DOCUMENT RESUME

ED 187 749 TM 800 251

AUTHOR Golub-Smith, Marna
TITLE The Application of Rasch Model Equating Techniques to the Problem of Interpreting Longitudinal Performance on Minimum Competency Tests.
PUB DATE Apr 80
NOTE 47p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Goodness of Fit; Mathematics; *Minimum Competency Testing; Reading Achievement
IDENTIFIERS Linear Models; New Jersey; New Jersey Minimum Basic Skills Program; *Rasch Model; *Test Equating

ABSTRACT

This study presents an application of the Rasch equating methodology on a minimum competency testing program in reading and mathematics. Common item equating was performed in two stages to link the 1978-1979 form of the New Jersey Minimum Basic Skills tests to the 1977-1978 form. Both fit to the Rasch model and stability of the common item pools were investigated prior to the actual equating. Equivalent raw scores derived from the Rasch methodology were compared with a prior linear equating. Special attention was paid to those raw score points around the state's cut-off score. Results from the study indicated moderate to good fit of the tests to the Rasch model. Several "unstable" equating items were found and the equating was therefore carried out in two ways: 1) using all twenty-five equating items, and 2) using just the "stable" equating items. Raw scores derived from both the Rasch and linear methods showed close though not perfect agreement. The equating using only the "stable" equating items changed six out of the seven equating tables. In only two of these six tables, did this change move the scores closer to that given by the linear method.
(Author/GSK)

THE APPLICATION OF RASCH MODEL EQUATING TECHNIQUES TO THE PROBLEM OF INTERPRETING LONGITUDINAL PERFORMANCE ON MINIMUM COMPETENCY TESTS

MARNA GOLUB-SMITH
NEW JERSEY STATE DEPARTMENT OF EDUCATION

Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts, April 1980.

THE APPLICATION OF RASCH MODEL EQUATING TECHNIQUES TO THE PROBLEM OF INTERPRETING LONGITUDINAL PERFORMANCE ON MINIMUM COMPETENCY TESTS

The New Jersey Minimum Basic Skills (MBS) program is a minimum competency testing program in reading and mathematics. Each spring both tests are administered to students in grades three, six, nine and eleven. Beginning with the 1977-1978 school year, the New Jersey State Board of Education adopted scores of 75 percent and 65 percent correct, as passing or "cut-off" scores on the reading and mathematics tests, respectively.

Because the law in New Jersey mandates the release of items after each test administration, new forms of the tests are developed each year. Although the test development process tries to insure comparability of items from one year to the next, present test construction techniques do not guarantee that two or more forms of a test--developed from the same set of specifications-- will be perfectly equivalent. In order then, to insure that the level of achievement required of students, as defined by the state cut-off score, would be the same on subsequent forms of the tests, it was decided to equate each new form to the previous one and hence to the original scale defined by the 1977-1978 form. Only with equated forms can one evaluate changes in student performance from one year to the next.

TEST DEVELOPMENT PROCESS

In order to understand the procedures used to equate the annual forms of the Minimum Basic Skills tests, it is necessary

to have an overview of the test development process. The development of a new form from item selection to final administration generally encompasses about nine months.

During early summer a statewide committee of reading and mathematics content specialists meets to review and edit items prepared by the test contractor.[1] A complete item-to-item replacement is made for each new test. In early fall these items are field-tested with school children in grades four, seven, ten and twelve (closest in ability to spring third, sixth, ninth and eleventh graders). The committee then reconvenes to review the results of the field test. Items are revised or replaced if necessary, and the new forms of the tests are ready for administration in the spring.

Several operational features of the Minimum Basic Skills program predetermined the method for equating future test forms:

1. A three-week-turnaround for reporting of results to local districts;

2. The inability to have a secure equating section on each year's final form test, as each year's test booklets remain in the local districts;

3. The use of previous year's items by classroom teachers to prepare students for the test.

Therefore, during the development of the second annual Minimum Basic Skills tests, it was decided to equate the two forms using an anchor test of twenty-five items selected from the 1977-1978 form and included on the field test of the 1978-1979 form. Angoff's design IV (1971, p. 579) was performed by the test

---

[1]The contractor for the Minimum Basic Skills program during 1977-1978 and 1978-1979 was Educational Testing Service.

contractor (Swineford, 1979).

PURPOSE OF THE STUDY

The present study was undertaken to answer the following two questions:

1. Could the Rasch model equating methodology be applied to a minimum competency test which was not specifically designed to fit the model?

2. How do the results of test form equating based on the traditional linear method compare with the results from the Rasch method?

METHODS OF EQUATING

Several methods are available to equate tests. Each provides a way of converting the system of units of one form to the system of units of another so that scores obtained after conversion will be equivalent. This notion of conversion implies two restrictions (Angoff, 1971):

- The two forms must be measures of the same characteristic;
- The conversion must be unique except for random error associated with the unreliability of the data and the method used for the transformation.

The two methods which were compared in this study were the linear and Rasch model methods of equating using a common set of anchor items.

The linear method of equating defines two scores as equivalent if they correspond to equal-standard score deviates. This method is based on the assumption that the shapes of the raw score distributions of two tests are identical. The use of linear equating with common items given to two separate non-random

groups is described by Angoff (1971, Design IV, p. 579). The major assumptions of this form of linear equating are:

- The regression system for the two groups of students would have been identical if the two groups had taken the same test;
- The common item set represents the same psychological function in both groups;
- The two groups do not differ very much in ability (they are not assumed to be equivalent).

The Rasch model, a one parameter latent trait model, provides a method of equating two tests using a common set of anchor items (Wright, 1977, Beard and Pettie, 1978). This method defines two scores as equivalent if they correspond to the same Rasch "log" ability values. These procedures are based on the estimation of equating constants which transform the item difficulties and ability estimates from one test on to the scale of the base test. The attractiveness of this method stems from the property of this model, labeled "Specific Objectivity" (Rasch, 1966), viz., the difficulty of the items are estimated independently of the ability of the calibrating sample and the estimates of the abilities are independent of the particular set of items.

METHODOLOGY

Description of the Tests

Each of the Minimum Basic Skills reading tests contains items which can be categorized into three major cluster areas: Word recognition, comprehension and study skills. The mathematics test items can be broken down into four major cluster areas: Computation, measurement/geometry,

number concepts and problem solving. The number of items in each test ranges from 90-110.

Selection of the Pool of Equating Items

Twenty-five items from each of the 1977-1978 Minimum Basic Skills tests were embedded into the 1978-1979 field test. This figure was approximately 22 to 27 percent of the total items on the original form. The number of items was selected from each of the cluster areas of the test in the same proportion in which they appeared on the original form. Other item characteristics, such as, p-values and biserials were considered when making the selection, in order to represent as accurately as possible, the parameters of the original form. Since items were selected solely for the purpose of traditional equating, no Rasch item statistics were used to select the pool. The resulting twenty-five item sections were truly "Minimum Basic Skills mini tests" in that they mirrored very closely the original form in both content and level of difficulty.

For the purpose of the equating, this section of items was placed at the end of the test. While in retrospect this was probably not the best place for these items due to such factors as fatigue, boredom, etc., operational simplicity of the field test took precedence in this first equating of the Minimum Basic Skills tests.

Description of the Samples

Three separate samples of test data were used to equate

each of the eight Minimum Basic Skills tests. They were:

1. 1978-1979 FIELD TEST SAMPLE

   This sample consisted of students in the fourth, seventh, tenth and twelfth grades who took one of the tests, either reading or mathematics. Approximately 400 to 500 students took each of the eight tests. A systematic sample of twenty-five students was selected by each participating school. Schools were asked to participate on a volunteer basis, however, they did have to fit into a pre-arranged stratified sampling matrix of geographic region and socioeconomic status. The sampling matrix was designed to provide a truly representative sample of the state.

