ED 243 965 TM 840 288

AUTHOR	Livingston, Samuel A.
TITLE	Item Selection and Pre-equating with Empirical Item Characteristic Curves.
PUB DATE	16 Apr 84
NOTE	11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Cutting Scores; *Equated Scores; Higher Education; *Item Analysis; *Latent Trait Theory; Minimum Competency Testing; *Predictor Variables; *Responses; Student Reaction; Test Construction
IDENTIFIERS	Common Item Effect; New Jersey College Basic Skills Placement Test

ABSTRACT

An empirical item characteristic curve shows the probability of a correct response as a function of the student's total test score. These curves can be estimated from large-scale pretest data. They enable test developers to select items that discriminate well in the score region where decisions are made. A similar set of curves can be used to predict the equating relationship between two forms of the test. These curves are based on the common-item score instead of the total score. They make it possible to estimate raw-to-scale score conversions before the test is administered. (Author)

April 16, 1984

Item Selection and Pre-equating with Empirical Item Characteristic Curves*

Samuel A. Livingston

Educational Testing Service

An empirical item characteristic curve shows the probability of a correct response as a function of the student's total test score. These curves can be estimated from large-scale pretest data. They enable test developers to select items that discriminate well in the score region where decisions are made. A similar set of curves can be used to predict the equating relationship between two forms of the test. These curves are based on the common-item score instead of the total score. They make it possible to estimate raw-to-scale score conversions before the test is administered.

*A paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April, 1984.

Item Selection and Pre-equating with Empirical Item Characteristic Curves

Samuel A. Livingston
Educational Testing Service

The New Jersey College Basic Skills Placement Test is a battery of tests used by state and county colleges in New Jersey to place students into or out of remedial courses in reading, writing, and mathematics. Educational Testing Service (ETS) develops a new form of the test each year. The colleges give these tests throughout the year. Six times a year they send the students' answer sheets to ETS to be scored. At the first of these six scoring "cycles", ETS does a common-item equating of the new test to the previous year's test. The test scores are reported on a score scale ranging from 135 to about 190. The statewide mean is about 165 and the statewide standard deviation is about 11. The standards for placement vary from college to college, but they tend to be in the region of about 160. The equated scaled scores for the first scoring cycle go out to the colleges about June 10, but some colleges need to make placement decisions in April and May. They can compute the students' raw scores themselves, but how can they adjust for possible year-to-year differences in the difficulty of the test?

The items for the test are selected on the basis of large-sample pretest data. Each test item is pretested by embedding it as an unscored item in the previous year's test. There are eight versions of the test, each with the same scored items but different unscored items. These eight versions are "spiraled", i.e., packaged in repeating numerical order. The effect is to divide the test-taker population into eight stratified samples. The item analysis is done after the second scoring cycle. By this time about 18,000 students have taken the test, so that each new item has been pretested on more than 2,000 students. But what is the best way to use this pretest data to select items for the next year's test?

In the New Jersey program at ETS, we are using the same basic technique to solve both of these problems -- pre-equating and item analysis. The technique is to use the pretest data to plot a curve for each item, showing the probability of a correct response to the item as a function of the student's score on the test.* We call these curves "empirical item characteristic curves," or "EICC's." (Some people prefer to call them "item-test regression curves.") Figure 1 is a sample EICC graph. When the test committee meets to select items, each committee member receives a set of these graphs. (The committee members' graphs also have conventional item analysis statistics printed at the top.) The graph shows how well the item discriminates in any particular portion of the score range. Notice that the total score, on the x-axis, is shown both as a raw score and as a scaled score.

To make this graph, the computer groups the students according to their total scores and plots the proportion correct for each group. This is the proportion of correct answers among those who respond to the item. Students who omit the item or who do not reach the item are not included in this calculation.

*This technique was introduced to the New Jersey program by Dr. Charles Pine, professor of physics at Rutgers University, Newark and chairman of the mathematics committee for the test.

The data points tend to line up fairly well in the higher score re‿ions, where there are many students at each score level. But at the low end, where there are fewer students, there is a lot of scatter. Some sort of smoothing is necessary. We use a three-stage procedure. First, we transform the percent-correct into a variable with constant variance. Then we fit a polynomial to the transformed values. Finally, we transform the smoothed values back into percent-correct.

The fitting of the polynomial is by weighted least squares. The weights are the sample sizes at each score level. We decided to use a third-degree polynomial after looking at some experimental results. Fourth-degree polynomials produced some questionable-looking curves, and any higher degree polynomial produced some very obvious cases of overfitting.

We estimate these curves not only for the new items, but also for the old items, since some of these will be selected to appear on next year's test. There is some spuriousness in the estimates for the old items, since the item is included in the total score, but most of the resulting bias is at the high and low ends of the ability scale. In the score range we are most concerned about -- roughly 60 to 80 percent correct -- the bias is small.

