DOCUMENT RESUME

ED 228 315 TM 830 241

AUTHOR Rubin, Lois S.; Mott, David E. W.
TITLE Comparison of Three Different Item Banking Methods for Longitudinal Test Equating.
PUB DATE Apr 83
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Difficulty Level; *Equated Scores; Grade 10; High Schools; *Item Banks; *Latent Trait Theory; Longitudinal Studies; Maintenance; Methods Research; Reading Tests; Test Construction; Test Items; *Test Theory
IDENTIFIERS *Linkage Analysis; One Parameter Model; *Rasch Model; Virginia

ABSTRACT

Three different item-bank maintenance procedures were compared for the purpose of investigating their cumulative effects on a series of longitudinal equatings of tests measuring the same characteristic. Item data used resulted from Rasch one-parameter latent trait test calibrations. Comparison of the results in the various final banks were made. These included comparing mean item difficulty values, ability scores, and the number of correctly answered items necessary for passing the tests. These comparisons were made on the following data: a test comprised of the original 54 items and their difficulty values both in the initial bank and in the final banks, and, also, the difficulty values of items used as a test in the final administration of this study. No significant differences were found in the cumulative effects resulting from the various procedures. (Author)

Comparison of Three Different Item Banking Methods for Longitudinal Test Equating

Lois S. Rubin
David E. W. Mott

Virginia Department of Education

Presented at the Annual Meeting of the American Educational Research Association in Montreal, April 1983.

COMPARISON OF THREE DIFFERENT ITEM BANKING METHODS FOR LONGITUDINAL TEST EQUATING

Introduction

Classical test theory has been found to have a number of deficiencies (Hamilton et al., 1978). One such shortcoming is in the area of item difficulty and item discrimination statistics. In classical test theory these statistics are not invariant across groups of examinees of different ability. That is, the item statistics gathered are useful only for populations similar to those on which the statistics were gathered. In addition, comparisons of examinees on an ability trait that can be measured by a test are limited to situations where the tests used are either the same or parallel. Also, examinee performance on a test item can not be predicted; and testing problems such as test design, item bias, and test equating found no adequate solutions.

In contrast to classical test theory, the Rasch one-parameter latent-trait model has item parameters which are said to be sample invariant (Wright, 1977). Because item parameters from different calibrations are linearly related, they can be put onto a single common scale and "the measures implied by scores on all such tests are automatically equated and no further collection or analysis of data is needed" (Wright, 1977, p. 106). Moreover, because all the items are on a common scale, they can be used to make up new tests which would be equated on the common scale.

However, with each equating there is a standard error of equating which, although small in comparison to the standard error of measurement (Angoff, 1971; Bryman, 1976), is transmitted to every score. The cumulative effect of this error can become quite large in a longitudinal testing program. In addition, when equating tests using items calibrated by means of the Rasch latent-trait model, the items used as the common core in the linking process will have two difficulty parameter values--one from the previous testing (its value in the item bank prior to testing) and a new value from the calibration on the form to be equated to the existing bank.

The procedure that should be utilized in maintaining an item bank has been the subject of some controversy. Researchers from 1978 to the present disagree as to which difficulty calibrations should be in the bank--the original, the most recent, or some combination. The reason given to keep the original calibrations in the bank (Mead, 1981) is that these calibrations act as an anchor. On the opposite side, the argument presented (Rentz, 1978; Cook, Eignor, Peterson, 1982) is that if changes in learning/teaching have occurred, the item calibrations should reflect them. Thus, this viewpoint advocates that item difficulty calibrations should be updated after each use. Ridenour (1980) has recently argued that one should consider as much information as possible when maintaining the underlying scale in an item bank. One should not follow a set routine, always either using the original item difficulty or updating the calibration to the most recent value, but should consider all information and choose the most appropriate value.

The purpose of this study is to investigate the cumulative effects of three different item-bank maintenance procedures on a series of longitudinal equatings of tests measuring the same characteristic. Many testing programs are run on the basis of future tests being equal in difficulty to the first test administered. Thus, it is important that the item-bank maintenance procedure chosen not cause the difficulty calibrations of the items to be artificially inflated or deflated resulting in succeeding tests being, in reality, easier or harder than the first.

Methodology

Data. During the fall of 1978, the Commonwealth of Virginia began its Minimum Competency Testing Program with the administration of statewide reading and mathematics tests to all tenth-grade students. To date, there have been twelve administrations of these tests either using new equated editions or, in two cases, reusing earlier editions. The data for this study came from selected administrations of the reading test. The one-parameter latent-trait model (Rasch) has been used throughout the program to obtain the item calibrations. BICAL-II (Wright, Mead, Bell, 1977) was employed for the first administration and BICAL-III (Wright, Mead, Bell, 1979) for all subsequent administrations.

The calibrations on the Fall 1978 test (0001*) were made on a 10,000 student random sample drawn from a tested population of approximately 80,000 students. These sixty items and their difficulty value calibrations on this administration constitute the initial item bank.

The subsequent administrations of the test chosen for the study are the ones in which these sixty items appear again. For the purpose of monitoring their calibrations, these items have been placed on operational tests as experimental items. They were spread over seven forms on the March 1980 administration (0003), over five forms on the March 1981 administration (0005), and over four forms on the March 1982 administration (0008). Using the operational items as the common core, a number of test forms were used during the March administrations to monitor the difficulty values of some older items and to gather statistical information on newly acquired items. These forms are packaged in sequential order, with packages beginning with as many different form numbers as forms being administered. The tests are given to the students in the order in which they appear in the package. There were eight test forms in the March 1980 administration and twenty test forms in each of the March 1981 and 1982 administrations. The calibrations on these three administrations (0003, 0005, 0008) were on a 10,000 student random sample drawn

* Administration Number

from a tested population of approximately 19,000 on the March 1980 administration and approximately 80,000 tenth-grade students on the March 1981 and March 1982 administrations. The entire original, Fall 1978, test was reused for the October 1981 administration (0007). The calibrations for this administration were on the entire population, approximately 3,300, made up of twelfth-graders who either had failed the test previously or were transfer students.

In addition to multiple calibrations of the items on the Fall 1978 test, the sixty items that were used as operational items on the March 1980 test have been used on March 1981 and 1982 tests, a few as operational items but the majority as experimental items, also for the purpose of monitoring their calibrations.

Procedure. The sixty items from the first administration of the reading test (0001), and the BICAL-II calibrations of these items constituted the initial bank in this study. The items on subsequent administrations (0003, 0005, 0007, 0008) were linked to the bank in the order of their administration. Items included in all linking procedures were those whose difficulty value calibration on the form to be equated was not significantly different from the difficulty value in the bank as tested by the evaluation statistic given in Best Test Design (Wright and Stone, 1979). A computer program (BLINK), developed in-house, has been used to link the items to the existing bank. This program is based on the linking procedures for a complete web discussed in the chapter "Constructing a Variable" in Best Test Design (Wright and Stone, 1979).

The item-bank maintenance procedures used in this investigation are the following: In the first procedure (Method I), the difficulty values of the items previously in the bank were updated to the new values after each test administration wherein these items appeared. In the second procedure (Method II), item bank difficulty values were not changed after use on tests, but those items whose difficulty values were sufficiently modified that they were not included on the link were permanently removed from the bank.

In the third procedure, item bank difficulty values were not altered after use in tests if these items were used on the link. If the items were not employed in calculating the linking constant, because there was a significant change between the difficulty value in the bank and the one on the test to be equated, a decision concerning which difficulty value would be put into the bank was then made. In Method IIIa, (automatic update), the difficulty values of the items taken off the link were automatically updated to difficulty values on the test to be equated and put into the bank. In Method IIIb, (studied/considered update), a determination concerning the difficulty value to be put into the bank was made in the following manner: If this calibration was the second calibration for the item, then the average of the difficulty value in the bank and the present calibration was the value placed in the bank for future use. If the item was taken off the link on the third or subsequent calibration, the difficulty values of the item obtained in each previous calibration was examined. If two or more difficulty values were close, the average of these was used as the new difficulty value placed in the bank. If all calibrations were significantly different from each other, the item was then checked to see if it had been rewritten and, if so, the latest calibration was placed in the bank. If the item had not been modified, the fit statistics in the BICAL calibrations for each administration were inspected. If the item had large fit statistics on one calibration, then the average of the difficulty values for the other two administrations was the value placed in the bank. If the item had large fit statistics on two calibrations, then the difficulty value of the third calibration was placed in the bank. If no basis could be found for the large difference in difficulty values on these administrations, the most recent difficulty value calibration was placed in the bank for future use.

Data analysis. Several methods were employed to compare outcomes of the different item-bank maintenance procedures. Comparisons were made on the difficulty values of the items in the various banks resulting from the linking procedures stated above.

Graphs of the difficulty value parameters of the items in the original bank were made by plotting the values estimated in the original bank against those in the bank at the end of the procedures.

Comparisons were made in the correlations of the difficulty value parameters of the items in the original bank and the same items at the end of the banking procedures. The means and standard deviations of these items were also calculated and compared. Moreover, the means of the difficulty values of all the items in the banks resulting from these different procedures were compared, adjusting for the deletions made during the Method II procedure.

Using the items that were placed on the March 1982 (0008) test and their difficulty parameter values in each of the banks at the completion of the above procedures, calculations were made to estimate ability parameters. These calculations are similar to those in the BICAL-III computer program except that the item difficulty values are held constant while the maximum likelihood calculations are made for the ability values only. Since the ability values of the group who actually were administered the test (administration 0008) were known, comparisons of the various effects for the different procedures could be made.

Results

Table 1 presents the means and standard deviations of the item difficulty values of fifty-four out of the original sixty items that comprised the original bank (Six of the original items have been rewritten and, for that reason, were not used in this study). The procedure used in Method II required that the item difficulty values remain approximately the same and, if an item was not used in calibrating the linking constant because there was a significant change in its value on the test being equated and its value in the bank, the item was deleted from the bank for all future use. Fourteen out of the original fifty-four items were deleted from the bank in the linking processes in this method. The mean (in logits) of the difficulty values of the fifty-four items was .004 with a standard deviation of 1.38 in the original bank, .083 with a standard deviation of 1.27 at the end of the linking process using Method I, .112 with a standard deviation of 1.15 using Method IIIA, and .022 with a standard deviation of 1.25 using Method IIIB.