2. 1977-1978 FINAL FORM SAMPLE

   This sample consisted of a two percent systematic sample generated by computer from the total 400,000 students who took the Minimum Basic Skills tests during the spring of 1978.

3. 1978-1979 FINAL FORM SAMPLE

   This sample consisted of approximately 1,500 to 2,100 students who took the Minimum Basic Skills tests in the spring of 1979. This sample was developed by randomly taking a selection of the earliest returns from the districts after the spring administration. From past experience, these districts provided a good indication of the performance of the state.

Rasch Common Item Equating

Wright (1977) presents a methodology for equating two tests by the use of common items. A pair of separate and independent estimates of difficulty are produced for each item that is common to a pair of tests. When these items are calibrated, the origin is normally set by fixing the average item difficulty to zero, which coincidentally, fixes the origin of the ability scale. According to the Rasch model, these common items should have the same average difficulty for both tests. Any difference in average difficulty of the common item set between the two tests

indicates a difference in scale origins of the two tests. This difference in average item difficulty of the common items can be used as an equating constant to adjust the scale of either test on to the scale of the other.

The formula for this constant which translates all item difficulties from the calibration of test b on to the scale of the base test a is given by Wright (1977, p. 107).

$$T_{ab} = \sum_{i=1}^{K} (d_{ia} - d_{ib})/K$$

where,

T_{ab} is the equating constant to transform the difficulties of test b onto the scale of test a.

d_{ia} are the common item difficulties on test a.

d_{ib} are the common item difficulties on test b.

K is the number of common items.

To place any item from test b onto the scale of test a one simply adds the constant, i.e., D_{ib} on scale of a = $D_b + T_{ab}$. Likewise, to transform the ability table from the scale of one test onto the scale of the other the addition of the constant is also made.

Procedures

For each of the eight tests the three samples of data outlined above were calibrated separately using a version of BICAL (Wright and Mead, 1977).

Prior to the actual Rasch equating, bivariate plots of the item difficulties for the twenty-five common items between the 1978-1979 field test and the 1977-1978 final form

test were made and analyzed for item stability, as suggested by Rentz (1978) and Beard and Pettie (1979). In addition, since these were tests which were not built using Rasch test development procedures, fit to the model was examined.

The actual equating proceeded in two phases. First, using twenty-five common items, the 1978-1979 field test was put on the scale of the 1977-1978 form, by the addition of a constant to the item difficulties of the field test form. Since there were some changes in items from field to final form, a second adjustment or "fine tuning" was necessary. Using the approximately eighty to ninety common items between the 1978-1979 field test and the 1978-1979 final form, the 1978-1979 final form item difficulties were placed on the scale of the 1977-1978 base form.

In order to derive equivalent raw scores, the ability tables for the 1978-1979 final form were also adjusted by the same constant and put on the scale of the 1977-1978 form. Equivalent raw scores were assigned and compared to those from the linear equating method. Special attention was given to those scores at and around the state standard.

DISCUSSION OF RESULTS

In order to answer the first question posed in the study, viz., could the Rasch model equating methodology be applied to tests which were not originally developed to fit the model, two kinds of data were analyzed:

1. Mean square fit statistics and discrimination indices from the BICAL runs.

2. Graphic displays of the difficulty estimates of the twenty-five common equating items.

Analysis of Item Fit

The mean square fit statistics used in this analysis were the total mean square fit values from the BICAL output. Items with a mean square fit of 1.5 or greater were flagged and considered of questionable fit to the model.

The discrimination index or slope is the regression of "item log odds" on "test log odds." Values should be near one for fitting items. A value less than one indicates that the item characteristic curve for that item is flatter than the test characteristic curve. Values greater than one indicate that the item characteristic curve is steeper than the test characteristic curve. The interval $1.0 \pm .20$ (Cartledge, 1975; Rentz and Bashaw, 1975; Beard and Pettie, 1978) was used as a yardstick by which to evaluate this type of item fit.

Tables 1, 2 and 3 present an analysis of the total mean square fit statistics and discrimination indices from the calibrations of each of the three samples used for the equating. Analyses were made for both total test and common items alone. In the calibration of the 1978-1979 field test sample (Table 1) the percent of items with a mean square fit less than 1.5 ranged from 88 to 96 for the total test and 92 to 100 for the common items. The percent of items whose discrimination index was within the recommended $1.0 \pm .2$ interval ranged from 52 to 75 for the total test and 56 to

84 for the common items.

In Table 2, this same analysis is provided for the calibration of the 1977-1978 final form sample. In this sample, the percent of items with total mean square fits less than 1.5 ranged between 91 to 100 for the total test and between 96 and 100 for the common items. Sixty-one to 87 percent of the items on the total test had slopes within the acceptable range, while 56 to 92 percent of the common items were within this range.

The analysis of the 1978-1979 final form calibration sample, given in Table 3, indicates similar findings. In this sample, between 92 and 100 percent of the items on the total test showed acceptable mean square fit statistics. The breakdowns for the common items were very like those for the total test. Similarly, the percent of items with slopes in the acceptable region was 61 to 79 for both total test and common item pool. For this sample both the total test and the common item pool showed the same proportion of items in the two types of item fit categories, i.e., mean square fit and discrimination index.

In summary, the three analyses above show moderate to good fit to the model. The percent of fitting items vis à vis the mean square fit statistic was high in all samples. On the other hand, while the item slopes showed less conformity to the model, the percent of items in the acceptable range was similar if not better than those reported by Rentz and Bashaw (1975) in their Rasch analysis of the Anchor Test

Study data. Rentz and Rentz (1978) state that the inolusion of a few non-fitting items would probably do no serious damage in test equating applications.

Common Item Stability

Figures one through eight plot the twenty-five common item difficulty estimates from the calibration of the 1978-1979 field test against those from the calibration of the 1977-1978 final form for the four reading and four mathematics tests. According to the model, the estimates in each pair should be statistically equivalent, except for a single constant of translation that is the same for [redacted] tems. The difficulty estimates in these types of plots should be located along a unit slope straight line. The intercept of this line is equal to the average difference in item difficulty estimates (equating constant) between these two calibrations. These plots provide an indication of the relative stability of the common item pool. Certainly when equating one wants a set of items whose calibrations are stable over time to be able to use them to adjust the difficulties of the other items to a particular scale. What one looks for then in these plots are the outliers, i.e., the items which fall away from the unit slope line.

Examining the plots in figures one through eight, there were a few items which seemed to fall further from the unit slope line then the others. (They have been blackened in.) In one test, sixth grade reading (figure 3), fit of the items to the line was very good and no outliers were detected.

The worst case was evident in third grade mathematics (figure 2), where the items seemed to fall all over the place and seven outliers were observed.

In order to try to determine the reasons for these items' seeming instability, their fit to the model was analyzed. However, there appeared to be no consistent pattern of explanation. On some tests, every outlier showed acceptable fit to the model (Math 6 and Read 3) using both mean square fit and slope statistics. In others, a few of the outlier items, but not all, had steeper slopes than the model predicts (discrimination indices greater than 1.2). In some of the tests, the outliers had slopes which were flatter (discrimination indices less than .8). In most cases, the outliers showed these aberrant slope statistics for the field test calibration sample only. The field test calibration samples were generally much smaller than the final form samples; however, it is unknown whether their size affected them. For only two out of the twenty-five items detected as outliers (out of 200 common items), one item in Math 9 and one item in Math 11 had both mean square fit and discrimination indices out of the range of acceptable fit to the model.

In one test, Math 11, there was also detected a strange clump of common items in the lower left corner of the plot (figure 8). However, on closer inspection of both content and format of the items, no clue was found as to the meaning of this clump.

In summary, while most of the common items seemed to uphold the model's predictions, there were still a few unstable items. Recommendations have been made to delete these items from testing applications (Rentz, 1978), as the effect of a small difference in the computation of the equating constant could mean a shift in the resultant equating of several raw score points. However, as there are really no objective rules for deleting these types of items, i.e., how "unstable" is unstable, and because all of the items in the common item pool were used in the linear equating, and one of the purposes of this paper was to compare these two methods, the equating was carried out using all the common items.

