DOCUMENT RESUME

ED 346 163	TM 018 554
AUTHOR	Noble, Julie; Sawyer, Richard
TITLE	A Comparison of Two Approaches for Measuring Educational Growth from CTBS and P-ACT+ Scores.
PUB DATE	Apr 92
NOTE	38p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 20-24, 1992).
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Academic Achievement; Achievement Gains; Comparative Analysis; *Educational Change; Grade 10; High Schools; *High School Students; Instructional Effectiveness; Mathematical Models; Predictive Measurement; *Regression (Statistics); Sample Size; *Scores; Statistical Significance
IDENTIFIERS	*Comprehensive Tests of Basic Skills; Educational Indicators; Linear Models; Mean Residuals; Outliers; *Preliminary American College Test Plus; Tennessee

ABSTRACT

The purpose of the study was to compare two regression-based approaches for measuring educational effectiveness in Tennessee high schools: the mean residual approach (MR), and a more general linear models (LM) approach. Data were obtained from a sample of 1,011 students who were enrolled in 48 high schools, and who had taken the Comprehensive Tests of Basic Skills (CTBS) in grade 8 and the Preliminary American College Test Plus (P-ACT+) in grade 10. Regression models were developed for predicting P-ACT+ scores from CTBS scores. In the MR approach, mean residual P-ACT+ scores were calculated for each high school from a pooled model. The LM approach included effect-coded high school dummy variables as predictors. Indicators of educational change, called Achievement Index (AI) values, were dveloped for each school; AI values were also calculated where race, coursework and grades, or entering achievement level was included in the models. The results show that the LM approach is more sensitive in identifying outlier schools and was less hampered by sample size constraints than was the MR approach. Statistical significance and identifying violations of the assumption of parallel slopes were also easier to determine. Five graphs and one table illustrate the discussion, and there are two references. (Author/SLD)

A COMPARISON OF TWO APPROACHES FOR MEASURING EDUCATIONAL GROWTH FROM CTBS AND P-ACT+ SCORES

Julie Noble and Richard Sawyer

ACT

A paper presented at the 1992 meeting of the National Council on Measurement in Education, San Francisco, CA

ABSTRACT

The purpose of this study was to compare two regression-based approaches for measuring educational effectiveness in Tennessee high schools: the mean residual approach (MR) and a more general linear models (LM) approach. Data were obtained from a sample of 1,011 students who were enrolled in 48 high schools, and who had taken the Comprehensive Tests of Basic Skills in Grade 8 and the P-ACT+ in Grade 10. Regression models were developed for predicting P-ACT+ scores from CTBS scores. In the MR approach, mean residual P-ACT+ scores were calculated for each high school from a pooled model. The LM approach included effect-coded high school dummy variables as predictors. Indicators of educational change, called Achievement Index (AI) values, were developed for each school; AI values were also calculated where race, coursework and grades, or entering achievement level was included in the models. The results showed that the LM approach was more sensitive in identifying outlier schools than was the MR approach and was less hampered by sample size constraints. Statistical significance and identifying violations of the assumption of parallel slopes were also easier to determine.

A COMPARISON OF TWO APPROACHES FOR MEASURING EDUCATIONAL GROWTH FROM CTBS AND P-ACT+ SCORES

Julie Noble and Richard Sawyer, ACT

There is a major push in education today to determine educational effectiveness of high schools and school districts. This pressure is reflected most recently in Raising Standards in Education by the National Council on Education Standards and Testing (1992), and in the strong orientation by accrediting agencies towards outcomes assessment. More and more, schools and districts are being held accountable for the educational achievement of their students, and for developing ways for improving their educational effectiveness. In addition, schools are continually striving to meet their own standards of excellence, both in meeting their students' needs in maximizing their students' educational experiences.

Testing has become a major tool and focus of outcomes assessment, and has frequently been criticized for its misuse. Test scores are outcome-oriented, but they do not directly measure educational process. Moreover, no validated theoretical model exists linking test scores to the myriad of possible process variables. Therefore, test results are best used, along with other indicators, as a "fire alarm" for identifying issues for further review and discussion, and not as a single indicator of educational process.

Effectiveness in teaching academic skills can be measured in two broad areas: students' current academic proficiencies (i.e., what do they know now?), and "growth" in their academic proficiencies over time (i.e., what have students learned over time?). Students' current academic proficiencies can be evaluated in terms of specific criteria or standards of performance; for example, what students should know and be able to do by Grade 10 in order to accomplish specific goals. Problems with this approach tend to be more political than methodological in nature. Showing growth in achievement over time not only can be highly political, but can also be problematic from a methodological standpoint.

Schools, districts, and states have sought equitable and practical methods for measuring the educational effectiveness of schools with respect to growth. Such methods often involve a test-retest approach using the same or similar tests over two or more time periods. Using parallel forms of tests over time, and thus having a constant score scale, enables the school to show mean gain over time in the skills and knowledge measured by the test.

With similar assessments, effectiveness cannot be shown by computing mean gain scores, due to the differences in scaling. School effectiveness can be shown, however, through using regression methods, in which students' predicted scores on the second test are compared with their initial scores.

A major limitation of the methods used to show school effectiveness is that students' academic and cultural backgrounds differ greatly among schools. For example, regression models may statistically control for student achievement at entry, but may not statistically control for school and student characteristics. Thus, differences in educational achievement or growth among schools might be confounded with students' background characteristics, and not an assessment of actual growth over time.

Mean gain scores and the use of regression models are both based on observed scores. They do not account for measurement error in either the first or second test score obtained by students, and thus do not measure "true gain" or relative improvement. Structural equation modeling is one possible method for determining "true gain" by estimating the amount of measurement error in test scores. This study was intended, however, to develop a meaningful and useful *indicator* of areas in need of further investigation, rather than to estimate true gain. With such a "fire alarm" system, structural equation models are not needed. Measurement error could, however, be minimized by using large sample sizes and maximizing prediction accuracy when developing such indicators.

Measuring Educational Effectiveness in Tennessee

ACT, in collaboration with the Tennessee State Department of Education, sought to address the issue of accountability for the high schools in Tennessee. Tennessee desired an assessment system where the growth in higher-order thinking skills by students and schools could be measured and compared to that of appropriate reference groups.

The project was designed to build on the existing assessment system in Tennessee. Tennessee currently administers the Comprehensive Tests of Basic Skills (CTBS) each spring to elementary, Grade 8, and Grade 10 students. The CTBS academic tests consist of three primary tests in Language, Reading, and Mathematics. The Total score is the average of the three tests. Each of these tests has two subscores: Language Mechanics and Language Expression for Language, Vocabulary and Comprehension for Reading, and Math Computation and Math Concepts for Mathematics. There are additional tests in Science, Study Skills, Social .dies, and Spelling. Scores range from 0 to 999. The CTBS is used to describe the current academic proficiency of students and schools, as well as to show academic growth over time. As implied by its title, the CTBS is a basic skills test; it does not assess higher-order thinking skills. Students' scores, and mean gain scores for schools, were used to show achievement in basic skills over time, and could not address the need for information about students' achievement in higher-order thinking skills.

