DOCUMENT RESUME

ED 171 786 TM 009 257

AUTHOR Hill, Richard K.
TITLE Use of the Rasch Model to Solve Data Problems Encountered by the California Assessment Program.
PUB DATE Apr 79
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Black Students; Caucasian Students; *Complexity Level; Cultural Differences; *Data Analysis; Item Analysis; *Mathematical Models; Primary Education; Racial Differences; Reading Achievement; *Reading Tests; Spanish Americans; State Programs; *Test Bias; *Testing Problems; Testing Programs; Test Items
IDENTIFIERS California; California Assessment Program; Rasch Model; *Rasch Scaled Scores

ABSTRACT

Four problems faced by the staff of the California Assessment Program (CAP) were solved by applying Rasch scaling techniques: (1) item cultural bias in the Entry Level Test (ELT) given to all first grade pupils; (2) nonlinear regression analysis of the third grade Reading Test scores; (3) comparison of school growth from grades two to three, using the Reading Test; and (4) analysis of growth from grades two to three, based on the Reading Test, in the areas of word identification, vocabulary, comprehension, and study locational skills. Solution of the problems demonstrated that existing Rasch Models have practical significance and should be more widely used by educators. (MH)

Use of the Rasch Model to Solve Data Problems Encountered by the California Assessment Program*

Richard K. Hill

RMC Research Corporation

This paper presents four problems that the staff of the California Assessment Program (CAP), the statewide testing program for the state of California, found difficult to solve. Each of these problems proved to be readily solvable when techniques being developed by advocates of the Rasch model were applied. The purpose of this paper is not to present new approaches for using the Rasch model, but to demonstrate that the approaches that have been developed already have great practical significance and should be disseminated and used more widely by practitioners.

Item Cultural Bias in the Entry Level Test

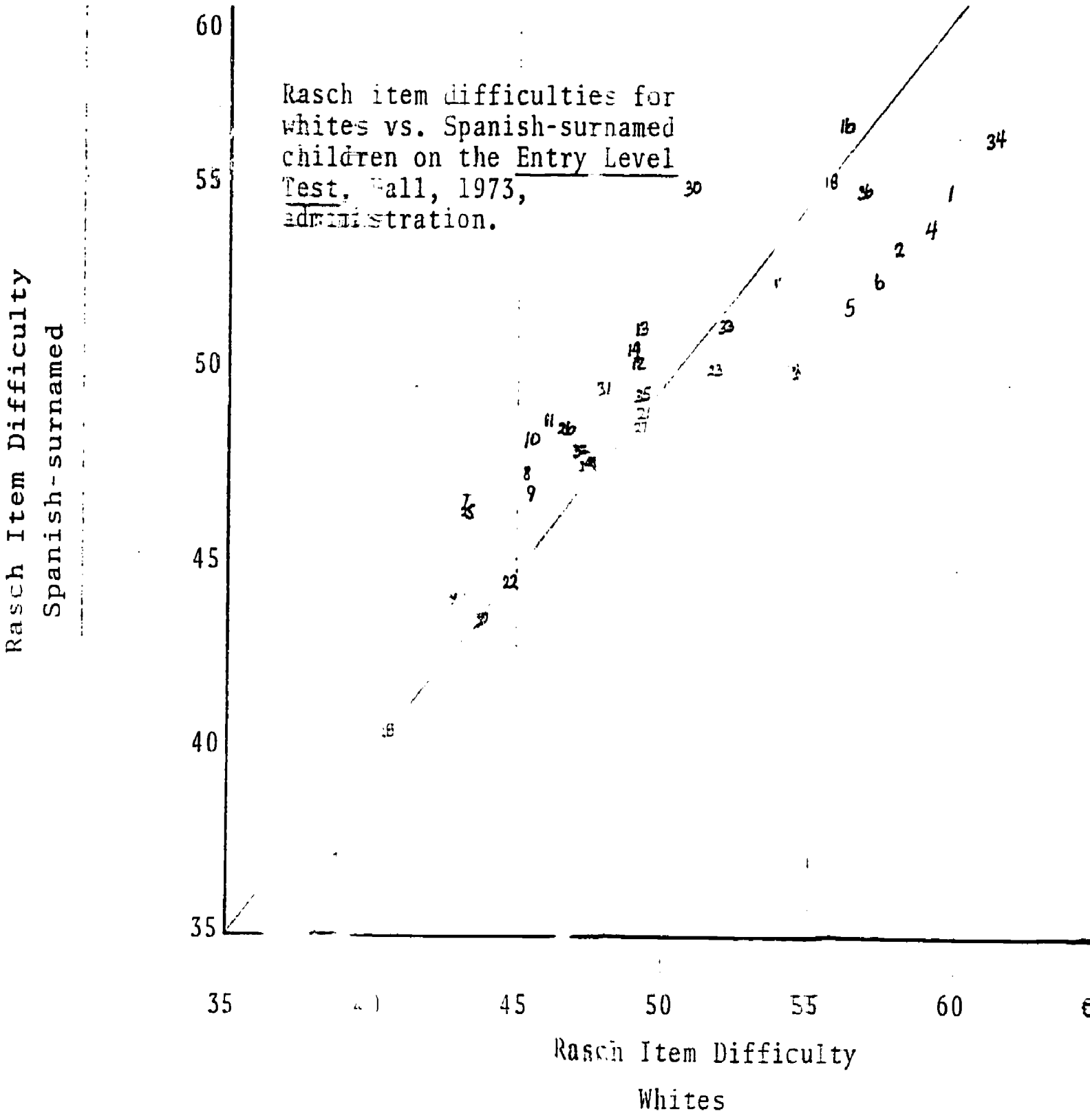
In order to collect baseline data, the California Assessment Program annually administers the Entry Level Test (ELT) to every first grader in California each September. At the time of the first administration of the ELT in September, 1973, it consisted of 36 items covering five subtests. Along with item data, the ethnic group of each child was reported. A one percent systematic sample of the state resulted in a file of 3,010 pupils available for analysis. The problem was to determine which, if any, of the test items contained cultural bias.