Equating Results

Table 4 presents a summary of the statistical characteristics of the twenty-five common item sets for each of the eight tests. In most cases, performance on these items during the previous spring administration was better than on the following fall field test, although the differences are quite small. This may be attributed to the fact that these twenty-five items were all placed at the end of the field test form and such factors as fatigue, boredom and lack of motivation may have interferred with the students' performance on the field test.

Table 5 presents mean item difficulties standard deviations and equating constants for the twenty-five items common to the 1978-1979 field test and the 1977-1978 final

form. The scale of the 1977-1978 form was chosen as the base form, as the state standards were empirically developed from this scale. The equating constants were formed by subtracting the field test mean item difficulty from the final form mean item difficulty. The equating constants are the values which are added to the 1978 field test difficulties to put all the items on the field test on to the scale of the base test.

Table 6 presents mean item difficulties, standard deviations and equating constants for the common items between the 1978-1979 field test, adjusted to the scale of the 1977-1978 test, and the 1978-1979 final form. This second "fine tuning" adjustment was necessary because of the changes in items from field test to final form in the 1978-1979 test. The equating constants in this table were used to adjust both the item difficulties and ability estimates from the 1978-1979 final form to the scale of the 1977-1978 form.

Assignment of Raw Scores

The last step in the equating process is the development of a table of equivalent raw or scaled scores. Equivalent scores are defined by the Rasch model as those which give rise to the same "log" ability estimates (Rentz and Bashaw, 1975). For the present analysis, adjusted log ability estimates from the 1978-1979 final from were matched with those from the 1977-1978 final form and equivalent raw scores were assigned for each of the eight tests.

Table 7 presents the equivalent raw scores derived from both the linear[2] and Rasch methods of equating. Because of the nature of these tests, i.e., minimum competency, the tables are presented for those raw score points at and around the state cut-off score. It is in this range of the distribution where the results of the equating are most crucial. While a difference in one or two raw score points at the extremes of the distribution may not have any practical consequences, at the cut-score this could mean a major difference in the numbers of students who pass the test.

The equivalent raw scores derived from the linear method and the Rasch method using all twenty-five common items (non-edited) show considerable similarity. In most cases the equivalent raw scores given by the two methods differ by only one raw score point. In eleventh grade reading, the two sets of equivalent raw scores are identical.

In the Rasch model, the resultant equivalent raw scores depend very heavily on the accuracy of the estimated constant. While the results obtained using all twenty-five items were good, it was decided to investigate the actual impact the editing of unstable items might have on the resultant Rasch equivalent raw scores. Therefore, the equating was reworked with the "unstable" items (represented by black dots on figures one through eight) eliminated.[3]

---

[2]These were derived from analyses prepared by the contractor.
[3]Tables providing the constants for the equating based on the edited item pool are not included in this paper. They are available upon request from the author.

Table 7 presents equivalent raw scores from this equating based on the edited initial item pool.

Table 8 provides a comparison of the two sets of Rasch equivalent raw scores. Since no items were deleted from the sixth grade reading test, only the results from seven tests were compared. In six out of the seven tests, the equivalent raw scores at the state's cut-off score were different. In only two of these six tests were the raw scores from the equating based on the edited item pools closer to those given by the linear method. Overall, the non-edited item pool generally gave the closest results to the linear method. In one test, eleventh grade reading, no differences were observed between the equivalent scores arising from the equating based on the edited and non-edited item pools. Coincidently, this was the same test where the Rasch and linear equivalent raw scores were identical.

In summary, the results from this comparison of the two methods of equating, viz., linear and Rasch indicate:

a) There is a good, though not perfect, match between the raw scores derived from the linear and Rasch methodologies;

b) The editing of a common item pool for "unstable" items changed the resultant assigned equivalent raw scores. However, no findings indicated that the editing provided a "better" table of raw scores. Unfortunately, the only criterion of "better" that was used in this application was the match with the raw scores defined by the linear method.

Further research is needed which uses independent criteria by which to evaluate these two methods.

REFERENCES

Angoff, W. H. Scales, Norms and Equivalent Scores. In R. L. Thorndike (ed.). Educational Measurement. Washington: American Council on Education, 1971.

Beard, J. and Pettie, A. A Comparison of Linear and Rasch Equating Results for Basic Skills Assessment Tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, 1979.

Cartledge, C. M. A Comparison of Equipercentile and Rasch Equating Methodologies. Unpublished doctoral dissertation, University of Georgia, 1974.

Rasch, G. An Item Analysis that takes Individual Differences into Account. British Journal of Mathematical and Statistical Psychology. 1966, 19, 49-57.

Rentz, R. and Bashaw, W. L. Equating Reading Tests with the Rasch Model, Volume I, Final Report. Athens, GA: University of Georgia, Educational Research Laboratory, 1975.

Rentz, R. Monitoring the Quality of an Item-Pool Calibrated by the Rasch Model. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Ontario, 1978.

Rentz, R. and Rentz, C. Does the Rasch Model really Work? A discussion for practitioners. Princeton: ERIC Clearinghouse on Tests, Measurement and Evaluation, No. 67, 1978.

Swineford, F. Test Analysis: New Jersey Educational Assessment Program, Minimum Basic Skills Test, 3BNJ. Princeton: Educational Testing Service, SR-79-55, June 1979.

Wright, B. D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D. and Mead, R. J. BICAL: Calibrating Items and Scales with the Rasch Model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1977.

Table 1

1978-1978 Field Test Fit and Slope Statistics
(Size of calibration samples ranged from 401 to 577)

Test	Mean Square Fit[1]	Slope[2]	Total Number of Items
Read 3	88	69	100
Read 3	*100*	*84*	*25*
Math 3	91	71	100
Math 3	*92*	*60*	*25*
Read 6	95	52	95
Read 6	*100*	*56*	*25*
Math 6	95	62	100
Math 6	*96*	*60*	*25*
Read 9	94	67	110
Read 9	*100*	*72*	*25*
Math 9	96	63	95
Math 9	*100*	*56*	*25*
Read 11	89	75	110
Read 11	*96*	*68*	*25*
Math 11	90	71	90
Math 11	*96*	*72*	*25*

[1]Percent of items with total MSF < 1.5
[2]Percent of items with discrimination indices within the interval (.8-1.2)
[3]Values in italics are for the twenty-five common equating items

Table 2

1977-1978 Final Form Fit and Slope Statistics
(Size of calibration samples ranged from 1757 to 2133)

Test	Mean Square Fit[1]	Slope[2]	Total Number of Items[3]
Read 3	95	83	75
Read 3	*100*	*92*	*25*
Math 3	95	84	75
Math 3	*100*	*92*	*25*
Read 6	100	66	70
Read 6	*100*	*72*	*25*
Math 6	96	61	75
Math 6	*100*	*64*	*25*
Read 9	95	74	85
Read 9	*100*	*80*	*25*
Math 9	91	73	70
Math 9	*100*	*56*	*25*
Read 11	98	87	85
Read 11	*96*	*80*	*25*
Math 11	92	76	55
Math 11	*96*	*88*	*25*

[1]Percent of items with total MSF < 1.5
[2]Percent of items with discrimination indices within the interval (.8-1.2)
[3]Values in italics are for the twenty-five common equating items

Table 3

1978-1979 Final Form Fit and Slope Statistics
(Size of calibration samples ranged from 1483 to 2110)

Test	Mean Square Fit[1]	Slope[2]	Total Number of Items[3]
Read 3 *Read 3*	92 *93*	75 *76*	100 *88*
Math 3 *Math 3*	93 *93*	78 *78*	100 *98*
Read 6 *Read 6*	97 *98*	62 *61*	95 *83*
Math 6 *Math 6*	98 *93*	66 *66*	100 *94*
Read 9 *Read 9*	100 *100*	62 *61*	110 *74*
Math 9 *Math 9*	98 *98*	64 *61*	95 *85*
Read 11 *Read 11*	97 *96*	76 *79*	110 *82*
Math 11 *Math 11*	93 *93*	66 *63*	90 *76*