The EICC's we use for pre-equating the scores are somewhat different from the ones we use for selecting items. The independent variable is not the full test score. Instead, it is the student's score on the "anchor test" made up of the items appearing on both the old and new forms of the test. Also, the students who omit the item are included in computing the curves used for pre-equating. Figure 2 shows an example of one of these curves.

The pre-equating is done by determining the regression of the full test score on the common-item score, for both the old form and the new form. For each possible common-item score, we estimate an expected full-test score on the old form and an expected full-test score on the new form. We then equate these expected full-test scores on the two forms.

The logic of the procedure is similar to the logic of item response theory (IRT) equating. We are equating expected scores at several ability levels. The expected score is the sum, over all items on the test, of the probability of a correct answer for students at the specified ability level. The procedure differs from IRT equating in that we cannot condition on the student's actual ability level, but only on an estimate of it. Both the old-form and new-form expected scores are conditioned on this same ability estimate - the common-item score. Since we have a large, representative sample of students responding to each item, the method produces good results.

But are we really equating the tests? The definition of equating states that two tests are equated if, for a student at any ability level, it is a matter of indifference which test the student takes. This definition implies that at every ability level, the conditional distribution of the equated scores should be the same for the two tests. What we have is a first-order approximation to that condition. At every ability level, the conditional means of the equated scores are the same for the two tests. The measure of ability is the student's score on the common items. The use of the common-item score as a measure of ability introduces some noise into the process but preserves the symmetry of equating, since the common items make up the same proportion of the new test as of the old test.

The only EICC's we actually need for pre-equating are those for the new items. Table 1 shows an example of the procedure. The conditional expected scores on the old form are estimated directly by averaging the scores of the students at each common-item score level. The expected score on the new form is the sum of the score on the common items and the expected score on the new items. We get the expected score on the new items by summing the probability estimates from the EICC's.

In the example shown in Table 1, the new form was considerably harder than the old form, especially for students in the middle ability range. Consider, for example a student who answered ten of the twenty common items correctly. This student would have an expected total score of 19.55 on the old form but only 17.79 on the new form.

This method produces a point-to-point correspondence, which we approximate very closely by a series of connected straight line segments. The data points do not quite cover the full range of possible scores on the test; we have to extend the highest and lowest segments slightly beyond the data. However, the critical area for placement decisions is in the middle of the score range; the accuracy of the equating at the upper and lower ends of the scale is not critical.

How well does the procedure work? We will know for sure in June, when we compare the preliminary equating results with the results of the regular equating. However, we used a "quick and dirty" version of this procedure last year, predicting the equating on the basis of the curves used for item analysis. Although that procedure is somewhat biased, we were still able to predict the raw-to-scale score conversions for all four tests within one scaled-score point, even though one of the tests was much more difficult than it had been the previous year. This year we expect to do even better.

Handout to accompany "Item Selection and Pre-Equating with Empirical Item Characteristic Curves". Samuel A. Livingston NCME 1984.

Figure 1: Sample graph of empirical item characteristic curve used for item analysis