In contrast, the P-ACT+ academic tests assess higher-order thinking skills in four areas: English, Mathematics, Reading, and Science Reasoning. The Composite score is the arithmetic mean of the four test scores. The English and Mathematics tests have two subscores: Usage/Mechanics and Rhetorical Skills for English, and Pre-Algebra/Algebra and Geometry for Mathematics. Test scores range from 1 to 32 and subscores range from 1 to 16. Linking the

CTBS at Grade 8 and the P-ACT+ at Grade 10 would provide information about the schools effects on student achievement over time (Grade 9), and, in particular, about students' acquisition of higher-order thinking skills, while controlling for their basic skills in Grade 8.

Two primary goals were identified for the Tennessee project that were directly related to measuring growth:

1. To determine the accuracy with which the P-ACT+ scores of sophomores from Tennessee could be predicted from their Grade 8 CTBS scores.
2. To determine the extent to which the relationship between P-ACT+ and CTBS scores could be used to compare the educational effectiveness of high schools between Grades 8 and 10. Regression models were used to determine a level of expected performance by schools. Schools were then be categorized according to whether they performed as expected, above expectation, or below expectation.

During the course of this project, we determined that accurate predictions of P-ACT+ scores could be made from CTBS scores, and that the predictions could be used to calculate measures of institutional effectiveness. The equitability and effectiveness of these measures was still in question, however. Therefore, a more comprehensive goal was developed:

3. Using the results for Goal 1, to compare two approaches for calculating measures of educational effectiveness in terms of their predictive accuracy, their sensitivity to intra-school differences, their ease of computation, and their equitability with regard to individual and school characteristics.

The purpose of this study was to compare two approaches for measuring educational effectiveness in Tennessee high schools: the mean residual (MR) approach and a more general linear models (LM) approach. The MR approach categorizes schools by the degree to which each school's mean residual, derived through multiple regression methods, deviates from the mean residual for all schools. The LM approach also uses multiple regression, but includes the high school as a dummy variable in the model, and categorizes schools according to the regression weights corresponding to the dummy variables. The two approaches were used to develop indicators of educational growth, or Achievement Index values, for describing each high school's growth relative to that of other schools. The results for both approaches were compared for the total group of high schools and by high school characteristic, CTBS score quartile, course work, and race, in order to determine potential insensitivities in the models for these groups.

As previously noted, the indicators of educational effectiveness, or Achievement Index values, developed for this study are intended to be used as a first look or "fire alarm" for school effectiveness. They can provide indicators of areas in need of further study, and are not intended to be used as an all-inclusive, stand-alone measure for program evaluation.

Data

The P-ACT+ was administered in fall, 1990 to nearly 1400 high school sophomores from Tennessee who had completed the CTBS as eighth graders in 1989. In taking the P-ACT+, students completed a Course/Grade Information Section that collects information about 26 high school courses in English, mathematics, social studies, natural sciences, foreign languages, and arts. Students were asked to indicate whether or not they had taken or were currently taking each course, and the grade they received. They also provided the number of years they planned to take in English, mathematics, social studies, natural sciences, and foreign languages.

The CGIS data were used to calculate variables representing course work taken or planned. The variables selected for inclusion in the study were those that have been shown in other research (P-ACT+ Supplement to the Technical Manual, in progress) to be statistically significantly associated with P-ACT+ performance ($p < .001$). For the MR models, four categorical core curriculum variables were developed:

1. Plan to take a college preparatory core curriculum (4 years of English, 3 years of mathematics, social studies, and natural sciences)
2. Taken or currently taking any English course; Algebra 1, Algebra 2, or Geometry; any social studies course; general science, physical science, or biology
3. Taken or currently taking any English course; Algebra 1, Algebra 2, or Geometry; any social studies course; physical science or biology
4. Taken or currently taking any English course; Algebra 1, Algebra 2, or Geometry; any social studies course; physical science or biology; any foreign languages course

For the LM models, three continuous course work variables were used:

1. The number of mathematics courses taken
2. The planned number of years of high school natural sciences course work
3. High school average using grades from all 26 courses

Demographic information about the students and the schools was collected from the Student Information Section of the P-ACT+ and from the Market Data Retrieval (MDR) files. The student information included race and grade in school. School variables included community type, lowest and highest grade in the school, school type (senior high school, vocational/technical, etc.), school enrollment, per-pupil expenditure, percent below federal poverty level in the district, percent of white students in the district, and percent of black students in the district. All schools were public schools.

The initial sample consisted of 1,349 student records from 56 high schools representative of all Tennessee high schools. A minimum sample size of 15 was required in order for a school to be included in the analyses; the final sample consisted of 1,011 records from 48 high schools. Of these, 885 records from 41 schools had complete P-ACT+/CTBS/CGIS information, and 787 records from 39 schools had complete P-ACT+/CTBS/CGIS and race information.

Method

The MR and LM approaches both involve regression methodology. To identify parsimonious models for predicting P-ACT+ scores, P-ACT+ scores were first regressed on CTBS scores using student data pooled across schools. All significant ($p < .001$) CTBS scores within each model were retained for further analysis. Scatterplots were then developed to illustrate the P-ACT+/CTBS score relationships, which showed curvilinearity in the relationships between P-ACT+ and CTBS scores. Pooled quadratic term models, which included both linear and quadratic CTBS test score terms were then estimated; all quadratic terms were significant ($p < .001$). The models listed below included both the linear and quadratic terms for each predictor variable:

E1. P-ACT+ English = f{CTBS Language, Reading}

E3. P-ACT+ English = f{CTBS Total}

M1. P-ACT+ Mathematics = f{CTBS Total Mathematics}

M2. P-ACT+ Mathematics = f{CTBS Total}

R1. P-ACT+ Reading = f{CTBS Total Reading, Study Skills}

R3. P-ACT+ Reading = f{CTBS Total}

S1. P-ACT+ Science Reasoning = f{CTBS Science, Total}

S2. P-ACT+ Science Reasoning = f{CTBS Total}

C1. P-ACT+ Composite = f{CTBS Science, Total, Vocabulary}

C2. P-ACT+ Composite = f{CTBS Total, Study Skills, Vocabulary, Science}

C3. P-ACT+ Composite = f{CTBS Total}

Achievement Index

The Achievement Index (AI) provides a means for comparing educational growth at an individual school against a reference group of schools. For this study, three AI values were possible for a given school: the school is achieving educational growth as expected, achieving growth below expectation, or achieving growth above expectation. A P-ACT+ scale score range of ± 1 was used to identify schools achieving above or below expectation for this study. Other categories or ranges of P-ACT+ score values could also be used to identify "outlier" schools. The number and types of categories would be determined by the precision required by the schools involved. The implications of statistical significance are discussed on p. 12.

