ED 392 847 TM 024 695

AUTHOR Baghi, Heibatollah; And Others

TITLE A Comparison of the Results from Two Equating Designs for Performance-Based Student Assessments.

PUB DATE Apr 95

NOTE 39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Comparative Analysis; *Constructed Response; Correlation; *Educational Assessment; Elementary Secondary Education; *Equated Scores; Mathematics Tests; *Norm Referenced Tests; Raw Scores; Sample Size; State Programs; *Test Content; Testing Programs

IDENTIFIERS *Anchor Tests; Delaware; Linking Metrics; Performance Based Evaluation; *Single Group Design

ABSTRACT

Issues related to linking tests with constructed response items were explored, specifically by comparing single-group and anchor-test designs to link raw scores from alternate forms of performance-based student assessments in the context of Delaware's assessment program using performance-based assessment. This study explored use of the two test designs for the mathematics assessments administered in 1993 and 1994. In the single-group design the equating study was conducted with 300 to 460 out-of-state students from 16 schools at grades 3, 5, 8, and 10. To equate the performance-based assessments using the anchor-test design (a sample of about 13,000 students each year), each student's 1993 performance-based assessment raw score was equated to the norm-referenced scale score from the Iowa Tests of Basic Skills for grades 3, 5, and 8 and the Tests of Achievement and Proficiency for grade 10. The same procedure was used for the 1994 performance-based assessment. Results suggested that, based solely on considerations of sample size, correlations, and sample matching, the anchor-test design would be expected to produce more accurate results than the single-group design. Commonality of content coverage between the performance assessment and the norm referenced test is essential. (Contains 3 figures, 18 tables in an appendix, and 10 references.) (SLD)

A Comparison of the Results from Two Equating Designs for Performance-Based Student Assessments

Heibatollah Baghi

Patricia Bent

Marshá DeLain

Delaware Department of Public Instruction

Sara Hennings

The Riverside Publishing Company

Paper presented at the NCME annual conference held in San Francisco, April 1995. Requests for reprints should be sent to Dr. Patricia Bent, Delaware Department of Public Instruction, P.O. Box 1402, Dover, Delaware 19903.

Objectives

Performance-based assessments have become popular in response to educational reform initiatives in many statewide student assessment programs. The items in performance-based assessments require that a student construct a response rather than choose from a set of given responses. The answers required in performance-based assessments are open ended and range from short answer to extended response item formats. Because the results of performance-based assessments are frequently compared from year to year, alternate forms are used in each year of testing. The purpose of this study is to investigate the issues related to linking tests with constructed response items.

More specifically, the purpose of this study is to compare two data collection designs used to link raw scores from alternate forms of performance-based student assessments. These designs are the single-group design and the anchor-test design. In the context of Delaware's assessment program that uses performance-based assessments, it is of considerable interest that the two designs produce highly comparable scores on alternate forms. A criterion for evaluating and recommending one of these designs for linking two alternate forms is the consistency of results from year to year. This study explores the use of the single-group and anchor-test designs for the mathematics assessments that were administered in 1993 (Form A) and 1994 (Form B).

Perspectives

In Delaware, performance standards for grades 3, 5, 8, and 10 in reading and mathematics were established for the performance-based assessments administered in 1993. The tests included both open-ended items and some selected response questions. These standards classified student performance into one of three proficiency levels: meets or exceeds the standard, approaches the standard, and considerably below the standard. During the standard-setting process, teachers accomplished three tasks: (1) the establishment of performance standards or cut scores, (2) the development of proficiency level descriptions or expectations at each level for each grade and content area, and (3) the selection of actual student work to illustrate each proficiency level. The proficiency level descriptions also describe what aspects characterize student work at each level. For example, in grade eight mathematics, some aspects of student work at the Meets or Exceeds the Standard level include the ability to analyze material and reason from it and to support ideas with well-developed responses, suitable evidence, and appropriate terminology. The proficiency level descriptions for the Approaches the Standard level indicate that students possess much of the knowledge and skills typical of the curriculum but have difficulty applying that knowledge and their responses lack support and/or adequate explanation. Students in the Considerably Below the Standard level possess some knowledge of the basic facts and procedures contained in the curriculum but have difficulty connecting and extending specific content. The student responses selected to illustrate the proficiency level descriptions were used in inservice workshops following the release of the 1993 assessment results.

To maintain consistency in performance standards from year to year when administering different forms of a test, the assessment forms need to be equated. Equating would establish a relationship between possible scores on the Delaware 1993 test forms and subsequent test forms. Once this relationship is known, scores can be determined for the 1994 test forms that are comparable to the scores on the 1993 test forms. The test results are reported in terms of the percent of students in each proficiency level at the school, district, and state level every year.

Angoff (1971); Peterson, Kolen, and Hoover (1989); and Skaggs and Lissitz (1986) describe three designs for collecting the data needed to equate the scores from two test forms: the single-group design, the anchor-test design, and the equivalent-groups design. The first two are used here to determine the relationship between the possible scores on the 1993 (Form A) test forms and the 1994 test forms (Form B). The single-group design requires that the same group of students take both tests to be equated. One advantage of this design is that student ability is the same for both assessments; hence, differences in performance will reflect differences in the test characteristics of those assessments (e.g., difficulty level). The practical problems are that it requires a large equating sample; it involves additional testing time and cost; possible test security issues arise if both assessments are administered in-state; and the second assessment can be affected by students' fatigue and practice, which may in turn distort the linkage between the two assessments.

The anchor-test design requires that two assessments be administered to two different groups of students. In addition, all students take a common set of anchor items. The anchor-test can be either an internal set of test items embedded in both assessment forms to be equated or an external set of test items administered as an additional test to all students. The common items are used to adjust for performance differences (e.g., ability) between the groups who took both assessments. The stronger the correlation between the anchor-test and the assessments being equated, the more useful are the anchor-test data for establishing the relationship between the possible scores on the two forms and hence, for linking the scores from the two assessments. Yen, Green, and Burket (1987) suggest that the anchor items should be a representative sample of the content covered by the assessments to be equated.

