity comparable to that of nonhierarchical (maximum likelihood) estimates from samples of fifty students.

Method

Data

Grades of students in college algebra courses were obtained from forty colleges. Colleges are technically the Level 1 units in this study because all of the data within a college is treated the same. The outcome variable, Y, was coded 1 if a student received a B or higher in the course, 0 if the grade was lower. Y was coded as missing if the student withdrew or received an incomplete (see Ang & Noble, 1993). The unweighted, across-college average of the proportion of successful students in each college ($\hat{p}_j$, j=1,2,..., 40) was .46. The average ACT Mathematics score, pooled over colleges, was 21.3. [The average ACT Mathematics score of all students in the graduating class of 2001 who took the ACT Assessment was 20.7.]

Within each college, random halves were created by random assignment, as illustrated in Figure 2 with the last student being dropped if the halves were already of equal size.

To evaluate the effects of sample size in this study, colleges were classified into four groups according to lower limits of 0, 20, 50, and 100 for half counts. A half count is the number of students in each half of a course's data. Group 1, for example, contained colleges with half counts less than 20. The groups are summarized in Table 1. There were seven colleges in Group 1. Half counts in Group 1 ranged from 5 to 14 and averaged 10. The largest sample size group, Group 4, contained eleven colleges with half counts ranging from 171 to 563 and averaging 307.

Table 1

Sample Size Groups based on Half Counts

Sample Size Group	Number of Colleges in Group	Half Counts	
		Range	Average
1	7	5 to 14	10
2	10	20 to 46	31
3	12	51 to 95	69
4	11	171 to 563	307

Logistic Regression Analyses

In all logistic regression analyses, the data were centered by subtracting 21.3, the across-college mean, from each student's ACT score. This constant was then added to (2) when regression weights were used to estimate cutoff scores. Hierarchical logistic regression analyses were performed with the WinBUGS computer program (Spiegelhalter, Thomas, & Best, 2000). The Level 2 model was specified as:

$$\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv N(\begin{bmatrix} -.152 \\ .22 \end{bmatrix}, \begin{bmatrix} 0.314 & -.16 \\ -.16 & 0.002 \end{bmatrix}) \qquad j=1,2,\ldots,J \qquad (10)$$

[$\boldsymbol{\Sigma}$ was specified as a precision matrix in WinBUGS.]. The values in (10) were obtained through a complete Bayesian analysis (Seltzer, Wong, & Bryk, 1996) that used the data of all forty colleges to estimate the Level 2 parameters (Schulz, Betebenner, & Ahn, 2001). Convergence of the Markov chains in WinBUGS was monitored using the Gelman-Rubin convergence diagnostic provided in the program (Gelman & Rubin, 1992.). Iterations 3001 through 5000 were used for sampling posterior distributions. Parameter estimates of the logistic regression weights were the means of posterior distributions. The notation for the estimates is given in Figure 2.

Nonhierarchical regression analyses were performed with the SAS LOGISTIC procedure (SAS, 1990). This procedure uses an iteratively reweighted least squares algorithm (SAS, 1990, p 1088). The notation for the estimates produced by these analyses is given in Figure 2.

Outliers

Limits were established for identifying and replacing outliers in some of the computations described below. For example, a cutoff score estimate of 13 would be replaced with a 16 in computing accuracy rates because, in practice, no cutoff score lower than a 16 would be recommended. Limits for the cutoff score were the lowest (16) and highest cutoff (28) scores found for algebra courses in ACT's course placement analyses (ACT, 1997). Limits for identifying intercept and slope outliers were based on the Level 2 distributions specified in (10). These were $\mu_0 \pm 4\sigma_0$ (-2.4 to 2.1) for intercepts and $\mu_1 \pm 4\sigma_1$ (.045 to .395) for slopes.

Nonhierarchical analyses produced 5 intercept outliers, 16 slope outliers, and 8 cutoff score outliers. All but one slope outlier and one cutoff score outlier were in Groups 1 and 2, representing sample sizes less than fifty. There were no hierarchical outliers.