Mean Residual (MR) Approach

The MR approach uses a pooled regression model to compute mean residual values for each high school. This approach assumes parallel regression slopes and intercepts across high schools. From the pooled quadratic model for each P-ACT+ score, mean residuals were calculated for each school:

$$\overline{R}_i = \sum_{j=1}^{n_i} \left(\frac{Y_{ij} - \hat{Y}_{ij}}{n_i}\right) \qquad (1)$$

where $\hat{Y}_{ij} = a + bX_{ij}$

i = school, $1 \leq i \leq 48$

j = student in school i

n_i = number of students in school i

X_{ij} = CTBS score for student j in school i

$\hat{Y}_{ij}$ = predicted P-ACT+ score for student j in school i

Outlier schools were then identified as those with mean residuals $\overline{R}_i$ exceeding ± 1 P-ACT+ scale score unit.

Outlier schools with positive values exceeding 1 were identified as achieving growth above expectation; outlier schools with negative values less than -1 were identified as achieving growth below expectation. Schools with values within ± 1 P-ACT+ score unit were identified as achieving growth as expected.

Linear Models (LM) Approach

The LM approach uses the same basic regression model as the MR approach, but includes effect-coded dummy variables in the model. It allows the intercepts to vary by high school, and provides a mechanism to test for possible interaction effects. For the 48 schools in this study:

$$\hat{Y}_{ij} = a + c_i d_i + bX_{ij} \text{ for school } i \ (1 \leq i \leq 47) \tag{2}$$

$$\hat{Y}_{ij} = a - \sum_{i=1}^{47} c_i + bX_{ij} \text{ for school 48} \tag{3}$$

where a = overall intercept (unweighted mean of school intercepts)

d_i = effect-coded dummy variable such that

d_i = 1, if the student is enrolled in school i ($1 \leq i \leq 47$)

= 0, if the student is enrolled in another school, $j \neq i$ ($1 \leq j \leq 47$)

= -1, if the student is enrolled in school 48

c_i = regression coefficient associated with variable d_i, for school i ($1 \leq i \leq 47$)

(intercept for school i - unweighted mean of intercepts)

Schools with dummy variable regression coefficients exceeding ± 1 P-ACT+ scale score unit were identified as outlier schools:

$$| c_i | \geq 1 \text{ for school } i \ (1 \leq i \leq 47) \tag{4}$$

The same category descriptions were used for this approach as were used for the MR approach:

$$| \sum_{i=1}^{47} c_i | \geq 1 \text{ for school 48} \tag{5}$$

Both the MR and LM approaches used assume parallel slopes across schools. To test this assumption, quadratic regression models with high school effect variables and high school by CTBS test score interactions were developed for all P-ACT+ scores. The most parsimonious model for each P-ACT+ score was used. Interaction terms were included only for the primary CTBS test score in each model. Statistically significant ($p < .05$) interactions (CTBS Total score by high school) were found for P-ACT+ Mathematics and the Composite, but not for the other three P-ACT+ test scores. A less stringent significance level of .05 was used for the interaction effects than for the original model development ($p < .001$), in order to include any possible factors that could influence the Achievement Index values for a given school.

Additional parsimonious regression models were developed for Mathematics and the Composite that included statistically significant ($p < .05$) high school by CTBS Total score interaction terms. The models took the following general form:

$$\hat{Y}_{ij} = a + c_i d_i + (b + f_i)X_{ij'} \text{ for students from schools with significant interaction terms} \tag{6}$$

$$\hat{Y}_{ij} = a + c_i d_i + bX_{ij'} \text{ for students from schools with nonsignficant interaction terms} \tag{7}$$

$$Y_{Kj} = a - \sum_{i=1}^{47} c_i + (b - \sum_{i=1}^{5} f_i)X_{Kj'} \tag{8}$$

where f = is the regression coefficient associated with the school by test score interaction term for school i.

Achievement Index values were calculated for each P-ACT+ test at the first, second, and third quartiles of the distribution of CTBS Total scores for all schools (750, 772, and 793, respectively).

Comparisons by Population Subgroup

Mean Residual (MR) Approach. Mean residuals were calculated by populatic subgroup for each school with a minimum of 15 students in a given subgroup. Subgroups used in this analysis included students whose CTBS scores were in the top and bottom quartiles,students who planned to take core curriculum courses, black students, and white students.

Linear Models (LM) Approach. Using the quadratic and quadratic/interaction models described on p. 9, additional models were developed that included course work taken or race as predictors of P-ACT+ performance. Inclusion of CTBS score main effects and interaction terms was analogous to analyzing the top and bottom quartiles using the MR approach. Using the same quartile values for CTBS Total, three AI values were developed for Mathematics and the Composite, one at each quartile. Schools without students scoring at or beyond the quartiles were not given an Achievement Index value for that analysis.

Another set of models was developed for the course work variables. Three course work variables were used: number of mathematics courses taken, the number of years of natural sciences courses planned, and high school GPA. Only the linear form was included for each variable. In adding the course work variables to the models, all nonsignificant ($p < .001$) linear and quadratic terms were dropped (selected CTBS scores were eliminated only from the Composite models). A significance level of .05 was retained for all interaction terms. The final course work models included the following predictor variables:

E1. P-ACT+ English = f{Total Language, Reading, high school average}

E3. P-ACT+ English = f{CTBS Total, high school average}

M1. P-ACT+ Mathematics = f{CTBS Total Mathematics, number of mathematics courses taken, high school average}

M2. P-ACT+ Mathematics = f{CTBS Total, number of mathematics courses taken, high school average}

Ri. P-ACT+ Reading = f{CTBS Total Reading, Study Skills, high school average}

R3. P-ACT+ Reading = f{CTBS Total}

S1. P-ACT+ Science Reasoning = f{CTBS Science, Total}

S2. P-ACT+ Science Reasoning = f{CTBS Total}

C1. P-ACT+ Composite = f{CTBS Science, Total, high school average}

C2. P-ACT+ Composite = f{CTBS Total, Study Skills, number of mathematics courses taken, high school average}

C3. P-ACT+ Composite = f{CTBS Total, high school average}

A third set of models was developed in which race was included as a dummy variable. These models allowed us to determine whether race was a statistically significant predictor of P-ACT+ score after accounting for CTBS performance and high school.

Achievement Index values were calculated for each school using the quadratic/course work and quadratic/race models. Further, significant interaction terms were included for Mathematics and the Composite for both sets of models. Three AI values were calculated for Mathematics and the Composite (one per quartile) using both the course work and race models.

Statistical Significance

To this point, all AI values were calculated using a P-ACT+ score range of ± 1, which is equal to approximately one standard error of measurement for the Composite. A major concern, particularly with the relatively small samples of students available for this study, is the extent to which an outlier school would be flagged on the basis of chance alone, and not due to any real growth. This would also be a concern when examining population subgroups, where even smaller sample sizes could be found. Given the potential political uses of this methodology, it is critical that decisions about schools be made, as much as possible, on real and meaningful outcomes.

Using the LM quadratic, quadratic/interaction, and quadratic/race models as examples, Achievement Index values were calculated where both the criterion of ± 1 P-ACT+ scale score unit and statistical significance ($p < .05$) were required to identify an outlier school. The regression coefficients (intercepts) of the schools flagged as outlier schools were required to differ significantly from 0. [For schools with significant interaction terms, the sum of the intercept and interaction (slope) terms must differ significantly from 0.]