The approach taken by the staff at that time was to run a factor analysis, considering responses to each of the items and membership in each of the ethnic groups to be a variable. The factor structure of the test itself was quite clean; most of the 36 items loaded into only one factor, and loaded jointly only with all the other items in their own subtest. The loadings of the ethnic groups were much less definitive, and it was not known if the problem was one of statistics, such as the restriction of range of interitem correlations when items are extremely easy (several items in the test had p-values greater than .9), or if the items in the tests were truly unbiased. The results of this analysis were reported by Lorrie Shepard at the 1975 NCME Annual Meeting in a paper entitled Developing the California Entry Level Test: Construct Validity of the Subtests. The opinion of the staff was that the amount of cultural bias in the test was unclear, although there was belief that it was relatively unbiased.

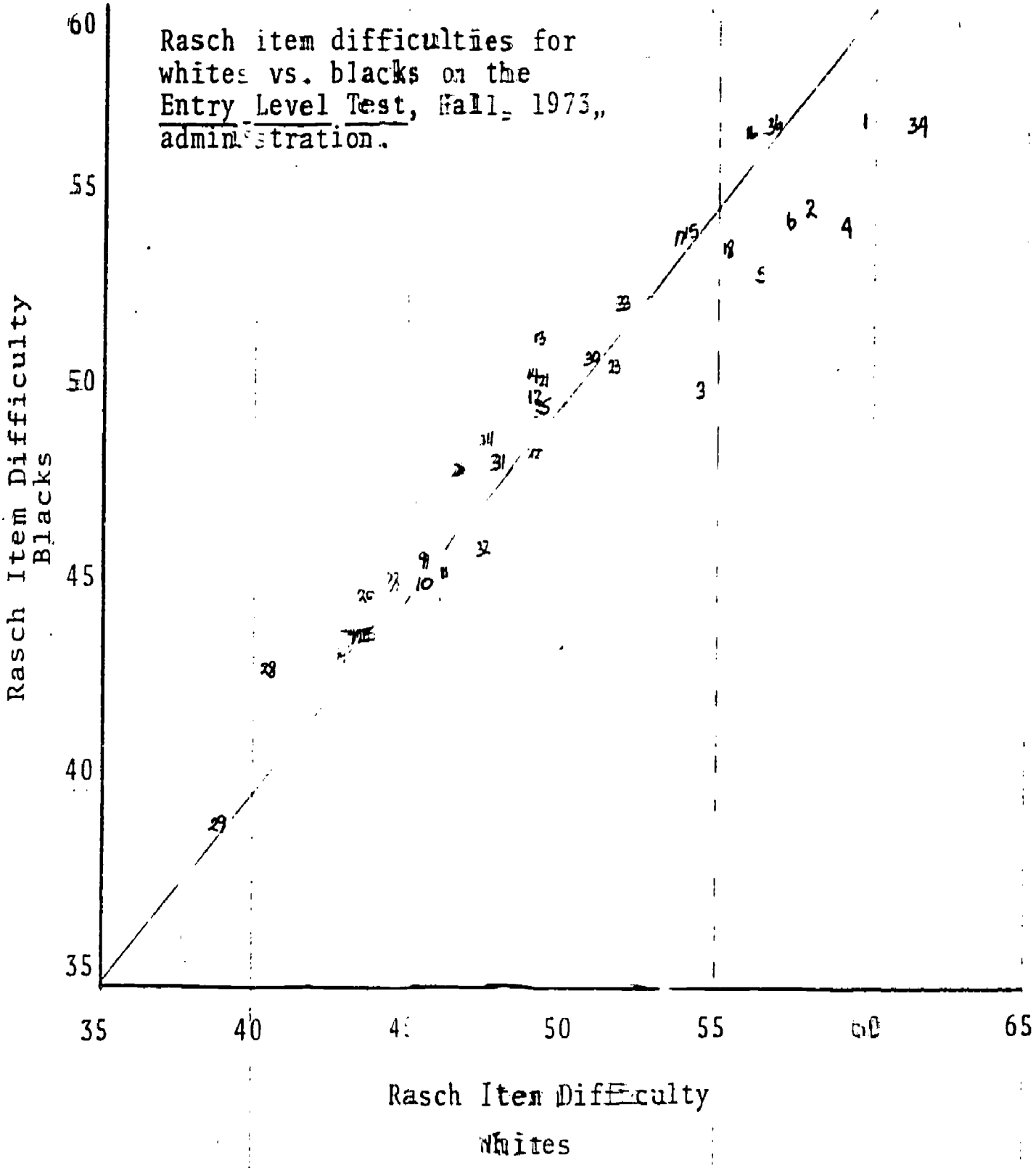
The data were reanalyzed last spring using the Rasch model. Rasch item difficulties were computed separately for whites, blacks, and Spanish-surnamed children. Plots of Rasch item difficulties for whites versus blacks and whites versus Spanish-surnamed were constructed. These two plots are shown as figures 1 and 2. Each plot demonstrated two distinct patterns of straight lines - one line consisting of the first six items on the test, and one very difficult item from the end of the test, and a second line drawn from the remaining items.

*Paper presented at the 1979 Annual Meeting of the National Council on Measurement in Education, San Francisco.

Figure 1


