DOCUMENT RESUME

ED 250 342 TM 840 624

AUTHOR Ziomek, Robert L.; Wright, Benjamin D.

TITLE A Procedure for Estimating a Criterion-Referenced Standard to Identify Educationally Deprived Children for Title I Services. Final Report.

INSTITUTION Des Moines Public Schools, Iowa.

SPONS AGENCY National Inst. of Education (ED), Washington, DC.

PUB DATE Mar 84

GRANT NIE-G-82-0030

NOTE 69p.; Small print in appendices.

PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS *Cutting Scores; Difficulty Level; Educationally Disadvantaged; Intermediate Grades; Item Analysis; *Latent Trait Theory; Mathematics Achievement; *Remedial Instruction; Remedial Teachers; *Standards; Teacher Attitudes; Test Items; Test Results; *Test Use

IDENTIFIERS Des Moines Public Schools IA; *Education Consolidation Improvement Act Chapter 1; *Rasch Model; Standard Setting

ABSTRACT

Techniques such as the norm-referenced and average score techniques, commonly used in the identification of educationally disadvantaged students, are critiqued. This study applied latent trait theory, specifically the Rasch Model, along with teacher judgments relative to the mastery of instructional/test decisions, to derive a standard setting procedure for Chapter 1 programming. Locally designed objectives-based tests, carefully matched to the mathematics curriculum and administered to students in the Des Moines Independent Community School District, were used to identify students in need of remedial assistance in mathematics as well as assessing students' overall mathematics achievement. Twenty-one Chapter 1 teachers participated in the study, 11 at the 4th and 5th grade levels and 10 at the 6th grade level. Teacher judgments of item difficulty were recorded. The study found that results of item calibration analyses confirmed stability of the item difficulty estimates for the tests and the groups of students studied. It further found that the lack of continuing remediation when the results so indicate, can potentially compound their deficiences at the succeeding grade level. Adhering to a predetermined "fixed" standard as opposed to a "floating" standard can affect a student's future learning experiences. (Author/EGS)

NIE-G-82-0036

ED250342

A PROCEDURE FOR ESTIMATING A CRITERION-REFERENCED STANDARD TO IDENTIFY EDUCATIONALLY DEPRIVED CHILDREN FOR TITLE I SERVICES

Robert L. Ziomek
Des Moines Independent Community School District
and
MESA Psychometric Laboratory
University of Chicago

Benjamin D. Wright
MESA Psychometric Laboratory
University of Chicago

Final Report of the
National Institute of Education

March, 1984

TM 840 629

ABSTRACT

That standards are set and utilized to aid in decision-making is indisputable. The determination to set standards, the type of information collected and the adoption of a standard setting model are all judgmental decisions. One can only ask that the process be reasonable and explicit, and those involved in establishing a standard are aware of what they are doing.

An area in which standard setting is of crucial importance is the identification of educationally disadvantaged students. Each year hundreds of thousands of elementary and secondary students are selected on the basis of need to receive remedial instruction funded under Chapter I legislation. The processes engaged in to make these determinations are as varied as the school districts and people making the decisions. Commonly employed techniques, such as the norm-referenced procedure, suffer from many technical problems associated with the measurement of the ability of low achieving students; inappropriate norming populations; in-and-out of level testing, to name a few.

Latent trait theory has evolved to the point of feasibly employing the techniques developed to study the achievement of students at varying ability levels. This study was directed at applying latent trait theory, specifically the Rasch Model, along with teacher judgments relative to the mastery of instructional/test decisions to derive a standard setting procedure for Chapter I programming.

That standards are set and utilized to aid in decision-making is indisputable. The assignment of course grades, the admission of students into graduate programs, the selection of a candidate for a job, being licensed to practice law, architecture or medicine, and the assignment of students to remedial educational programs are everyday examples of situations where the careful consideration of standards of performance or ability comes into focus. The process of standard setting, by its very nature, is a subjective undertaking, depending almost entirely on human judgment. Standards are set because it is believed that imperfect standards are better than none, and in order to introduce a degree of objectivity into the decision-making dilemma. The determination to set standards, the type of information to be collected, the adoption or creation of a standard setting model are all judgmental decisions. One can only ask that the standard-setting process be reasonable and explicit, and that those who are involved in establishing or making decisions on the basis of a standard are aware of what they are doing and why.

An area in which standard setting is of crucial importance is the identification of educationally disadvantaged students in need of remedial assistance. Each year hundreds-of-thousands of elementary and secondary students are selected on the basis of "need" to receive remedial instruction in a variety of subject area programs funded under Chapter 1 legislation. The processes engaged in to make these determinations are as varied as the school districts and people making the decisions. Federal regulations call for a selection procedure that is systematic, uniform and as objective as possible.