Comparing the Mean Residual and General Linear Models Approaches

For each approach, model, and subgroup, the percentage of schools identified as achieving educational growth as expected, below expectation, and above expectation were computed. Comparisons were made between the models shown in Table 1.

Table 1. **Mean Residual and Linear Models Used to Calculate Achievement Index Values**

Model		AI subgrouped by:
MR 1.	Quadratic	------
MR 2.	Quadratic	First quartile Third quartile
MR 3.	Quadratic	Core taken Core not taken
MR 4.	Quadratic	Black students White students
LM 1.	Linear	------
LM 2.	Quadratic	------
LM 3.	Quadratic, interactions (Mathematics, Composite)	First quartile Second quartile Third quartile
LM 4.	Quadratic, course work	------
LM 5.	Quadratic, course work, interactions (Mathematics, Composite)	First quartile Second quartile Third quartile
LM 6.	Quadratic, race	------
LM 7.	Quadratic, race, interactions (Mathematics, Composite)	First quartile Second quartile Third quartile
LM 8.	Quadratic ($p < .05$)	------
LM 9.	Quadratic, interactions ($p < .05$) (Mathematics, Composite)	First quartile Second quartile Third quartile
LM 10.	Quadratic, race ($p < .05$)	------

Comparisons among the different models and analytic methods were made. In addition, the schools identified as achieving growth above or below expectation for any approach or model were compared in terms of consistency of Achievement Index values across models and approaches. Further, the characteristics of these schools (community type, school enrollment, per-pupil expenditure, percent below federal poverty level in the district, and percent of black students in the district) were examined to determine the types of schools identified as outliers, and whether or not specific types of schools were being systematically identified as outliers using these approaches and models.

Results

For each content area, P-ACT+ score was regressed on selected CTBS scores using the models listed on pp. 11-12. The results showed slightly greater prediction accuracy, as measured by multiple R and SEE, for one specific model within each content area. Only the models with the greatest prediction accuracy for each P-ACT+ score are discussed here; information pertaining to the other models may be obtained from the author.

The models on pp. 11-12 with the greatest prediction accuracy, as measured by multiple R and SEE, were:

E3. P-ACT+ English = f{CTBS Total}

M2. P-ACT+ Mathematics = f{CTBS Total}

R1. P-ACT+ Reading = f{CTBS Total Reading, CTBS Study Skills}

S1. P-ACT+ Science Reasoning = f{CTBS Science, CTBS Total}

C2. P-ACT+ Composite = f{CTBS Total, CTBS Study Skills, CTBS Vocabulary, CTBS Science}

The results comparing the MR and LM approaches are shown in Table 2. The analytic method is specified in the first column of the table, followed by the relevant model. For the purpose of clarity, the models have been identified as MR 1-MR 4 and LM 1-LM 10, and will be referred to as such in this paper. The number of schools used in each analysis is provided in parentheses. The Subgroup column identifies the relevant population subgroup used to categorize students. In subsequent columns, the percentage of schools identified as achieving educational growth below expectation, as expected, and above expectation are provided for each set of predictor variables and criterion, as well as the multiple correlation (R) and standard error of estimate (SEE). For example, the first row pertains to P-ACT+ English score predicted from CTBS Total score. The MR 1 model resulted in 8% of the schools being identified as achieving growth below expectation and 15% as achieving growth above expectation. The linear (LM 1) and quadratic (LM 2) models resulted in identical percentages. However, prediction accuracy, as measured by multiple R and SEE, was greatest for the quadratic model (LM 4; multiple R = .78, SEE = 2.57).

Total Group Comparisons

As shown in Table 2, the linear (LM 1) and quadratic (LM 2) models identified slightly greater percentages of schools as achieving less than expected educational growth than did the MR 1 model for Mathematics, Reading, and the Composite. For example, 17% of the schools were identified as achieving less than expected growth in Mathematics using LM1 and LM2, compared to 10% for MR1. The results for English using these two models were identical.

The linear model (LM 1) also identified slightly higher percentages of schools as achieving growth above expectation than did either the MR 1 or LM 2 models for Science Reasoning and the Composite.

In general, the quadratic (LM 2) model resulted in somewhat greater prediction accuracy then did the MR 1 or LM 1 models, as measured by both multiple R and SEE. For all three models, the greatest prediction accuracy, as measured by multiple R and SEE, was found for the Composite, with multiple R ranging from .81 to .84, and SEE ranging from 1.66 to 1.72.

In terms of the specific schools identified as achieving growth above or below expectation, the three methods were relatively consistent in identifying the same schools. The models tended to identify two to four schools (4-8% of the schools) differently within each subject area except for Reading. For Reading, AI values for eight schools (17%) differed between the linear and quadratic models.

Comparisons of Population Subgroups

Achievement Index results were compared using the MR and LM approaches for the top and bottom quartiles of the distribution of CTBS Total scores, by course work, and by race (black and white students). In examining the course work variables, the core curriculum variable that required students to have taken foreign language (# 4) resulted in the largest number of schools of sufficient size for analysis (K = 22; n ≥ 15). The results using this course work variable are therefore reported here; information about the other course work variables may be obtained from the author.

MR Models. As shown in Table 2 for the MR approach, there were very few, if any, schools for which sufficient (N >= 15) sample sizes were available for calculating AI values for most population subgroups. Even when the minimum sample size was reduced to 10 for the racial/ethnic groups, the numbers of schools did not increase substantially. The analysis for the top and bottom quartiles was initially conducted using the top and bottom quartiles for the total group of students. Because the sample sizes for most schools were less than 15 using this approach, this analysis was also conducted using the quartile values for each school. The results of this analysis still showed very small sample sizes (N < 15), too small to make statements about the level of student achievement at these schools.

The results by core curriculum revealed that the percentages of schools flagged as outlier schools differed for all P-ACT+ scores when course work taken was considered. For all P-ACT+ scores except the Composite, for students identified as having taken the core curriculum, no schools were identified as achieving growth below expectation. In contrast, when noncore curriculum students were examined, no schools were identified as achieving growth above expectation for English and Mathematics. When considering only the noncore curriculum group, the results for Reading, Science Reasoning, and the Composite were similar to those for the total group (MR 1).

An additional analysis was conducted to examine the distribution of AI values for each P-ACT+ test obtained for each racial/ethnic group, regardless of the school attended. Figures 1 through 5 depict the distributions of AI values for each racial/ethnic group. As the graphs show, the shapes of the distributions are similar, except that the white distributions show greater spread in AI values than do the black distributions.

The LM Approach. The results using the LM approach (see Table 2) showed AI differences for the top and bottom quartiles for Mathematics and the Composite, as shown by the significant ($p < .05$) high school by CTBS Total score interaction terms in the models for these P-ACT+ scores. Percentages for the quartiles were based on only those schools with valid CTBS Total scores ranging from less than 750 to than 793. School without students scoring above or below the quartiles were not included. A maximum of two schools were deleted for any given percentage. Slightly fewer schools were identified as achieving growth below expectation for Mathematics at the top quartile than for the total group quadratic model (LM 2). For the Composite, the percentage of schools identified as achieving growth below expectation tended to decline from the bottom to top quartiles, with 13% of the schools identified as such for the bottom quartile, and 4% identified for the top quartile. SEE tended to decrease slightly for both P-ACT+ Mathematics and Composite scores; multiple R was similar. It may be assumed, due to the nonsignificant interaction terms for English, Reading, and Science Reasoning, that the within-school AI values for these scores did not differ across CTBS Total score quartiles.