Stability of Estimates

The stability of an estimate within a given sample size group was measured by the mean absolute difference. The mean absolute difference between hierarchical estimates of the intercept in Group 1 was:

$$\frac{\sum_{j \in G_1} \left| {}_1\hat{\beta}_{(\mathrm{H})0j} - {}_2\hat{\beta}_{(\mathrm{H})0j} \right|}{7},$$

where G_1 is the set of 7 colleges in Group 1. Similar computations were performed for slope and cutoff score estimates, both hierarchical and nonhierarchical, within each sample size group. Outliers were replaced in these computations.

Cross Validity of Intercept and Slope Estimates

Let ${}_1\mathrm{LL}_{(\mathrm{H})j}$ represent the log likelihood of ${}_1\hat{\boldsymbol{\beta}}_{(\mathrm{H})j}$ conditional on ${}_2\mathbf{z}_j$ and let ${}_2\mathrm{LL}_{(\mathrm{H})j}$ represent the log likelihood of ${}_2\hat{\boldsymbol{\beta}}_{(\mathrm{H})j}$ conditional on ${}_1\mathbf{z}_j$. The cross-validity log likelihood of hierarchical estimates in Group 1 was:

$$\sum_{j \in G_1} \sum_{k=1}^{2} {}_k\mathrm{LL}_{(\mathrm{H})j} .$$

Similar computations were performed for each sample size group using nonhierarchical and hierarchical estimates. Outliers were not replaced in these computations. For comparison, the log-likelihood of the Level 2 means ($\boldsymbol{\mu}$ in (10)) were obtained in like manner by computing the log likelihood of $\boldsymbol{\mu}$ separately using each half of the data for course j and summing these log likelihoods over courses within sample size group.

Cross-Validity of Cutoff Scores

Accuracy rates. Using Equation (5), the estimated cross-validity accuracy rate of hierarchical cutoff scores in Group 1 was:

$$\frac{\sum_{j \in G_1} \left(\hat{A}({}_1K_{(H)j} |_2 \mathbf{z}_j) + \hat{A}({}_2K_{(H)j} |_1 \mathbf{z}_j) \right)}{14}.$$

The denominator enumerates the number of half data sets in Group 1 (7 colleges times 2 halves per college). Similar computations were performed for each sample size group using nonhierarchical and hierarchical cutoff scores. Outliers were replaced in these computations.

Two additional indices were computed for comparison to the cross-validated accuracy rates: 1) the accuracy rate of using a "common" cutoff score of 22, and 2) the accuracy rate floor. The common cutoff, 22, was the average estimated cutoff score across the colleges used in this study. One also obtains a common cutoff of 22 (after rounding) if one substitutes the hypermeans of the regression coefficients (elements of $\boldsymbol{\mu}$ given in (10)) into Equation (2) and adds the centering constant, 21.3. The accuracy rate of the common cutoff score represents the possible practice of using the average cutoff score across colleges when sample size is judged to be too small to estimate a college-specific cutoff score.

The accuracy rate floor for a sample size group was either the proportion of successful students or the proportion of unsuccessful students in that group, whichever was higher. The floor corresponds to placing either all or none of the students into the course—whichever produces the higher accuracy rate. Although this action is impractical in most cases, accuracy rates are not strictly comparable across groups with different floors. There is no reason to believe that the accuracy rate floor is systematically related to sample size, but the floor could show quite large variation across sample size groups, especially the smaller ones. It might therefore be important to take the accuracy rate floor into account when interpreting trends in accuracy rates with sample size.

Conditional accuracy rates. Differences in the accuracy rates of two, alternative cutoff scores were assessed by counting the number of students accurately placed by each cutoff score, among students differently placed. For example, if the hierarchical and nonhierarchical cutoff scores estimated from half 1 of a college's data were 20 and 23 respectively, students in half 2 of the college's data with ACT Math scores ranging from 20 to 22 would have been differently placed. Of these students, those who were successful would have been placed accurately by the hierarchical cutoff score and those who were unsuccessful would have been placed accurately by the nonhierarchical cutoff score. Counts according to this description were made for the following contrasts:

1. Hierarchical versus nonhierarchical cutoff scores,
2. Hierarchical cutoff scores versus twenty-two (the common cutoff score), and
3. nonhierarchical cutoff scores versus twenty-two.