In large scale testing, such as the current Delaware Interim Assessment Program that will span five years, alternate forms of assessments must be administered in different years, regardless of whether the test consists of short answer items or extended-response items. Although these assessments are not considered high-stake tests, the State desires that the results of the annual assessment of students be reported across years in terms of the percent of students who meet, approach, or fall considerably below the standard. These forms generally assess the same broad content domains that are detailed in proficiency level descriptions, but use different tasks to measure them. Test lengths also differ across forms. Assessments in the same broad content area but focusing on different sub-domains violate some of the conditions required for strict equating. This situation is most obvious in mathematics when different tasks are used to measure the same performance standards or levels of proficiency.

When performance-based assessments are developed, they usually contain a group of items or activities that pertain to steps involved in accomplishing a large task, such as planning the landscape design for a private residence or evaluating advertising claims for the possible misuses of statistics, graphs, and data. Given this situation or context, students are expected to solve a series of problems and to explain or justify the procedure used for solving the problems. Even though the assessment tasks included in each year's test forms differ in content as well as some of the conditions assumed for equating are not satisfied (such as the test forms not being parallel), educational policy makers would like the test forms to be statistically linked. The reason behind this decision is that when the State releases the assessment results, school districts and newspapers make comparisons across years. Since the performance standards (or cut scores) were set on the 1993 forms, comparable cut scores must be determined for alternate forms to be administered in subsequent years. Thus, the question is raised whether results obtained from statistical designs and procedures developed for equating are indeed valid for practical purposes. This study is concerned with procedures that will produce scores that approximate the product of the equating process. Hence, in the text that follows, the terms "linking" and "equating" are used interchangeably to refer to the process of obtaining highly comparable scores on different assessment forms.

As Linn and Kiplinger (1994) have noted, it has long been a common practice to link results of different forms of a test and then treat the results from administrations of these different forms as interchangeable. Test publishers routinely publish alternate forms of an achievement test that are equated to a common scale so that users can obtain comparable results using a particular form in one year and another form in the next year. Mislevy (1991) and Linn (1993) discuss types of linking that have less stringent requirements but yield weaker results that support comparisons in more limited circumstances. These methods are discussed under the headings of calibration, projection, statistical moderation, and social moderation. For example, Linn (1993) defines the calibration process as a linkage that produces scores that are comparable but differ in their reliability (i.e., equal precision throughout the range of levels of student achievement). This approach involves calibrating the results from two assessments to a common scale. If the assumption that the two assessments measure the same construct is relaxed, a procedure called statistical moderation or social moderation can be used. Another approach that can be used when the two assumptions of comparable validity and reliability are relaxed is called projection. This approach uses a regression technique to predict scores on one test from scores on a second assessment. Several researchers have raised concerns about the applicability of existing designs for equating scores from tests that are comprised of constructed response items. Dunbar, Koretz, and Hoover (1991) have strongly recommended that additional work be done to define methods for establishing the comparability of scores from performance-based student assessments.

Because performance-based assessments are new, insufficient research has been done at present to indicate which of the data collection designs is more appropriate for linking such assessment results. Loyd, Englehard, and Crocker (1993), who have been working on designs for equating the assessments of the National Board of Professional Teaching Standards (NBPT), concluded that they could not identify appropriate equating designs until the NBPT's tests were more clearly defined.

The purpose of this study is to compare two data collection designs used to link scores from alternate forms of performance-based student assessments. One criterion for evaluating and recommending one of these designs for linking is the consistency of results from year to year.

Methods and Data Source

Equating Method. The equating procedure used in this study is equipercentile equating (Peterson, Kolen, and Hoover, 1989). The equipercentile procedure requires that the two test forms measure a common trait. This procedure makes no assumptions about the test score distributions.

Data Collection Design and Instruments Used. The data collection designs used in this study were the single-group design and the anchor-test design. In the single-group design, the equating study was conducted with 300 to 460 out-of-state students from 16 schools at grades 3, 5, 8, and 10. The data were collected by The Riverside Publishing Company. The order of administration was counterbalanced, that is, some students took the 1993 (Form A) assessment first and others took the 1994 (Form B) test first.

To equate the performance-based assessments using the anchor-test design, each student's 1993 performance-based assessment raw score was equated to the norm-referenced test (NRT) scale score (the standard score scale) obtained on the Iowa Tests of Basic Skills® (ITBS® Survey Battery levels 9, 11, and 14) for grades 3, 5, and 8 and the Tests of Achievement and Proficiency™ (TAP® Survey Battery level 16) for grade 10. The same procedure was used for the 1994 performance-based assessment to find the comparable NRT scale score for each performance-based assessment raw score.

Both the *ITBS* and the *TAP* focus on the content and skills in the standards developed by the National Council of Teachers of Mathematics (NCTM). Specifically, the *ITBS* Survey Battery tests in mathematics are comprised of four parts: concepts, estimation, problem solving, and data interpretation. The results for these four sections are combined to produce a mathematics total score. In grade three, there are 30 items, in grade five there are 35, and 45 in grade eight. The *TAP* Survey Battery includes concepts and problem solving subtests which assess number and numeration theory, arithmetic operations and procedures, algebraic functions and procedures, geometric operations and procedures, probability and statistics, and mathematical reasoning. There are 36 items in the *TAP* Survey Battery.

Although both the Form A (1993) and Form B (1994) performance-based assessments also reflect the NCTM standards, the tasks, activities, and sub-domains differ. For example, in Form A for grade 10, students read and interpreted statistical data and used experimental observations to estimate population parameters whereas in Form B, students identified and explained the misuse of statistics, identified and explained misrepresentations or distortions of data, and wrote mathematical explanations based on empirical evidence.

In the anchor-test design, approximately 40 percent of the Delaware students who had actually taken the assessments were selected. This resulted in samples of about 13,000 students for each of the two years. The raw scores on the 1993 performance-based assessments were linked to the 1994 performance-based assessments via the NRT scale. This process yields 1994 raw scores (Form B) that are equivalent to the 1993 (Form A) raw scores.

Smoothing Method. After determining the equated scores, the raw scores on the performance-based assessment are plotted against the scale scores on the NRT, and a smooth curve is drawn using an analytical technique described by Kolen (1984) and Zeng (1993). As explained by Zeng (1993), there are two analytic smoothing methods: presmoothing and postsmoothing. In the former method, the score distributions for the assessments to be equated are smoothed before the equipercentile equating is performed. In the latter method, equated scores obtained from the unsmoothed score distributions are smoothed. In this study, the postsmoothing method recommended by Zeng (1993) was used.

