DOCUMENT RESUME

ED 259 028 TM 850 402

AUTHOR Livingston, Samuel A.
TITLE Large-Sample Pre-equating: How Accurate?
PUB DATE Apr 85
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1-3, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Basic Skills; Cutting Scores; Diagnostic Tests; *Equated Scores; Higher Education; *Item Analysis; *Latent Trait Theory; Raw Scores; *Regression (Statistics); Remedial Instruction; *Scores; Statistical Studies; Student Placement; Test Items; Timed Tests
IDENTIFIERS Educational Testing Service; *New Jersey College Basic Skills Placement Test; *Pre Equating (Tests)

ABSTRACT

Four tests were pre-equated on the basis of item-test regressions computed from a large sample providing pretest data. The tests were two mathematics tests (math computation and elementary algebra) and two verbal tests (sentence sense and reading comprehension) from the New Jersey College Basic Skills Placement Test. Regular equipercentile equating results showed that pre-equating was highly accurate in three of the four tests. The fourth test (reading comprehension) showed a small but systematic inaccuracy in predicting the equating relationships--the difficulty level was underestimated, especially for students in the lower middle ability range. An investigation of this inaccuracy suggested that time limits may have affected the students' item responses. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED259028

Large-Sample Pre-Equating: How Accurate?

Samuel A. Livingston

Educational Testing Service

A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April, 1985.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- ✓ This document has been reproduced as received from the person or organization originating it
- Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Livingston, Samuel A.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Large-sample Pre-equating: How Accurate?

Samuel A. Livingston

Educational Testing Service

Four tests were pre-equated on the basis of item-test regressions computed from large-sample pretest data. Regular equipercentile equating results showed that the pre-equating had been highly accurate in three of the four cases. The fourth case showed a small but systematic inaccuracy in predicting the equating relationship. An investigation into the causes of the inaccuracy suggested that time limits may have affected the students' item responses.

Large-sample Pre-equating: How Accurate?

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", we perform a common-item equating of the new test to the previous year's test. 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.

For the benefit of these colleges, we perform an unofficial pre-equating, based on 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 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 regression of the full test on the equating anchor test can be estimated directly for the old form, which the students have already taken; we simply compute the conditional means. But, at the time of the pre-equating, no students have taken the entire new form. Therefore we have to estimate its regression synthetically, by the equation:

$$\begin{array}{c}\text{Expected score}\\ \text{on full-test}\end{array} = \begin{array}{c}\text{Score on}\\ \text{common items}\end{array} + \begin{array}{c}\text{Expected score}\\ \text{on new items}\end{array}$$

We estimate the expected score on the new items by computing a separate regression for each new item, and then summing over the new items. We base the individual item regressions on the pretest data. The regression for each new item takes the form of an item response curve, with the anchor test (common-item) score on the x-axis and the percent of students answering the new item correctly on the y-axis. The key assumption is that the response curve will be the same this year, when the item is included in the students' scores, as it was last year, when the item was being pretested. We refer to these response curves as "EICC's", short for "empirical item characteristic curves", and we call the method the "EICC method"*.

How well did the method work? That is, how well did it predict the actual equating results? Table 1 shows the equated scaled scores for selected raw scores on the 1983 and 1984 test forms. For the 1984 form, the table shows scaled scores from the EICC equating based on pretest data

---

* For a more detailed description of this method, see Livingston (1984).

and from the equipercentile equating* used to determine the official raw-to-scale score conversion. Although the scaled scores are rounded to the nearest integer for score reporting, Table 1 shows them to a tenth of a scaled-score point. A scaled-score point on these tests is slightly less than one-tenth of the population standard deviation. For the two math tests the EICC equating predicted the actual equating quite closely. However, the difficulty of these tests changed very little from 1983 to 1984. Even if we had used the 1983 conversion as the preliminary conversion for these tests, the results would have been adequate.

For the two verbal tests, the story is different. The difficulty of the Sentence Sense test did increase from 1983 to 1984, particularly for students in the lower ability range. The EICC equating predicted this change quite accurately. But the Reading Comprehension test increased in difficulty by an even greater margin, and the EICC equating underestimated the increase, especially for students in the lower middle ability range, where some of the community colleges have set their placement cutoff scores. Although the inaccuracy is less than a scaled-score point, it appears to be quite systematic.

To try to find the source of the problem, we conducted a series of analyses. One of these was to repeat the pre-equating procedure, but to use data from the regular administration instead of the pretest data. This

---

* The equipercentile equating was based on approximately 12,000 students taking each form (1983 and 1984) and was done by two methods: "direct" (Angoff, 1984, p. 116) and "frequency estimation" (Angoff, 1984, p. 113). Scaled scores produced by the two methods differed by less than 0.2 scaled-score points in most cases, and nowhere by more than 0.5 points. Only the frequency estimation equipercentile results are shown in Table 1.

analysis, we reasoned, would tell us how much of the discrepancy was due to differences in the data and how much to differences in the method. Table 2 shows the results of this analysis. When we applied the EICC method to the regular equating data, the results were much closer to the regular equating. Something was wrong with the pretest data -- but what?