[1]Percent of items with total MSF < 1.5
[2]Percent of items with discrimination indices within the interval (.8-1.2)
[3]Values in italics are for the common items between the 1978-1979 Field Test and the 1978-1979 Final Form

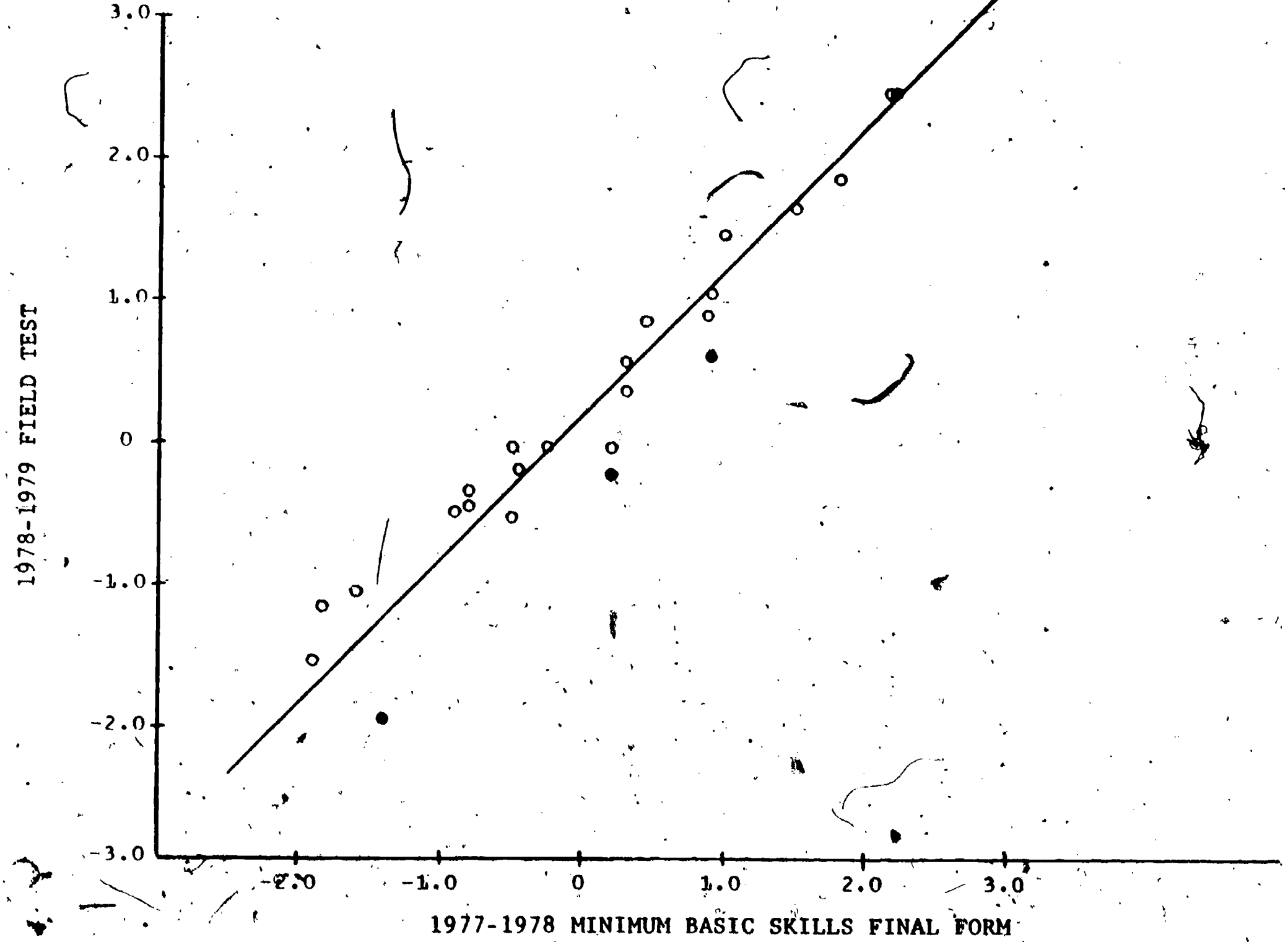

FIG. 1: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the third grade reading test.

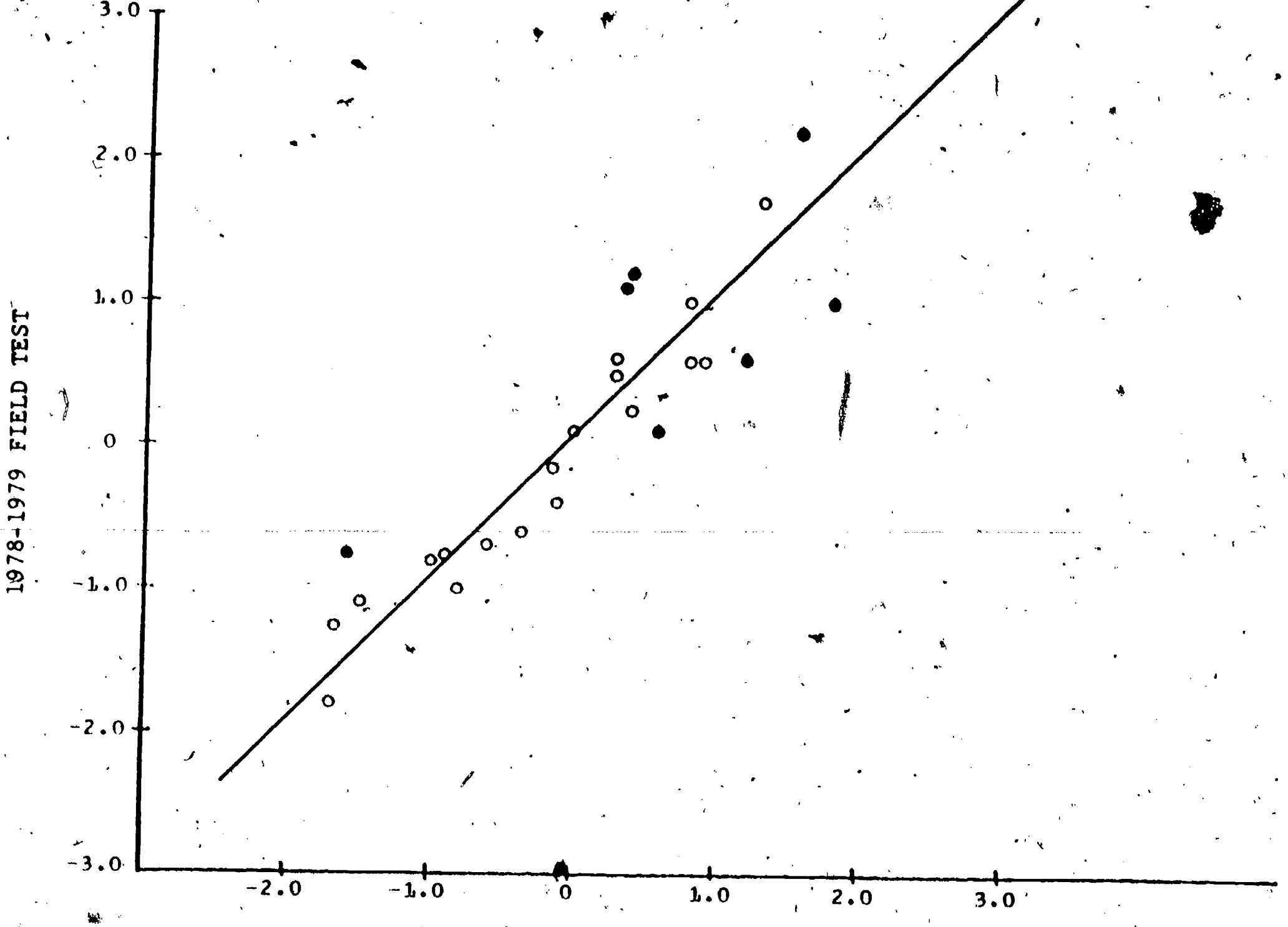

FIG. 2: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the third grade mathematics test.

