DOCUMENT RESUME

ED 281 857 TM 870 236

AUTHOR	Hills, John R.; And Others
TITLE	Equating Minimum Competency Tests: Comparison of Methods.
INSTITUTION	Florida State Univ., Tallahassee.
SPONS AGENCY	Florida State Dept. of Education, Tallahassee.
REPORT NO	FSU-2215-526-32
PUB DATE	13 Feb 87
CONTRACT	086-135
NOTE	42p.
PUB TYPE	Reports - Research/Technical (143) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Comparative Testing; *Cost Effectiveness; *Equated Scores; Feasibility Studies; High Schools; *Latent Trait Theory; Mathematical Models; *Minimum Competency Testing; State Programs; Test Theory
IDENTIFIERS	Florida; Florida State Student Assessment Test; Florida Statewide Assessment Program; *Linear Equating Method; *Rasch Model; Three Parameter Model

ABSTRACT

The 1986 scores from the Statewide Student Assessment Test-II, a minimum-competency test required for high school graduation in Florida, were placed on the scale of the 1984 scores from that test using five different equating procedures: (1) linear method; (2) Rasch model; (3) three-parameter item response theory (IRT)--concurrent method; (4) three-parameter IRT--fixed-parameter method; and (5) three-parameter IRT--formula method. The results were compared, as well as the computer costs. Also, the results from six different lengths of anchor items were compared. The different equating methods yielded very similar results. They would be essentially equally satisfactory in this situation in which the tests were made parallel in difficulty and content item by item, and the groups of examinees were population cohorts separated by only two years. Computer costs for the linear method were the least--one-tenth the costs of the most expensive, the concurrent IRT method. An anchor of 10 items provided equating as effective as 30 items using the concurrent IRT method. (Author/GDC)

EQUATING MINIMUM COMPETENCY TESTS: COMPARISON OF METHODS

Final Report

Contract/Grant # 086-135

FSU Number 2215 526 32

Between Florida Department of Education and
Florida State University

February 13, 1987

by

John R. Hills, Raja G. Subhiyah, and Thomas M. Hirsch

Florida State University

Abstract

The 1986 scores from the SSAT-II, a minimum-competency test required for high-school graduation in the State of Florida, were placed on the scale of the 1984 scores from that test using five different equating procedures. The results were compared, as well as the computer costs. Also the results from six different lengths of anchor items were compared. The different equating methods yielded very similar results. They would be essentially equally satisfactory in this situation in which the tests were made parallel in difficulty and content item by item and the groups of examinees were population cohorts separated by only two years. Computer costs for the linear method were the least--one tenth the costs of the most expensive which was the concurrent IRT method. An anchor of 10 items provided equating as effective as 30 items using the concurrent IRT method.

EQUATING MINIMUM COMPETENCY TESTS: COMPARISON OF METHODS

In this research project, which comprises the first two investigations described under Phase Two on page 38 of the report by Hills and Beard of 1984 entitled An Investigation of the Feasibility of Using the Three Parameter IRT Model in Florida's Student Assessment Program, we studied whether available equating procedures differ appreciably in their results and which procedures should be recommended for routine use based on quality, efficiency, and economy. We also evaluated whether any anomalies arise due to the fact that the Statewide Assessment Tests are minimum competency tests and thus have extremely skewed distributions of scores.

Of the many equating methods that might have been examined, the linear method has previously been used with this test (Beard, Julian, & Subhiyah, 1985), and was included here. Equipercentile equating, in which scores are considered equivalent if they represent the same percentile rank, was not included in this study for three reasons. First, if moments of the score distributions above the second (i. e., skew and kurtosis) are equal, this method gives the same general equating relationship as the linear method, which is based on the equation of a straight line. For these data, the skew and kurtosis were very similar on the two forms. Second, Lord (1982) has shown that for data sets which have similar score distributions the equipercentile method has a considerably larger standard error than the linear method. Third, the linear method does not involve the subjectivity of curve fitting which is part of the equipercentile method. Thus, when both the linear method and the equipercentile method are appropriate, the linear method would automatically be preferred.

Use of the one-parameter (Rasch) IRT model for equating has been studied by several authors (Skaggs & Lissitz, 1986). A synthesis of these studies indicates that procedures based on the Rasch model are effective in horizontal equating where "the data are reasonably reliable, tests are nearly equal in difficulty, and samples are nearly equal in ability" (p. 523). Furthermore, this method is the simplest of the IRT based procedures. Thus, it was included in this study.

Several procedures based on the three-parameter IRT model were included for contrast with those based on the linear and the one-parameter IRT models. One procedure, IRTCON, estimated the item parameters of both test forms in a single analysis. Another, IRTFIX, fixed the b parameters on the anchor items at their 1984 values when the 1986 item parameters were estimated. The third, IRTFOR, used b parameters from the anchor items in 1984 and 1986 in a formula to transpose the 1986 scale to the 1984 scale.

A number of possible complications might arise when using IRT based equating methods with tests such as SSAT-II. One might question whether the data from the SSAT-II test are sufficiently

unidimensional for any IRT model to be useful. The mathematics and communications tests were not designed to be homogeneous. Each of them is composed of separate small sets of items on applications of supposedly distinct basic skills. However, recent study (Hills, Beard, Yotinprasert, Roca, & Subhiyah, 1985) seems to indicate that the communicatons and mathematics items each measure one global dimension to a sufficient degree that parameter estimates are not seriously distorted.

Another problem that could arise from using the three-parameter IRT model is the difficulty in estimating the probability that a person of very low ability will answer very easy items correctly. Data for people of such low ability are extremely scarce for very easy items because the score distribution is negatively skewed. In such situations, computer programs such as LOGIST 5 (Wingersky, Barton, & Lord, 1982) often set the c or "psuedo-guessing" parameters for many items to a common value instead of estimating each independently. This may cause the three-parameter model to be ineffective for equating these data.

In spite of these possible difficulties, there may be advantages in using IRT related methods for equating provided IRT is also routinely used in item analysis and item banking. For example, to have a bank of items with known discrimination, guessing, and difficulty parameters all based on the same scale, should be a tremendous asset in test construction. Not only could one select items for a new test which measure the same skill and are of approximately the same difficulty, but items of approximately the same discrimination and susceptibility to guessing could be chosen. Further, one might take the approach of developing a new form which has the same test information curve as the previous form. Or one might use IRT parameters in building a test with an information curve optimal for a specific function or property (e.g., high accuracy at the cutting score). Thus, by providing substantially more sophisticated test construction possibilities than classical methods permit, IRT methods may offer advantages that outweigh any difficulties associated with them.

Methods

Three kinds of comparisons were made in this study. First methods for equating were compared. Five methods for equating the 1986 subtests with the 1984 subtests were used. One of the methods used is based on classical test theory, while the other four are based on item response theory. Each equating method yielded an equating chart which listed the 1984 scores opposite their equivalent (equated) 1986 scores in a dictionary-like manner. These equating charts or "dictionaries" were translated into graphs by plotting the 1986 scores versus the 1984 scores to obtain equating curves. The curves obtained by the various methods were then examined in order to compare differences in equated scores across methods, particularly at the cutoff point. Finally, the scores on the charts were rounded off to the nearest

whole number, and are presented in Appendix 1 for detailed study.

Second, results obtained from using anchors of different sizes were compared. In order to compare the sets of equated scores that were yielded by using different anchor test sizes, only the items unique to each test were placed on the charts. This kept the length of the tests that were being equated constant at 45 items, while the number of common items was manipulated.

Five common items were chosen randomly from the pool of 30 available items and analyzed using the IRTCON method with the unique items. Subsequently, another five randomly-chosen common items were added to the next analysis to make a total of ten common items, and so on until all the common items were included.

Each analysis resulted in a set of equated true scores for the 45-item tests. The equating chart obtained by using all 30 common items was the standard against which the other equatings were judged. In order to determine how closely an equating agreed with the standard, the average absolute difference and the standard difference were calculated as follows.

1. Each score on the 1986 test was subtracted from its equivalent 1984 score. This step was repeated for all the analyses, resulting in the primary differences, D.

2. The D values of each run were subtracted from the corresponding D values of the standard run (with 30 common items). This yielded a set of secondary differences, M.

3. The absolute values of the secondary differnces (Ms) were averaged to yield a statistic that indicates to what extent any equating differed from the standard 30-common-item equating.

4. The secondary differences (Ms) were squared, summed, and divided by the number of scores to yield a "variance" of these differences. The square root of this variance was called the standard difference between the equating methods.

5. The maximum M value in each case was also reported to indicate the greatest discrepancy that was found in each comparison.

Third, the costs of the different equating methods were compared. The costs being compared here were based solely on the average computer expenditure for the various runs.

Equating Methods

Following is a brief description of each of the equating methods used in this study:

The Linear Method (LINEAR).

This method is widely known as Angoff's Design IV (Angoff,1984). It involves administering two tests with a set of common items (anchor test) to two nonrandom groups of respondents. The mean and standard deviation of the common item scores for each group are calculated, and the raw scores of the two groups are equated using the formula:

$$Y = \frac{X - \hat{M}_x}{\hat{S}_x} \hat{S}_y + \hat{M}_y$$

where $\hat{M}_x$ and $\hat{M}_y$ are estimated means, and $\hat{S}_x$ and $\hat{S}_y$ are the estimated standard deviations for tests Y and X for the total population. This method is based on classical test theory and uses the difficulties (p values) of the common items to equate raw scores.

Rasch Model (RASCH).