When course work was added to the quadratic model (LM 4), the percentages of schools identified in each AI category were similar to those for the quadratic model (LM 2) for English, Reading, and Science Reasoning, but not for Mathematics or the Composite. For Mathematics, the percentage of schools identified as achieving growth below expectation was 5% lower for the quadratic/course work model. Multiple R was slightly higher for LM 4 (.74), and SEE was smaller (2.33). For the Composite, no schools were identified as achieving growth below expectation, as compared to 8% for the LM 2 model. Ten percent were identified as achieving growth above expectation, as compared to 4% for LM 2. Multiple R and SEE were similar to those for LM 2.

The quadratic/course work models for Mathematics and the Composite were also evaluated by CTBS Total score quartile (LM 5). The results showed variability across quartiles in the percentages of schools identified as performing below expectation, with percentages ranging from 12% to 22% for Mathematics, and 2% to 10% for the Composite. In general, the percentages of schools identified as achieving growth above expectation for the LM 5 models tended to be similar to those obtained for LM 4.

When race was added to the quadratic model (LM 6), it was found to be not statistically significant ($p > .05$) using these models. Though not statistically significant, AI values were calculated using the quadratic/race models to determine the influence of race on the categorization of schools (shown as LM 6 and LM 7). In comparison to LM 2, the addition of race to the quadratic model (LM 6) resulted in similar percentages of schools in each category for all P-ACT+ tests except Reading, where fewer schools were identified as achieving growth above and below expectation. When considered by quartile (LM 7), some variability was found in the percentages of schools identified as achieving growth below and above expectation for Mathematics, with percentages ranging from 15 to 21%. In general, the results by quartile were similar to those for the quadratic/race model for Mathematics. The percentages of schools identified as achieving growth below expectation using the Composite varied somewhat across quartiles, with percentages ranging from 5 to 13%.

Statistical significance. Three models were selected to determine the impact of requiring statistical significance ($p < .05$) on the percentages of schools identified in each AI category. There is no easy method for testing statistical significance for the MR approach; therefore, three LM models were selected (LM 2, LM 3, and LM 6). The results for these models are reported in Table 2 as LM 8-LM 10.

For models LM 8 and LM 10, requiring statistical significance in order to be identified as an outlier school tended to result in similar or slightly smaller percentages of outlier schools, as compared to those for LM 2 or LM 6, for English, Mathematics, Science Reasoning, and the Composite. Substantially smaller percentages of outlier schools were identified for Reading. For LM 9, the percentages were generally slightly smaller than those for LM 3; the results for the Composite were identical.

Summary. The percentages of schools identified as achieving educational growth below expectation across all models (MR 1-LM 7) ranged from 43% to 67%, depending on subject area. For schools flagged as achieving growth above expectation, the percentages of schools ranged from 38% to 88%, with English models being the most consistent and Reading models being the least consistent.

Inconsistencies across approaches most frequently occurred when course work or race was added to the models. On average, the AI values for 4 schools differed from those for LM 2 when course work or race was added to the model. In most cases, the differences in the regression coefficients corresponding to these AI values exceeded ± .20 P-ACT+ scale score units.

School Characteristics and the Achievement Index

Initial examination of the characteristics of the schools identified as outliers showed a tendency for certain types of schools to be flagged as outlier schools for certain P-ACT+ tests.

In order to evaluate possible systematic AI/school characteristic relationships, community type, per-pupil expenditure, enrollment, percentage of students in the district below federal poverty level, and percent of black students in the district were correlated with AI values for models MR 1 and LM 1 through LM 7.

The results showed statistically significant ($p < .05$) positive correlations between per-pupil expenditure and AI values calculated from all MR and LM Science Reasoning models, with values ranging from .31 to .37. Schools with higher per-pupil expenditures were more likely to be flagged as achieving growth above expectation for Science Reasoning. Percent of students in the district below federal poverty level was negatively correlated with the AI values from all Mathematics models, with correlations ranging from -.61 to -.35. Schools with high percentages of students in the district below federal poverty level were more likely to be identified as achieving growth below expectation for Mathematics. Finally, community type was statistically significantly ($p < .05$) related to Composite AI values for model LM 7, but only for the top and bottom quartiles. Urban and suburban schools were more likely to be identified as achieving growth above expectation than rural schools, using this model.

Conclusions

Goal 1 of the Tennessee project was to determine the accuracy of predictions of P-ACT+ scores using CTBS scores as predictors. As measured by multiple R and SEE, these findings show accurate predictions of P-ACT+ scores, with multiple R ranging from .59 to .84, and SEEs ranging from 1.63 to 3.28 across P-ACT+ test scores. Further, these models were used to address and meet Goal 2 by developing Achievement Index values for each high school. The results showed that the models were able to differentiate among institutions in terms of the skills and knowledge measured by both tests, thus illustrating school effectiveness in those areas.

The results of this study also support the use of the LM approach for calculating AI values for the purpose of evaluating educational growth (Goal 3), rather than using the mean residual approach. The LM approach, particularly the quadratic model, resulted in greater prediction accuracy than the MR approach, which is not a unexpected result. Further, the LM approach identified similar or greater percentages of schools as achieving educational growth above or below what might be expected.

The major advantage of the LM approach was shown in the results for selected population subgroups. Using the mean residual approach severely limited the number of schools that could be studied, and thus the subgroups that could be compared. In order to use the MR method effectively, much larger samples would be needed within each school. Using the LM approach allowed for subgroups within schools with very small sample sizes.

In examining the results by race subgroup, the distributions of AI values from the mean residual approach for black and white students showed very similar results. However, as shown by the LM approach, even though race was not a statistically significant predictor variable, including it in the models tended to change the AI values for some schools. This result might suggest a greater sensitivity to the racial/ethnic composition of the schools using the LM approach. However, even with the LM approach, AI values differed by race/ethnicity, which should be considered in evaluating educational growth.

The importance of including interaction terms in the models (i.e., taking into consideration the entering CTBS achievement level of the student), in particular for Mathematics and the Composite, was shown by the differences in AI percentages between the bottom and top quartiles. In some cases, schools were identified as achieving educational growth below expectation (or above expectation) for the bottom quartile, but the opposite for the top quartile.

Adding statistical significance to the criteria for identifying outlier schools did not substantially reduce the percentages of outlier schools identified, with the exception of Reading.

These results suggest that the relatively large percentages of outlier schools for Reading might be a function of sampling error, and not as a re^ult of true educational growth. Further analysis with larger sample sizes is needed to substantiate this hypothesis, however. Given the strong potential for sampling error in these models, we recommend the inclusion of statistical significance in identifying schocls achieving educational growth above or below expectation.