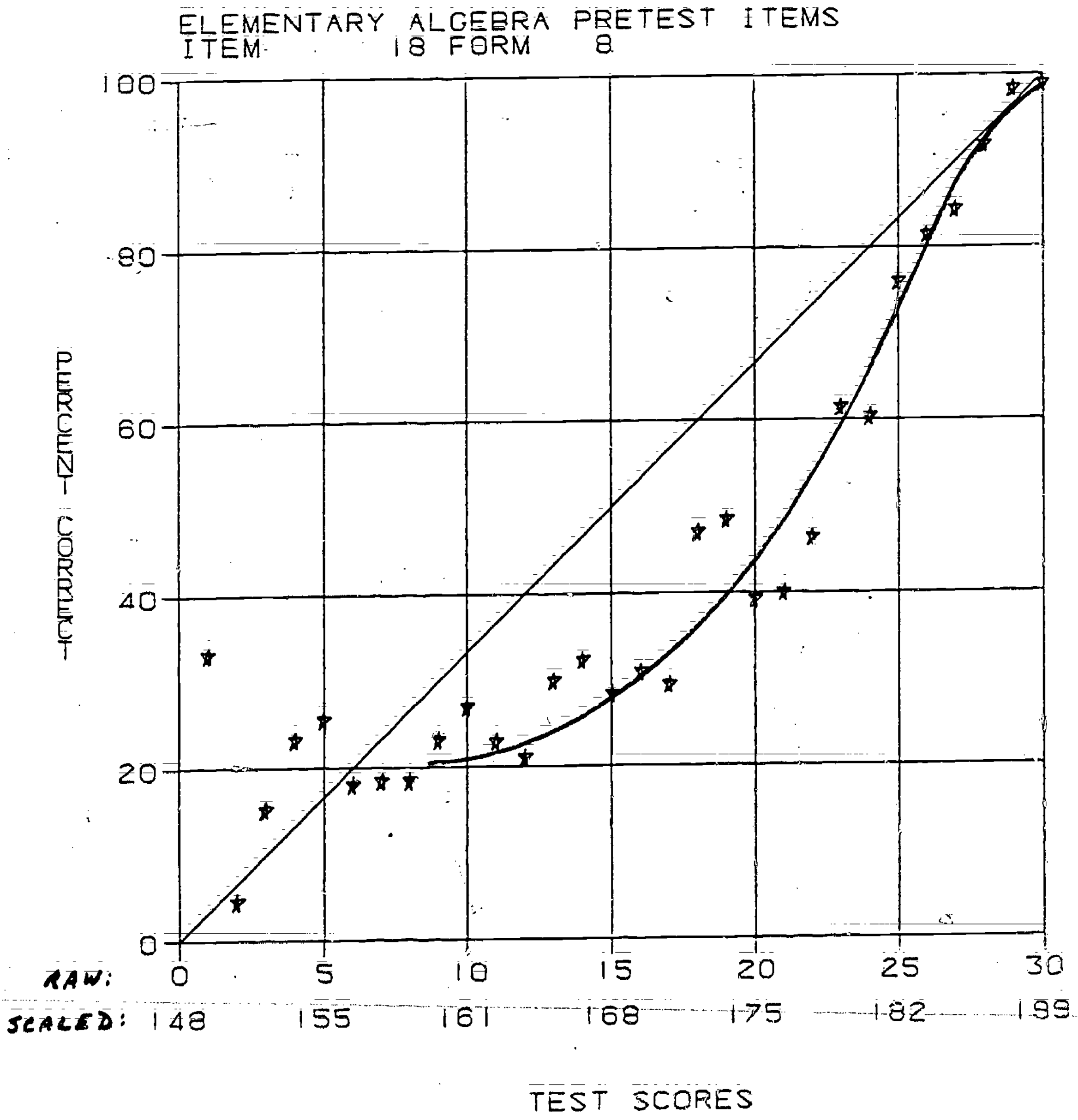

FIGURE 2. SAMPLE GRAPH OF EMPIRICAL ITEM CHARACTERISTIC CURVE USED FOR PRELIMINARY EQUATING.

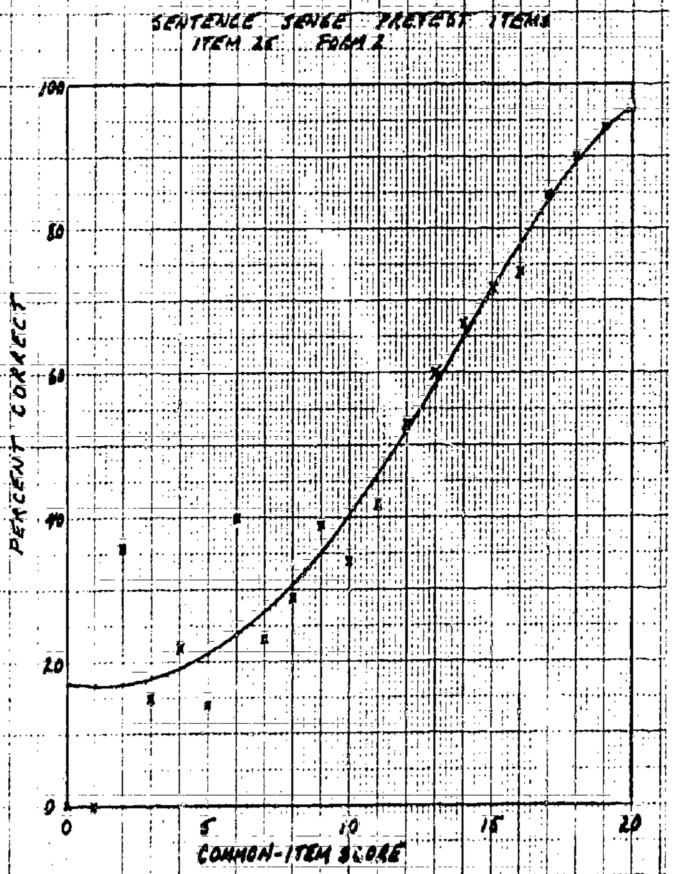

Table 1. Preliminary equating table

Common-item score	+	Sum of Smoothed P's (from EICC's)	=	Expected new form score	Expected old form score
0		2.28		2.28	3.80
1		2.67		3.67	5.57
2		3.14		5.14	6.80
3		3.67		6.67	8.01
4		4.23		8.23	9.91
5		4.81		9.81	11.32
6		5.39		11.39	12.91
7		5.99		12.99	14.46
8		6.58		14.58	16.16
9		7.19		16.19	17.96
10		7.79		17.79	19.55
11		8.40		19.40	21.06
12		9.02		21.02	22.67
13		9.64		22.64	24.14
14		10.27		24.27	25.61
15		10.90		25.90	26.98
16		11.53		27.53	28.38
17		12.15		29.15	29.73
18		12.77		30.77	31.08
19		13.35		32.35	32.40
20		13.87		33.87	33.71