In this method of equating, BICAL (Wright, Mead, & Bell, 1980) was used. First, the difficulty parameters (b values) of all the items on each test were determined. Then, the mean b values of the common items on each test were calculated. Subsequently, the mean b value of the 1984 test was subtracted from that of the 1986 test to obtain the additive constant. This additive constant is then added to the log abilities of the 1986 test to obtain equated log abilities for the 1986 test (i.e., the 1986 test is put on the scale of the 1984 test). The raw scores corresponding to the equated log abilities were considered equivalent, and put on the equating curves and charts.

Three-parameter IRT: Concurrent Method (IRTCON).

The data were treated as if there were 6000 respondents to whom all the items in both the 1984 and the 1986 tests were administered. The items unique to the 1986 test were coded "not reached" for the 1984 respondents, while the items unique to the 1984 test were coded "not reached" for the 1986 respondents. This set of data was analyzed using LOGIST 5 (Wingersky, Barton, & Lord, 1982), automatically placing the 1984 and 1986 ability and item parameter estimates on the same scale (Cook & Eignor, 1985).

True scores that correspond to selected ability (theta) levels were calculated for each year by using the formula:

$$\Omega_j = \sum_{i=1}^{n} P_i(\theta_j)$$

where Ω_j is the true score, P_i is the probability of the respondent getting the item right, and θ_j is the estimated ability level (Hambleton & Swaminathan, 1985, p.212). The two sets of true scores are then considered equated approximations of raw scores (Lord & Wingersky, 1983), and plotted onto the equating curves. Subsequently, they were rounded off to the nearest whole number and placed on the equating charts.

Three-parameter IRT: Fixed-parameter Method (IRTFIX).

In this method the 1984 data, alone, were analyzed using LOGIST 5 and item parameter estimates were obtained. Then, the 1986 data were analyzed using LOGIST 5, but with the difficulty parameters for the common items fixed to the values obtained in the first analysis i.e., to their "bank" values. This procedure fixes the 1986 scale onto the 1984 scale. It should be noted that this method relies heavily on the assumption that the estimated item parameters are invariant across groups. True scores are then obtained and treated as in IRTCON.

Three-parameter IRT: Formula Method (IRTFOR).

The 1984 and 1986 data were separately analyzed using LOGIST 5. For each year, the mean and standard deviation of the b-values for the common items were calculated. Then the two scales were equated by using the following formulas (Hambleton & Swaminathan, 1985, p. 222) for fixing the discrimination (a) and difficulty (b) parameters:

$$b_y = \alpha b_x + \beta$$

$$a_y = a_x/\alpha$$

where: $\alpha = s_y/s_x$ and $\beta = \bar{y} - \alpha\bar{x}$

True scores were obtained and treated as in IRTCON to obtain the equating curves and charts.

Instruments

The communications and mathematics subtests of the SSAT-II for the years 1984 and 1986 which were used in this investigation consist of 75 items each. All of the items were multiple choice with four alternatives. No attempt had been made to construct these tests so that they would be unidimensional. However, the fact that separate scores were reported for communications and mathematics could be expected to produce reasonably

unidimensional tests (Hills, Beard, Yotinprasert, Roca, and Subhiyah, 1985). For a detailed description of the tests, see Florida Statewide Assessment Program (1985).

As described earlier, the yearly versions of each subtest were rather similar in content as well as in statistical properties. Table 1 presents descriptive statistics for each of the subtests. Thirty items were common to the 1984 and the 1986 versions of each subtest.

Table 1. Descriptive Statistics of SSAT-II Tests

Test	Mean	SD	Kurtosis	Skew	Relative Mean
Communications, 1984	67.2	8.6	4.7	-2.0	.896
Communications, 1986	67.6	8.0	7.4	-2.4	.901
Mathematics, 1984	60.0	11.1	0.1	-0.9	.800
Mathematics, 1986	62.0	10.5	1.1	-1.2	.827

* Relative Mean = Mean divided by maximum possible raw score.

Sample

For each year, a random sample of 3000 respondents who had taken both the communications and the mathematics subtests was selected. These subjects were enrolled in grades nine to eleven of schools in Florida. The two samples represented cohort populations who passed through the educational system in Florida two years apart and were thus expected to be very similar in performance patterns on the SSAT-II. The descriptive statistics of the raw scores obtained by the two samples on the common items of each subtest presented in Table 2 illustrate this similarity between the two groups.

Table 2. Statistics of Raw Scores on Common Items of the SSAT-II

Test	Mean	SD	Relative Mean*
Communications, 1984	26.3	3.8	.877
Communications, 1986	26.6	3.6	.887
Mathematics, 1984	24.5	4.3	.817
Mathematics, 1986	24.8	4.4	.827

* Relative Mean = Mean divided by number of common items.

Results

Because of the lack of an absolute for evaluating the different methods, the linear equating results were used as the base for comparison. These results are presented in a series of graphs comparing the various equating methods for both the communications and the mathematics tests. More detailed lists are provided in Appendix 1 with each table containing a chart of the equated values.

Communications Equating

First we shall examine the results from the communications tests. In Figure 1, the graph shows virtually no difference between the RASCH method and the LINEAR method especially in the region of the cutoff score, where they give identical results. The only apparent discrepancy between the two methods was at the lower extreme.

Figure 2, which contrasts the results of three IRT based equating methods, IRTCON, IRTFIX, and IRTFOR, shows that two of these methods of equating provide very similar results. The IRTFOR and IRTCON methods are within one point of each other throughout the range of scores, as can be verified in Appendix 1. In contrast the IRTFIX method deviates considerably from the other two methods between the scores 25 and 60 on the communications 84 test.

Figures 3, 4, and 5 show the relationship of the LINEAR method to each of the three IRT based methods. Both the IRTFOR and IRTCON results are extremely close to the LINEAR equating results, with the IRTCON results being virtually indistinguishable from those of the LINEAR. At the cut-off score of 56 on

Figure 1. Comparison between the RASCH and the LINEAR methods in equating communications tests

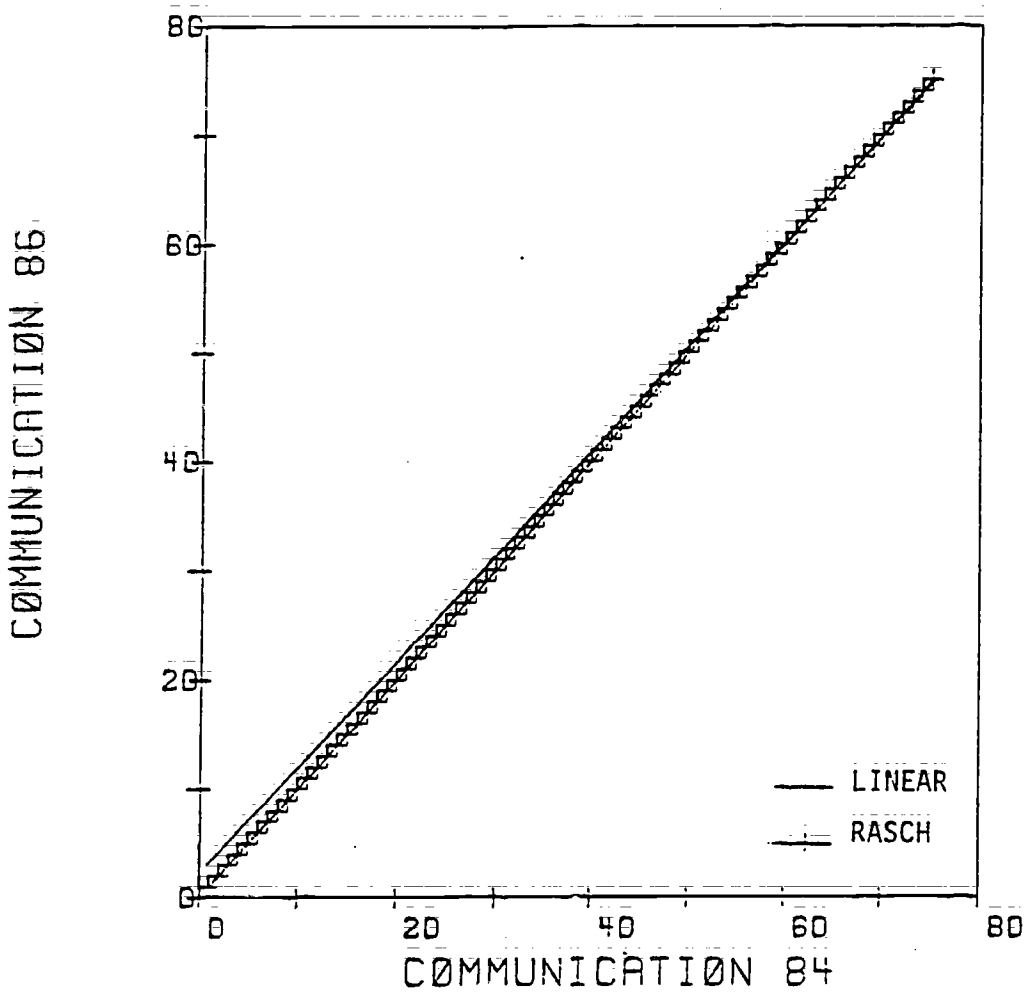


Figure 2. Comparisons among IRTCON, IRTFIX, and IRTFOR methods in equating communications tests