Results and Discussions

As stated previously, this study compares the results of two designs to collect the data needed to equate scores from alternate forms of two years of performance-based assessments administered in the Delawarc statewide student assessment program. The two designs are the single-group design and the anchor-test design. Since mathematics performance-based assessments often present more challenges in achieving comparable content coverage across forms, the results for mathematics only will be presented and discussed in this section.

The Single-Group Design

Table 1 summarizes the descriptive statistics for the Form A (administered in 1993) and the 1994 Form B mathematics tests. Using the data in this table, a measure of difficulty can be computed by dividing the mean score for each test by the maximum possible score for the test. The resulting percent of maximum scores range from 44% (grade 10, Form B) to 71% (grade 3, Form A). Both forms have similar difficulties with differences between the two years of 5% or less except for grade 10 that has a percent of maximum of 63% for Form A and 44% for Form B, the latter being the more difficult test.

The data in Table 1 also indicate that Forms A and B are very similar in terms of skewness, a descriptive statistic that reflects the shape of the raw score distribution. Likewise, the standard deviations for both test forms are very much the same. Again, the one exception is the values for the grade 10 tests. For both forms in grades 3, 5, and 8, the skewness ranges from -0.676 to -0.123 and the standard deviation ranges from 5.16 to 6.78. In grade ten, the skewness for Form A is -0.729 and 0.240 for Form B while the standard deviation is 4.60 for Form A compared to 6.30 for Form B. In situations in which the distributions are essentially the same, either equipercentile equating or linear equating could be used. Finally, the internal consistency (alpha) of both forms ranges from 0.715 to 0.847 and thus is acceptable for the purposes of equating.

Table 2 presents data about the results of the counterbalanced design for the single-group study. For grades 3, 5, and Form A in grade ten, there were minimal test order effects, that is to say, the mean scores for the test form administered first were not substantially different from the mean scores for the tests administered second. Thus, for these grades and forms, the unequal numbers of students in each group may not be critical. By looking at Table 2, one notes that 61% of the students in grade 3 and 68% of the students in grade 5 took Form A before Form B. Only in grade 8 was there a 50-50 split in the two groups of students in the counterbalanced design. The lack of balance was most marked in grade 10, with 84% of the students assigned to the group that took Form A before Form B. In grade eight and for Form B in grade 10, the means were lower for the students who took Form A before taking Form B, thus, a test order effect seems to be present. The test order effect would be expected to have an impact on the equating especially for grade 10 that also had great disparity in the number of students assigned to each group. Thus, the confidence in the counterbalanced design is limited because of the lack of balance in students assigned to the two groups and the test order effects.

Table 3 shows the correlations between Form A and Form B raw scores that range from 0.586 (grade 3) to 0.671 (grade 5). These values indicate that the two test forms differ in content with less than 50% common variance. The third column refers to the correlation coefficients corrected for attenuation.

Using the equipercentile procedure, the equated scores that have been smoothed are reported for grades 3, 5, 8, and 10 in Tables 4 through 7. The first column lists the possible raw scores on the 1994 Form B assessment and the second column indicates the scores equated to the 1993 Form A. For example, in grade 3, a raw score of 20 on Form B is equated (or comparable) to a score of 22.3 on Form A. An example of the scatterplot for the single-group design is shown in Figure 1.

The results presented in the tables for the single-group design indicate that the sample sizes were small and the counterbalanced design did not work the way it was planned. All cases for each grade level were less than 500 because of unanticipated circumstances. These sample sizes cannot be expected *a priori* to produce accurate equating results. A larger sample size is typically recommended for the single-group design.

The Anchor-Test Design

Tables 8 and 9 summarize the descriptive statistics for Form A administered in 1993, Form B administered in 1994, and the NRT that was administered in both years. Utilizing the percent maximum values again to gauge the difficulty of the assessments (dividing the mean by the number of maximum points on a test form), one sees that the grade 3 and grade 5 tests were very similar in difficulty. For grades 8 and 10, however, a substantial difference in difficulty emerges between Form A and Form B (percent maximum is .36 versus .64 for grade 8 and .44 versus .64 for grade 10). Form B is the more difficult test. The standard deviations are similar between forms for all grades except grade 10 that shows a value of 4.67 for Form A and 6.43 for Form B. Hence, Form B scores indicates greater difficulty and more dispersion in the distribution.
In this design, each grade had a sample size of at least 2400.

Tables 8 and 9 also display the mean and standard deviation for the NRT scale scores for students included in the anchor-test design. These students, selected at random from Delaware students who actually took the assessments, appear to have been well matched in terms of their overall performance on the tests. The mean scale scores on the NRT differ by no more than 2.69 points (in grade 5) and the standard deviations differ by no more than 1.01 (in grade 10). Such comparability between the two groups of students contributes to the accuracy of the equating process.

Correlations between the performance-based assessments and the NRT are displayed in Tables 10 and 11 and range from 0.642 (grade 10, Form A) to 0.788 (grade 5, Form B). Although these values are only modest, overall they were higher than the direct correlations between the two forms of the performance assessments. Although not ideal, the modest correlations with the NRT are probably about as high as could be expected, given the psychometric characteristics of the performance assessments.

Again, the equipercentile procedure was used for equating the results of Form B to those of Form A via the norm-referenced test. The first step in equipercentile equating is to determine the percentile ranks for the scores on each of the two distributions to be equated. This procedure is used to equate the 1993 raw scores on the performance assessments to the scale scores on the *ITBS/TAP* (NRT Scale). The procedure is duplicated for the 1994 data. Then the raw scores on the 1993 performance-based assessments are linked to the 1994 assessments through the NRT scale. After determining the equated scores, the raw scores on the performance-based assessments are plotted against the scale score on the NRT and a smoothed curve is drawn.