Also in Table 1 are the correlations of the initial difficulty values of these original fifty-four items with the difficulty values existing in the banks after equating by means of Methods I, IIIA, and IIIB. The slopes of the regression lines are also shown in Table 1.

An output of the computer program used in this study to obtain ability scores is the pairing of the number of items correct score with an ability score in logits. For these tests the passing score was set at attaining at least 1.12 logits. Table 2 contains the ability scores necessary for passing calibrated from the difficulty parameter values of the original fifty-four items as they were in the original bank and as they were in the banks at the end of the linking processes using Method I, IIIA, and IIIB. This ability score necessary for passing for the values in the original bank was 1.158, at the end of Method I was 1.178, at the end of Method IIIA was 1.156, and at the end of Method IIIB was 1.222. Thus to obtain a passing score of at least 1.12 logits, Methods I and IIIA would give the same attainment of number correct items as was true in the original bank, but Method IIIB would require one more correct item.

Figures 1-3 are graphs of the difficulty values of the original fifty-four items with the values in the original bank plotted against the values in the bank at the end of Methods I, IIIA and IIIB. Figures 4-6 are graphs of the ability scores calibrated using the difficulty values of the original fifty-four items derived at the end of Methods I, IIIA and IIIB plotted against the ability values derived from using the original difficulty values. Figures 7-9 are graphs of the ability scores derived from the difficulty values at the end of Methods I, IIIA, and IIIB plotted against each other.

The means of the difficulty values of all 804 items in the final banks by Methods I, IIIA, and IIIB and the means of the difficulty values of all items in the final banks minus the values of the thirty-two items that were deleted by Method II are presented in Table 3.

The items that comprised the test for administration 0008 were selected from each of the banks resulting from Methods I, II, IIIA and IIIB. The means and standard deviations of the difficulty values of these items in the corresponding banks are shown in Table 4. Also shown are the ability passing scores, in logits, given that a value of at least 1.12 logits is required. The number of items correct for the corresponding logit score is also shown. Figures 10-15 are the graphs of the ability scores, calibrated from the difficulty values of the 0008 administration items after they were equated to the banks resulting from each of the methods, plotted against each other. In Table 5 are the equating constants needed to equate the test given in administration 0008 to the existing banks by each of the methods. These constants range from -.617 to -.569.

In Table 6 are the number of items removed from the calculations for deriving the linking constants for each of the methods in each of the links. Since the first link was the same for all methods, the items removed from the calculations were the same. However, for Method I the number of items taken off the second link include nine items previously removed from the calculations in the first link, those taken off the third link include seven items previously deleted from the first two links, and those taken off the fourth link include one item which was removed from the first two links and one item removed from the third link. During the Method II process, if an item was removed from the linking constant calculations, it was permanently removed from the bank and thus entered into no further calculations. During the calculations of the linking constants in Method IIIA there were nine items taken off the second link that had also been removed from the first link, six items taken off the third link that had been removed from the first two links, and one item removed from the fourth link that had been removed from the first and second links. For Method IIIB in the second link two items were removed that had been taken off the first link, in the third link one item that had been removed from the first two links and one item that had been taken off the first link were removed, and in the fourth link one item that had been removed from the first link and one item that had been taken off the first three links were taken out of the linking calculations.

Discussion

The purpose of this study was to investigate the cumulative effects of three different item-bank maintenance procedures on a series of longitudinal equatings of tests measuring the same characteristic. One way to compare these cumulative effects was to choose a set of items comprising a test and to make various comparisons using both the difficulty values in the banks resulting from the different linking processes and the ability scores calibrated from these difficulty values.

For this study, the items used on the test for administration 0008 were chosen. The means of the difficulty values of these items in the different banks at the end of the linking processes (Table 4) were compared. These ranged from -.798 for Method IIIA to -.846 for Method IIIB. Ability scores were calculated using the difficulty values found in the banks resulting from the different processes. Since an ability score of at least 1.12 logits must be attained in order to pass, the ability values needed for a passing score ranged from 1.148 for Method IIIB to 1.195 for Method IIIA. For each number of items correct score there is a corresponding ability score. Although, for this test, there was variability in the ability logit score, the number of items correct needed for passing was identical for all methods. Figures 10-15 are graphs of the ability scores (at each number of items correct value) calibrated from the difficulty values of the items in this test at the end of the procedures. There was some variation in the low and high ability scores but very little variation in the middle scores. Although the procedures used in Methods IIIA and IIIB were the most similar, the difficulty value means and the ability scores derived from these difficulty values show the greatest difference.

The differences in the means of the difficulty values among the methods and the differences in ability score logits necessary for the passing score seem to indicate that, although at this stage there seems to be no difference among the methods in the determination of a passing score on a test, significant differences may show up in subsequent longitudinal equatings. However, since the difficulty values of the items chosen for the test produce passing ability scores which are equated to the same number correct score, no comparisons can be made of the effects of using each of the different procedures.

Another way to compare the cumulative effects of the different item-bank maintenance procedures was to use the original fifty-four items and compare the difficulty values and ability score values in the original bank to those values at the end of Methods I, IIIA, and IIIB. Method II was not included in these comparisons because, in this procedure, an item either retained the difficulty value in the original bank or was deleted from the bank permanently (as fourteen original items were), if the difficulty value of the item on the test to be equated differed significantly from the bank value.

The means of the difficulty values of these fifty-four items in the original bank and at the end of the procedures are given in Table 1. The means of the difficulty values at the end of the procedures range from .022 to .112 compared to .004 in the original bank. The values in the bank following the Method IIIB procedure have a correlation of .990 with the original bank values, while the correlation of the Methods I and IIIA resulting values with the original bank values are .956 and .954 respectively. Figures 1-3 are graphs of the difficulty values of the original fifty-four items at the end of Methods I, IIIA and IIIB plotted against the difficulty values of these items in the original bank. In Figure 1 except for a few outliers the difficulty values of the items are close to a line which has a slope of 1 and a vertical axis-intercept of 0. In Methods IIIA and IIIB the difficulty value of an item retained its original value unless the difficulty value in the original bank and that on the test to be equated differed so significantly that the value was not used in the link. In that case, the item was given a new value. Comparing Figures 2 and 3 it is seen that Method IIIA resulted in the new difficulty values being further from a line with slope 1 and vertical axis-intercept of 0 than those given in Method IIIB.

As seen in Table 2 Methods I and IIIA require the same number of items correct score as was needed in the original bank, while Method IIIB requires one additional item correct. Moreover, the ability passing score resulting from the values in Method IIIA is the closest to that of the original bank (1.156 to 1.158). Figure 4 shows that the ability scores (at each number of items correct value) are less when derived from the original bank difficulty values than they are when calibrated from those difficulty values at the end of Method I. Figures 5 and 6 show that for ability score values less than 1 and 0 respectively, the ability scores derived from the original bank difficulty values are less than those derived from the results of Method IIIA and IIIB, but for larger ability scores, the opposite situation is true.

Figures 7-9 show the graphs of ability scores derived from the difficulty values of the original fifty-four items at the end of Methods I, IIIA, and IIIB plotted against each other. For ability scores below 0, those scores derived from Method I difficulty values are less than those calibrated from Method IIIA values; however, for ability scores above 0, the reverse is true. The ability scores derived from the Method I process are greater than those calibrated from the difficulty values in Method IIIB, while the ability scores derived from Method IIIA are greater than those from Method IIIB.

Although the Method IIIA difficulty value mean differs more from the original bank difficulty value than those resulting from Methods I and IIIB, the Method IIIA ability score necessary for passing was the closest to the original value. Method IIIB item difficulty values correlated to a greater degree with the item difficulty values in the original bank than did those resulting from Methods I or IIIA. However, the Method IIIB ability score necessary for passing differed the most from that calibrated for the original item bank and required a number correct score one greater than Methods I, IIIA, and the original item bank. These results are not as consistent as those obtained using the items of the test in administration 0008.

Table 3 contains both the mean item difficulty values of all the items in the various banks and the mean item difficulty of all except the thirty-two items that were deleted in the Method II process. The values resulting from the Method IIIB procedure have the lowest mean. Those items that were deleted in the Method II procedure had a smaller effect on Method IIIA than on Methods I and IIIB.

From the preceding discussion it is clear that there are slight differences in the item banking procedures for the original fifty-four items. However, these differences did not manifest themselves significantly when the difficulty values in the different banks of the items chosen in such a manner as to comprise a test equated to the original Fall 1978 test were compared. For this reason it must be concluded that on the basis of this study there are no significant differences in the different item-bank maintenance procedures studied for longitudinal equating. However, it may be that four longitudinal links are insufficient and that differences in the procedures may become significant as the number of links increase.