The next step in the investigation was to plot the response curves for the items in both analyses. We wanted to see whether the response curves based on the pretest data were similar to those based on the regular equating data. What we found was very interesting. For most of the items, the two response curves looked similar, as they should. But for some of the items, the curve for the actual administration was clearly below the curve based on the pretest data, indicating that these items had somehow become more difficult. And most of these items belonged to the last two new item sets on the test. Figure 1 shows the two response curves for one of these items.

At this point we realized that the problem might be the result of speededness. When these items were pretested, they were in the middle of the test, where all students had ample time. When they became part of the test, they were near the end, where some students would not reach them and others would be forced to work hurriedly. Yet, our usual speededness statistics indicated that the test was only slightly speeded. Even the last item was reached by 88 percent of the students. Why were the response curves different?

In this case, the usual speededness statistics do not tell the whole story, for two reasons. First, they do not tell us how many students were

forced to hurry through the last several items. Second, these ties are not conditional on the students' ability level. The distribution of scores on this test is highly skewed, with most of the students piled up at the top end. For these more able students, the test may well have been unspeeded. And indeed, the pre-equating worked quite well in this range. But for the less able readers, the test may have been speeded enough to make the last few items substantially harder than they would have been if these students had been given more time.

Now that we know what the problem is, what can we do about it? The inaccuracy due to speededness was not large - less than one scaled score point. Any attempt to adjust for speededness might create a greater inaccuracy, since we cannot predict in advance how much of a speededness effect we will have. The wisest course of action may be to let well enough alone - to use the method as it is, with no adjustment. Even without a speededness adjustment, EICC pre-equating is a vast improvement over the alternative of using the raw-to-scale conversions from previous years to make the early placement decisions that must be made.

Reference

Angoff, W. H. Scales, Norms, and Equivalent Scores. Princeton, NJ: Educational Testing Service, 1984.

Livingston, S. A. Item selection and pre-equating with empirical item characteristic curves. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April, 1984.

Table 1

Raw-to-Scale Score Conversions: 1983, 1984 EICC equating based on pretest data and 1984 equipercentile equating

Math Computation					Elementary Algebra				
	Scaled Scores					Scaled Scores			
Raw Score	1983	1984 EICC	1984 equi %	error of prediction	Raw Score	1983	1984 EICC	1984 equi %	error of prediction
30	180.2	180.2	180.2	0.0	30	188.7	189.3	189.1	+0.2
25	173.3	173.4	173.7	-0.3	25	181.9	182.8	183.1	-0.3
20	166.5	166.6	167.1	-0.5	20	175.1	176.0	176.0	0.0
15	159.6	159.9	159.8	+0.1	15	168.3	168.5	168.6	-0.1
10	152.8	152.8	152.5	+0.3	10	161.5	160.7	160.6	+0.1
5	145.9	145.6	144.9	+0.7	5	154.7	154.2	153.7	+0.5
Range of placement cutoff scores: 160 to 172					Range of placement cutoff scores: 161 to 178				

Sentence Sense					Reading Comprehension				
	Scaled Scores					Scaled Scores			
Raw Score	1983	1984 EICC	1984 equi %	error of prediction	Raw Score	1983	1984 EICC	1984 equi %	error of prediction
35	180.5	180.5	180.4	+0.1	50	181.5	181.8	181.5	+0.3
30	171.8	172.5	172.4	+0.1	45	174.2	176.5	176.7	-0.2
25	163.1	165.0	165.1	-0.1	40	167.9	171.	171.5	-0.2
20	154.5	157.3	157.5	-0.2	35	161.9	166.0	166.5	-0.5
15	145.8	148.5	148.6	-0.1	30	155.4	160.3	160.9	-0.6
10	137.1	139.7	139.6	+0.1	25	147.6	153.3	154.1	-0.8
					20	139.8	144.7	145.1	-0.4
Range of placement cutoff scores: 153 to 173					Range of placement cutoff scores: 155 to 171				

Table 2

Raw-to-Scale Conversions for Reading Comprehension Test

Raw Score	Scaled Scores		
	Pretest Data	Regular Administration	
	EICC method	EICC method	Equipercentile
45	176.5	176.5	176.7
40	171.3	171.5	171.5
35	166.0	166.5	166.5
30	160.3	161.0	160.9
25	153.3	154.2	154.1
20	144.7	145.6	145.1

FIGURE 1. ITEM RESPONSE CURVES FOR AN ITEM APPEARING LATE IN THE TEST