Tables 12 and 13 for grade 3 mathematics will serve to illustrate the steps used in the anchor-test design. As an example, in Table 12, a raw score of 12 on Form A (1993) has a percentile rank of 2.36 (the first two columns in the table). Through interpolation, a NRT scale score of 148.4 has a percentile rank of 2.36. These results are in the third and fourth columns of Table 12. Thus, a score of 12 on the performance assessment and a NRT scale score of 148.4 are equivalent in that they both have the same percentile rank. One would then look at the 1994 results in Table 13 and repeat the procedure. Here, a raw score of 9 has a percentile rank of 2.45 (the first two columns) and a NRT percentile rank of 2.45 is associated with a scale score of 148.4 (the third and fourth columns of the table). The corresponding 1994 performance-based scores have a percentile rank of 2.45 that is associated with a raw score of 9 (the first two columns). Each of these distributions would be smoothed and the corresponding smoothed NRT scale score is 147.6 for both the 1993 and 1994 assessment results. (The smoothed NRT scale scores are shown in the last column of the two tables.) Thus, through the equating process, a score of 12 on the 1993 Form A performance-based assessment is comparable to a 9 on the 1994 Form B. One can see this by looking in the last three columns of both tables. In Table 12, a smoothed NRT scale score of 147.6 is equated to a performance- based assessment score of 12 for 1993 whereas in Table 13, the same NRT score of 147.6 in 1994 is associated with a performance-based assessment score of 9.

The final smoothed NRT scale scores and the corresponding values for the performance-based assessment scores from 1993 (Form A) and 1994 (Form B) are displayed in Tables 14 through 17. Again, looking at Table 14 for third grade mathematics, one sees that a 1994 score of 9 is comparable to a 1993 score of 12. The resulting scatterplot with the lines plotted is depicted in Figures 2 and 3 for grade 3 mathematics.

In conclusion, based solely on the considerations of sample size, correlations, and sample matching, the anchor-test design would be expected to produce more accurate results than the single-group design.

Comparison of the Two Data Collection Designs

In reviewing Tables 1, 8, and 9, the sample of students in the single-group and the anchor-test design were very similar in their average performance, with the exception of grade 8 Mathematics for Form B. The results for this assessment show a mean of 16.75 for the single-group design sample and 10.66 for the anchor-test samples. Such disparities can contribute to differences in equating results. Additionally, the data in Tables 3, 10, and 11 indicate that the correlations between Form A and Form B were consistently lower than the those between the performance-based assessments and the norm-referenced tests. Thus, there is more statistical common ground between the norm-referenced tests and either Form A or Form B than there is between Forms A and B themselves. Other strengths of the anchor-test design in the Delaware experience was that the student samples were large and randomly selected from a larger pool of students. As shown in Tables 8 and 9, the scores from these samples were very similar and this finding lends support to the comparability of the two groups of students. Such a claim of comparability is more difficult to make in the counterbalanced design for the single-group study. Thus, the decision was made to use the NRT data from the anchor-test design to statistically adjust the scores on Form B so that they would be comparable with those of Form A.

Final evidence for using the anchor-test design can be found in Table 18. This table shows the percent of students placing at each of the three proficiency levels for grades 3, 5, 8, and 10 in 1993 and in 1994 using the results of both equating designs. One would expect very little change in the percent of students in each of these levels because the test administration was only one year apart. Form A was administered in May of 1993 and Forrr B in May of 1994. Additionally, the results of the 1993 administration were not released until the Fall of 1993. The results from the anchor-test design are more consistent with 1993 than those from the single-group design. This can be seen by reviewing Table 18 that shows that the anchor-test design resulted in no more than a 4% difference in the percent of students placing in the proficiency levels in 1994 compared to those from 1993. On the other hand, differences resulting from the single-group design were as high as 10%. In conclusion, this consistency of the anchor-test design is additional evidence for its greater accuracy and also supports its use in Delaware for equating performance-based assessment results using the anchor-test design.

Limitations, Issues, and Recommendations

This study utilized existing methodology for linking the results of performance-based assessments. Three issues arise when recommending a methodology (existing or new) for linking such results that uses NRT data. One involves the reason for giving performance-based assessments. Frequently, these tests are administered based on the belief that they measure attributes, such as higher order thinking skills and problem solving, not measured by NRT's. On the other hand, if NRT items are to be used in a linking design, what correlation is appropriate or required between the NRT and performance-based measures? A second, and related issue, also involves the acceptable correlation between the NRT and performance-based tests. If the correlation is high, what reasons can be advanced for administering the more resource intensive performance-based assessments? If the correlation is low, how can the use of NRT's be defended in linking procedures? In either case, how does the correlation influence the comparability of the obtained results? The third issue, although not a psychometric one, is also important. When performance-based assessments have been promoted as being preferable to NRT's, how does one defend the use of NRT's in linking the results of performance-based assessments?

Several recommendations arise from these issues and from the results of this study. First, and most obvious, more research is needed about procedures for linking performance-based assessments. If procedures use NRT data, the issues raised previously must be addressed. Second, for the third year of the Delaware Interim Assessment Program, the stability of linking results should be investigated for gender and race-ethnicity subgroups. These analyses would indicate if the linking procedures produce results for the subgroups that are comparable to statewide results. Third, the anchor items or anchor-test should have significant commonality with the performance assessments. In the test development cycle, both tests used in the Delaware statewide student assessment program (the NRT's and the performance-based assessments) should be reviewed for this commonality of content coverage.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in development and use of performance assessments. Applied Measurement in Education, 4, 289-303.

Linn, R. L. (1993). Linking results of distinct assessments. Applied Measurement in Education, 6, 83-102.

Linn, R.L. & Kiplinger (1994). Linking statewide tests to the National Assessment of Educational Progress: Stability of results. Center for Research on Evaluation, Standards, and Student Testing. (University of Colorado at Boulder:)

Lóyd, B., Englehard, G., & Crocker, L. (1993, April). Equity, equivalence, and equating: Fundamental issues and proposed strategies for the National Board for Professional Teaching Standards. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Mislevy, R. J. (1992). Linking educational assessments: Concepts, issues, methods, and prospects. Princeton, NJ: Educational Testing Service.

Peterson , N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.) Educational measurement, (3rd ed., pp. 221-262) New York: American Council on Education/Macmillan Publishing Company.

Skaggs, G., & Lissitz, R. (1986). Test equating. Review of Educational Research, 56, 495-529,

Yen, W. M., Green, D. R., & Burket, G. (1987). Valid normative information from customized achievement tests. Educational Measurement: Issues and Practices, 6, 7-13.