TABLE 1

Summary Statistics and Correlations of Difficulty Values for the Fifty-four Original Bank Items

	Original Bank	Method I	Method II	Method IIIA	Method IIIB
Mean	.004	.083	*	.112	.022
Standard deviation	1.38	1.27	*	1.15	1.25
Correlation with Original Bank Values	-	.956	*	.954	.990
Slope of Regression Line	-	.881	*	.795	.894

*No values here because 14 items were deleted from the bank, all other items retained their original bank difficulty values

TABLE 2

Fifty-four Original Bank Items

	Original Bank	Method I	Method II	Method IIIA	Method IIIB
Number of Items Correct Needed for Passing	38	38	*	38	39
Ability Passing Score (Logits)	1.158	1.178	*	1.156	1.222

TABLE 3

Mean Difficulty Values of Items in Final Banks

	Method I	Method II	Method IIIA	Method IIIB
All Items	-.238	*	-.230	-.264
All Items Except the 32 Deleted in Method II	-.213	-.228	-.211	-.238

TABLE 4

Summary Statistics of All Items Used in Administration 0008

	Method I	Method II	Method IIIA	Method IIIB
Mean of Difficulty Values	-.803	-.838	-.798	-.846
Standard Deviation of Difficulty Values	1.43	1.43	1.43	1.43
Number of Items Correct Needed for Passing	49	49	49	49
Ability Passing Score (Logits)	1.190	1.156	1.195	1.148

TABLE 5

Equating Constants in Final Linkage

Method I	Method II	Method IIIA	Method IIIB
-.574	-.609	-.569	-.617

TABLE 6

Number of Items Removed from Linking Calculations

Link	Method I	Method II	Method IIIA	Method IIIB
1	13	13	13	13
2	20	9	17	10
3	9	1	7	3
4	23	9	13	13
Total	65	32	50	39

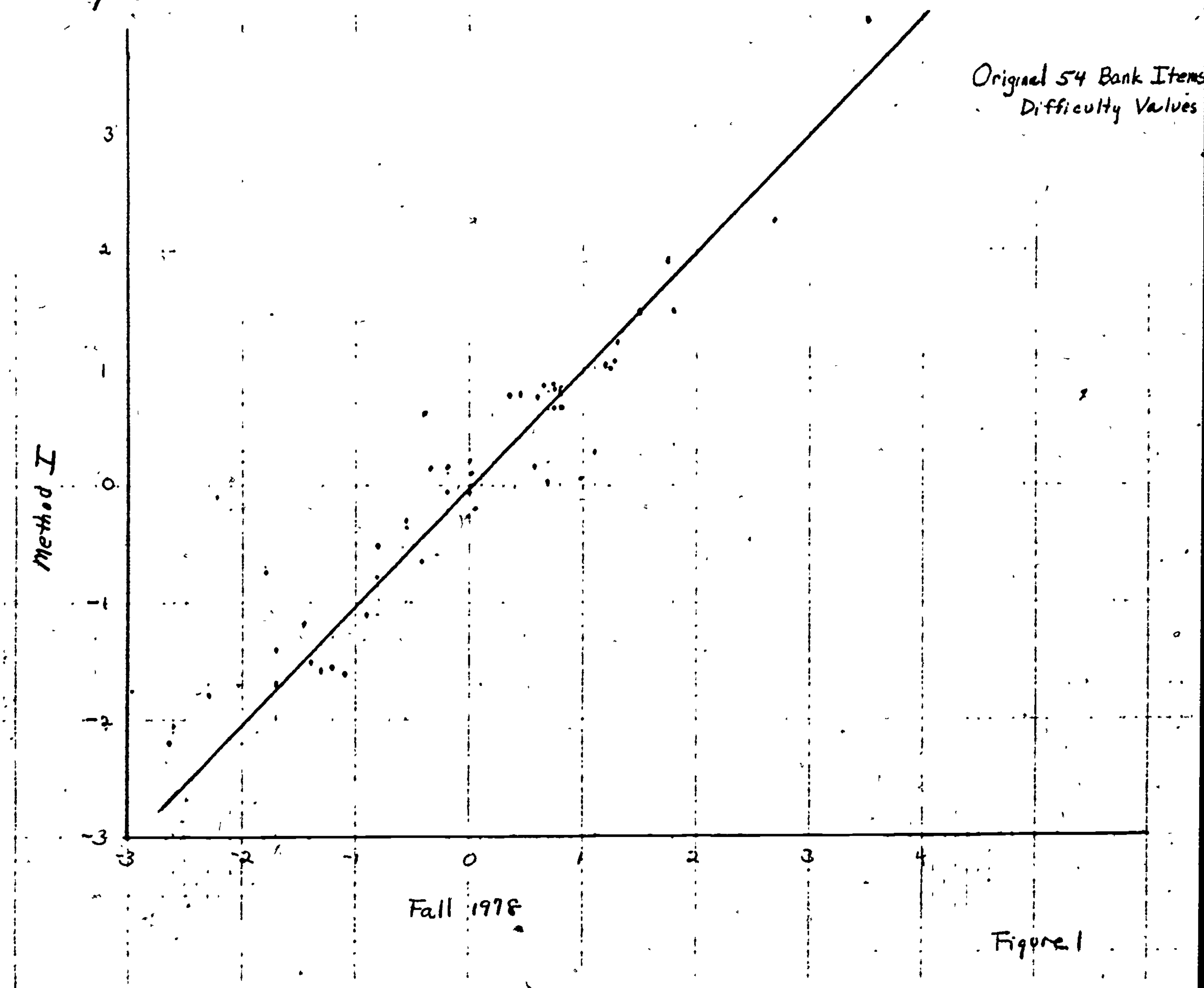
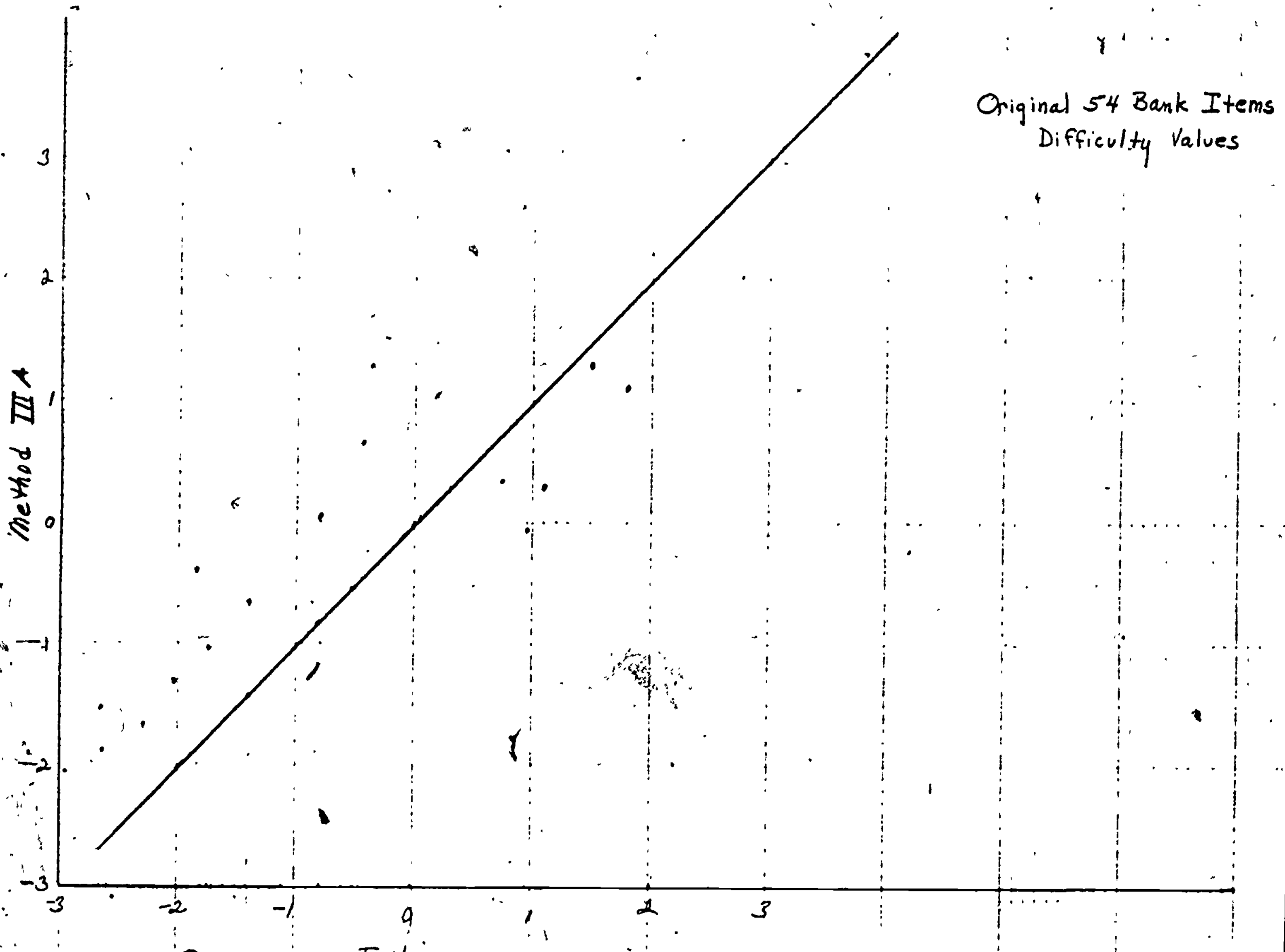

Original 54 Bank Items
Difficulty Values
Method I
Fall 1978
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
4

Figure 1



Original 54 Bank Items
Difficulty Values
Method IIIA
Fall 1978
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Figure 2.