The MR approach, compared to the LM approach, lacks the ease of computation of AI values, in particular when statistical significance is required. Standard statistical packages can be used to develop regression weights, interaction terms, and to calculate statistical significance for the LM models. Comparable requirements using the MR approach would require elaborate matrix programming, in addition to standard multiple regression modeling.

It should be noted here that to fully validate the use of these prediction models, other information is needed. Research would need to determine that schools achieving below expectation are also low on educational process variables such as per-pupil expenditure; the educational level of teachers or assessment scores of teachers; the attitude of the administration towards the teachers, students, and the school as a whole; and other non-test information. This study showed that per-pupil expenditure and percentage below the federal poverty level in the district are related to Mathematics and Science Reasoning Achievement Index values. Schools with low per-pupil expenditure, and those with higher percentages of students below the federal poverty level in the district, were more likely to have low Achievement Index values. These results would appear to support the use of the models for these purposes.

Program evaluation requires an intensive study of schools, and draws on several sources of information. The Achievement Index is one such source of information, and provides preliminary evidence for targeting areas for further review and study. It should always be accompanied with and followed by collection of additional information related to the students and to the school environment. It is not appropriate for the AI to stand alone as a measure for program evaluation.

REFERENCES

American College Testing. (in press). Supplement to the P-ACT+ Technical Manual. Iowa City, Iowa: Author.

National Council on Education Standards and Testing. (1992) Raising standards for American education. Washington, D.C.: Author.

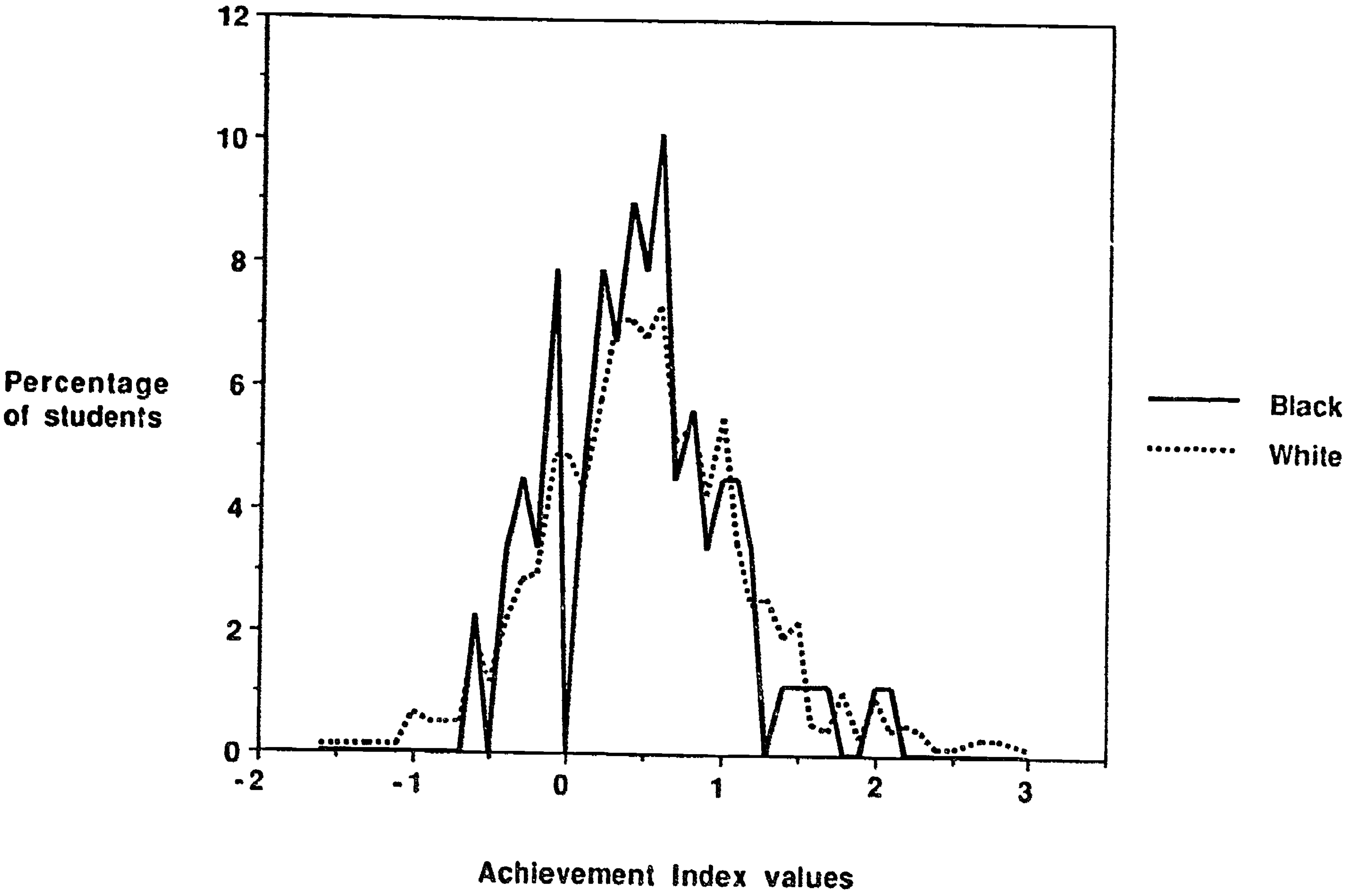

Figure 1. Student Achievement Index Values by Race (P-ACT+ English; Residual Model)