Zeng L. (1993, April). The optimal degree of smoothing in equipercentile equating with postsmoothing. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

APPENDIX

Table 1

Descriptive Statistics for Forms A and B: Single Group Design

	Mathematics												
	Form A						Form B						
Grade	# Items	Max. Score	Mcan	SD	Skewness	Alpha	# Items	Max. Score	Mean	SD	Skewness	Alpha	N
3	22	34	24.07	5.56	-0.676	0.775	21	33	21.84	5.16	-0.420	0.725	462
5	15	27	15.13	5.58	-0.123	0.715	14	23	12.68	5.19	-0.297	0.801	456
8	19	34	20.88	6.56	-0.538	0.836	18	30	16.75	6.78	-0.144	0.847	420
10	15	25	15.74	4.60	-0.729	0.763	12	27	11.87	6.30	0.240	0.793	414

Table 2

Descriptive Statistics for the Counterbalanced Design for Forms A and B: Single Group Design

Students Taking Form A Before Form B

		Form A				Form B			
Grade	r	Mean	SD	Skewness	N	Mean	SD	Skewness	N
3	0.608	24.342	4.920	-0.405	284	22.127	5.043	-0.402	284
5	0.678	14.853	5.538	-0.147	312	12.237	5.160	-0.181	312
8	0.628	18.962	6.925	-0.269	212	14.528	6.645	0.136	212
10	0.625	15.627	4.533	-0.653	346	11.338	6.162	0.288	346

Students Taking Form B Before Form A

		Form A				Form B			
Grade	r	Mean	SD	Skewness	N	Mean	SD	Skewness	N
3	0.564	23.640	6.447	-0.764	178	21.382	5.334	-0.427	178
5	0.650	15.729	5.637	-0.089	144	13.632	5.137	-0.580	144
8	0.550	22.841	5.528	-0.677	208	19.014	6.141	-0.378	208
10	0.613	16.324	4.940	1.120	68	14.574	6.337	-0.039	68

Table 3

Correlations Between Forms A and B Raw Scores: Single-Group Design

Mathematics			
Grade	r	r*	N
3	0.586	0.782	462
5	0.671	0.887	456
8	0.633	0.752	420
10	0.621	0.798	414

*corrected for attenuation

Table 4

Final Smoothed (Equated) Table: Single-Group Design

Grade 3

Mathematics	
B	A
Raw Score	Raw Score
0	0.0
1	0.6
2	1.2
3	1.9
4	2.5
5	3.1
6	4.2
7	5.3
8	6.8
9	8.3
10	9.9
11	11.5
12	13.0
13	14.4
14	15.7
15	17.0
16	18.2
17	19.3
18	20.3
19	21.3
20	22.3
21	23.3
22	24.3
23	25.3
24	26.3
25	27.4
26	28.5
27	29.6
28	30.7
29	31.7
30	32.5
31	33.2
32	33.5
33	34.0

Table 5

Final Smoothed (Equated) Table: Single-Group Design

Grade 5

Mathematics	
B	A
Raw Score	Raw Score
0	0.0
1	3.1
2	4.1
3	5.1
4	6.2
5	7.2
6	8.2
7	9.2
8	10.1
9	11.1
10	12.0
11	13.0
12	14.0
13	15.1
14	16.2
15	17.3
16	18.5
17	19.8
18	21.0
19	22.3
20	23.6
21	24.7
22	25.7
23	27.0

Table 6

Final Smoothed (Equated) Table: Single-Group Design

Grade 8

Mathematics	
B Raw Score	A Raw Score
0	0.0
1	1.6
2	2.7
3	4.1
4	5.7
5	7.4
6	9.1
7	10.8
8	12.5
9	14.0
10	15.4
11	16.6
12	17.7
13	18.7
14	19.5
15	20.3
16	21.0
17	21.6
18	22.2
19	22.9
20	23.6
21	24.3
22	25.1
23	26.1
24	27.1
25	28.2
26	29.4
27	30.6
28	31.9
29	33.0
30	34.0

Table 7

Final Smoothed (Equated) Table: Single-Group Design

Grade 10

Mathematics	
B	A
Raw Score	Raw Score
0	0.0
1	3.4
2	5.4
3	7.4
4	9.4
5	10.9
6	12.1
7	13.3
8	14.2
9	15.1
10	15.8
11	16.4
12	17.0
13	17.5
14	18.0
15	18.4
16	18.8
17	19.2
18	19.6
19	20.0
20	20.4
21	20.9
22	21.5
23	22.2
24	23.0
25	23.9
26	24.5
27	25.0

Table 8

Descriptive Statistics for Form A and NRT Scores: Anchor-Test Design

Mathematics

	Performance Assessment Raw Score				NRT Scale Scores		
Grade	# Items	Maximum Score	Mean	SD	Mean	SD	N
3	22	34	23.68	5.84	181.26	21.03	3,616
5	15	27	12.37	5.88	213.86	27.25	3,578
8	19	34	17.52	7.64	252.26	34.11	3,272
10	15	25	16.20	4.67	270.68	36.07	2,564

Table 9

Descriptive Statistics for Form B and NRT Scores: Anchor-Test Design

Mathematics

	Performance Assessment Raw Score				NRT Scale Scores		
Grade	# Items	Maximum Score	Mean	SD	Mean	SD	N
3	21	33	21.68	5.96	181.45	21.27	4.304
5	14	23	10.98	5.68	211.17	27.90	4.082
8	18	30	10.66	7.40	252.99	34.75	3.022
10	12	27	11.85	6.43	271.01	37.08	2.493

Table 10

Correlations Between Performance Assessment Raw Scores on Form A and NRT Scale Scores Obtained in 1993: Anchor-Test Design

Mathematics		
Grade	r	N
3	0.658	3,616
5	0.712	3,578
8	0.740	3,272
10	0.642	2,564

Table 11

Correlations Between Performance Assessment Raw Scores on Form B and NRT Scale Scores Obtained in 1994: Anchor-Test Design

Mathematics		
Grade	r	N
3	0.687	4,304
5	0.788	4,082
8	0.742	3,022
10	0.706	2,493

Table 12

Percentile Ranks on Two Assessments, Form A Performance-Based Assessments (PBA) and NRT (ITBS), Administered in 1993