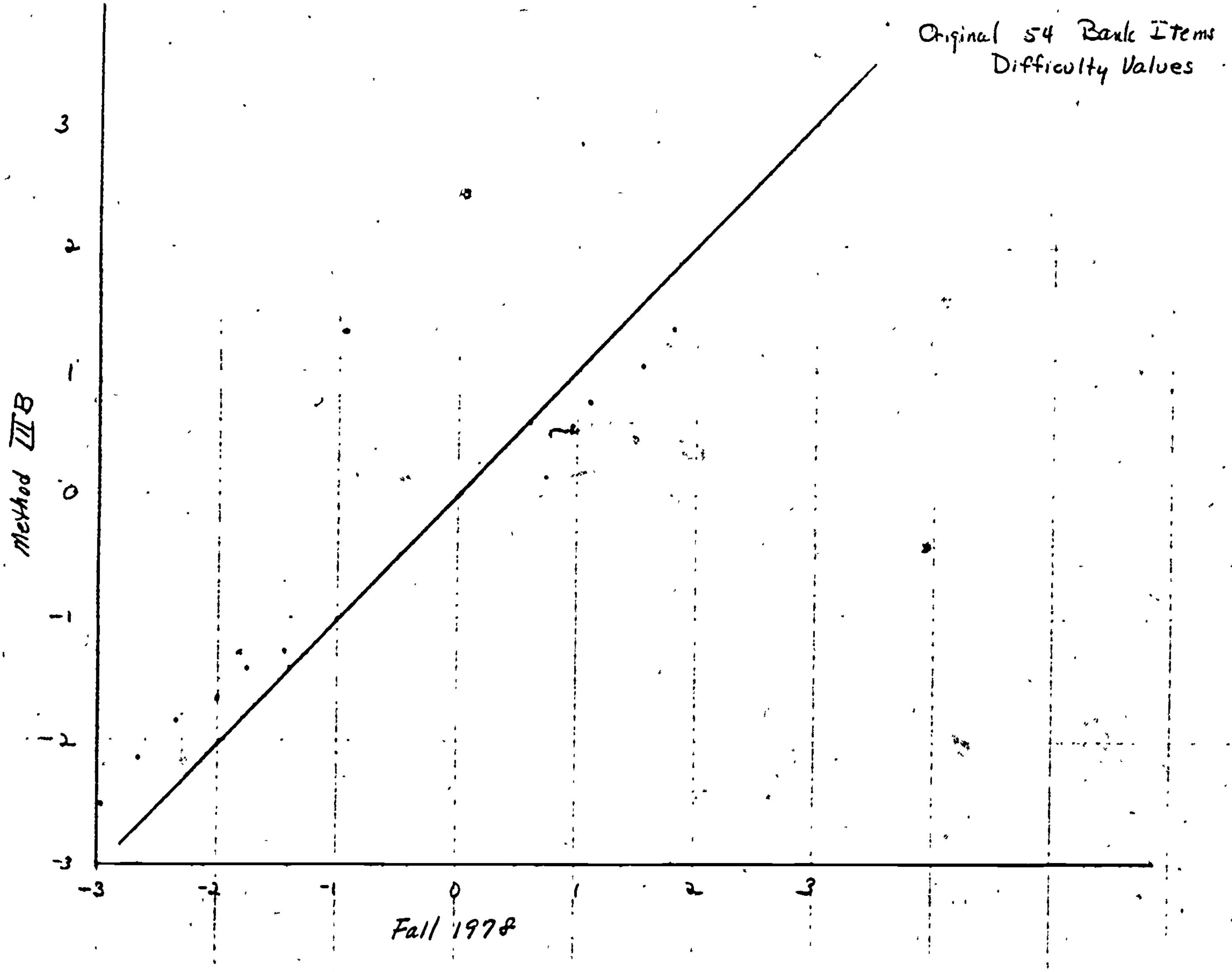

Original 54 Bank Items
Difficulty Values
Method IIIB
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Fall 1978

Figure 3

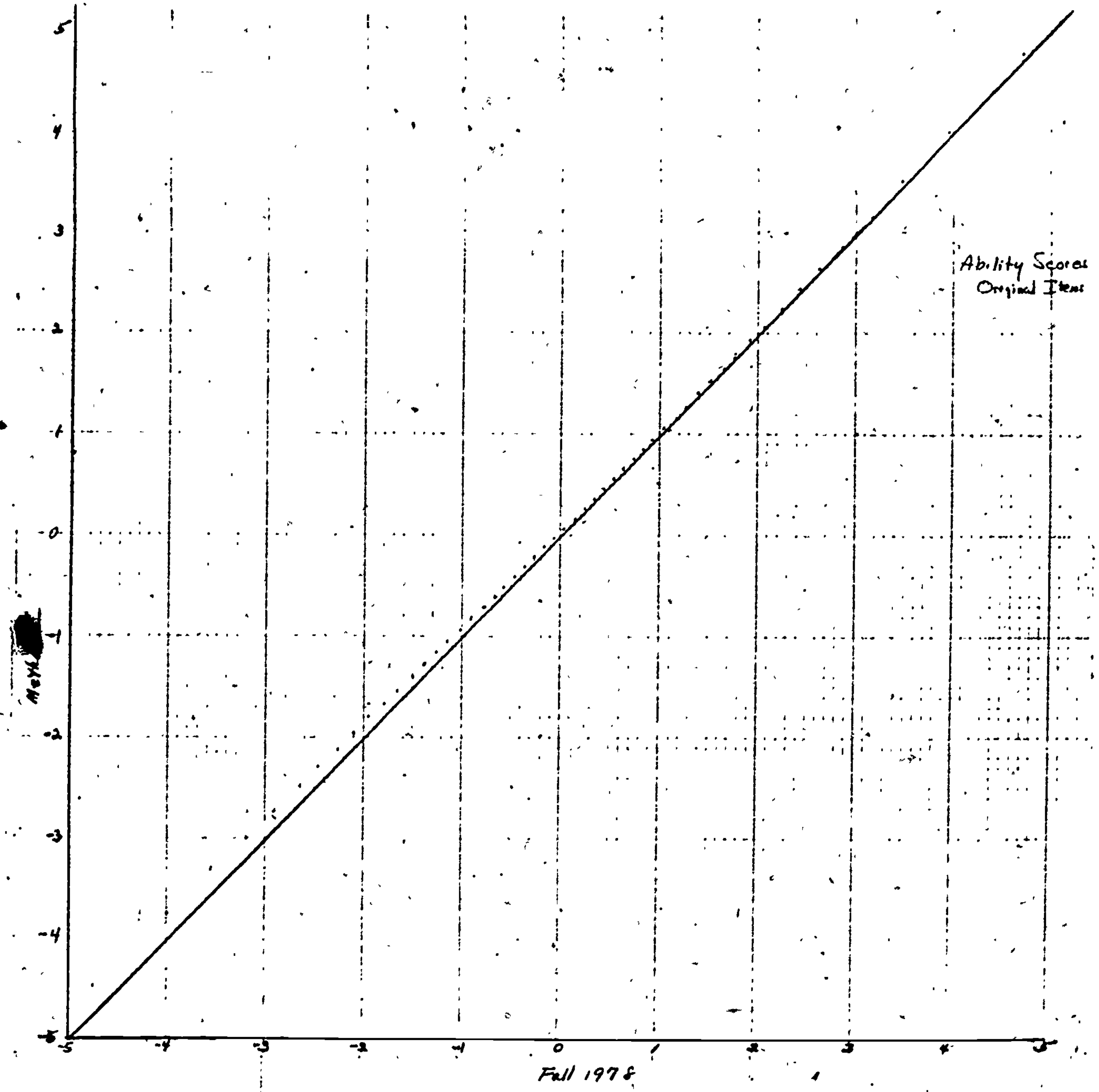


Figure 4

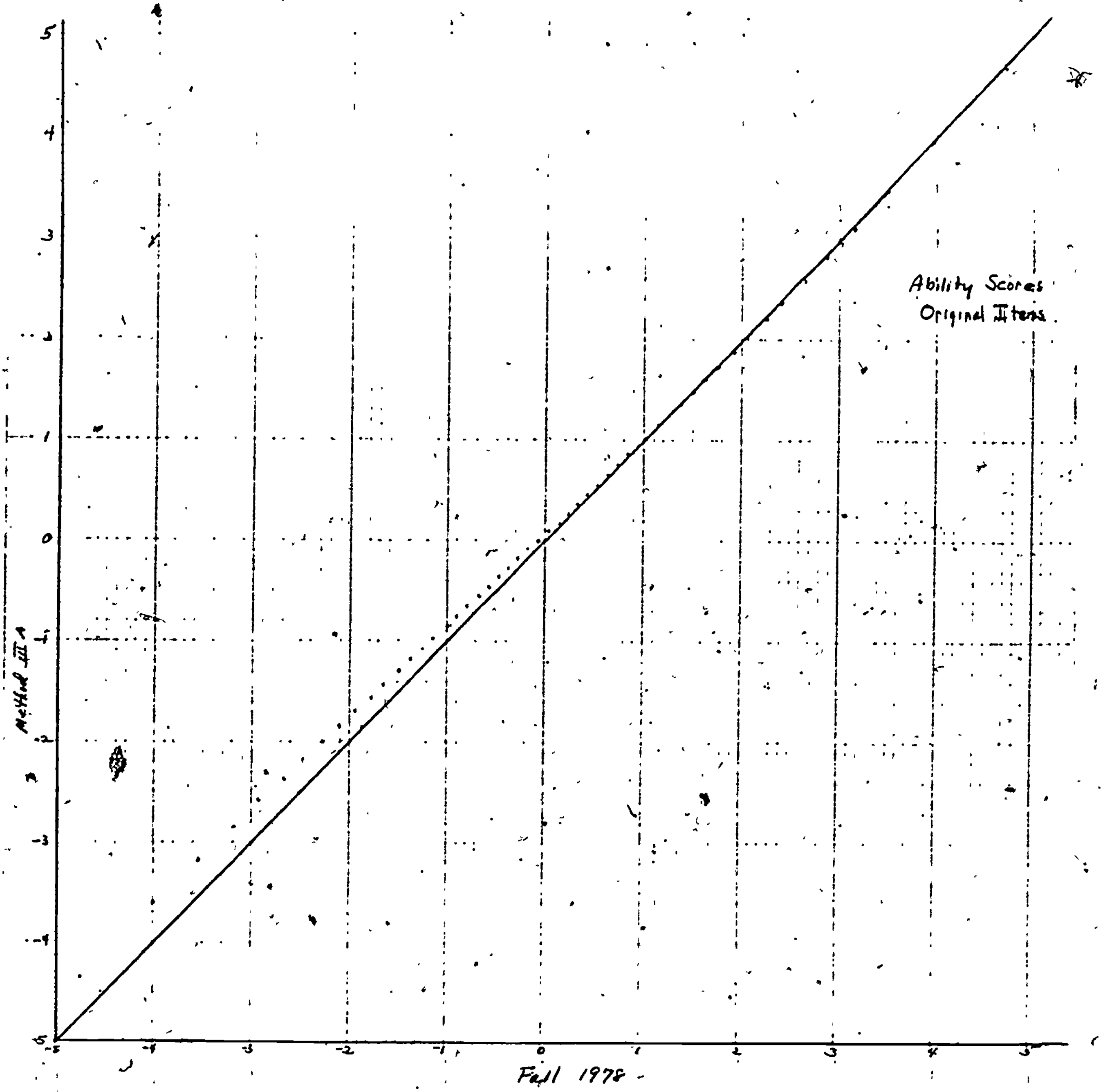

Ability Scores
Original II tems
Method III A
Fall 1978
5
4
3
2
1
0
-1
-2
-3
-4
-5