For each contrast, counts were summed over colleges within sample size group. A one-degree of freedom Chi-square test was performed on the difference in numbers of students accurately placed by the cutoff scores in a given contrast.

Results

Figures 3 through 5 show the mean absolute difference between random half estimates of, respectively, the intercept, slope, and cutoff score by model and sample size group. These figures show that hierarchical estimates are more stable than nonhierarchical estimates, and that this advantage increases as sample size gets smaller. The stability effect of the hierarchical model is strongest for the slope. The figures present a conservative picture of the stabilizing effect of the hierarchical model in Groups 1 and 2 because many of the nonhierarchical estimates in these groups were outliers and were replaced with a limit.

Figure 3: Stability of intercept estimates.

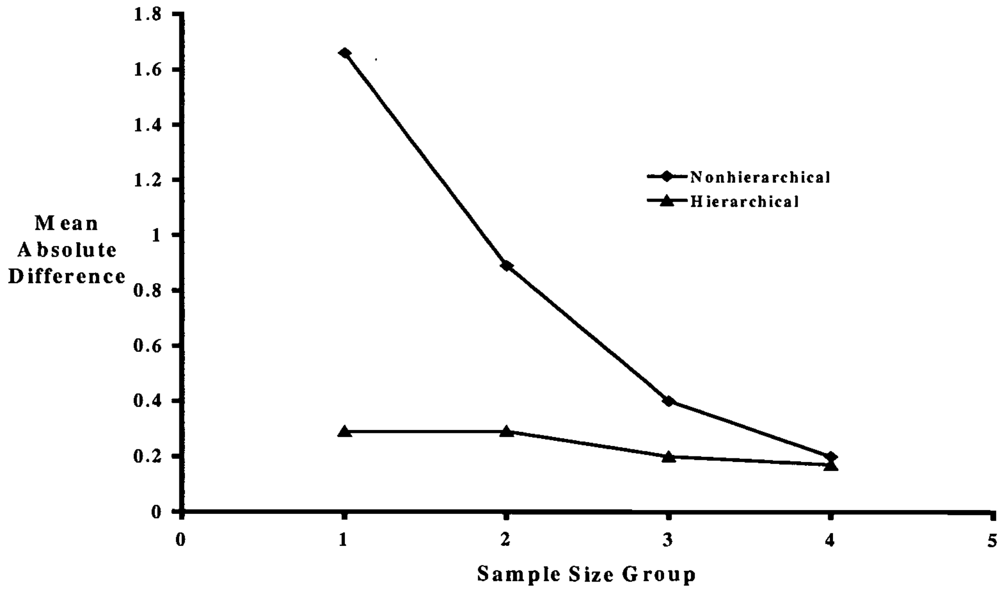


Figure 4: Stability of slope estimates.