RS PBA	PR PBA	SS NRT	PR NRT	RS PBA	Equated NRT Score	Smoothed NRT Score
				0		124.0
1	0.03	133	0.03	1	133.0	129.0
2	0.10	136	0.12	2	135.3	135.3
3	0.17	140	0.30	3	137.1	138.1
4	0.22	143	0.62	4	138.2	139.8
5	0.26	146	1.15	5	139.1	141.0
6	0.37	148	1.92	6	140.7	141.7
7	0.62	149	2.99	7	143.0	142.4
8	0.83	150	4.52	8	144.2	143.0
9	1.05	152	6.65	9	145.4	143.8
10	1.38	154	9.33	10	146.6	144.8
11	1.76	157	12.38	11	147.6	146.0
12	2.36	160	15.69	12	148.4	147.6
13	3.57	163	19.62	13	149.4	149.4
14	5.16	165	24.24	14	150.6	151.6
15	7.25	168	29.16	15	152.4	153.9
16	10.38	171	34.54	16	155.0	156.5
17	14.09	174	40.09	17	158.5	159.2
18	17.99	177	45.44	18	161.8	162.0
19	22.43	180	50.68	19	164.2	164.8
20	27.48	183	56.55	20	167.0	167.6
21	32.85	186	62.79	21	170.1	170.4
22	38.44	190	68.69	22	173.1	173.2
23	43.86	194	74.64	23	176.1	175.9
24	49.29	198	80.13	24	179.2	178.7
25	55.38	203	85.41	25	182.4	181.5
26	61.50	210	90.38	26	185.4	184.6
27	67.41	218	94.30	27	189.1	187.9
28	73.42	228	97.32	28	193.2	191.8
29	79.27	248	99.32	29	197.4	196.5
30	84.49	248	99.32	30	202.1	202.2
31	88.94	248	99.32	31	208.0	209.4
32	93.03	248	99.32	32	215.4	220.0
33	96.88	248	99.32	33	226.5	234.0
34	99.35	248	99.32	34	248.0	248.0

Table 13

Percentile Ranks on Two Assessments, Form B Performance-Based Assessments (PBA) and NRT (ITBS), Administered in 1994

RS PBA	PR PBA	SS NRT	PR NRT	RS PBA	Equated NRT Score	Smoothed NRT Score
				0		124.0
				1		127.3
2	0.02	124	0.01	2	125.3	130.6
3	0.07	129	0.05	3	131.0	133.9
4	0.15	133	0.09	4	135.6	136.9
5	0.35	136	0.16	5	139.8	139.9
6	0.66	140	0.36	6	142.6	142.5
7	1.01	143	0.70	7	144.6	144.6
8	1.53	146	1.27	8	146.6	146.2
9	2.45	148	2.20	9	148.2	147.6
10	3.71	149	3.32	10	149.3	149.0
11	5.05	150	4.70	11	150.3	150.5
12	6.73	152	6.92	12	151.8	152.1
13	8.96	154	9.72	13	153.5	154.0
14	11.72	157	12.79	14	156.0	156.1
15	14.82	160	16.33	15	158.7	158.4
16	18.31	163	20.68	16	161.4	160.9
17	22.33	165	25.49	17	163.7	163.6
18	26.80	168	30.22	18	165.8	166.5
19	31.91	171	35.11	19	169.0	169.5
20	37.45	174	40.16	20	172.4	172.6
21	43.10	177	44.90	21	175.9	175.8
22	48.69	180	49.64	22	179.4	179.0
23	54.45	183	55.07	23	182.7	182.4
24	60.83	186	60.71	24	186.1	185.9
25	67.52	190	66.87	25	190.4	189.6
26	73.36	194	73.12	26	194.2	193.6
27	78.87	198	79.00	27	197.9	198.1
28	84.28	203	84.64	28	202.7	203.3
29	88.88	210	90.06	29	208.5	209.4
30	93.16	218	94.44	30	215.7	216.8
31	96.55	228	97.44	31	225.0	225.7
32	98.73	248	99.42	32	241.0	236.7
33	99.81	248	99.42	33	248.0	248.0

Table 14

Final Smoothed (Equated) Table: Anchor-Test Design

Grade 3

Mathematics	
94 (B) Raw Score	93 (A) Raw Score
0	0.0
1	0.7
2	1.3
3	1.8
4	2.6
5	4.1
6	7.2
7	9.8
8	11.1
9	12.0
10	12.8
11	13.5
12	14.2
13	15.0
14	15.9
15	16.7
16	17.6
17	18.6
18	19.6
19	20.7
20	21.8
21	23.0
22	24.1
23	25.3
24	26.4
25	27.4
26	28.4
27	29.3
28	30.1
29	31.0
30	31.7
31	32.4
32	33.2
33	34.0

Table 15

Final Smoothed (Equated) Table: Anchor-Test Design

Grade 5

Mathematics	
94 (B)	93 (A)
Raw Score	Raw Score
0	0.0
1	1.6
2	2.5
3	3.6
4	4.6
5	5.7
6	6.7
7	7.7
8	8.7
9	9.7
10	10.7
11	11.7
12	12.7
13	13.7
14	14.8
15	15.8
16	16.8
17	17.9
18	19.0
19	20.3
20	21.5
21	23.1
22	25.0
23	27.0

Table 16

Final Smoothed (Equated) Table: Anchor-Test Design

Grade 8

Mathematics	
94 (B) Raw Score	93 (A) Raw Score
0	0.0
1	1.8
2	3.3
3	6.6
4	10.8
5	13.0
6	14.7
7	16.1
8	17.4
9	18.4
10	19.4
11	20.2
12	21.0
13	21.7
14	22.3
15	23.0
16	23.6
17	24.2
18	24.8
19	25.4
20	26.1
21	26.8
22	27.5
23	28.3
24	29.1
25	29.9
26	30.8
27	31.6
28	32.2
29	32.8
30	34.0

Table 17

Final Smoothed (Equated) Table: Anchor-Test Design

Grade 10

Mathematics	
B Raw Score	A Raw Score
0	0.0
1	1.8
2	4.7
3	7.5
4	10.2
5	11.6
6	12.7
7	13.6
8	14.4
9	15.2
10	15.9
11	16.6
12	17.2
13	17.8
14	18.4
15	19.0
16	19.5
17	20.0
18	20.5
19	21.0
20	21.5
21	21.9
22	22.3
23	22.6
24	23.0
25	23.6
26	24.3
27	25.0

Table 18

Percent of Students Falling in Each Performance Standard Category by Equating Method