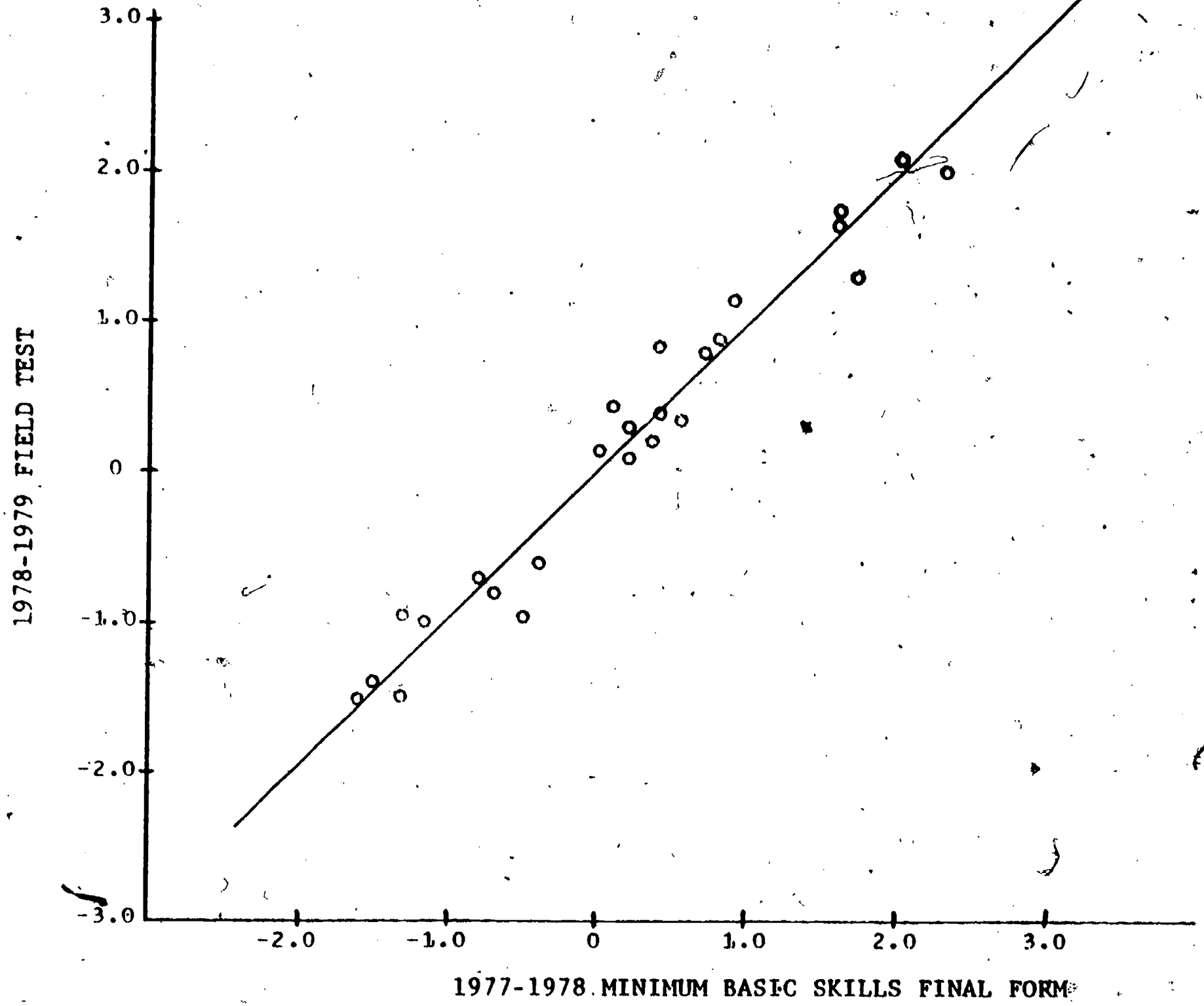

FIG. 3: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the sixth grade reading test.

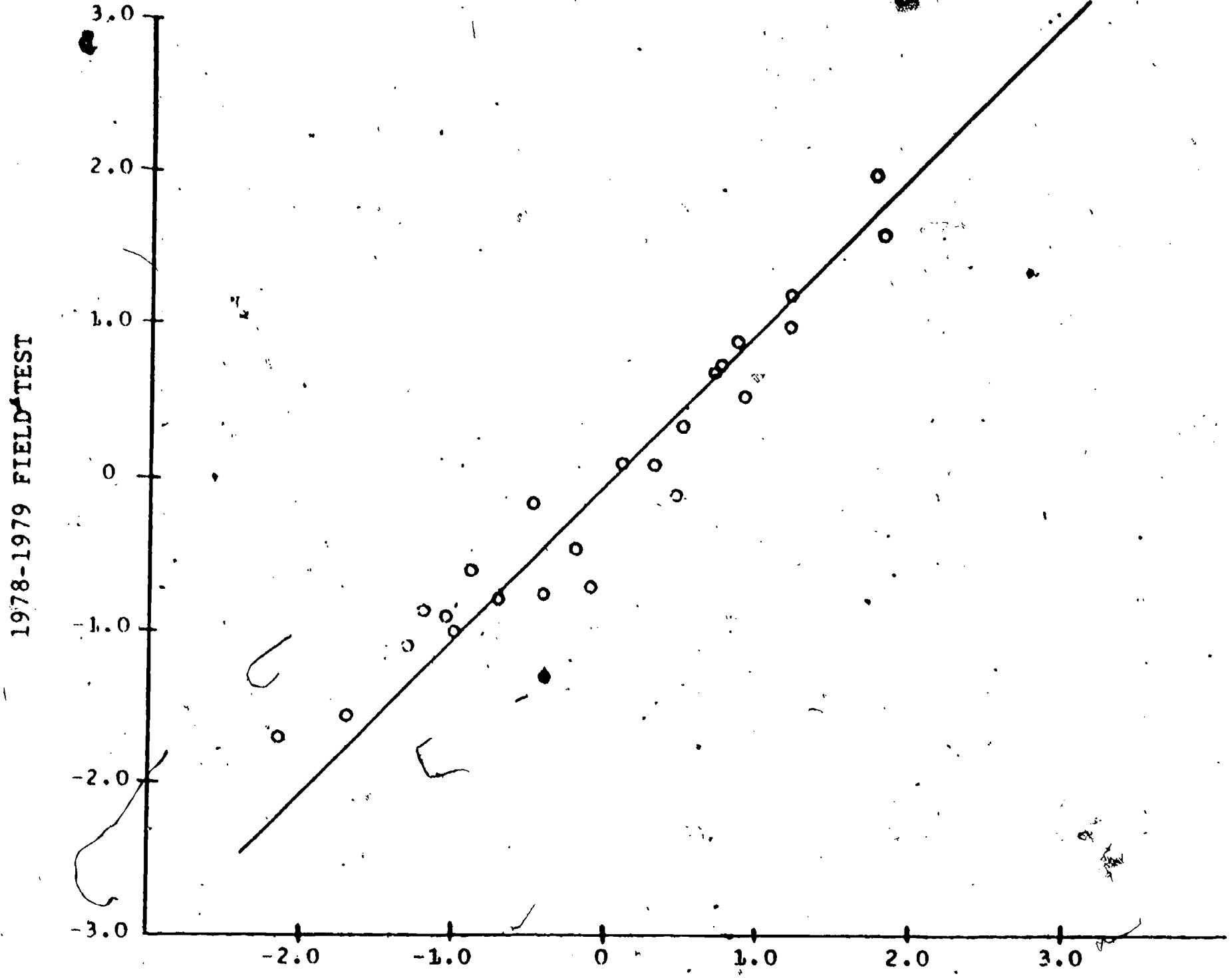

FIG. 4: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the sixth grade mathematics test.