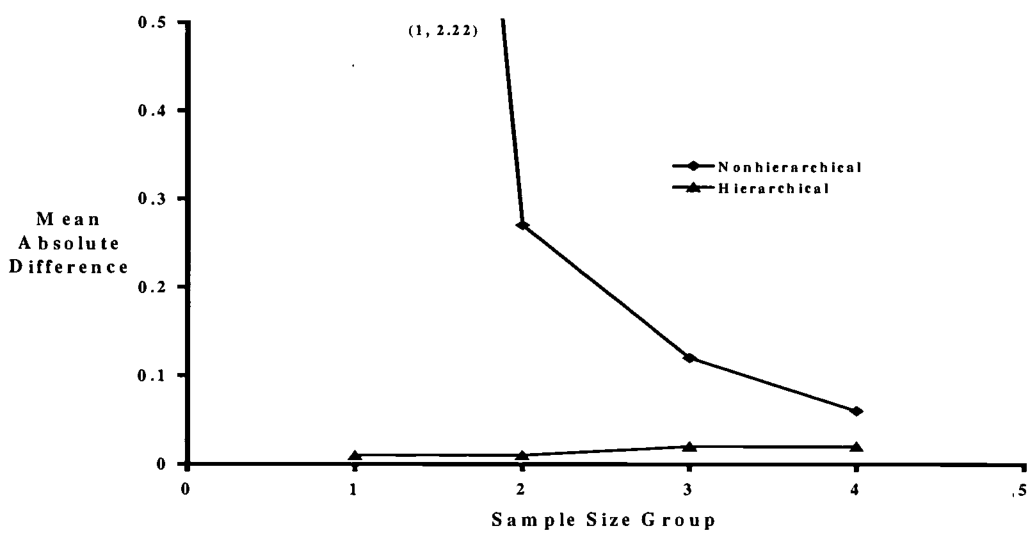
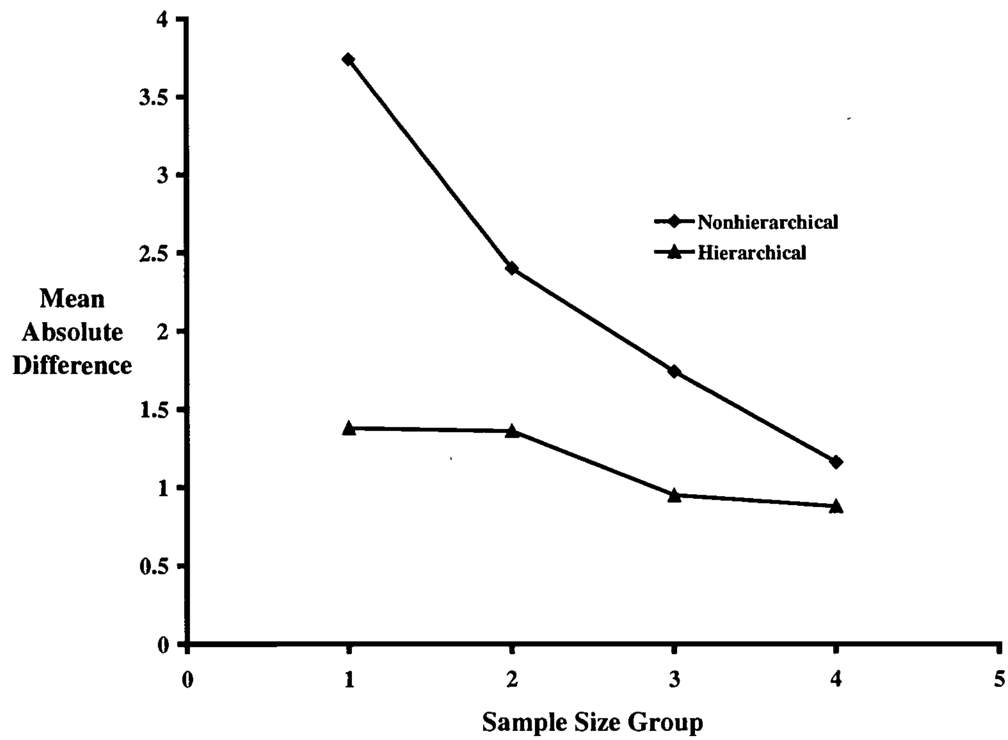
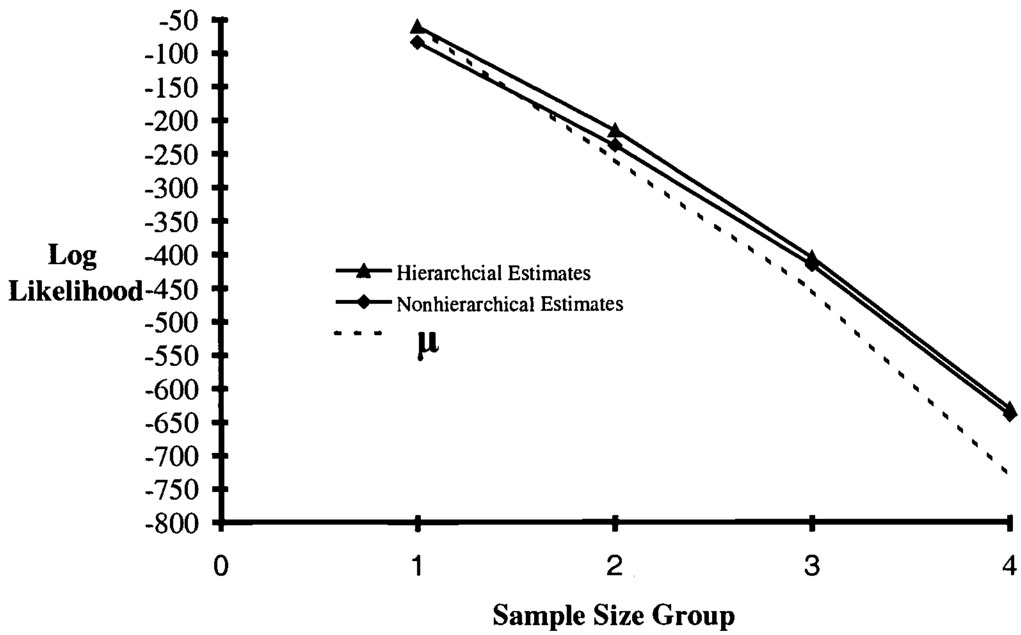