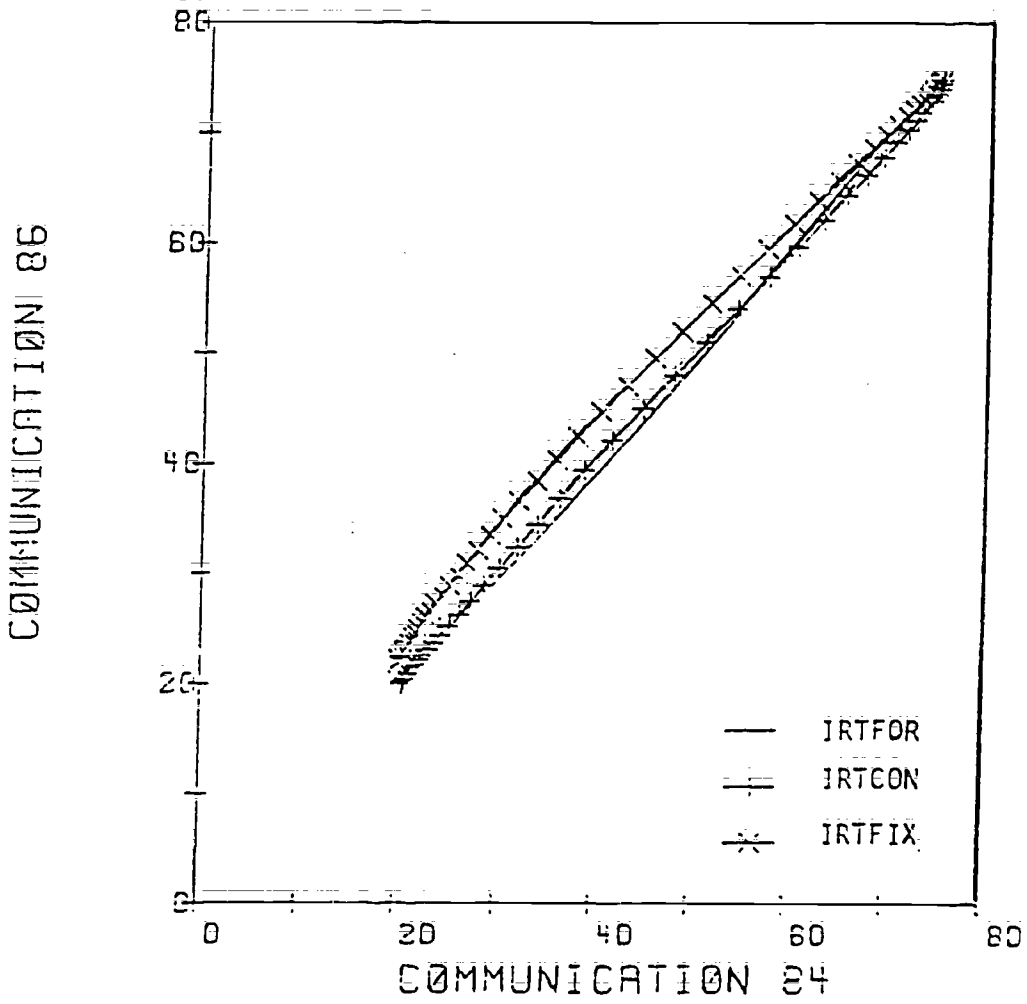


Figure 3. Comparison of LINEAR and IRTFOR methods in equating communications tests

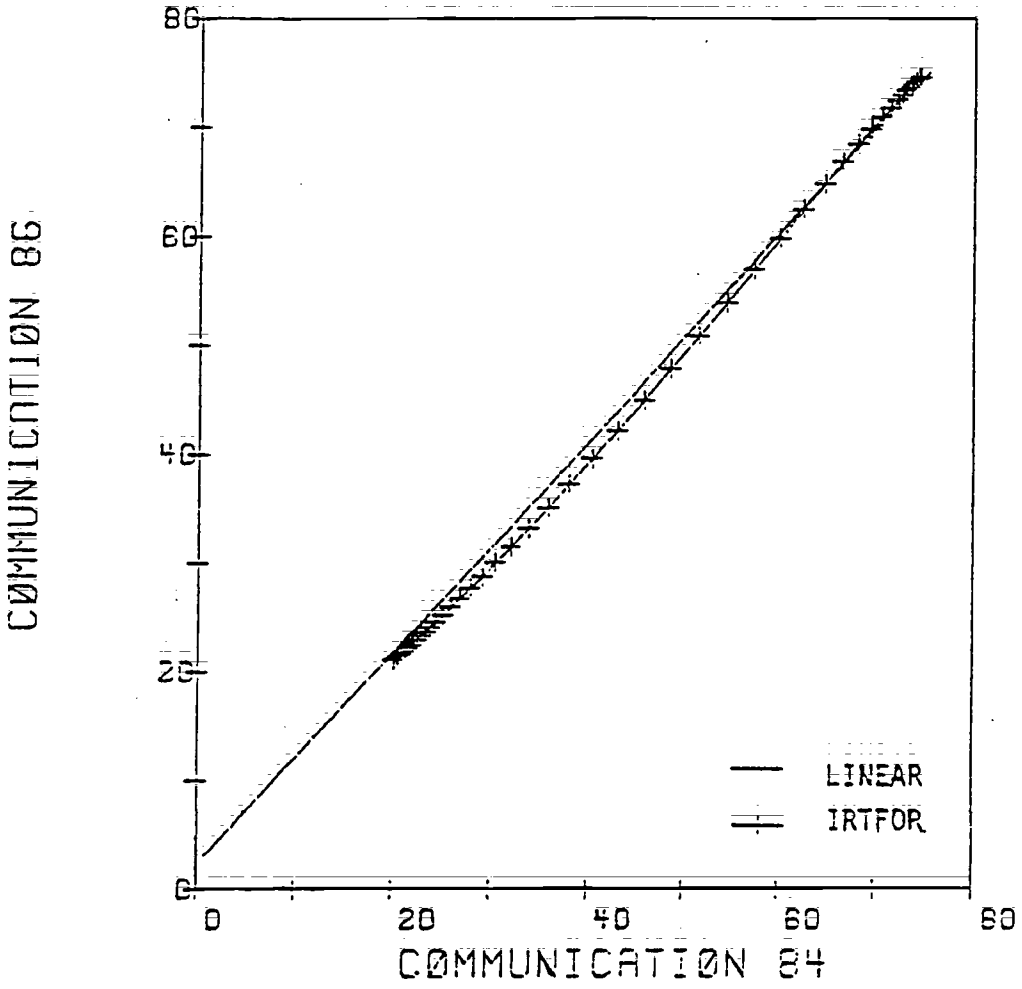


Figure 4. Comparison of LINEAR and IRTCON methods in equating communications tests