Figure 5

26

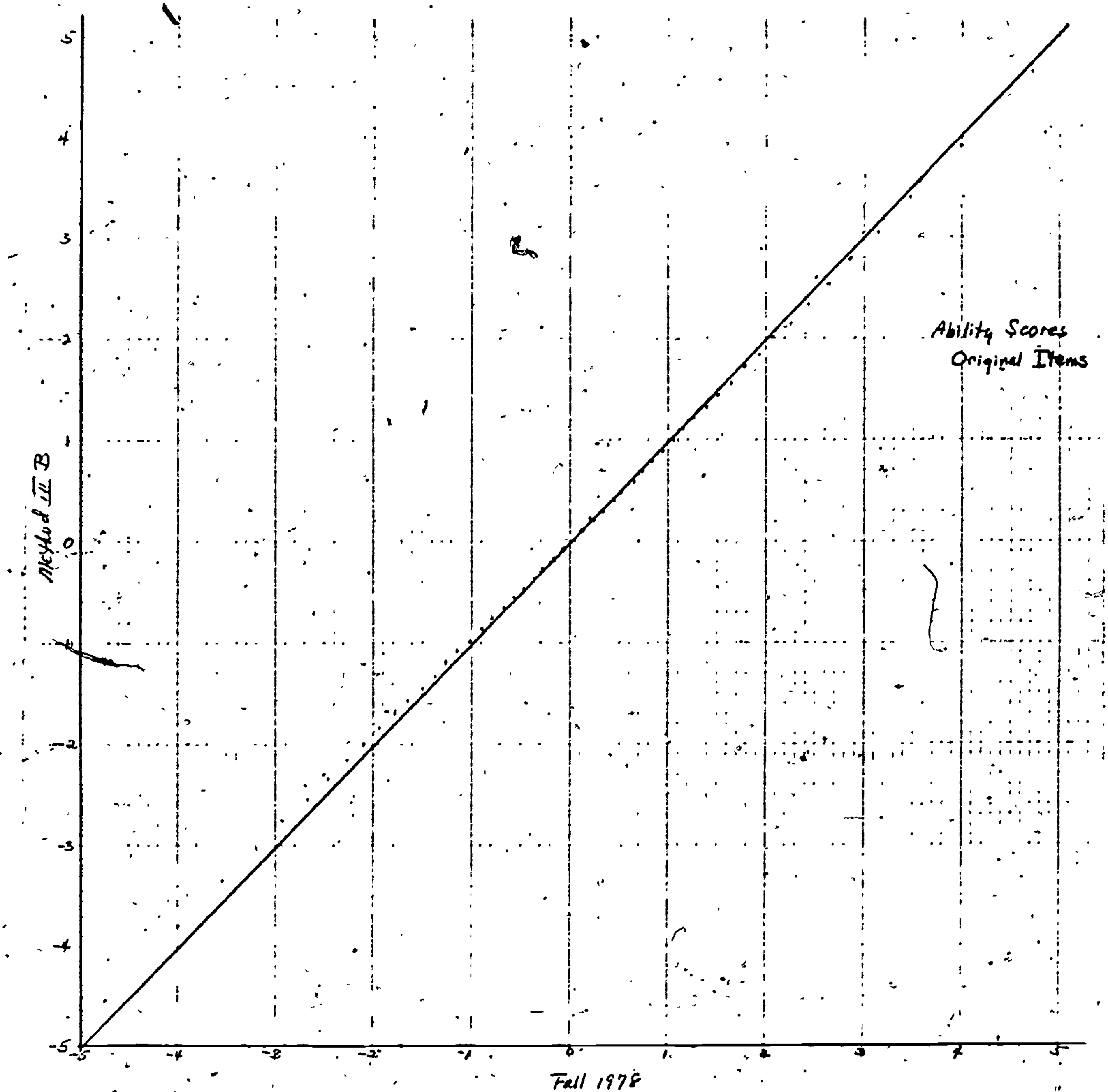

Ability Scores
Original Items
Method III B
Fall 1978
5
4
3
2
1
0
-1
-2
-3
-4
-5

Figure 6

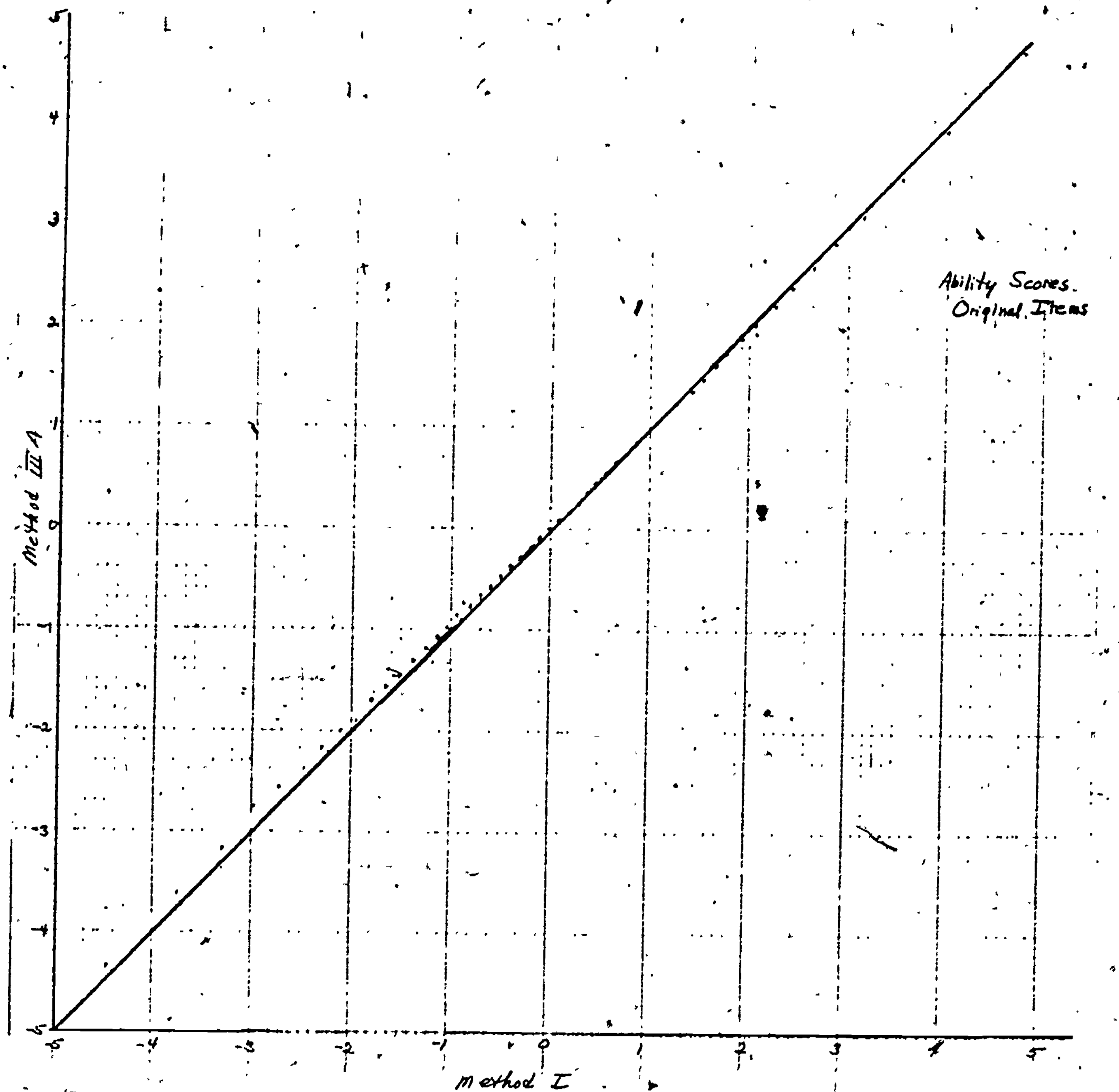


Figure 7

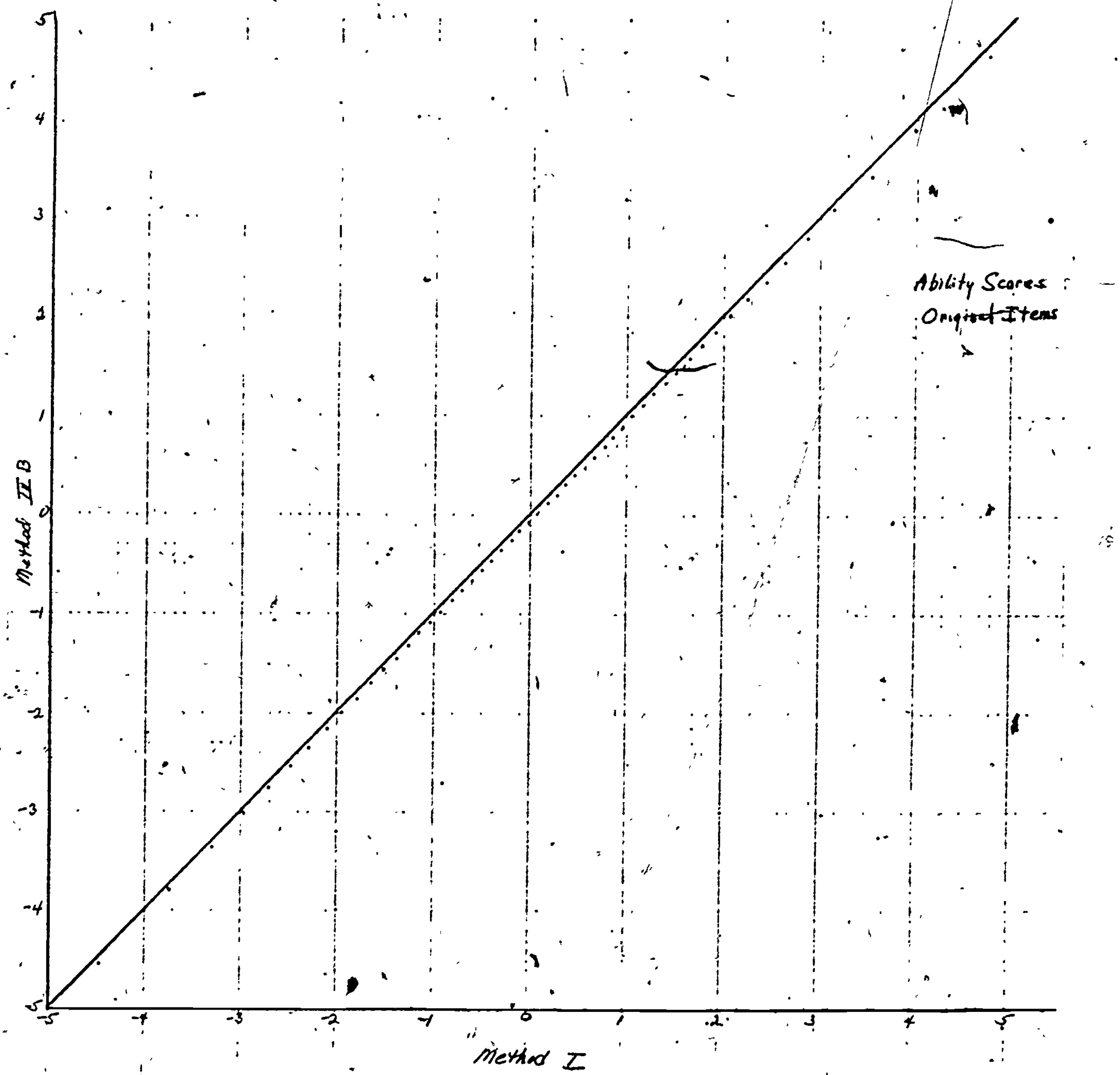


Figure 8