Figure 5: Stability of cutoff score estimates.



As shown in Figure 6 hierarchical estimates of the logistic regression parameters had greater log likelihood (cross-validity) than nonhierarchical estimates in every sample size group. This advantage also increases as sample size gets smaller.

Figure 6: Cross-validity of intercept and slope estimates.



Compared to the Level 2 means (μ), college-specific estimates, both hierarchical and nonhierarchical, had higher log likelihood in all but Group 1. In Group 1, the exact log likelihoods of μ, hierarchical estimates, and nonhierarchical estimates were, respectively, -61.5, -60.5, and -83.9. These values suggest that, compared to μ, nonhierarchical estimates are less valid when sample sizes are approximately 10, but hierarchical estimates are at least as valid, if not more.

Figure 7: Cross-validity of optimal cutoff score estimates.

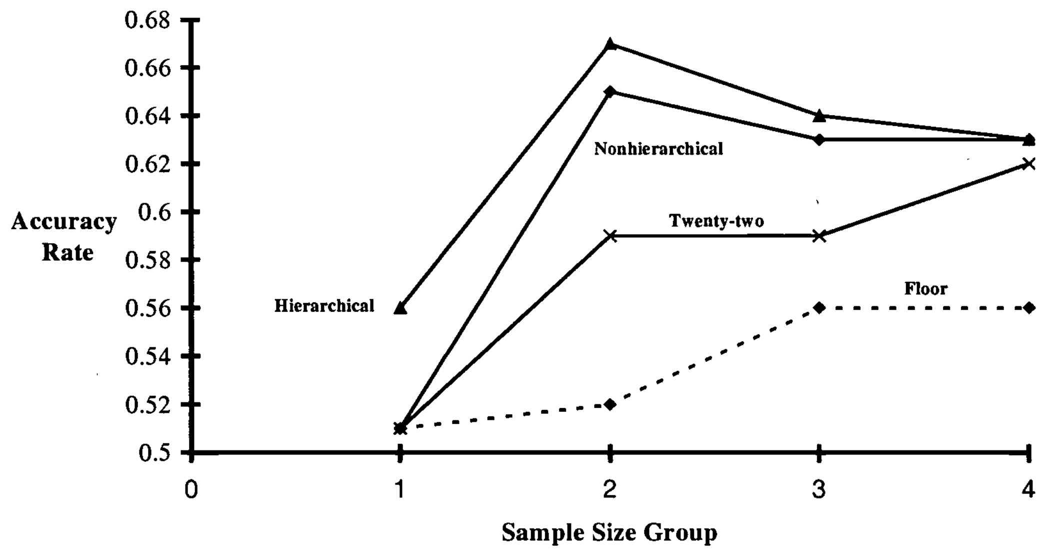


Figure 7 shows trends in accuracy rates and floors with sample size. Except from Group 2 to Group 1, there was no decrease in the accuracy rates of college-specific cutoff scores as sample size decreased (hierarchical and nonhierarchical). The accuracy rates actually appear to increase as sample size decreased from Group 4 to Group 2. Also, the accuracy rate floor and the accuracy rate of the common cutoff score, twenty-two, decrease unexpectedly as sample size group decreases. These rates should not vary with sample size. The unexpected trends in this figure may be within the range of the sampling error of points plotted (see discussion).