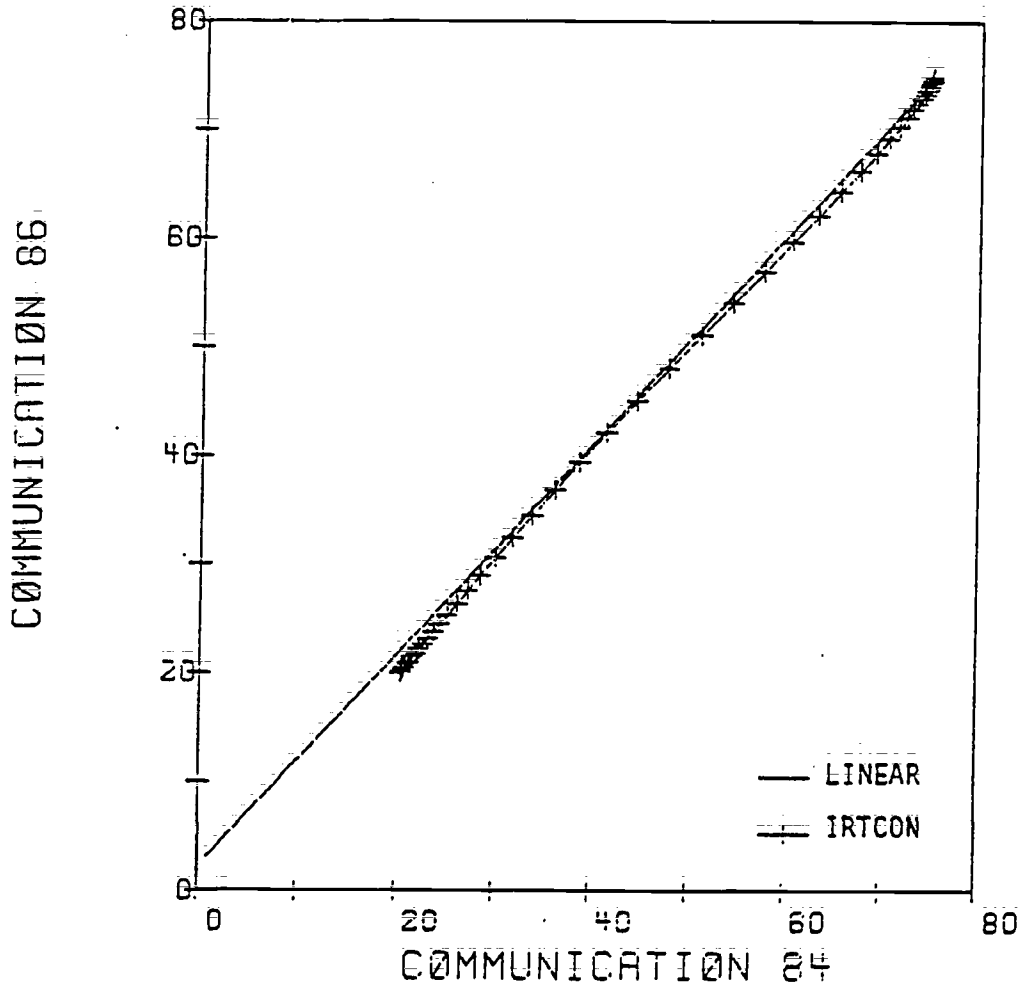
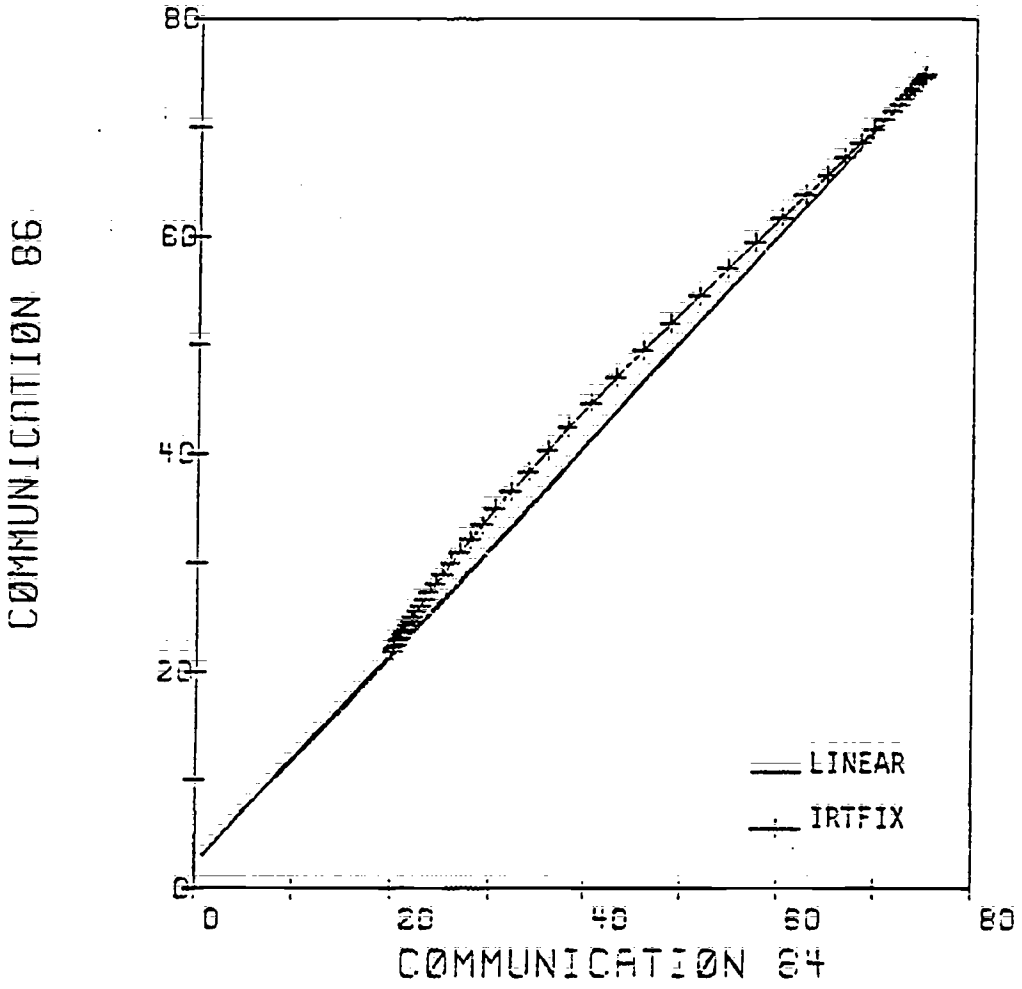


Figure 5. Comparison of LINEAR and IRTFIX methods in equating communications tests



the communication 84 test, both methods differ from the LINEAR method by one point. The IRTFIX method results deviate substantially from the LINEAR results between the scores 25 and 60. At the cutoff score they differ by only two points. A maximum discrepancy of slightly less than three points between the IRTFIX and the LINEAR methods occurs at the score of 41 on the 84 test.

Note that all three IRT equating methods using the three-parameter model do not equate the tests over the entire range of raw scores. This is because they equate true scores whose lower limit is restricted by the c-parameters. This lower limit (asymptote) is usually greater than zero, hence the discrepency between the ranges found in the LINEAR and IRT based methods.

Mathematics Equating

Turning to the mathematics tests, we find a different pattern in the behavior of the three IRT-related equating methods. In Figure 6, we see that it is the IRTFIX and IRTCON methods which coincide. As can be verified in appendix 1, between the scores 19 and 52 on the mathematics 84 test, these two methods provide essentially identical results. In contrast to the communications tests, here it is the IRTFOR method which deviates substantially from the other two between the scores 40 and 75.

In Figures 7 and 8 it can be seen that the results of the IRTFIX and IRTCON methods are in very close agreement with those of the LINEAR method throughout the range of equated scores. At the cutoff score of 47 on the mathematics 84 test, both methods differ from the LINEAR method by only one point. Although the IRTFOR results in Figure 9 agree with the LINEAR results between scores 25 and 40, they deviate appreciably for the remainder of the score range. At the cutoff point the IRTFOR and the LINEAR results differ by one point. They differ by a maximum of almost three points at the score of 60 on the 84 test.

Finally, it can be seen in Figure 10 that the one-parameter based (RASCH) method gives results consistent with the LINEAR method in the mathematics tests, repeating its performance on the communications tests, though with a little greater deviation as scores become lower.

The inconsistent behavior of the three IRT based methods clearly agrees with the findings of Cook and Eignor (1985). As in their study, it appears from these results that the IRTCON equating method provides the most consistent results when compared with LINEAR methods.

Figure 6. Comparisons among IRTCON, IRTFIX, and IRTFOR methods in equating mathematics tests