Rasch item difficulties for whites vs. Spanish-surnamed children on the Entry Level Test, Fall, 1973, administration.
Rasch Item Difficulty
Spanish-surnamed
60
55
50
45
40
35
35
45
50
55
60
Rasch Item Difficulty
Whites

igure 2


Rasch item difficulties for whites vs. blacks on the Entry Level Test, Fall, 1973, administration.
Rasch Item Difficulty Blacks
60
55
50
45
40
35
35
40
50
55
65
Rasch Item Difficulty Whites

The first six items should have been the most culture-fair items on the test. They all were classified as Immediate Recall, and consisted of the teacher showing the children pairs of objects, putting the pictures away, and then having the children choose the picture that matched the stimulus picture. If it is indeed true that these were culture-fair items, then it can be shown from the graphs that some cultural bias is contained in the vast majority of items in the ELT.

To compute the amount of bias, the correlations and associated regression lines were computed for whites vs. blacks and whites vs. Spanish-surnamed for both the 7 extraordinary items and the remaining 29 items. The results are displayed in Table 1.

Table 1

Correlations and Regression Equations

	Regression of Blacks on Whites	Regression of Spanish-surnamed on White
Correlation		
Items 1-6 and 34	.940	.996
Remaining 29 items	.968	.926
Regression		
Items 1-6 and 34	Y = .973 - 1.72	Y = .890X + 1.98
Remaining 29 items	Y = .925X + 4.40	Y = .902X + 5.77

The amount of bias can be defined as the difference between the scores predicted from the "unbiased" regression line (the one computed on items 1-6 and 34) and those predicted from the "biased" regression line. Using a typical score of 50, this translates into a bias of 3.72 points for the blacks and 4.39 points for the Spanish-surnamed.