One of the key results of this study, illustrated in Figure 7, is that hierarchical cutoff scores tend to have higher cross validity (accuracy rates) than nonhierarchical cutoff scores. This advantage, like that of the log likelihoods and stability, appears to increase as sample size gets smaller. In Group 4, there was no difference—the accuracy rate was 0.63 for both sources of cutoff score. But in Group 1, representing the smallest sample sizes, the accuracy rate was .56 for hierarchical cutoff scores and .51 for nonhierarchical cutoff scores.

It should also be noted that in Group 1, the accuracy rate of nonhierarchical cutoff scores was not higher than the accuracy rate floor. In other words, nonhierarchically-estimated cutoff scores in Group 1 made no positive contribution to the placement accuracy rate.

The counts in Table 2 are consistent with the information plotted in Figure 7. For example, in Group 1, thirty-five students would have been placed differently if hierarchical cutoff scores had been used instead of nonhierarchical cutoff scores. [There were approximately 140 students total in this group.] Of these thirty-five, 22 would have been accurately placed by the hierarchical cutoff scores for a conditional accuracy rate of .63. (Conversely, 13 would have been accurately placed by nonhierarchical cutoff scores for a conditional accuracy rate of .37.) Although the difference between these numbers, or rates, was not statistically significant in a one-degree of freedom Chi-square test, the difference based on the combined counts of Groups 1 and 2, which together represent sample sizes of less than fifty, was statistically significant ($p<.05$).

Table 2

Counts of Students Affected by Cutoff Score Differences

Sample Size Group	Number of Students Affected[§]	Counts: Number Accurately Placed		Proportion Accurately Placed
Hierarchical Versus Nonhierarchical				
		Hierarchical	Nonhierarchical	Hierarchical
1	35	22	13	.63*
2	65	39	26	.60*
3	90	50	40	.56
4	70	31	39	.44
Hierarchical versus Twenty-two				
		Hierarchical	Twenty-two	Hierarchical
1	8	6	2	.75
2	182	115	67	.63**
3	348	222	126	.64**
4	1299	720	579	.55**
Nonhierarchical versus Twenty-two				
		Nonhierarchical	Twenty-two	Nonhierarchical
1	39	18	21	.47
2	212	134	78	.63**
3	421	271	150	.64**
4	1398	807	591	.58**

§ These students would have been placed differently by the two cutoff scores.
*Proportion for Groups 1 and 2 combined differs significantly from 0.5 ($p<.05$).
**Proportion differs significantly from 0.5 ($p < .05$ or less).

In comparison to the common cutoff score of twenty-two, both hierarchical and nonhierarchical cutoff scores had higher conditional accuracy rates in Groups 2, 3, and 4 ($p<.05$ for each group separately). In Group 1, hierarchical cutoff scores, but not nonhierarchical cutoff scores, outperformed the common cutoff score, but neither difference was statistically significant due to the small numbers of students differently placed.

Discussion

This study provides some useful detail to the earlier demonstration that hierarchical regression weights in course placement are more stable than their nonhierarchical counterparts (Houston & Woodruff, 1997). The earlier demonstration showed the stabilizing effect of the hierarchical model in units of Euclidean distance between paired logistic regression parameter vectors. It did not show the stability of the intercept and slope separately, and did not include the stability of the cutoff score. Although it is not surprising to see that the slope is more stabilized than the intercept, this result is gratifying and the details of these separate trends with sample size may prove useful in implementing hierarchical analyses for course placement in the future.

Our results suggest that the stability of parameter estimates cannot be a criterion for establishing minimum sample sizes for hierarchical analyses. A reasonable benchmark for stability might be that of nonhierarchical estimates in Group 3, where sample sizes are fifty or more. Figures 3 to 5 indicate that there is no sample size below fifty (Groups 1 and 2) were hierarchical estimates will become as unstable as nonhierarchical estimates are with sample sizes of fifty or more (Group 3). In fact, below a certain sample size, hierarchical estimates appear to become more stable, or at least maintain the same level of stability, as sample size decreases.