		93 (Form A)	Single Group Design	Anchor-Test Design
Grade 3	Meets or Exceeds the Standards	17%	17%	17%
	Approaches the Standard	46%	48%	42%
	Considerably Below the Standard	37%	35%	41%
Grade 5	Meets or Exceeds the Standard	13%	16%	12%
	Approaches the Standard	39%	41%	39%
	Considerably Below the Standard	48%	43%	49%
Grade 8	Meets or Exceeds the Standard	12%	7%	12%
	Approaches the Standard	42%	32%	40%
	Considerably Below the Standard	46%	61%	48%
Grade 10	Meets or Exceeds the Standard	11%	7%	13%
	Approaches the Standard	54%	56%	55%
	Considerably Below the Standard	35%	37%	32%

Figure 1

Scatterplot for Form A and Form B Raw Scores: Single-Group Design

Grade 3 Mathematics

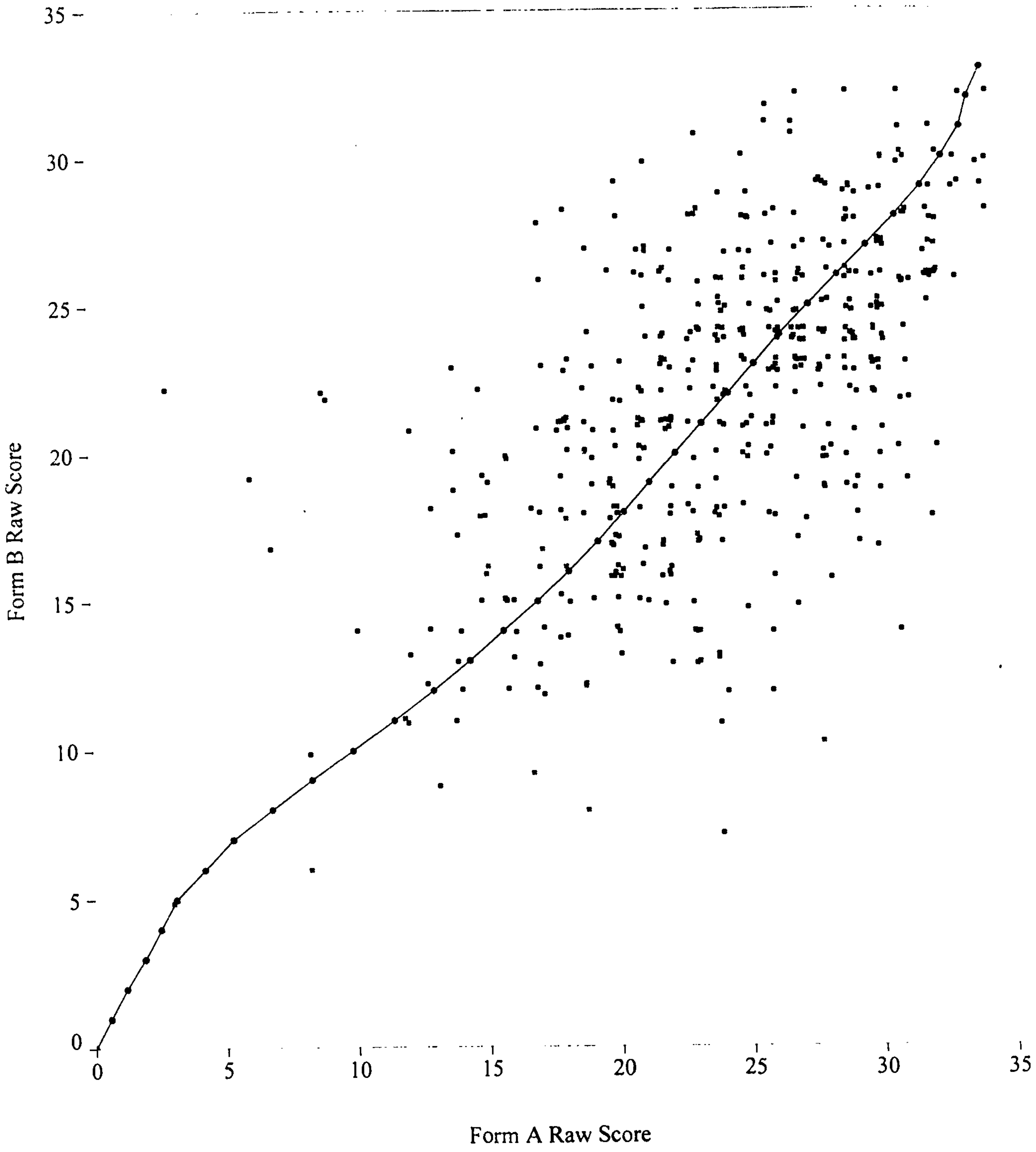


Figure 2

Scatterplot for Form A and NRT Scores Obtained in 1993

Grade 3 Mathematics

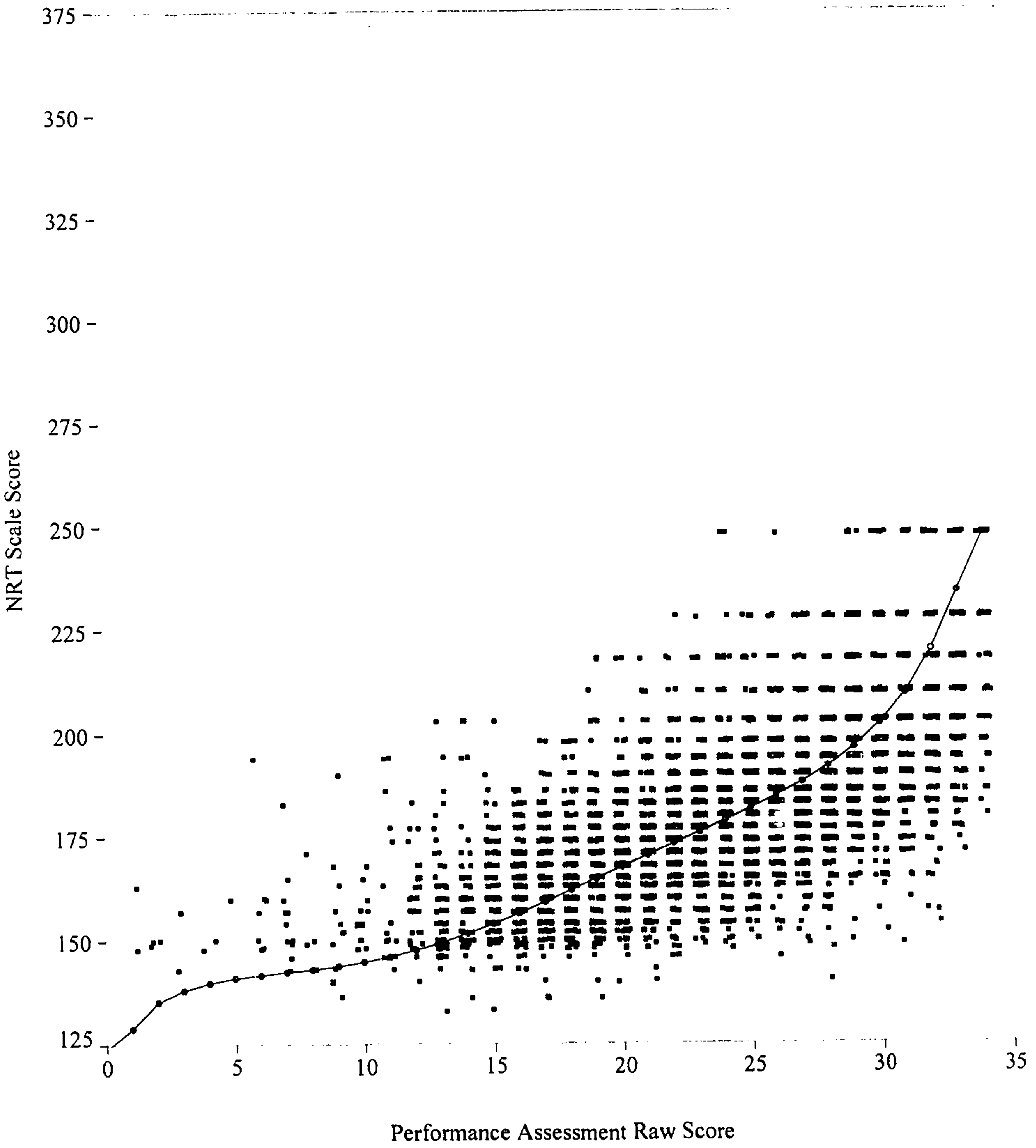
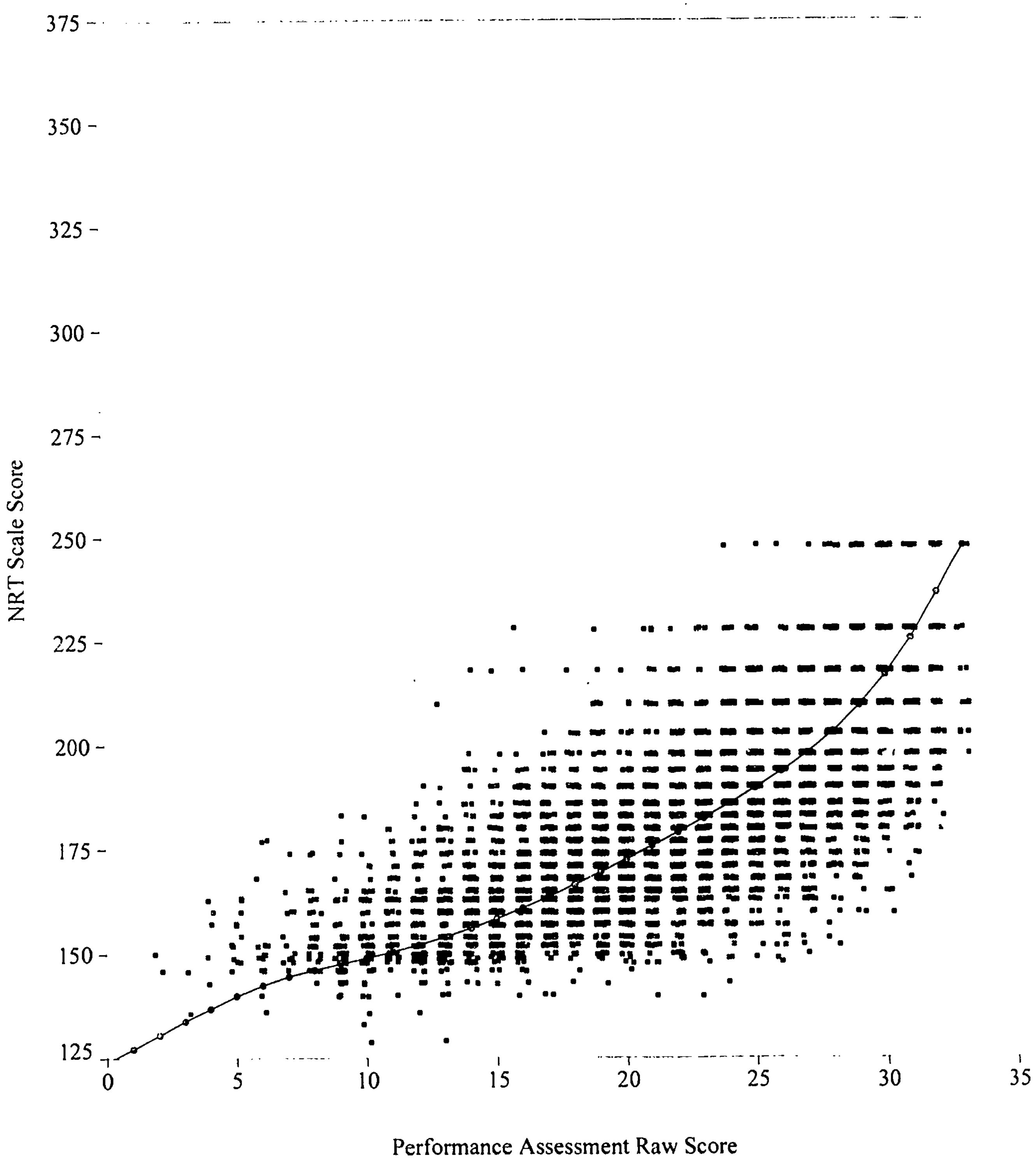


Figure 3

Scatterplot for Form B and NRT Scores Obtained in 1994

Grade 3 Mathematics


NRT Scale Score
Performance Assessment Raw Score