Nonlinear Regression Analysis

As part of the reporting system of the California Assessment Program, multiple linear regression analyses are done for each achievement test, relating a series of predictor variables to school mean achievement. While the several regression analyses done at most grades are linear, the third grade Reading Test analysis is not. A possible reason for this is because the Reading Test is quite easy for third graders, and the subsequent ceiling effect produces nonlinear relationships between the predictor variables and third grade Reading Test scores.

The solution to this problem for several years was to include second and third order moments of the principal predictor (ELT scores) in the regression equation. While this approach removed the nonlinearity, it clearly is a patchwork approach to a measurement problem.

As a potential alternative, the regression was rerun using transformed Reading Test scores. The scores were transformed by taking the natural log odds ratio-ln (P/(100-P)). This transformation was made, rather than transforming to Rasch scaled scores because a) the two transformations are highly similar and b) because of the matrix sampling procedures, no definitive method for obtaining a scaled ability estimate for a school has been agreed upon.

The regressions run on the transformed data were far more linear than the regressions that had been run on the untransformed Reading Test scores. This can be seen from the results reported in Table 2. With the untransformed data, the square and cube of ELT deviations loaded before mobility scores, added a total of .649 percent to the variance accounted for, and had beta weights of -.18 and -.10, respectively. After the Reading Test scores had been transformed, the square and cube of ELT deviations were the last variables to be loaded, added only .107 percent to the variance accounted for, and had beta weights of just .03 and -.11, respectively.

A negative finding that resulted from the transformation is that less variance is accounted for by the predictors (84.7% vs. 87.3% on the untransformed data). This may be occurring because some of the predictor variables need to be transformed as well. In particular, ELT, AFDC rate and the bilingualism rate all have skewed distributions. If they were transformed, it is possible that the regressions would become even more linear, while at the same time increasing the variance accounted for. In support of this hypothesis, note that when the Reading Test results were transformed, the simple correlations between Reading Test scores and the two relatively symmetrical variables (socioeconomic status and mobility) increased slightly, while the correlations to the three variables with skewed distributions (ELT, AFDC rate and bilingualism rate) declined.

Comparing School Growth from Grade Two to Grade Three

The same Reading Test is given both to all second and third graders in California. While the test is moderately difficult for second graders, it is quite easy for third graders. The ceiling effect that complicated the third grade regression analysis also presented problems when it was desired to examine growth from grade two to grade three.

This issue arose when, under a contract from the California State Department of Education, staff at SRI International wanted to compare the changes of test scores from grade two to grade three of different types of classes in schools in California. By drawing ogives on linear/normal graph paper, it could be demonstrated that higher scoring schools exhibited less percentage growth on the Reading Test from grade two to grade three than did lower scoring schools. The ogives are shown as Figure 3. Since this result is contrary to all expectation (it is usually observed that higher scoring children exhibit more achievement growth, not less, when advancing to higher grades), it was assumed to be caused by the ceiling effect of the third grade Reading Test.

Table 2

Results of Multiple Regression Analysis
on Third Grade Reading Test Scores

Untransformed Reading Test Scores

Variable	Multiple R Square	R Square Change	Simple R	Beta
ELT	.67482	.67482	.82143	.37258
AFDC rate	.72899	.05417	-.69321	-.24975
Bilingualism rate	.74575	.01676	-.63371	-.12376
Socioeconomic Status	.75516	.00940	.76847	.20835
(ELT-27.321)**2	.75998	.00482	-.[illegible]	-.17973
(ELT-27.321)**3	.76165	.00167	-.[illegible]	-.10481
Mobility	.76212	.00046	-.[illegible]09	-.02228

Transformed Reading Test Scores

Variable	Multiple R Square	R Square Change	Simple R	Beta
ELT	.62195	.62195	.78864	.39197
Socioeconomic Status	.69673	.07478	.77847	.28059
AFDC rate	.70937	.01264	-.66579	-.19341
Bilingualism rate	.71429	.00492	-.60463	-.12412
Mobility	.71572	.00143	-.16646	-.03750
(ELT-27.321)**2	.71677	.00104	-.43131	.03412
(ELT-27.321)**3	.71679	.00003	.46035	-.01325

The probability that the problem was due to ceiling effect also could be shown from Figure 3. If the distribution of scores was normal, the ogive on linear/normal paper would be a straight line. These ogives were fairly linear until school mean scores reached approximately 80 percent, and then began to curve away from the straight line. This indicates that the distribution of school mean scores was negatively skewed, a likely outcome when ceiling effect is encountered.