The stability of hierarchical estimates reflects their regression to μ or to the common cutoff score (twenty-two). In Group 1, where the regression effect is strongest due to the small sample sizes, hierarchically-estimated cutoff scores place very few students (eight) differently than the common cutoff score. [Nonhierarchically-estimated cutoff scores placed many more students (39) differently than the common cutoff.]. Also in this group, the log likelihood of hierarchical estimates (-60.5) was nearly equal to the log likelihood of μ (-61.5), indicating that the values of these estimates were very nearly the same.

Evidently, however, the advantage of stability outweighs the disadvantage of regression bias in hierarchical estimates when cross-validity is considered. The cross-validity log likelihoods, accuracy rates, and conditional accuracy rates in this study show that hierarchical estimates have greater cross-validity than nonhierarchical estimates, particularly with sample sizes less than fifty. The similarity of these results to those of Houston and Sawyer (1988) provide a wider basis for the notion that hierarchical models can generally reduce sample size requirements in applied settings.

In one respect, our results suggests that a sample of 30—the approximate average of Group 2—would be sufficient for either hierarchical or nonhierarchical analyses and that even smaller sample sizes would be acceptable for hierarchical analyses. Both sources of college-specific estimates outperformed μ and the common cutoff score in Group 2. Hierarchical estimates slightly outperformed μ and the common cutoff score in Group 1.

These results pertain to the very real possibility that a college might base placement decisions for a given course on the 'average' cutoff score for courses of the same title, if it could not provide a sufficiently large sample for obtaining a local estimate. Our results indicate the sample sizes of 30, or even 10 if hierarchical analysis is used, would be preferable to this practice. Additional considerations and benchmarks will probably also figure into the eventual establishment of minimum sample size requirements.

The absence of expected trends and the presence of unexpected trends with sample size in Figure 7 may be explained by the measurement error of accuracy rates, and possibly by sampling bias. Even if the optimal cutoff score for a college were known, estimates of the accuracy rate contain measurement error when based on a sample. With the small size of the samples in Groups 1 through 3 and the small number of colleges per sample size group, the average accuracy rates plotted in Figure 7 contain significant amounts of measurement error. This error alone might account for the difference between expected and observed trends in Figure 7. Sampling bias might also be a factor. Our samples include only students who took the placement test. These students may differ systematically from other students in the course, particularly in colleges with very small sample size. The small sample size of these colleges may be due more to attrition from lack of scores on the given placement test, than to the actual size of the class.

Specific recommendations about minimum sample sizes, such as the notion that thirty may be sufficient for a nonhierarchical analyses, might depend on specific characteristics of the data used in this study. The baseline success rates in this study were close to 0.5—a favorable condition for estimating the parameters of a logistic regression function. Success rates closer to 0 or 1 might require larger sample sizes. Success rates would have been closer to 1 if a "C or higher" success criterion had been used, or if courses that are traditionally easier than college algebra had been used.

The similarity of courses sharing the same Level 2 parameters might also influence sample size requirements. Houston and Woodruff (1997) classified algebra courses by whether they are offered in 2-year or 4-year institutions. The Level 2 parameters used in this study were estimated with a more diverse collection of colleges (Schulz, Betebenner, & Ahn, 2001). With more homogeneous course groupings, the Level 2 variance parameters might be smaller and the Level 2 means more specific. This condition would decrease the regression effect for a given sample size, and decrease the sample size needed for a given level of cross-validity of estimates from the hierarchical model.

More useful information concerning sample size requirements for hierarchical analyses might be also obtained through simulation (e.g., Houston, 1993). With simulation, estimates of the logistic regression parameters and cutoff score can be compared to known values. Repeated samples of a given size can be created to obtain empirical distributions of statistics such as the accuracy rate. Rather than sampling the placement population, known values of the conditional probability of success in the placement population can be used to compute accuracy rates.

The use of real data in the present study, however, shows that hierarchical analyses have practical advantages. The sample size requirement for course placement can be substantially less than fifty if the hierarchical model is used. This means course placement analyses can be performed for more courses and colleges and can be used to establish cutoff scores on placement tests that may have been taken by relatively few students prior to taking the course.