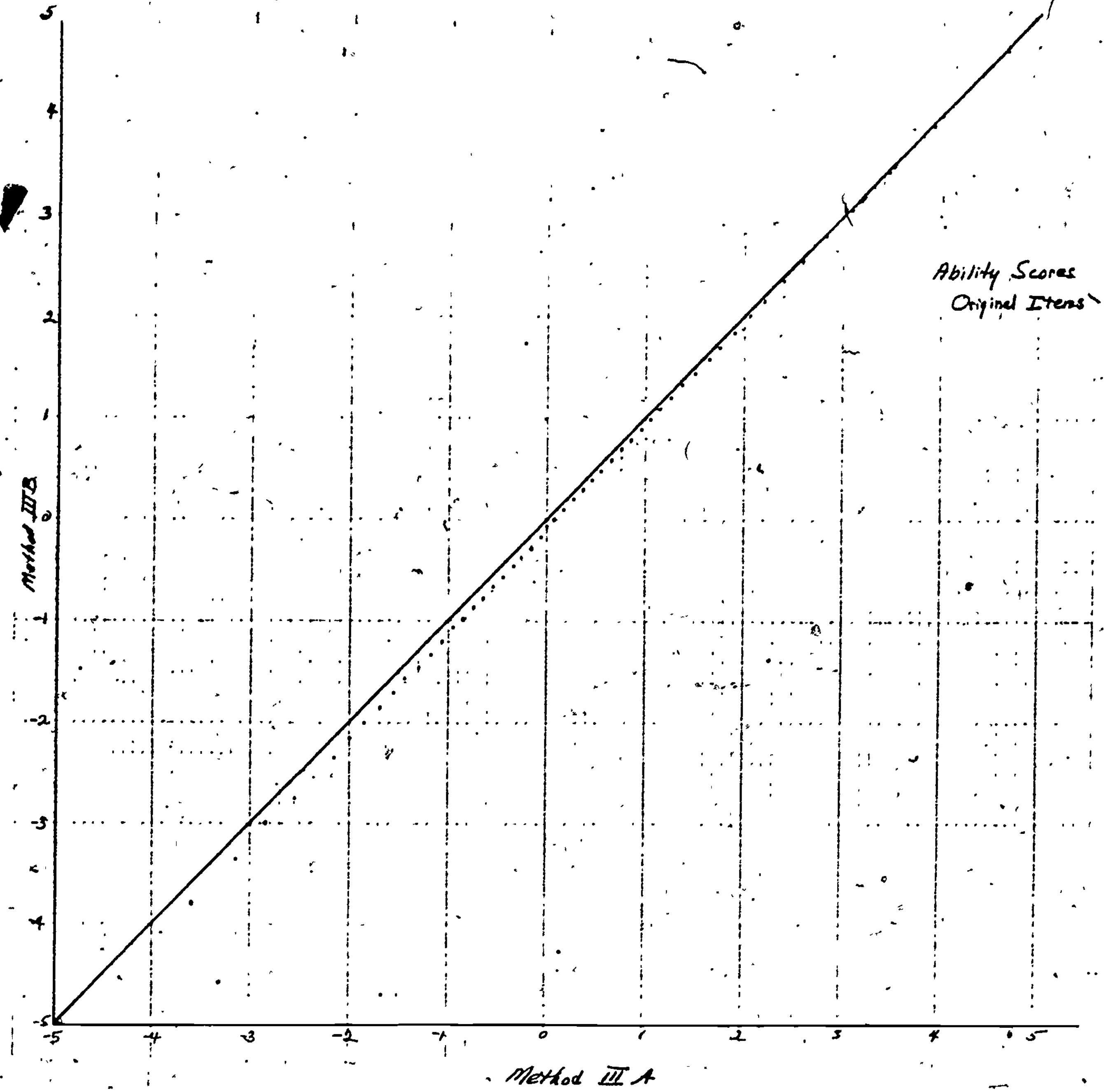

Ability Scores
Original Items
Method III B
Method III A
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 9

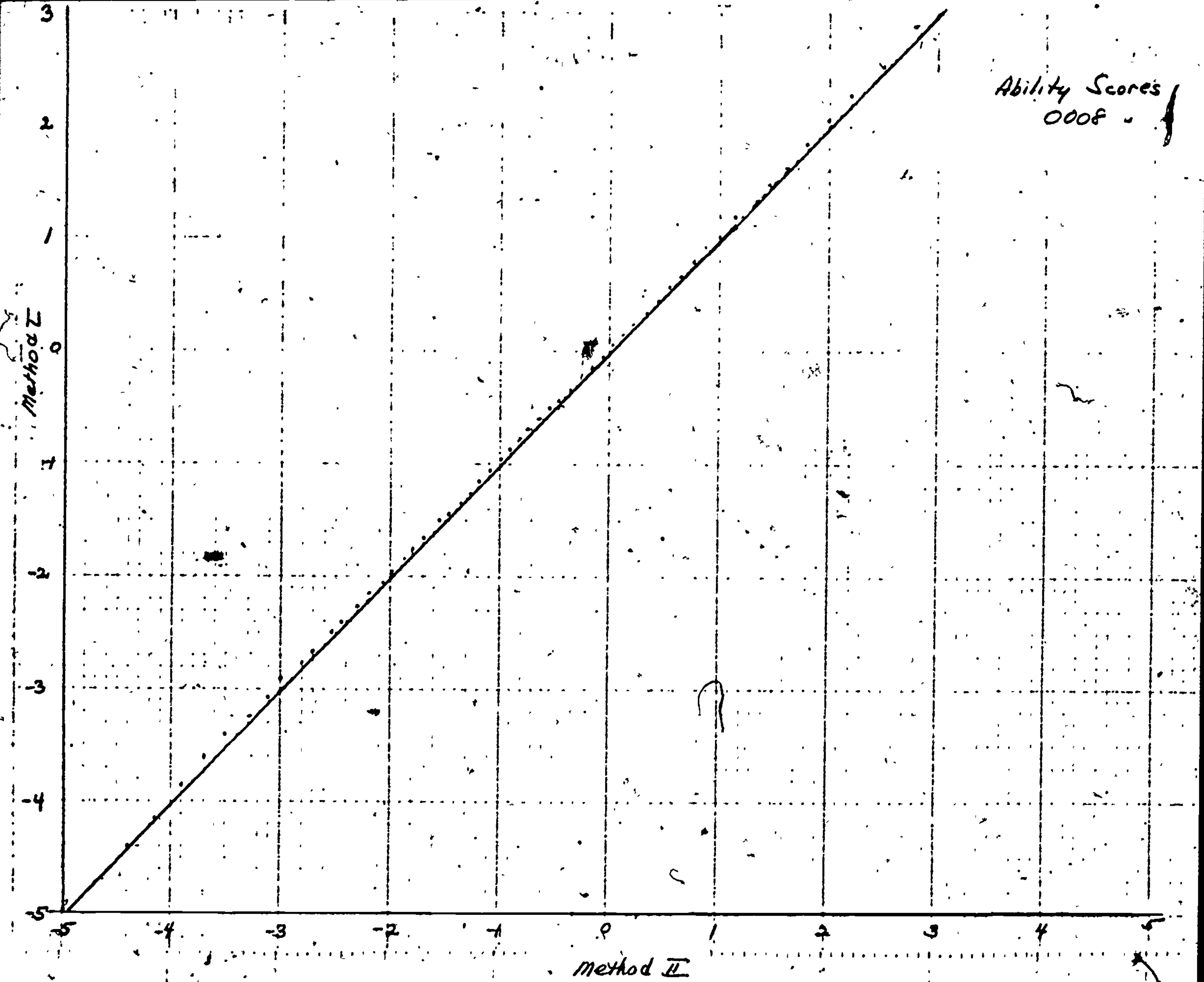

Ability Scores
0008
Method I
Method II
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 10