The school mean scores were scaled by converting them to a log odds ratio, and the ogives redrawn. These ogives are shown in Figure 4. This time, the curves were almost perfectly linear. The ogive for grade two was almost parallel to the third grade ogive, except that higher scoring schools showed slightly more growth from grade two to grade three than did lower scoring schools. This result was consonant with expectations, and it was concluded that the log odds ratio scaling presented a more accurate picture of change from grade two to grade three than did percentage correct scores.

Analyzing Growth in Content Areas

After each year's testing, the Reading Test results are presented to an advisory committee for their review. The test has four major content areas (word identification, vocabulary, comprehension, and study-locational skills), and one question the committee posed was, "In which of the four areas is the growth the greatest from grade two to grade three?" The changes in percent correct scores were greatest for comprehension and vocabulary, and lowest for word identification and study-locational skills, but this result was confounded by the fact that word identification and study-locational items were the easiest on the test, and comprehension items the hardest. Thus, it was likely that ceiling problems were having differential effects on the results for the four content areas.

Given the fact that reading scores, relative to national norms, declined in California after the third grade, the committee had expected to find growth to be poorest in comprehension. Since changes in percent correct scores were greatest in comprehension, the suspicion was that ceiling effect was confounding the interpretation.

Table 3 shows the results presented to the advisory committee. It shows that gains from grade two to grade three are negatively correlated with the difficulty of the content area.

To address this problem, Rasch difficulties were computed at each grade for the 250 items on the test. To do this, a two percent systematic sample of the students tested statewide was selected, and the analysis done on the samples of approximately 6,000 students per grade. The 250 items on the Reading Test had been divided into 10 parallel forms. Therefore, 10 analyses were done at each grade, each consisting of the responses of approximately 600 students to 25 items each. After Rasch difficulties were computed for each item at each grade, they were summed across items for each content area to get a mean difficulty for each area.