References

ACT, Inc. (1997). *ACT Assessment Technical Manual.* Iowa City, IA: Author.

ACT, Inc. (1995). *ASSET Course Placement Service Interpretive Guide.* Iowa City, IA: Author.

ACT, Inc., (1994). *ACT Assessment Course Placement Service Interpretive Guide.* Iowa City, IA: Author.

Algina, J. & Keselman, H. J. (2000). Cross-validation sample sizes. *Applied Psychological Measurement. 24*, 173-179.

Ang, C. H., & Noble, J. P. (1993). *The Effects of Alternative Interpretations of Incomplete Grades on Course Placement Validity Statistics.* ACT Research Report Series 93-3. Iowa City, IA: ACT, Inc.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical Linear Models.* Newbury Park: Sage Publications.

Crouse, J. (April, 1993). *Diagnosing Influential Observations in Estimating Logistic Regression Functions.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Education Week (April 13, 1994). *Colleges and Universities Offering Remedial Instruction.* (No. 29, p 6.). Author.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D. B. (1995). *Bayesian Data Analysis*, New York: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* Second Edition. New York: John Wiley & Sons, Inc.

Houston, W. M., (1993, April). *Accuracy of validity indices for course placement systems.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Houston, W. M., & Woodruff, D. J. (1997). *Empirical Bayes estimates of parameters from the logistic regression model.* ACT Research Report Series 97-6. Iowa City, IA: ACT, Inc.

Houston, W. M., & Sawyer, R. (1988). *Central Prediction Systems for Predicting Specific Course Grades*. ACT Research Report Series 88-4. Iowa City, IA: ACT, Inc.

McNabb, T. (1990). *Course Placement Practices of American Postsecondary Institutions*. ACT Research Report Series 90-10. Iowa City, IA: ACT, Inc.

Mosier, K. R. (1951). Symposium: The need and means of cross-validation. *Educational and Psychological Measurement*, *11*, 5-11.

Petersen, N. S. (1976). An expected utility model for "optimal" selection. *Journal of Educational Statistics*, *1*, 333-358.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, *13*, 3-29.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity and the use of equal weights in prediction. *Applied Psychological Measurement*, *21*, 292-305.

SAS (1990). *SAS/STAT User's Guide, Volume 2*. Cary, NC: SAS Institute, Inc.

Sawyer, R. L. (1996). Decision theory models for validating course placement tests. *Journal of Educational Mea*surement, 33, 271-290.

Schiel, J. L., & King, J. (1999). *Accuracy rates of course placement validity statistics under various soft truncation conditions*. Research Report 99-2. Iowa City: ACT Inc.

Schiel, J. L., & Harmston, M. (2000). *Validating two-stage course placement systems when data are truncated.* Research Report 2000-3. Iowa City: ACT, Inc.

Schiel, J., & Noble, J. (April, 1993). *The Effects of Data Truncation on Estimated Validity Indices for Course Placement*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Schulz, E. M., Betebenner, D. & Ahn, M. (2001). Hierarchical Analyses for Course Placement. ACT Research Report (in press). Iowa City: ACT, Inc.

Schulz, E. M. (1993, April). *Statistics for cross-validating course placement models*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Seltzer, M. H., Wong, W. H., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics*, *21*, 131-167.

Spiegelhalter, D., Thomas, A., & Best, N. (April, 2000). *WinBUGS Version 1.3 User Manual.* MRC Biostatistics Unit, Institute of Public Health. Cambridge, UK.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

TM033873

I. DOCUMENT IDENTIFICATION:

Title: A Comparison of Hierarchical and Nonhierarchical Logistic Regression for Estimating Cutoff Scores in Course Placement	
Author(s): E. Matthew Schulz, Damian Betebenner & Meeyeon Ahn	
Corporate Source: ACT, Inc.	Publication Date: April, 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☒	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: E M Schulz	Printed Name/Position/Title: Senior Statistician	
Organization/Address: P.O. Box 168, Iowa City 52243	Telephone: 319 337-1468	FAX:
	E-Mail Address: schulzm@act.org	Date: 4/10/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com