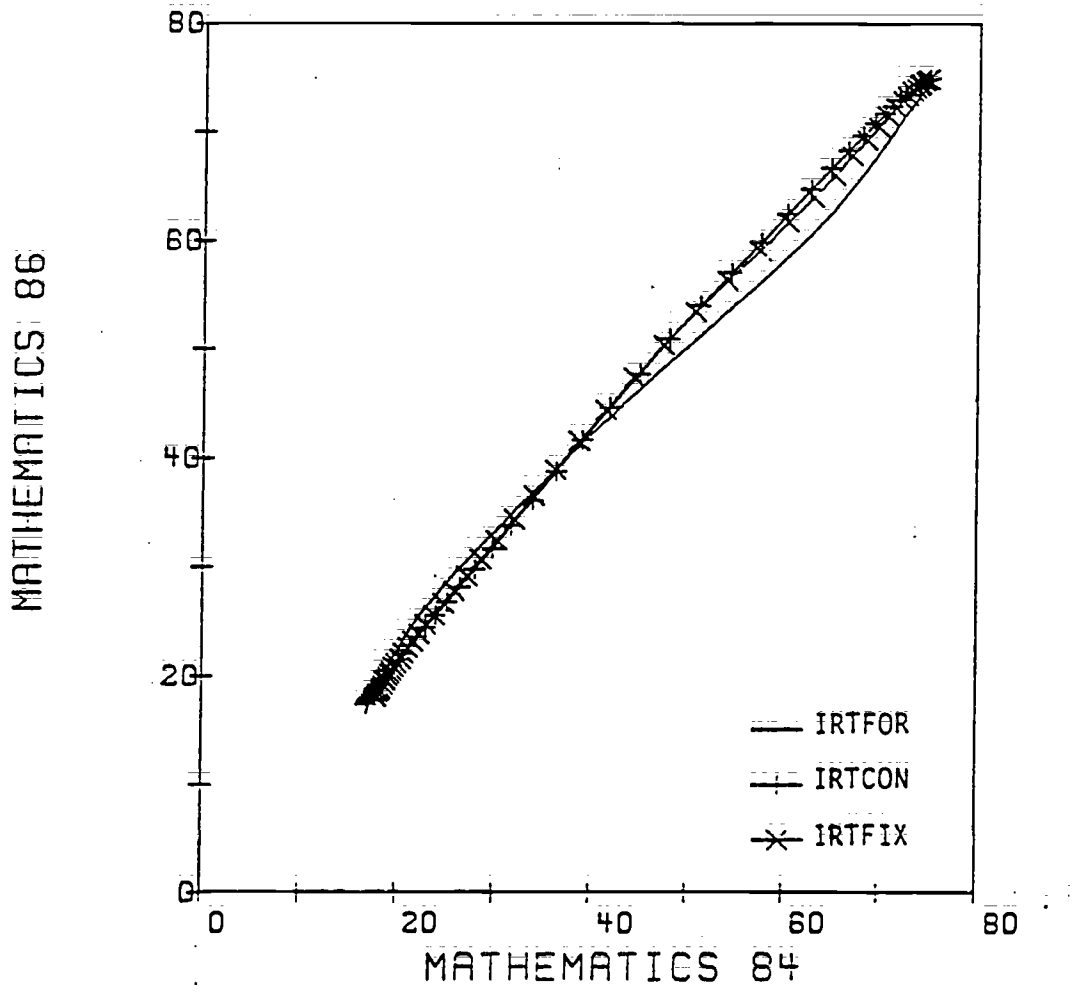


Figure 7. Comparison of LINEAR and IRTCON methods in equating mathematics tests

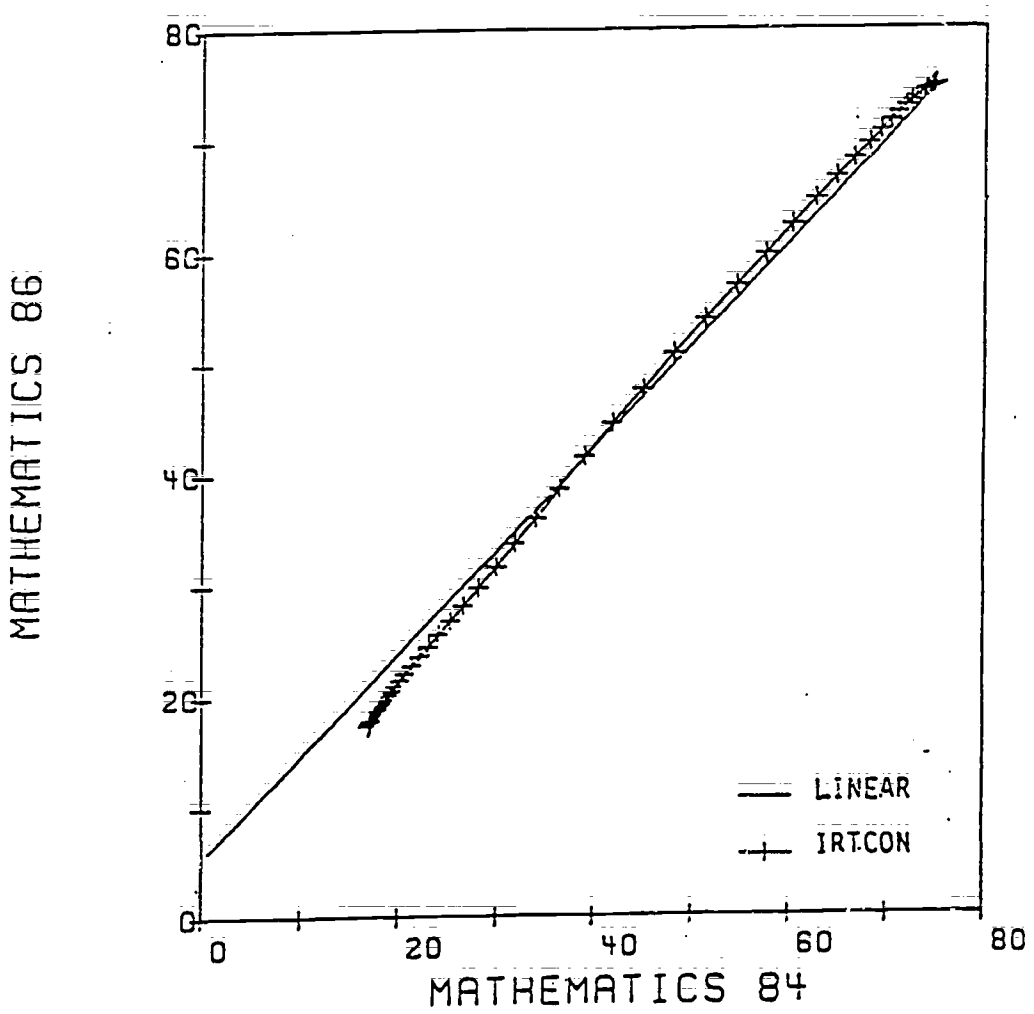


Figure 8. Comparison of LINEAR and IRTFIX methods in equating mathematics tests

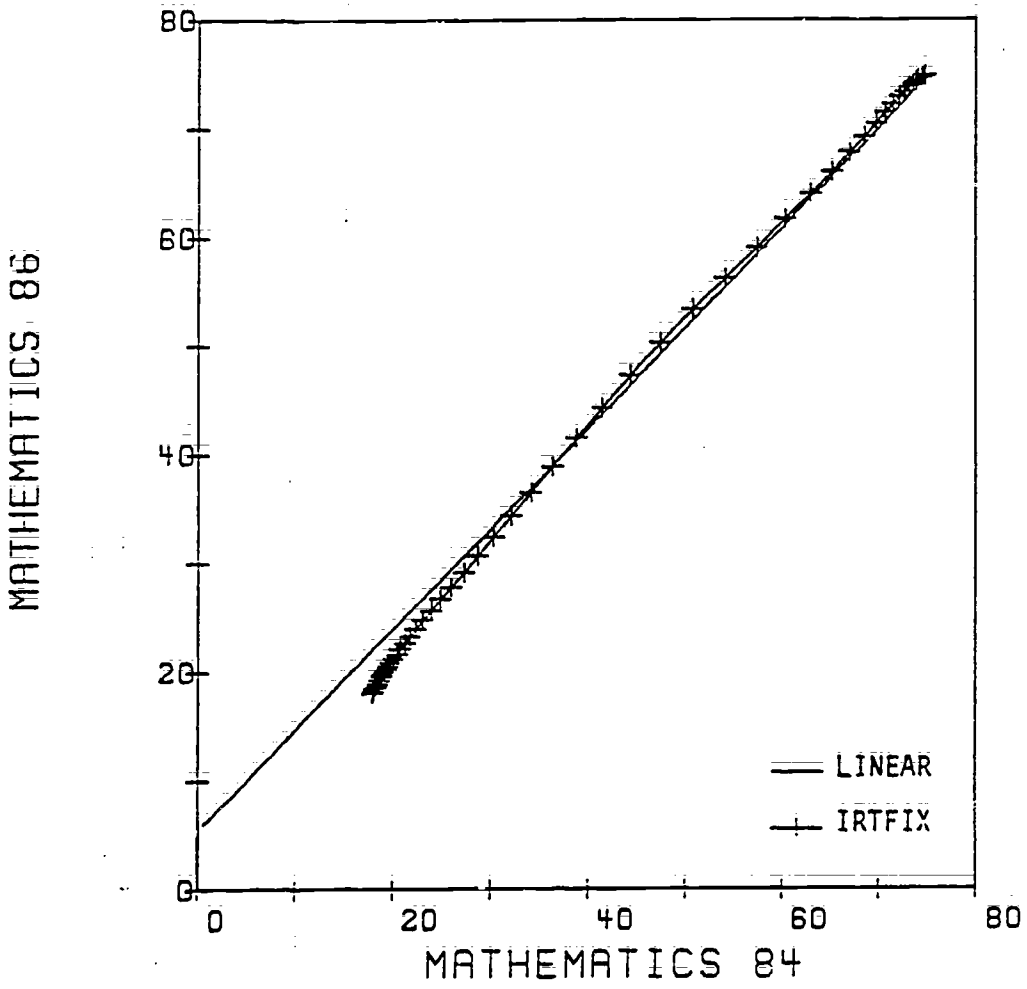


Figure 9. Comparison of LINEAR and IRTFOR methods in equating mathematics tests

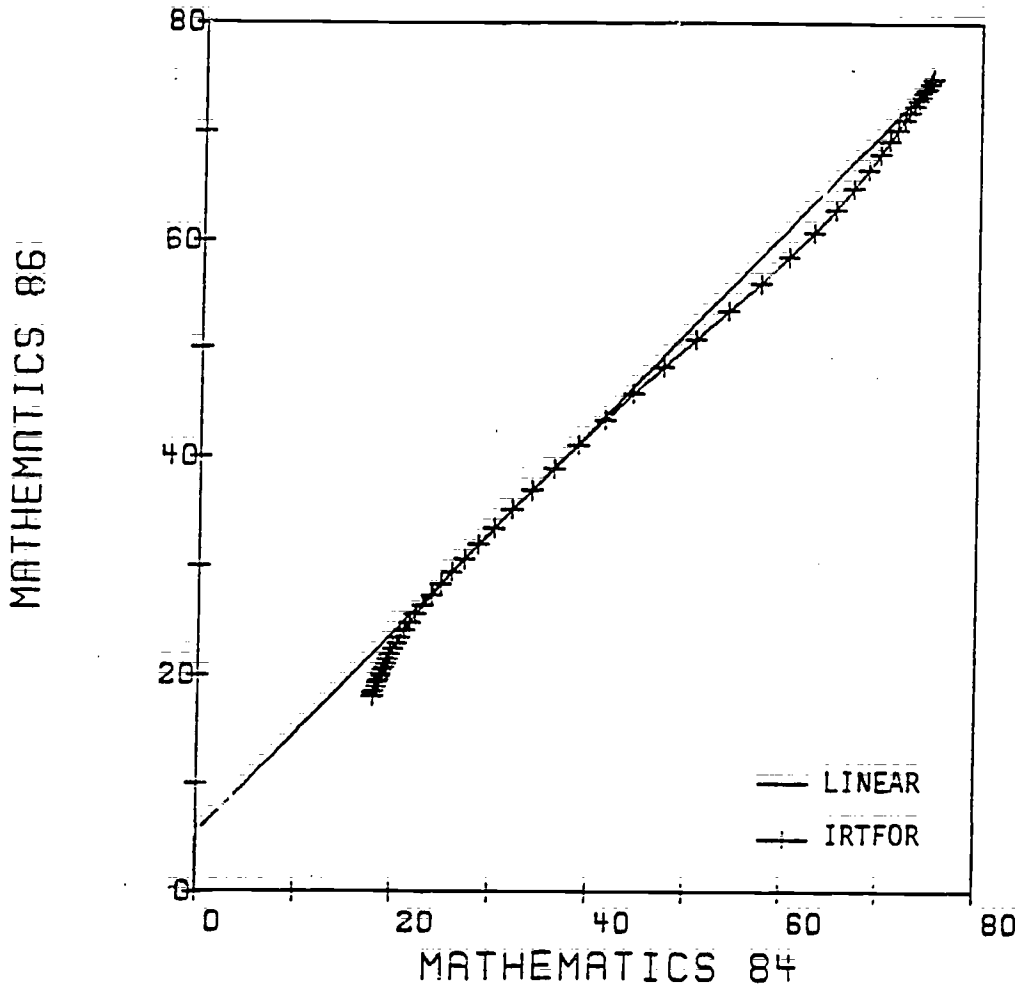
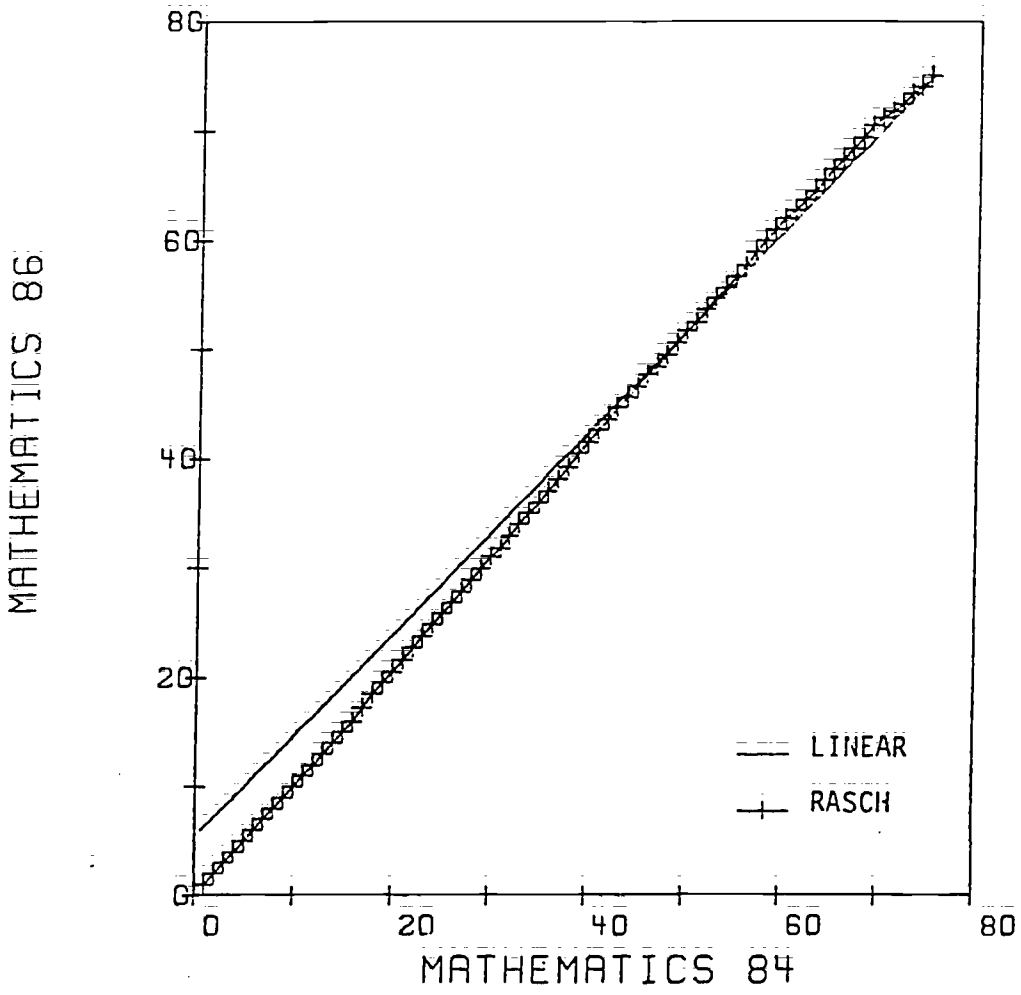


Figure 10. Comparison of RASCH and LINEAR methods in equating mathematics tests



Anchor Length

The results pertaining to anchor test length are displayed in Table 3 Using five anchor items to equate the tests results in a mean absolute difference of 0.934 and a maximum difference of 3.0. This maximum occurs at the theta value of - 0.2. The smallest mean absolute difference is found using ten anchor items. The mean is 0.197 with a maximum difference of 1.0. Surprisingly, as more than 10 anchor items are used, the mean absolute difference does not decrease. When twenty anchor items are used, the mean absolute difference increases to 0.377. The standard deviation of the absolute differences remains fairly constant for the different equatings using ten through twenty five anchor items.

Table 3. Statistics Comparing the Effect of Anchor Test Size on Equating SSAT-II Mathematics Scores using IRTCON.

Anchor Test Size	Mean Absolute Difference	SD of Absolute Difference	Maximum Absolute Difference
5	.934	.793	3
10	.197	.401	1
15	.344	.479	1
20	.377	.489	1
25	.246	.434	1

Computer Costs

The average costs of computer runs used to analyze the data are presented in Table 4. All of these analyses involved 3000 records with the exception of the IRTCON analyses which analyzed 6000 records in one run. Although IRTCON appears to be the most expensive procedure to run, it takes only one analysis to get the equating results, whereas IRTFOR and IRTFIX each requires two analyses, one for each year. If, however, equating is done every year, then results of previous years can be used, and only one new IRTFIX or IRTFOR analysis would be required each year.

Table 4. Computer Analysis Costs for Various Equating Programs.

Model	Program	Cost Range
LINEAR	LINE*	$ 2-5
RASCH	BICAL	$ 5-10
IRTFOR	LOGIST 5	$30-50
IRTFIX	LOGIST 5	$30-40
IRTCON	LOGIST 5	$50-70

* Program LINE was used with permission from Dr. J. G. Beard of the Florida State University.

Conclusions

As it turns out, none of the methods could be clearly chosen as "best" in this situation. Despite extremely skewed distributions, all the methods gave similar results. These tests are so nearly "classically parallel" (Gulliksen, 1950) that little or no equating is necessary. The linear method was least expensive, so it might well be used as the preferred method at least until such time as IRT technology is adopted for the entire test development and analysis procedure for this testing program. At that time it might be sound to use IRT equating even though it is relatively more expensive. The total amounts of money involved are very small compared to the other costs of the testing program. Among the IRT procedures, IRTCON was the most consistent and would be preferred.

The number of common or anchor items that must be used in equating these tests is interesting. The tradition, developed by Angoff in equating the College Board's Scholastic Aptitude Tests (Donlon and Angoff, 1971) is to use an anchor consisting of 20 items or 20% of the number of items in the test, whichever is larger. Wingersky and Lord (1984) and Raju, et al. (1986) have indicated that in using the three-parameter IRT model, as few as 5 items might be sufficient. We found that using the three-parameter model and the IRTCON procedure 10 randomly-chosen anchor items was an adequate number. This indicates that in using the three-parameter IRT method of equating, the designers of these tests could reduce greatly the number of common items and thereby increase the level of "security" of the test. (Notice that this result does not apply to any other method of equating.)