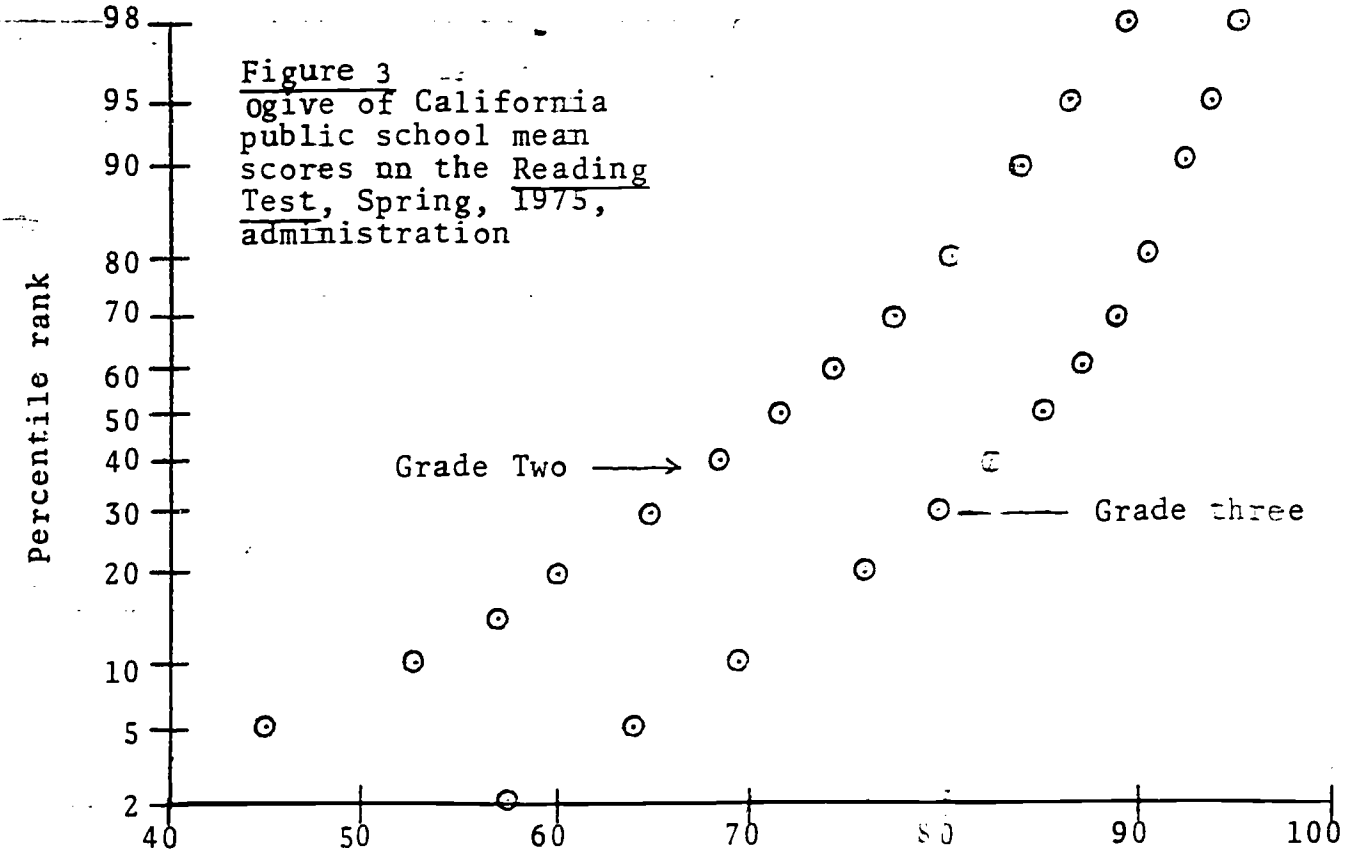


Figure 3
Ogive of California public school mean scores on the Reading Test, Spring, 1975, administration