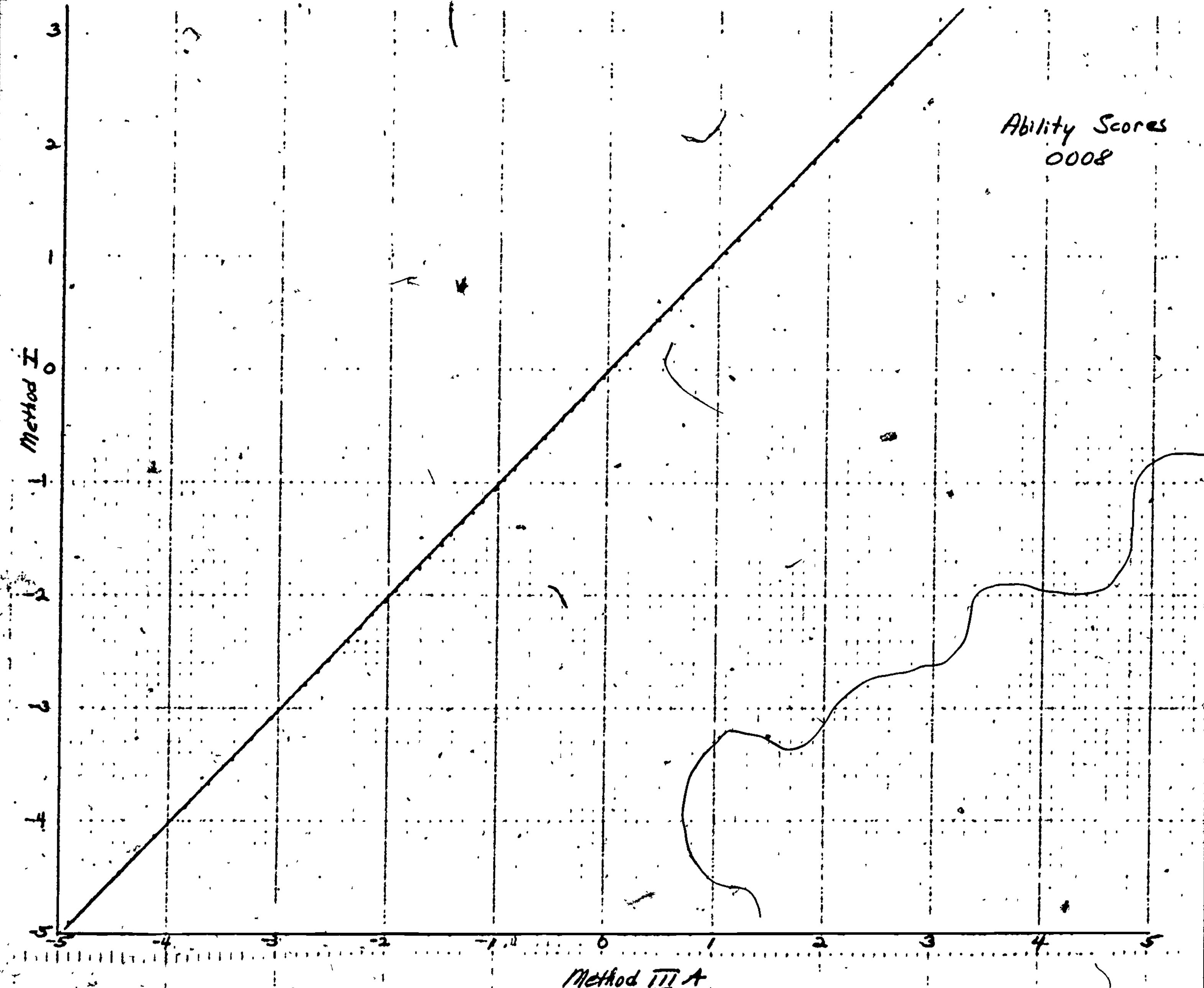

Ability Scores
0008
Method I
Method III A
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 11

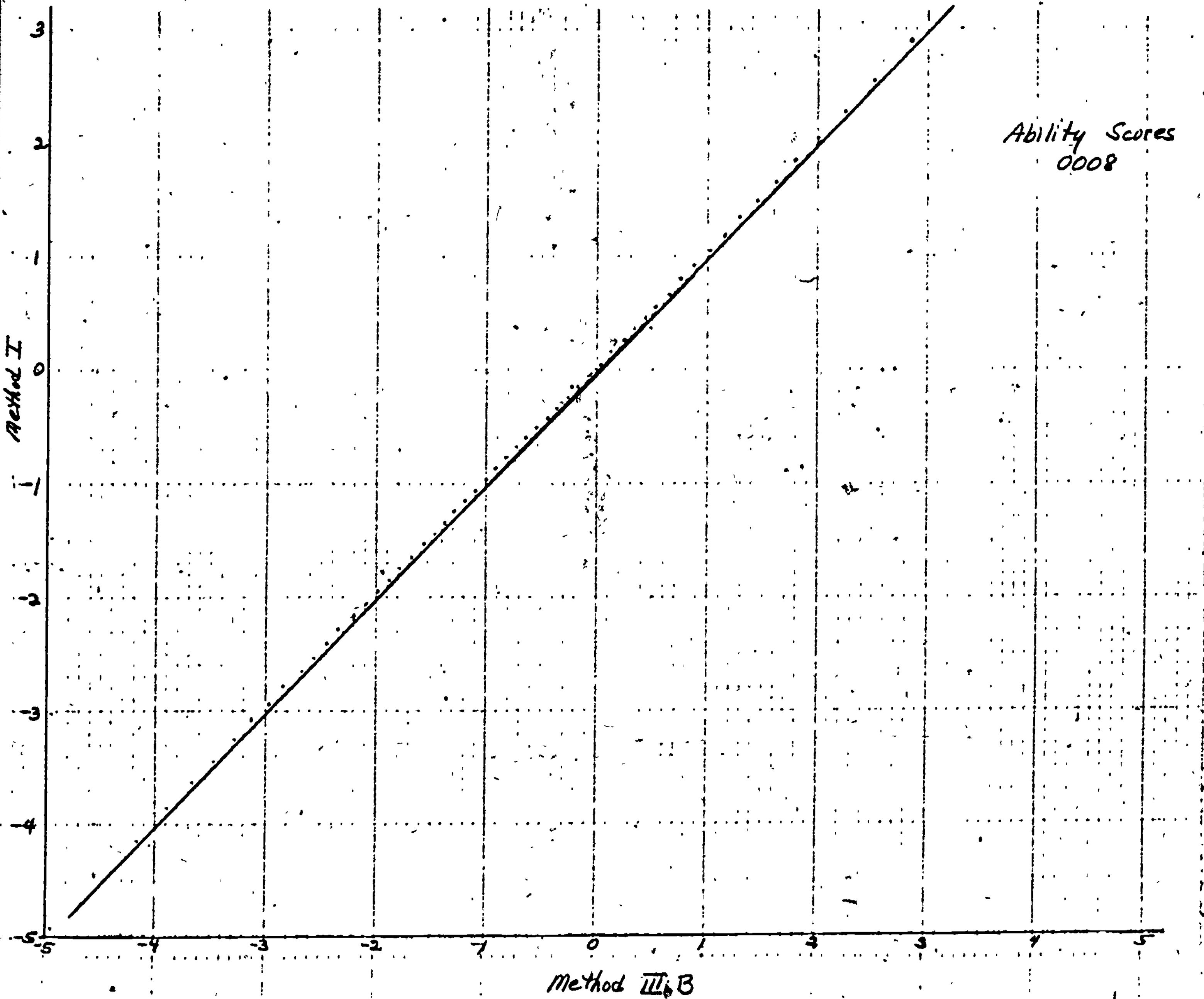

Ability Scores
0008
Method I
Method III B
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 12