The results of this study are encouraging to those who wish to equate minimum-competency tests which are given repeatedly, as in a state-wide testing program or in a situation involving pretesting and posttesting to evaluate quality of instruction.

However, it must be recognized that these minimum-competency tests were constructed in a somewhat unusual way. The content was divided into skill areas. Within each skill area each item for a new form was chosen to have as nearly as possible the same p value (proportion correct) as the item being replaced. Such a test construction procedure resulted in highly parallel tests, but at the cost of wasting many items that cannot ever be used because their p values do not match the p va ue of any item in the original test form. One must conside. whether such an extremely high level of parallelism might soundly be sacrificed to some extent in order to be able to use more of the available items, with the slack in parallelism being rectified by means of equating. In such a case, the equating procedure would have more to accomplish, but the consistency of the results from these varied methods suggests that the available methods are quite adequate to the task in a situation such as this in which the populations being tested are very similar from year to year.

References

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Beard, J. G., Julian, E. R., & Subhiyah, R. G. (1985). Equating Florida's March 1985 statewide assessment test to the 1978 scale: Results for grade 10 SSAT-I and SSAT-II. Tallahassee, FL: Florida State University. Unpublished paper.

Cook, L. L., & Eignor, D. R. (1985). An investigation of the feasibility of applying item response theory to equate achievement tests. Research Report 85-31. Princeton, NJ: Educational Testing Service.

Cook, L. L., Dunbar, S. B., & Eignor, D. R. (1981). IRT equating: A flexible alternative to conventional methods for solving practical testing problems. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Donlon, T. F. & Angoff, W. H. In Angoff, W. H. (Ed.)(1971) The College Board Admissions Testing Program, New York: College Board.

Florida Statewide Assessment Program. (1985). An investigation of the feasibility of merging the SSAT-I and SSAT-II. Tallahassee, Fl.: Florida Department of Education.

Gulliksen, H. O. (1950). Theory of mental tests. New York: Wiley.

Hambleton, R., K., & Swaminathan, H. (1985). Item response theory. Boston: Kluwer Nijhoff Publishing.

Hills, J. R., Beard, J. G., Yotinprasert, S., Roca, N. R., & Subhiyah, R. G. (1985) An investigation of the feasibility of using the three-parameter model for Florida's statewide assessment tests. Tallahassee, FL: Florida State University. Unpublished paper.

Kolen, M. J., & Whitney, D. R. (1981). Comparison of four procedures for equating the tests of general educational development. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020.

Lord, F. M. (1975). A survey of equating methods based on item characteristic curve theory. Research Bulletin 75-13. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1982). The standard error of equipercentile equating. Journal of Educational Statistics, 7, 165-174.