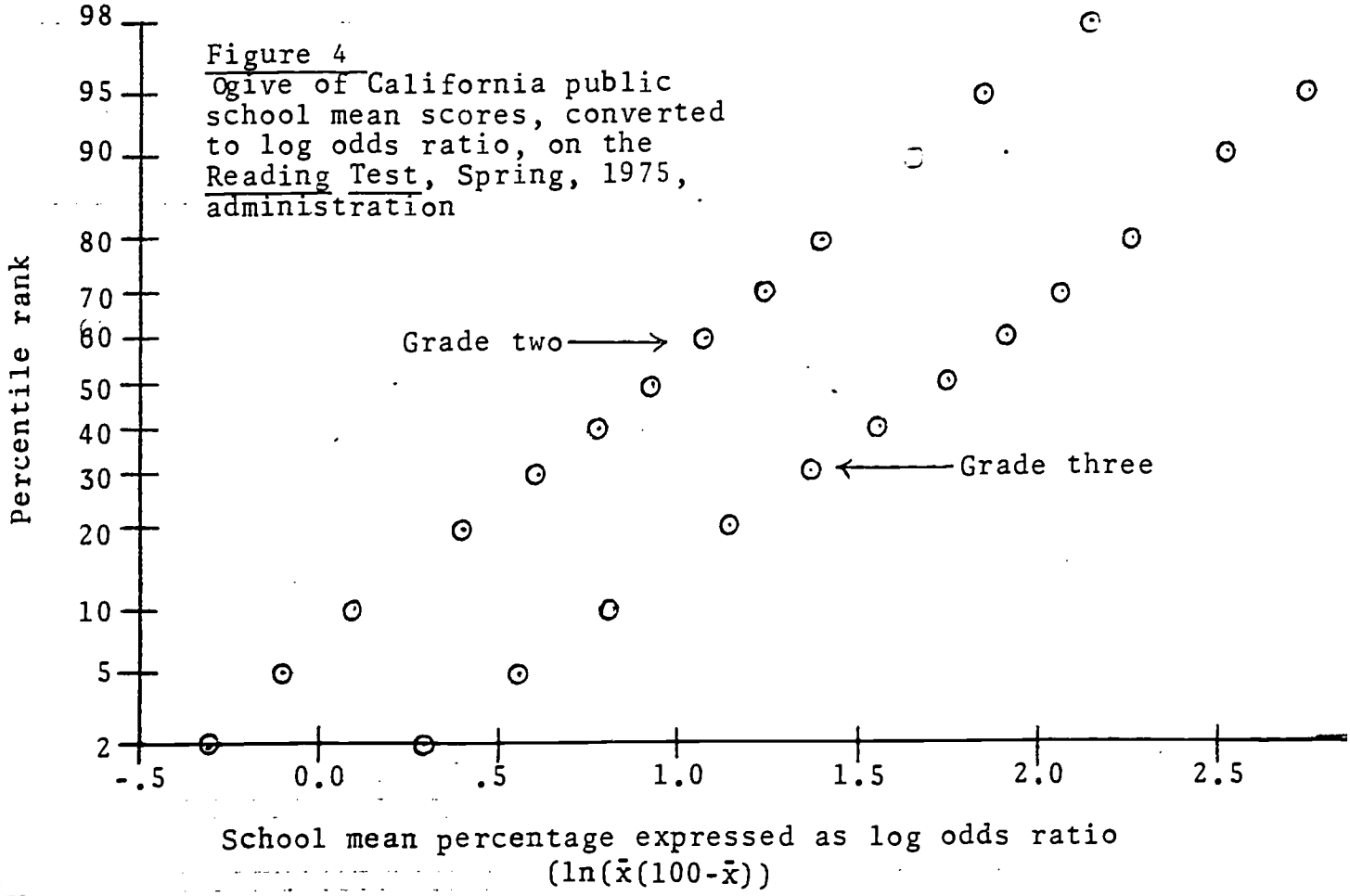


Figure 4
Ogive of California public school mean scores, converted to log odds ratio, on the Reading Test, Spring, 1975, administration

Table 4 shows the results presented in terms of Rasch scaled scores. Since Rasch scaled scores increase as relative difficulty increases, a gain in scores from grade 2 to grade 3 means that an area became relatively more difficult. Therefore, the content area in which the most gain was made was Study Locational Skills, the area which was rated third out of four on the basis of increase in percentage of correct responses. Similarly, comprehension, which appeared to have the largest increase on the basis of percentage of correct responses, was third of the four content areas when rated on the basis of the Rasch scaled scores.

Table 3

Results for the Reading Test, 1975

Content Area	Grade 2	Grade 3	Difference
Word Identification	75.4	85.8	10.4
Vocabulary	67.7	82.6	14.9
Comprehension	61.3	77.0	15.7
Study-Locational Skills	75.5	88.0	12.5
Total Test	67.6	81.3	13.7

Table 4

Results for the Reading Test, 1975, Reported by Rasch Scaled Scores

Content Area	Grade 2	Grade 3	Difference	Correlation Between Grade 2 and Grade 3 Scores, Computed over Items
Word Identification	47.18	47.53	+.35	.97
Vocabulary	50.18	49.90	-.28	.89
Comprehension	51.91	51.97	+.06	.95
Study-Locational Skills	47.43	46.88	-.55	.93
Total Test	50.00	50.00	.00	.95

To avoid overinterpretation of these results, a simple one-way analysis of variance was run on the data. (While a multivariate or repeated measures design might have been more appropriate and more powerful, the cost of analyzing the data in so complex a manner was not judged to be worth the return. The analysis of gain scores using four groups of items was judged sufficient to provide a ballpark figure concerning statistical significance). The data produced an F-ratio of 3.71, with 3 and 246 degrees of freedom, which is significant at the .05 level, but not at the .01 level. However, there was no contrast of pairs that were significantly different from each other. Consequently, it seems safe to assume that if growth from grade two to grade three is greater in some content areas than others, those differential gains are so small that they are not readily detectable with current CAP data.

Summary

The Rasch model has been around for over a decade now, but practical applications of it still are in their infancy. This paper demonstrates that the average measurement practitioner needs to be made more aware of its potential uses and power. Four problems faced by the California Assessment Program that were either solved in makeshift fashion or left completely unresolved were solved simply and straightforwardly by application of the Rasch model. If Rasch scaling can be used this effectively, it is important that more "frontliners" be instructed in its use.