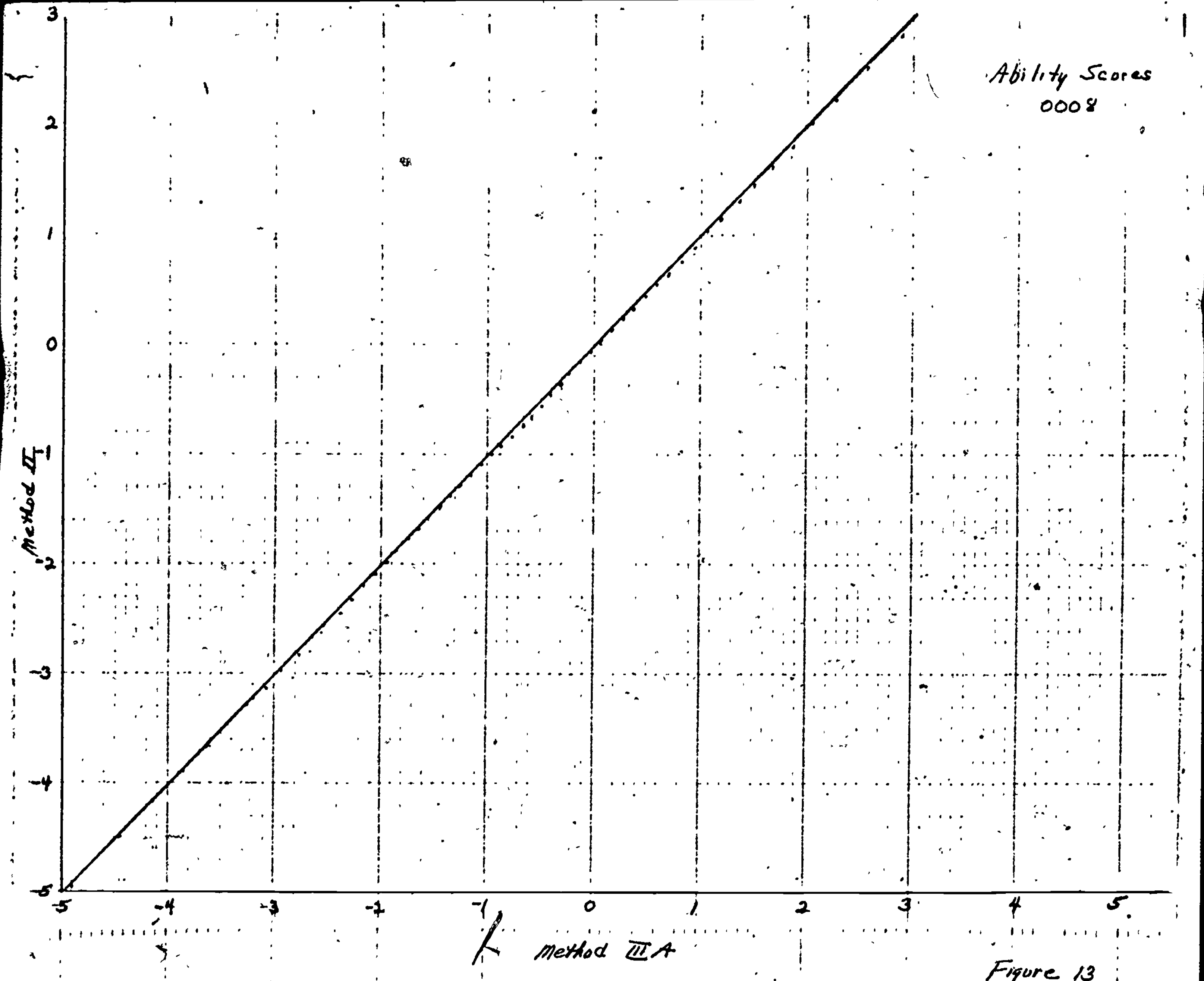
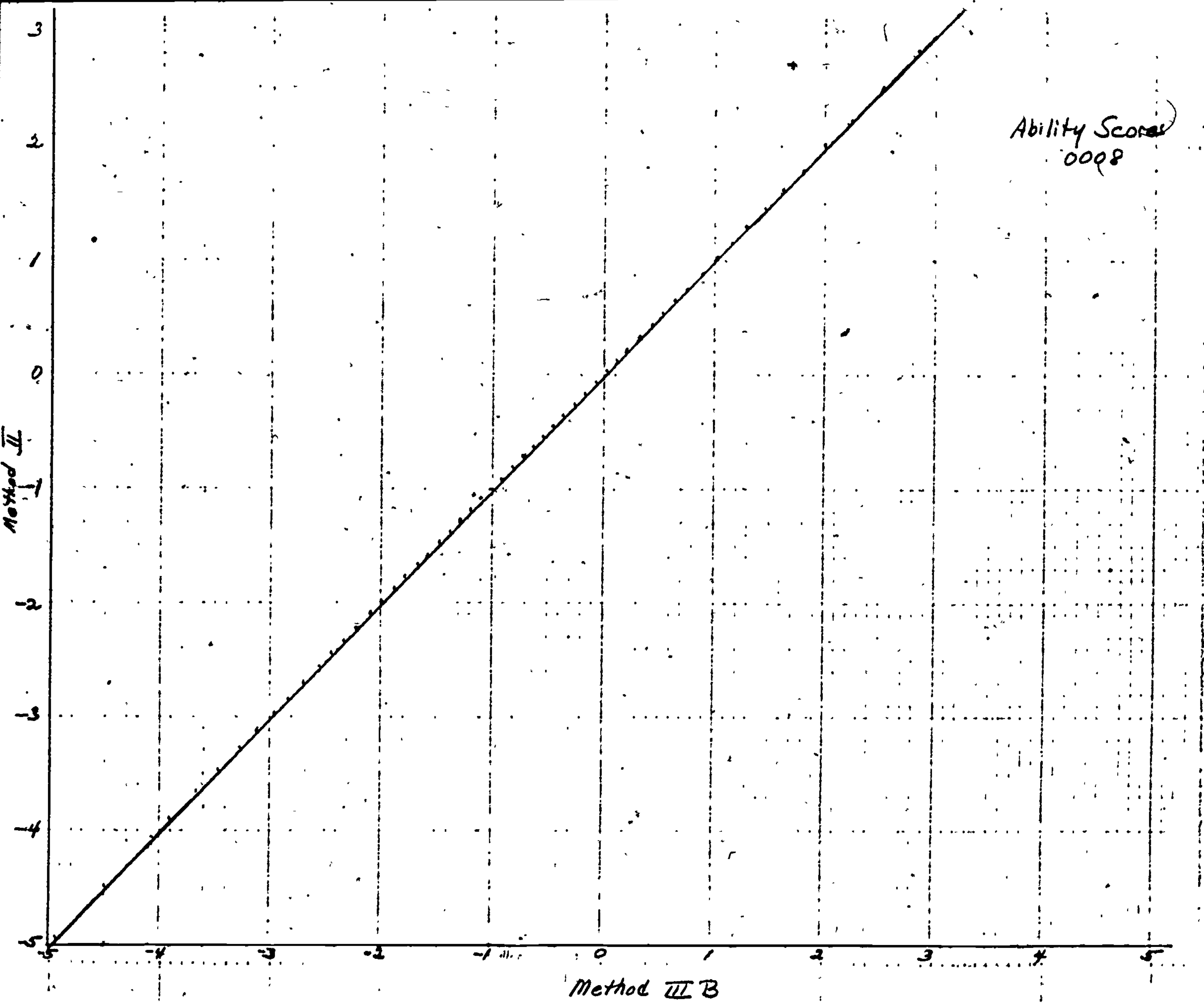
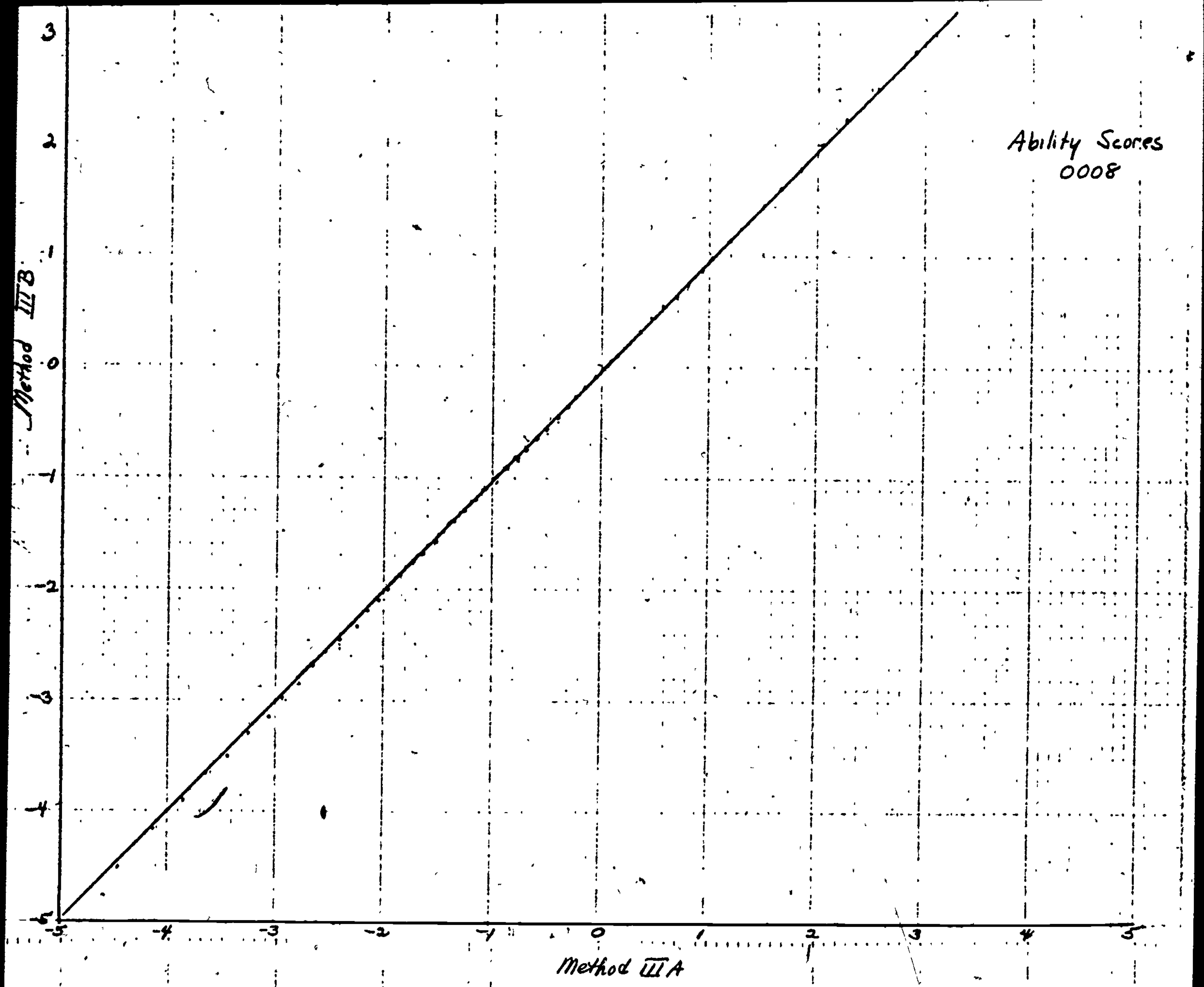

Ability Scores
0008
Method II
Method IIIA
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 13



Ability Scores
0098
Method III
Method III B
3
2
1
0
-1
-2
-3
-4
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 14



Ability Scores
0008
Method III B
Method III A
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 15

References

Angoff, W.M. Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational Measurement: Second edition. Washington D.C.: American Council on Education, 1971.

Brigman, S.L. and Bashaw, W.L. Multiple test equating using the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Cook, L.L., Eignor, D.R., and Peterson, N.S. A study of the temporal stability of IRT item parameter estimates. Paper presented at the annual meeting of the American Educational Research Association, New York, 1982.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., and Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Mead, R.J. Basic ideas in item banking. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, 1981.

Rentz, R.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, 1978.

Ridenour, S. Maintaining an item bank's underlying Rasch ability scale. Paper presented at the annual meeting of the American Educational Research Association, Boston, April, 1980.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D., Mead, R.J., and Bell, S.R. BICAL: Calibrating items with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, The University of Chicago, January 1977.

Wright, B.D., Mead, R.J., and Bell, S.R. BICAL: Calibrating items with the Rasch model. Research Memorandum No. 23B, Statistical Laboratory, Department of Education, The University of Chicago, June 1979.