Lord, F. M., & Wingersky, M. S. (1983). Comparison of IRT observed-score and true-score "equatings." Research Bulletin 83-26. Princeton, NJ: Educational Testing Service.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1981). IRT versus conventional equating methods: A comparative study of scale stability. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Raju, N. S., Bode, R. K., Larsen, V. S., & Steinhaus, S. (1986). Anchor-test size and horizontal equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Skaggs, G. & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Wingersky, M. S. & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8,347-364.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). Research Memorandum No. 23 c. Chicago: University of Chicago Department of Education Statistics Laboratory.

Appendix 1

Table 1. Equating True Scores of SSAT-II Communications: IRTCON

THETA	SCORE1	SCORE2	INFO1	INFO2
-3	32	32	7.92	8.55
-2.9	33	33	8.77	9.41
-2.8	34	34	9.70	10.29
-2.7	35	36	10.68	11.18
-2.6	36	37	11.72	12.06
-2.5	37	38	12.81	12.92
-2.4	39	39	13.95	13.75
-2.3	40	41	15.11	14.55
-2.2	42	42	16.29	15.32
-2.1	43	44	17.44	16.05
-2.	45	45	18.55	16.73
-1.9	46	46	19.58	17.36
-1.8	48	48	20.49	17.91
-1.7	50	49	21.25	18.38
-1.6	51	51	21.82	18.73
-1.5	53	53	22.18	18.95
-1.4	54	54	22.31	19.01
-1.3	56	55	22.20	18.91
-1.2	58	57	21.86	18.64
-1.1	59	58	21.32	18.22
-1.	61	60	20.59	17.65
-.9	62	61	19.70	16.96
-.8	63	62	18.68	16.17
-.7	64	63	17.57	15.29
-.6	66	64	16.39	14.35
-.5	67	65	15.17	13.37
-.4	67	66	13.94	12.37
-.3	68	67	12.71	11.37
-.2	69	68	11.51	10.39
-.1	70	69	10.36	9.44
0	70	69	9.26	8.54
.1	71	70	8.23	7.68
.2	71	70	7.27	6.89
.3	72	71	6.39	6.16
.4	72	71	5.60	5.49
.5	73	72	4.89	4.88
.6	73	72	4.25	4.34
.7	73	72	3.69	3.85
.8	73	72	3.20	3.41
.9	74	73	2.77	3.02
1.	74	73	2.39	2.67
1.1	74	73	2.07	2.37
1.2	74	73	1.79	2.10
1.3	74	73	1.55	1.86
1.4	74	74	1.34	1.64
1.5	74	74	1.16	1.46
1.6	74	74	1.00	1.29
1.7	75	74	.87	1.14
1.8	75	74	.75	1.02
1.9	75	74	.65	.90

Table 1. Continued

2.	75	74	.57	.80
2.1	75	74	.49	.71
2.2	75	74	.43	.64
2.3	75	74	.38	.57
2.4	75	74	.33	.51
2.5	75	74	.29	.45
2.6	75	74	.25	.41
2.7	75	74	.22	.36
2.8	75	75	.19	.33
2.9	75	75	.17	.29
3.	75	75	.15	.26

Table 2. Equating True Scores of SSAT-II Mathematics: IRTCON

THETA	SCORE1	SCORE2	INFO2	INFO2
-3	25	27	4.71	4.43
-2.9	26	27	5.20	4.91
-2.8	27	28	5.73	5.45
-2.7	27	29	6.29	6.06
-2.6	28	30	6.90	6.76
-2.5	29	31	7.54	7.55
-2.4	30	32	8.23	8.43
-2.3	31	33	8.95	9.40
-2.2	32	34	9.72	10.44
-2.1	33	35	10.54	11.53
-2.	34	36	11.39	12.64
-1.9	35	37	12.29	13.74
-1.8	36	39	13.22	14.79
-1.7	38	40	14.18	15.77
-1.6	39	42	15.16	16.68
-1.5	40	43	16.16	17.50
-1.4	42	45	17.15	18.23
-1.3	43	46	18.12	18.88
-1.2	45	48	19.04	19.43
-1.1	47	49	19.90	19.90
-1.	48	51	20.66	20.25
-.9	50	52	21.28	20.50
-.8	51	54	21.74	20.62
-.7	53	55	21.99	20.60
-.6	54	57	22.00	20.40
-.5	56	58	21.76	20.03
-.4	57	60	21.26	19.48
-.3	59	61	20.51	18.76
-.2	60	62	19.56	17.88
-.1	62	64	18.44	16.88
0	63	65	17.21	15.79
.1	64	66	15.92	14.65
.2	65	67	14.60	13.50
.3	66	68	13.31	12.36
.4	67	68	12.07	11.24
.5	67	69	10.89	10.18
.6	68	70	9.80	9.19
.7	69	70	8.80	8.25
.8	69	71	7.90	7.40
.9	70	71	7.10	6.61
1.	70	72	6.39	5.89
1.1	71	72	5.77	5.24
1.2	71	72	5.23	4.65
1.3	71	73	4.76	4.11
1.4	72	73	4.34	3.64
1.5	72	73	3.97	3.21
1.6	72	73	3.64	2.82
1.7	73	74	3.34	2.48

Table 2. Continued

1.8	73	74	3.07	2.18
1.9	73	74	2.81	1.91
2.	73	74	2.57	1.67
2.1	73	74	2.34	1.46
2.2	74	74	2.13	1.28
2.3	74	74	1.92	1.12
2.4	74	74	1.73	.98
2.5	74	74	1.55	.85
2.6	74	75	1.39	.75
2.7	74	75	1.24	.65
2.8	74	75	1.10	.57
2.9	74	75	.97	.50
3.	74	75	.86	.44

Table 3. Equating True Scores for SSAT-II Communications: IRTFOR

THETA	SCORE1	SCORE2	INFO1	INFO2
-3	34	33	7.81	7.29
-2.9	35	34	8.49	8.12
-2.8	36	35	9.19	9.01
-2.7	37	36	9.91	9.94
-2.6	38	37	10.65	10.90
-2.5	39	38	11.42	11.87
-2.4	41	40	12.20	12.84
-2.3	42	41	13.00	13.78
-2.2	43	42	13.81	14.67
-2.1	44	43	14.63	15.52
-2.	46	45	15.43	16.32
-1.9	47	46	16.21	17.09
-1.8	49	48	16.94	17.83
-1.7	50	49	17.59	18.56
-1.6	52	51	18.11	19.26
-1.5	53	52	18.49	19.91
-1.4	55	54	18.70	20.44
-1.3	56	55	18.71	20.82
-1.2	57	57	18.53	20.97
-1.1	59	58	18.16	20.87
-1.	60	60	17.62	20.51
-.9	61	61	16.94	19.90
-.8	63	62	16.15	19.08
-.7	64	64	15.27	18.09
-.6	65	65	14.34	16.97
-.5	66	66	13.37	15.77
-.4	67	67	12.38	14.54
-.3	67	68	11.40	13.29
-.2	68	68	10.42	12.06
-.1	69	69	9.47	10.85
0	70	70	8.56	9.68
.1	70	70	7.68	8.56
.2	71	71	6.86	7.52
.3	71	71	6.10	6.56
.4	71	72	5.40	5.71
.5	72	72	4.76	4.94
.6	72	72	4.19	4.28
.7	72	73	3.68	3.70
.8	73	73	3.23	3.20
.9	73	73	2.83	2.76
1.	73	73	2.48	2.40
1.1	73	73	2.18	2.08
1.2	73	74	1.91	1.81
1.3	74	74	1.68	1.57
1.4	74	74	1.48	1.37
1.5	74	74	1.30	1.20
1.6	74	74	1.15	1.05

Table 3. Continued

1.7	74	74	1.01	.92
1.8	74	74	.90	.81
1.9	74	74	.80	.72
2.	74	74	.71	.63
2.1	74	74	.63	.56
2.2	74	74	.56	.50
2.3	74	74	.50	.44
2.4	74	75	.45	.39
2.5	75	75	.40	.35
2.6	75	75	.36	.32
2.7	75	75	.32	.28
2.8	75	75	.29	.25
2.9	75	75	.26	.23
3.	75	75	.24	.21

Table 4. Equating True Scores of SSAT-II Mathematics: IRTFOR

THETA	SCORE1	SCORE2	INFO1	INFO2
-3	26	29	4.89	3.46
-2.9	27	30	5.39	3.71
-2.8	27	31	5.92	3.98
-2.7	28	31	6.47	4.29
-2.6	29	32	7.05	4.65
-2.5	30	33	7.65	5.05
-2.4	30	33	8.27	5.50
-2.3	31	34	8.92	6.00
-2.2	32	35	9.59	6.54
-1.9	35	38	11.84	8.37
-1.8	36	39	12.69	8.99
-1.7	38	40	13.61	9.61
-1.6	39	41	14.59	10.20
-1.5	40	42	15.65	10.76
-1.4	41	43	16.78	11.30
-1.3	43	45	17.97	11.81
-1.2	44	46	19.22	12.29
-1.1	46	47	20.49	12.74
-1.	48	48	21.75	13.16
-.9	49	50	22.96	13.56
-.8	51	51	24.04	13.92
-.7	52	52	24.92	14.26
-.6	54	53	25.50	14.55
-.5	56	55	25.71	14.79
-.4	57	56	25.48	14.96
-.3	59	57	24.81	15.03
-.2	60	58	23.72	14.99
-.1	62	60	22.31	14.80
0	63	61	20.66	14.46
.1	64	62	18.88	13.99
.2	65	63	17.06	13.40
.3	66	64	15.26	12.71
.4	67	65	13.55	11.97
.5	68	66	11.97	11.19
.6	68	66	10.54	10.40
.7	69	67	9.27	9.61
.8	70	68	8.17	8.85
.9	70	69	7.23	8.12
1.	71	69	6.45	7.43
1.1	71	70	5.81	6.78
1.2	71	70	5.28	6.17
1.3	72	71	4.84	5.61
1.4	72	71	4.46	5.08
1.5	72	71	4.12	4.60
1.6	73	72	3.80	4.16
1.7	73	72	3.49	3.75
1.8	73	72	3.18	3.38