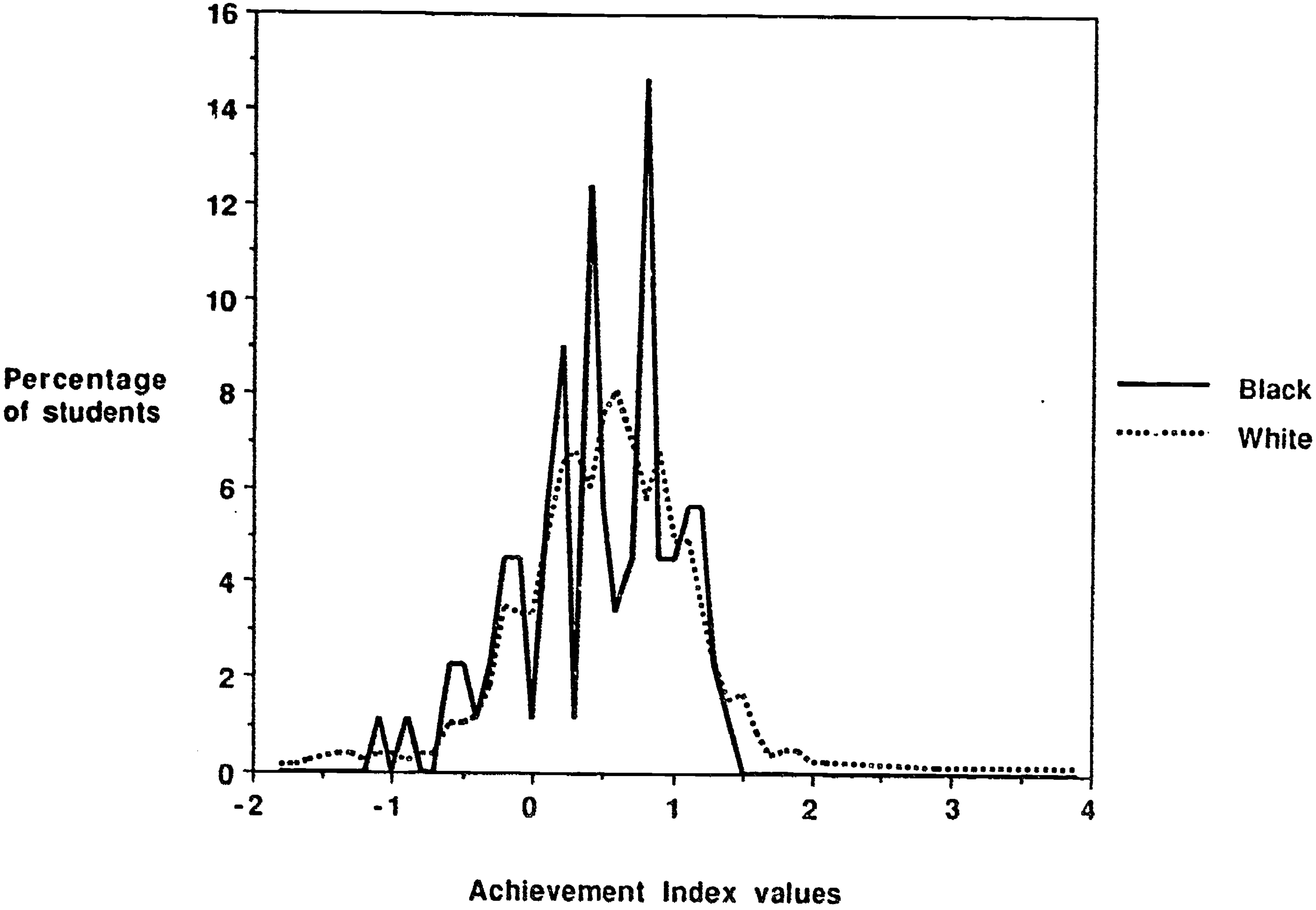

Figure 2. Student Achievement Index Values by Race (P-ACT+ Mathematics; Residual Model)

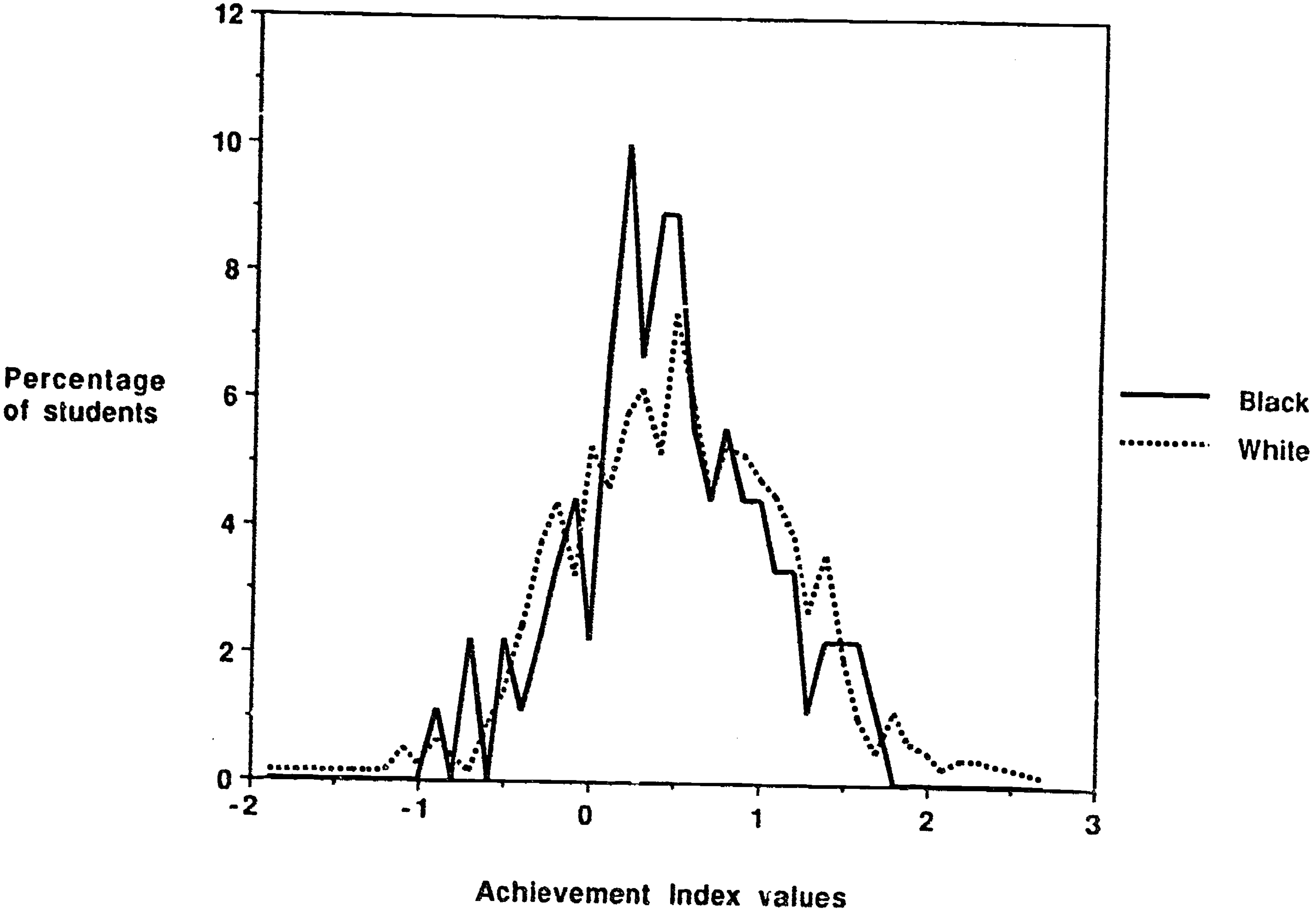

Figure 3. Student Achievement Index Values by Race (P-ACT+ Reading; Residual Model)