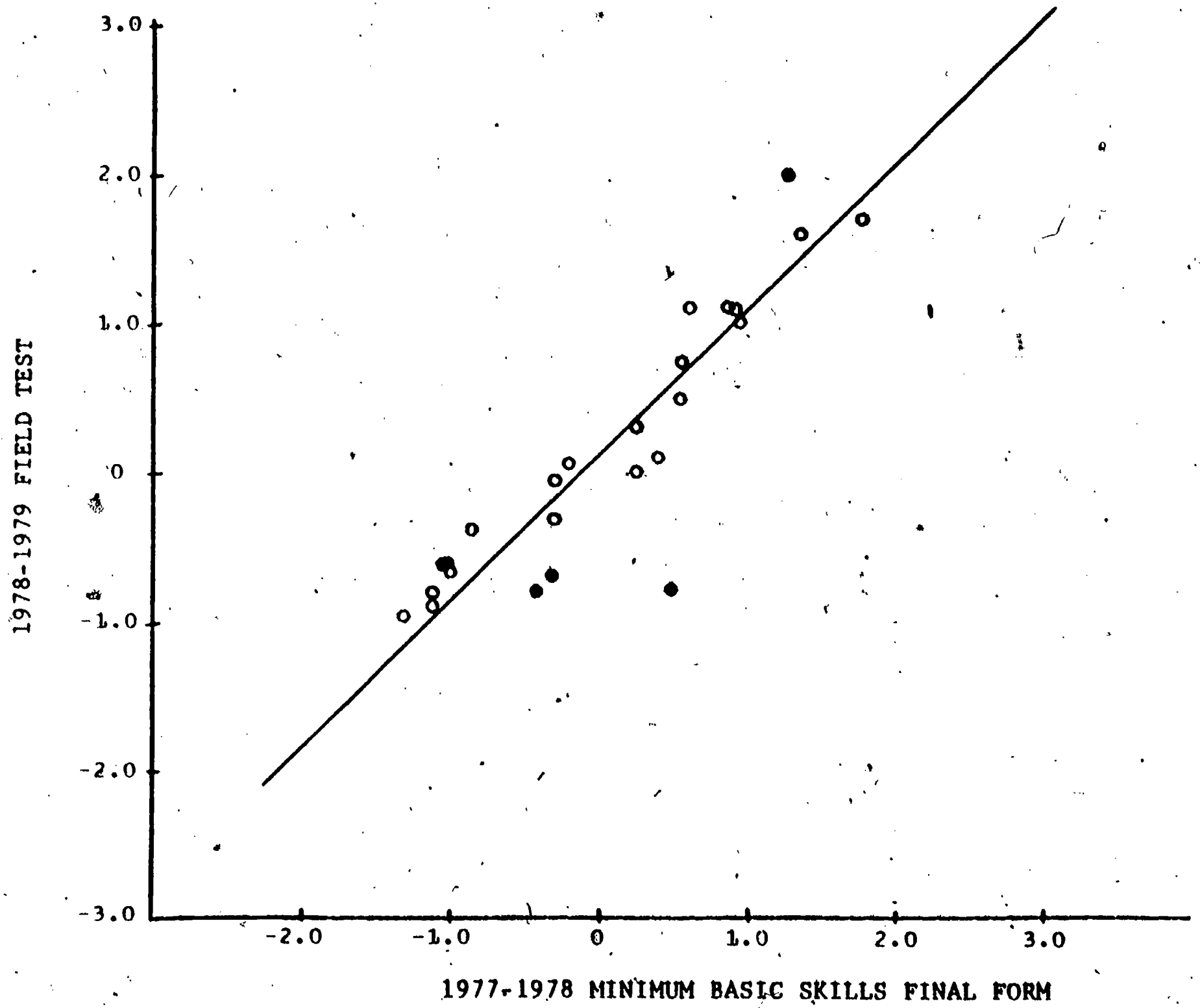

Figure 5: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the ninth grade reading test.

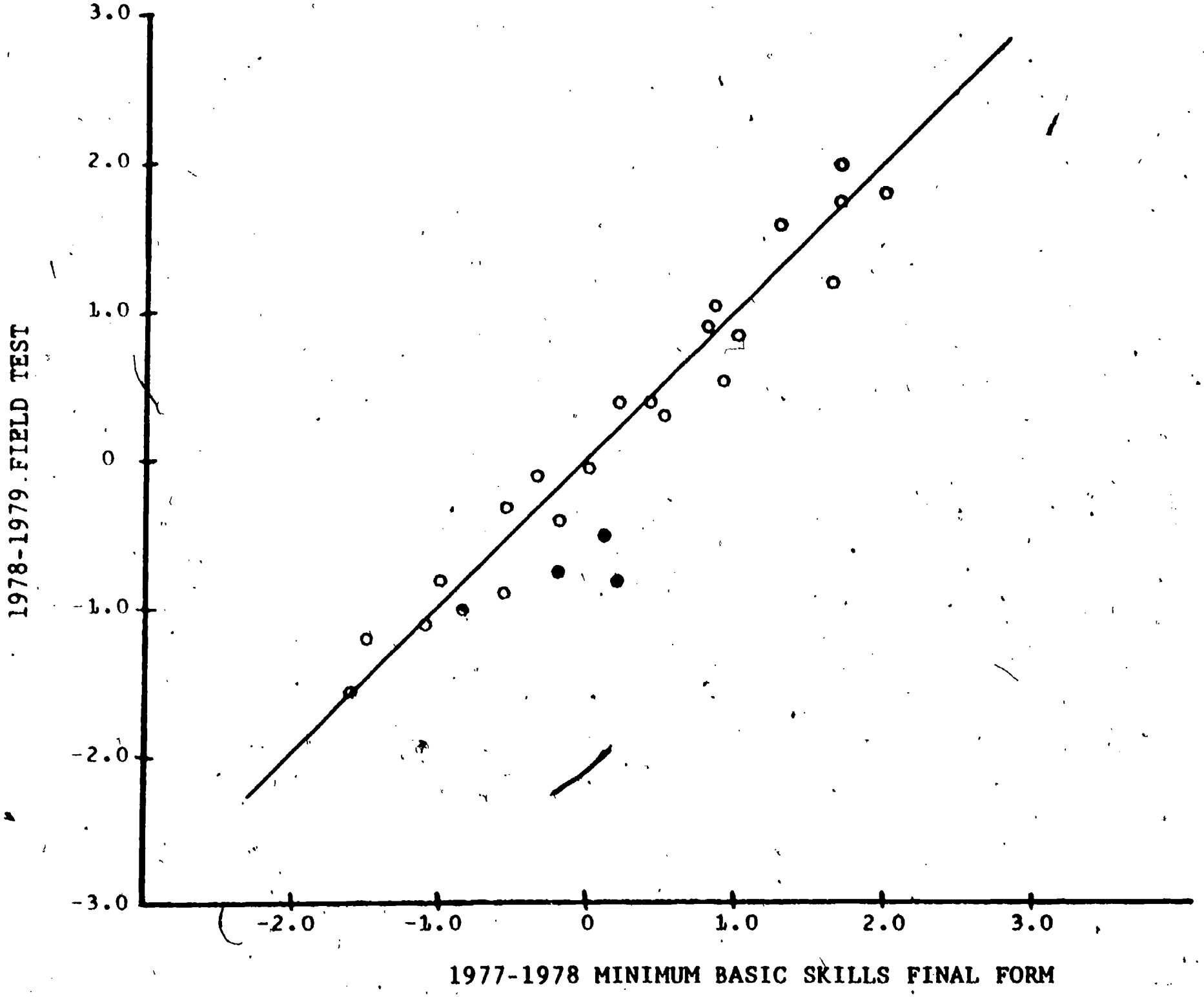

FIG. 6: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the ninth grade mathematics test.