Table 4. Continued

1.9	73	73	2.88	3.03
2.	73	73	2.59	2.72
2.1	74	73	2.31	2.43
2.2	74	73	2.06	2.17
2.3	74	73	1.82	1.94
2.4	74	74	1.61	1.73
2.5	74	74	1.42	1.54
2.6	74	74	1.24	1.37
2.7	74	74	1.09	1.22
2.8	74	74	.96	1.08
2.9	74	74	.84	.96
3.	74	74	.73	.86

Table 5. Equating SSAT-II Communications True Scores: IRTFIX

THETA	SCORE1	SCORE2	INFO1	INFO2
-3	34	38	7.81	7.35
-2.9	35	39	8.49	7.91
-2.8	36	40	9.19	8.48
-2.7	37	41	9.91	9.07
-2.6	38	42	10.65	9.66
-2.5	39	44	11.42	10.25
-2.4	41	45	12.20	10.83
-2.3	42	46	13.00	11.40
-2.2	43	47	13.81	11.96
-2.1	44	48	14.63	12.49
-2.	46	49	15.43	12.98
-1.9	47	51	16.21	13.43
-1.8	49	52	16.94	13.83
-1.7	50	53	17.59	14.16
-1.6	52	55	18.11	14.41
-1.5	53	56	18.49	14.57
-1.4	55	57	18.70	14.63
-1.3	56	58	18.71	14.58
-1.2	57	59	18.53	14.42
-1.1	59	61	18.16	14.15
-1.	60	62	17.62	13.76
-.9	61	63	16.94	13.27
-.8	63	64	16.15	12.70
-.7	64	65	15.27	12.06
-.6	65	66	14.34	11.37
-.5	66	66	13.37	10.66
-.4	67	67	12.38	9.93
-.3	67	68	11.40	9.19
-.2	68	69	10.42	8.45
-.1	69	69	9.47	7.72
0	70	70	8.56	6.99
.1	70	70	7.68	6.30
.2	71	71	6.86	5.64
.3	71	71	6.10	5.02
.4	71	71	5.40	4.46
.5	72	72	4.76	3.95
.6	72	72	4.19	3.49
.7	72	72	3.68	3.09
.8	73	73	3.23	2.73
.9	73	73	2.83	2.42
1.	73	73	2.48	2.14
1.1	73	73	2.18	1.90
1.2	73	73	1.91	1.68
1.3	74	73	1.68	1.50
1.4	74	74	1.48	1.33
1.5	74	74	1.30	1.19
1.6	74	74	1.15	1.06

Table 5. Continued

1.7	74	74	1.01	.95
1.8	74	74	.90	.85
1.9	74	74	.80	.76
2.	74	74	.71	.68
2.1	74	74	.63	.61
2.2	74	74	.56	.55
2.3	74	74	.50	.50
2.4	74	74	.45	.45
2.5	75	74	.40	.41
2.6	75	74	.36	.37
2.7	75	74	.32	.33
2.8	75	75	.29	.30
2.9	75	75	.26	.28
3.	75	75	.24	.25

Table 6. Equating SSAT-II Mathematics True Scores: IRTFIX

THETA	SCORE1	SCORE2	INFO1	INFO2
-3	26	27	4.89	5.53
-2.9	27	27	5.39	5.98
-2.8	27	28	5.92	6.47
-2.7	28	29	6.47	6.99
-2.6	29	30	7.05	7.57
-2.5	30	31	7.65	8.19
-2.4	30	32	8.27	8.88
-2.3	31	33	8.92	9.64
-2.2	32	34	9.59	10.46
-2.1	33	35	10.30	11.32
-2.	34	36	11.05	12.22
-1.9	35	37	11.84	13.13
-1.8	36	39	12.69	14.03
-1.7	38	40	13.61	14.91
-1.6	39	41	14.59	15.75
-1.5	40	43	15.65	16.56
-1.4	41	44	16.78	17.34
-1.3	43	46	17.97	18.09
-1.2	44	47	19.22	18.80
-1.1	46	49	20.49	19.48
-1.	48	51	21.75	20.09
-.9	49	52	22.96	20.63
-.8	51	54	24.04	21.04
-.7	52	55	24.92	21.28
-.6	54	57	25.50	21.30
-.5	56	58	25.71	21.05
-.4	57	60	25.48	20.52
-.3	59	61	24.81	19.72
-.2	60	62	23.72	18.69
-.1	62	64	22.31	17.49
0	63	65	20.66	16.20
.1	64	66	18.88	14.87
.2	65	67	17.06	13.55
.3	66	67	15.26	12.28
.4	67	68	13.55	11.09
.5	68	69	11.97	9.98
.6	68	70	10.54	8.95
.7	69	70	9.27	8.02
.8	70	71	8.17	7.17
.9	70	71	7.23	6.40
1.	71	72	6.45	5.71
1.1	71	72	5.81	5.08
1.2	71	72	5.28	4.52
1.3	72	73	4.84	4.01
1.4	72	73	4.46	3.55
1.5	72	73	4.12	3.14
1.6	73	73	3.80	2.77

Table 6. Continued

1.7	73	73	3.49	2.44
1.8	73	74	3.18	2.15
1.9	73	74	2.88	1.89
2.	73	74	2.59	1.66
2.1	74	74	2.31	1.45
2.2	74	74	2.06	1.27
2.3	74	74	1.82	1.12
2.4	74	74	1.61	.98
2.5	74	74	1.42	.86
2.6	74	74	1.24	.75
2.7	74	75	1.09	.66
2.8	74	75	.96	.58
2.9	74	75	.84	.51
3.	74	75	.73	.45

Table 7. Equated SSAT-II Communications Raw Scores: RASCH

1984	1986	1984	1986
1	1	41	41
2	2	42	42
3	3	43	43
4	4	44	44
5	5	45	45
6	6	46	46
7	7	47	47
8	8	48	48
9	9	49	49
10	10	50	50
11	11	51	51
12	12	52	52
13	13	53	53
14	14	54	54
15	15	55	55
16	16	56	56
17	17	57	57
18	18	58	58
19	19	59	59
20	20	60	60
21	21	61	61
22	22	62	62
23	23	63	63
24	24	64	64
25	25	65	65
26	26	66	66
27	27	67	67
28	28	68	68
29	29	69	69
30	30	70	70
31	31	71	71
32	32	72	72
33	33	73	73
34	34	74	74
35	35	75	75
36	36		
37	37		
38	38		
39	39		
40	40		

Table 8. Equating SSAT-II Mathematics Raw Scores: RASCH

1984	1986	1984	1986
1	1	40	42
2	2	41	43
3	3	42	44
4	4	43	45
5	5	44	46
6	6	45	47
7	7	46	48
8	8	47	48
9	9	48	50
10	10	49	51
11	11	50	52
12	12	51	53
13	13	52	54
14	14	53	55
15	15	54	56
16	16	55	57
17	17	56	58
18	19	57	59
19	20	58	60
20	21	59	61
21	22	60	62
22	23	61	63
23	24	62	64
24	25	63	65
25	26	64	66
26	27	65	67
27	28	66	67
28	29	67	68
29	30	68	69
30	31	69	71
31	32	70	71
32	33	71	72
33	34	72	73
34	35	73	73
35	36	74	74
36	37	75	75
37	38		
38	39		
39	41		

Table 9. Equating SSAT-II Communications Raw Scores: LINEAR

1984 SCORE	1986 SCORE	1984 SCORE	1984 SCORE
75	75	34	35
74	74	33	34
73	73	32	33
72	72	31	32
71	71	30	31
70	70	29	30
69	69	28	29
68	68	27	28
67	67	26	27
66	66	25	26
65	65	24	25
64	64	23	
63	63	22	24
62	62	21	23
61	61	20	22
60	60	19	21
59	59	18	20
58	58	17	19
57	57	16	18
56	56	15	17
55	55	14	16
54	54	13	15
53	53	12	14
52	52	11	13
51		10	12
50	51	9	11
49	50	8	10
48	49	7	9
47	48	6	8
46	47	5	7
45	46	4	6
44	45	3	5
43	44	2	4
42	43	1	3
41	42		
40	41		
39	40		
38	39		
37	38		
36	37		
35	36		

Table 10. Equating SSAT-II Mathematics Raw Scores: LINEAR

1984 SCORE	1986 SCORE	1984 SCORE	1986 SCORE
75	75	34	37
74	74	33	36
73	73	32	35
72	72	31	34
71	71	30	33
70	70	29	32
69	69	28	31
68	68	27	30
67	67	26	29
66	66	25	
65	65	24	28
64		23	27
63	64	22	26
62	63	21	25
61	62	20	24
60	61	19	23
59	60	18	22
58	59	17	21
57	58	16	20
56	57	15	19
55	56	14	18
54	55	13	17
53	54	12	16
52	53	11	
51		10	15
50	52	9	14
49	51	8	13
48	50	7	12
47	49	6	11
46	48	5	10
45	47	4	9
44	46	3	8
43	45	2	7
42	44	1	6
41	43		
40	42		
39	41		
38			
37	40		
36	39		
35	38		