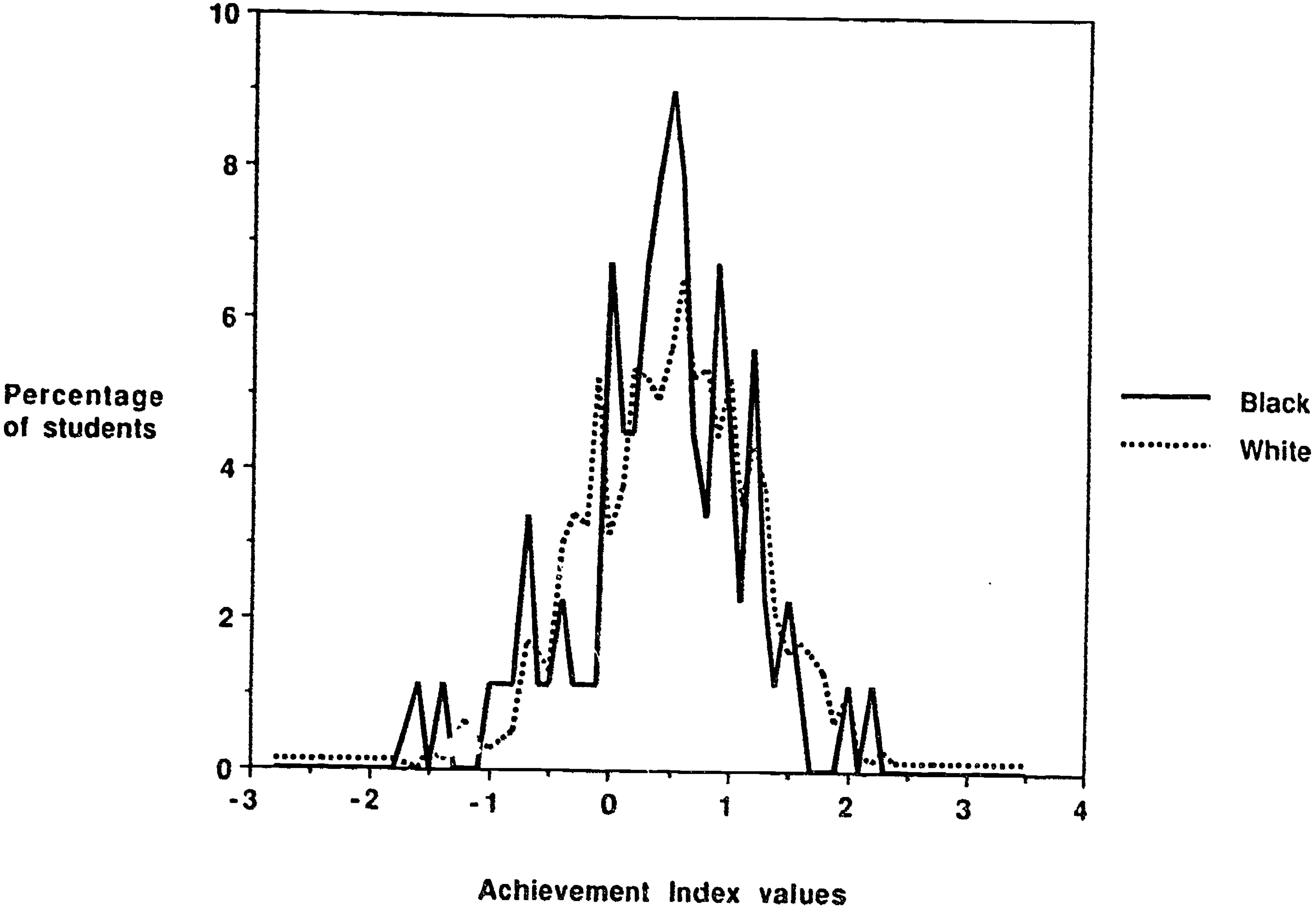

Figure 4. Student Achievement Index Values by Race
(P-ACT+ Science Reasoning; Residual Model)

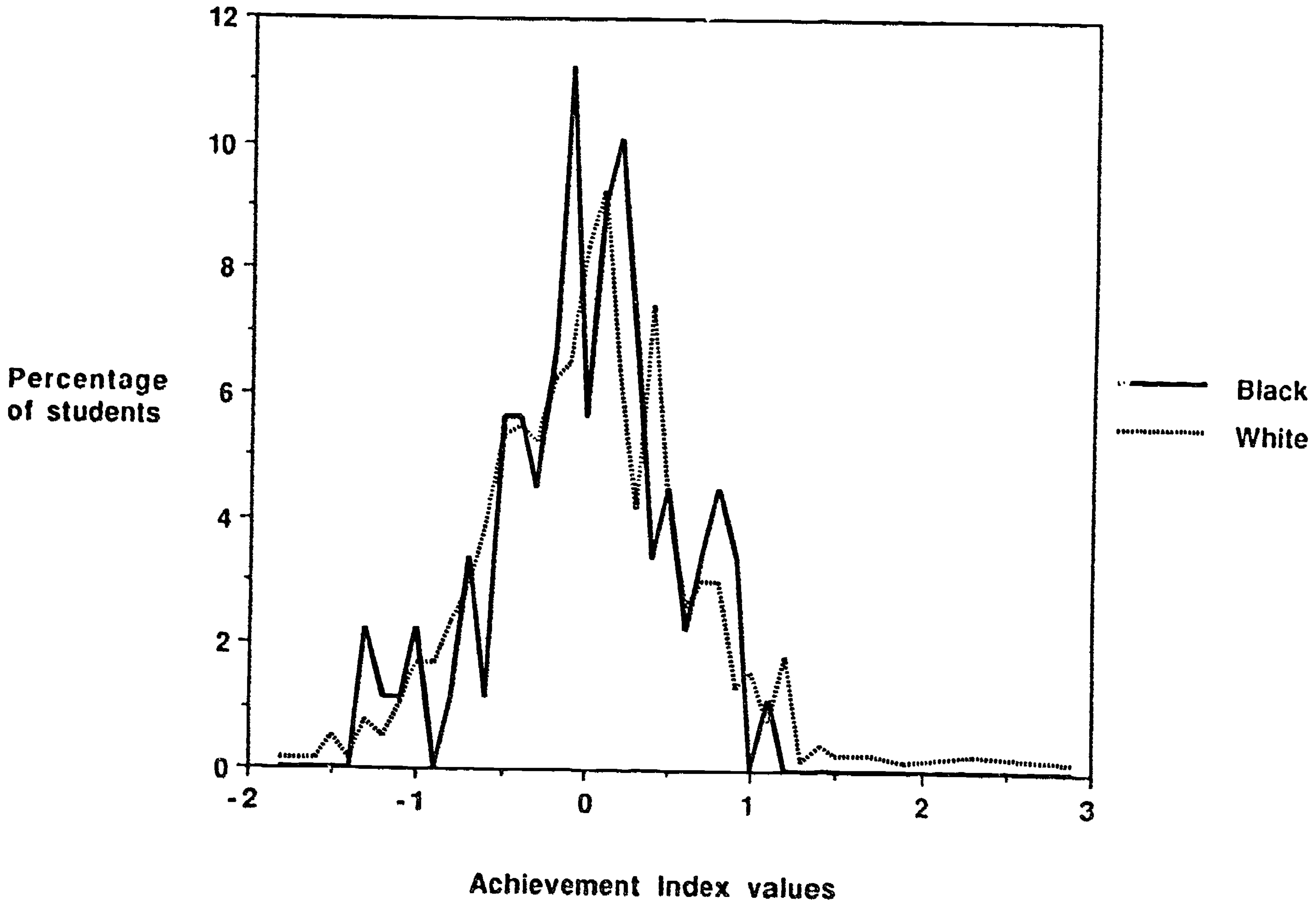

Figure 5. Student Achievement Index Values by Race (P-ACT+ Composite; Residual Model)