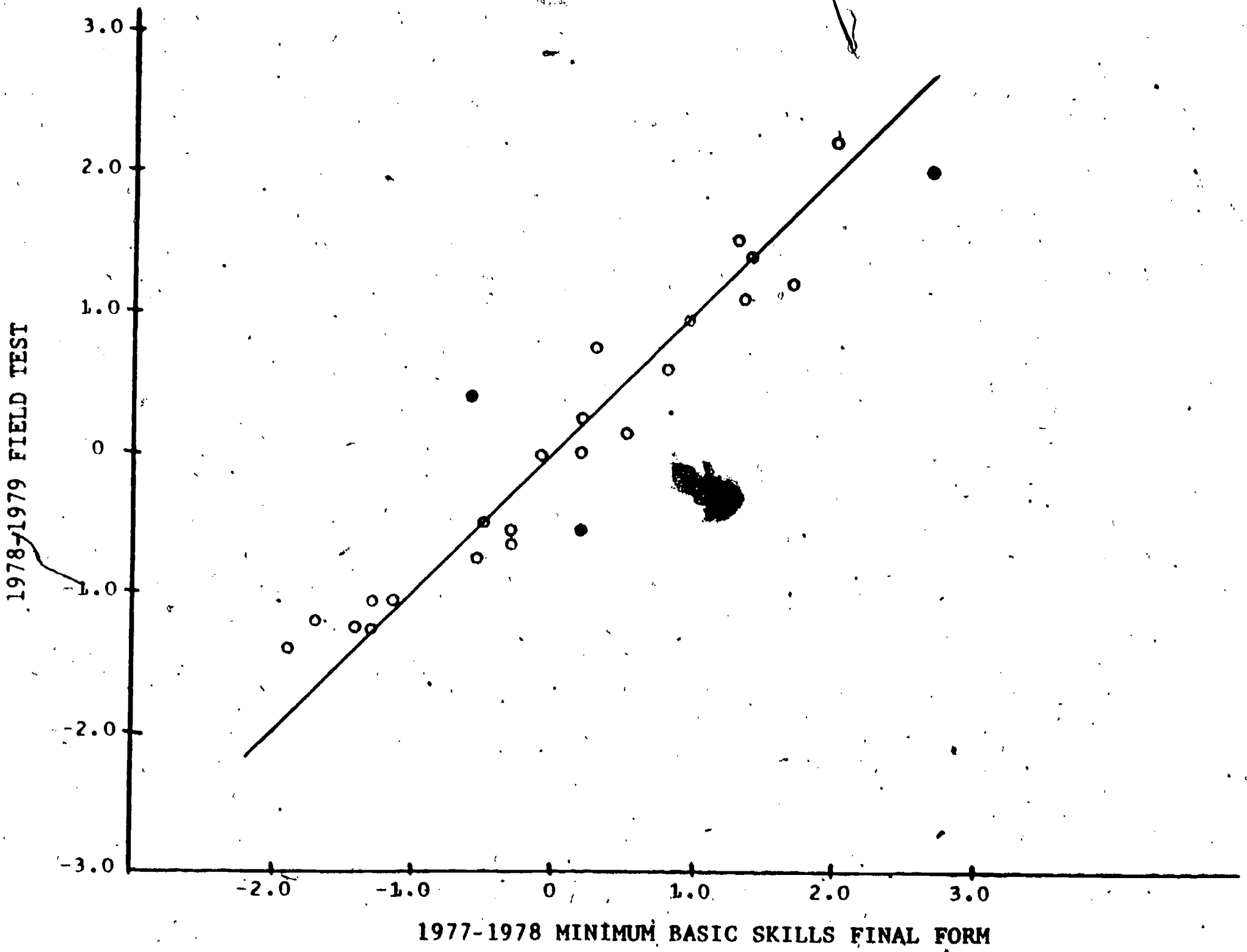

FIG. 7: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the eleventh grade reading test.

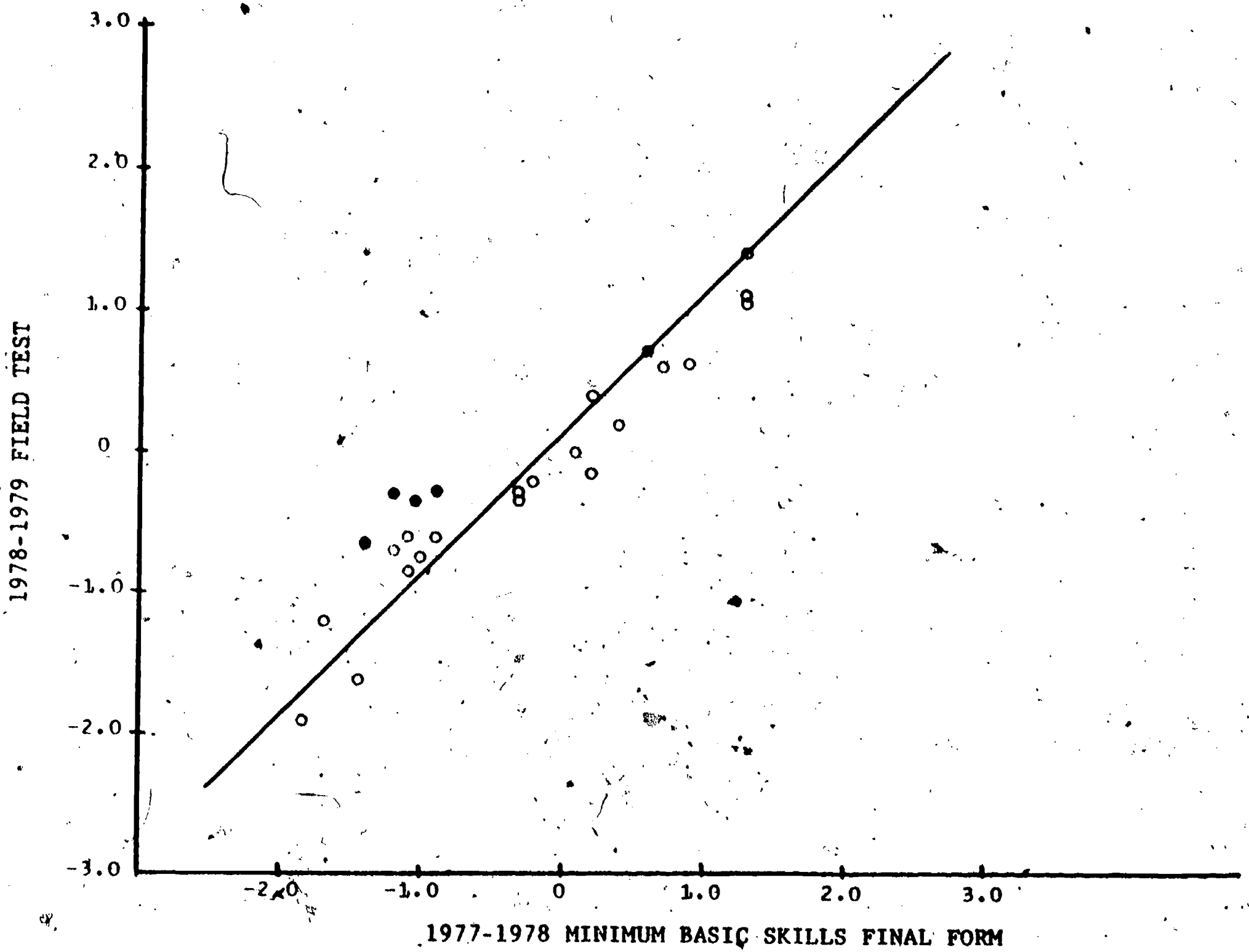

FIG. 8: Plot of the item difficulty estimates for the twenty-five common items used to equate the two forms of the eleventh grade mathematics test.

Table 4

Statistical Characteristics of the Twenty-Five Common Items

Test	Sample	N	Mean	Variance	Standard Deviation	Mean P
Read 3	F	577	21.67	13.82	3.71	.866
Read 3	A	1757	21.74	12.76	3.57	.868
Math 3	F	401	18.82	27.17	5.21	.755
Math 3	A	1782	18.97	22.98	4.79	.759
Read 6	F	405	19.61	16.21	4.02	.784
Read 6	A	1906	20.12	14.83	3.85	.804
Math 6	F	446	17.57	24.42	4.94	.704
Math 6	A	1899	18.29	22.22	4.71	.667
Read 9	F	408	20.87	18.74	4.32	.834
Read 9	A	2137	20.84	16.64	4.07	.834
Math 9	F	460	17.32	26.74	5.17	.692
Math 9	A	2133	18.44	22.39	4.73	.735
Read 11	F	454	21.89	12.71	3.56	.876
Read 11	A	1881	22.21	10.03	3.16	.886
Math 11	F	406	20.98	21.06	4.58	.838
Math 11	A	1894	21.30	13.52	3.67	.848

Legend

F - 1978 Field Test

A - 1977-1978 Final Form

Table 5

Rasch Item Difficulty Estimates for the Twenty-Five Common Items between the 1978-1979 Field Test and the 1977-1978 Final Form

Test	Sample	Mean Item Difficulty	Standard Deviation	Equating Constant
Read 3	F	.173	1.16	.131
Read 3	A	.042	1.19	
Math 3	F	.084	.99	-.068
Math 3	A	.016	1.02	
Read 6	F	.204	1.12	-.017
Read 6	A	.187	1.12	
Math 6	F	-.113	1.01	.059
Math 6	A	-.054	1.07	
Read 9	F	.153	.91	-.104
Read 9	A	.049	.88	
Math 9	F	.170	1.04	.050
Math 9	A	.220	1.03	
Read 11	F	.091	1.09	.003
Read 11	A	.094	1.22	
Math 11	F	-.189	.82	-.123
Math 11	A	-.312	.98	

Legend

F - 1978-1979 Field Test

A - 1977-1978 Final Form

Table 6

Rasch Item Difficulty Estimates for the Common Items between the 1978-1979 Field Test and 1978-1979 Final Form

Test	Sample	Mean Item Difficulty	Standard Deviation	Equating Constant	Number of Items
Read 3	A	-.083	1.24	-.208	88
Read 3	AF	-.291	1.28		88
Math 3	A	.016	1.05	-.089	98
Math 3	AF	-.073	1.03		98
Read 6	A	-.039	1.30	-.096	83
Read 6	AF	-.135	1.34		83
Math 6	A	-.090	.97	.040	94
Math 6	AF	-.050	1.03		94
Read 9	A	-.002	1.03	-.178	74
Read 9	AF	-.180	1.00		74
Math 9	A	-.034	1.19	-.036	85
Math 9	AF	-.070	1.15		85
Read 11	A	.048	1.17	-.020	82
Read 11	AF	-.068	1.19		82
Math 11	A	-.051	1.15	-.141	76
Math 11	AF	-.192	1.12		76

Legend

AF - Adjusted 1978-1979 Field Test

A - 1978-1979 Final Form

Table 7

A Comparison of the Results of
Linear and Rasch Equating for Selected Raw Score Intervals
Near the State Cut-Off Scores

THIRD GRADE READING			
Raw Score	Linear	Rasch[1]	
		Non-Edited	Edited
80	83	83	84
79	82	82	83
78	81	81	82
77	80	80	81
76	79	79	80
*75	79	78	79
74	78	77	78
73	77	76	77
72	76	75	76
71	75	74	75
70	74	73	75

THIRD GRADE MATHEMATICS			
Raw Score	Linear	Rasch[1]	
		Non-Edited	Edited
70	72	72	71
69	71	71	70
68	70	70	69
67	70	69	68
66	69	68	67
*65	68	67	66
64	67	66	65
63	66	65	64
62	65	64	63
61	64	63	62
60	63	62	61

[1]The results of the Rasch equating are being reported for both the non-edited and edited item pools.

*Denotes the cut-score on the 1977-78 form of the test.

Table 7 (Cont.)

SIXTH GRADE READING			
Raw Score	Linear	Rasch[2]	
		Non-Edited	Edited
77	78	77	
76	77	76	
75	76	75	
74	75	74	
73	75	73	
*72	74	72-73	
71	73	72	
70	72	71	
69	71	70	
68	70-71	69	
67	70	68	

SIXTH GRADE MATHEMATICS			
Raw Score	Linear	Rasch[1]	
		Non-Edited	Edited
70	69	69	70
69	68	69	69
68	67	68	68
67	66	67	67
66	65	65-66	66
*65	64	64-65	65
64	63	63	64
63	62	62	63
62	61	61	62
61	60	60	61
60	59	59	60

[1]The results of the Rasch equating are being reported for both the non-edited and edited item pools.

[2]There were no items edited for this test.

*Denotes the cut-score on the 1977-78 form of the test.

Table 7 (Cont.)

NINTH GRADE READING			
Raw Score	Linear	Rasch[1]	
		Non-Edited	Edited
88	91	90	92
87	90	90	91
86	90	88-89	90
85	89	88	89
84	88	87	88
*83	87	86	87
82	86	85	86
81	85	84	85
80	84	83	84
79	84	82	84
78	83	81	83

NINTH GRADE MATHEMATICS			
Raw Score	Linear	Rasch[1]	
		Non-Edited	Edited
67	67	68	69
66	66	67	68
65	65	66	67
64	64	65	66
63	63	64	65
*62	62	63	64
61	61	62	63
60	60	61	62
59	59	60	61
58	58	59	60
57	57	58	59

[1]The results of the Rasch equating are being reported for both the non-edited and edited item pools.

*Denotes the cut-score on the 1977-78 form of the test.

Table 7 (Cont.)

ELEVENTH GRADE READING			
Raw Score	Linear	Rasch[1]	
		Non-Edited	Edited
88	89	89	89
87	88	88	88
86	87	87	87
85	86	86	86
84	85	85	85
*83	84	84	84
82	83	83	83
81	82	82	82
80	81	81	81
79	80	80	80
78	79	79	79

ELEVENTH GRADE MATHEMATICS			
Raw Score	Linear	Rasch[1]	
		Non-Edited	Edited
64	67	67	66
63	66	66	65
62	65	65	64
61	64	64	63
60	63	63	62
*59	63	62	61
58	62	61	60
57	61	60	59
56	60	59	58
55	59	58	57
54	59	57	56

[1]The results of the Rasch equating are being reported for both the non-edited and edited item pools.

*Denotes the cut-score on the 1977-78 form of the test.

Table 8

A Comparison of the Rasch Equating Constants from the Edited and Non-Edited Twenty-Five Common Item Pools

	Read 3	Math 3	Read 6[1]	Math 6	Read 9	Math 9	Read 11	Math 11
Non-edited item pool constant	-.131	-.068	-.017	.059	-.104	.050	.003	-.123
Edited item pool constant	-.215	-.035		.024	-.192	.004	-.016	-.047
Difference in constant	.084	-.033		.035	.088	.046	.019	-.076
Approximate differences in log ability values between raw score points at cut-score	.06-.07	.05-.06		.06	.06	.06	.06-.07	.06-.07
Does the Rasch equating with the edited item pool change the equivalent raw scores at cut-point?	Yes	Yes		Slightly	Yes	Yes	No	Yes
Is this change closer to the value given by the linear results?	Yes	No		No	Yes	No	N/A	No
Which Rasch equating is closer to the linear results?	Edited	Non-Edited		Non-Edited	Edited	Non-Edited	Both Same	Non-Edited

[1]No items were edited from sixth grade reading.