At present, the two commonly employed techniques to establish a cut-off score (criterion score or standard for selection) are the norm-referenced and average-score procedures. The norm-referenced procedure involves administering a nationally standardized test and designating a score (percentile) as the standard below which students will be provided remedial assistance. The average-score technique, often used by school districts testing with locally developed instruments, establishes the district's test average, or a score slightly below that point, as the criterion.

The most disturbing drawbacks associated with the norm-referenced procedure are: 1) different "standardized" tests produce different results; 2) inappropriate norming populations; 3) inappropriate content relative to a district's curriculum; and 4) problems associated with the measurement of low ability students (because standardized tests are targeted at average ability levels).

A fundamental problem associated with the average-score procedure is the variability of a district's average test score measured from year to year. As a district's average test score increases, that is, as students appear to be getting smarter (or the test easier), the criterion goes up. Similarly, as a district's average score decreases, students appear to be getting dumber (or the test has gotten harder), the standard goes down. Consequently, the standard is dictated by the interaction of the varying ability of the student population and the varying difficulty of the test (without cognizance of or control over this confounding interaction) rather than being controlled by instructional priorities. The standard is, in essence, established opportunistically after-the-fact, instead of representing any predetermined guidelines.

Traditional psychometric methods for constructing and analyzing assessment tests encounter problems with the mesurement of achievement for both low and high ability groups. The achievement of differing ability groups cannot be assessed with tests which are not population invariant (i.e., tests that do not have the same psychometric characteristics across the populations being studied) in a way which preserves the consistency of a standard setting process.

The necessity of test invariance has been recognized in the pschometric literature for 60 years. The calibration of test item difficulty must be independent of the ability distribution of the population used to establish item difficulty estimates. The measurement of person ability must be independent of the particular test items used for the ability measurement. Consequently, the measure estimated for a test score must be an estimate of a person's ability which is freed from the difficulty distribution of the items comprising the test (i.e., an easy or hard test must lead to statistically equivalent estimates of a student's ability). Likewise, an estimate of an item's difficulty must be freed from the ability distribution of the calibrating population. Advances in latent trait theory, specifically the Rasch model, have evolved to the point where it has become feasible to accurately assess the achievement of students of varying levels of ability and concomitantly to make an objective explication of the standard setting problem.

A second phase of the standard setting process is the establishment of a priori criteria based on the expert judgments of qualified professionals and the translation of these judgments into test outcomes. In standard setting situations where either a norm-referenced or locally developed test is administered, decision makers specify a score criterion after the results of testing are in hand. This specification is dominated by concern over how many students can be served given fiscal and staffing resources and whether the truly needy students have been identified. Fixation on any particular total score, as if it were interpretable (without giving careful consideration to the test items passed or objectives mastered) often leads to confusion and even misidentification of students. This is because there exists no clear idea in the decision makers' mind as to what the score means and what its use implies with respect to instruction. This line of reasoning does not dismiss fiscal concerns and program resources, but it suggests that a better accounting of the dollars can be provided if the standard setting process is coupled with the interpretability of a test score in a way that is well understood by those who establish the criterion and allocate remedial instructional services.

Although the utilization of subject matter experts is not unique to the standard setting process to be discussed (the Nedelsky, Angoff, Ebel and Contrasting Groups procedures are techniques used to set standards on minimum competency tests employing expert judgment), the population free item difficulties provided by a Rasch analysis enable the experts to focus their judgments on a standard expressed in terms of item content in a way which is independent of population idiosyncracies. The explanation and utilization of a standard can be enhanced, if those who are to employ it understand what it means, how to use it, and have contributed to its establishment.

STANDARD SETTING UTILIZING

THE RASCH MODEL

Locally designed objectives-based tests, carefully matched to the mathematics curriculum of kindergarten through eighth grade have been administered each spring to all kindergarten through eighth grade students in the Des Moines Independent Community School District. These tests, in addition to assessing overall mathematics achievement, have been utilized as screening instruments in all Chapter I schools to identify students in need of remedial assistance in mathematics. Each year a score at or slightly below the school district's average has been identified as the cut-off score over the period of test utilization, and set after the results of testing have been examined. The final choice of a standard has been based upon the monies available and consequently the number of children that could be provided service. This standard is dictated by student results and not deduced from prior judgment.

The involvement of teachers in the determination of an a priori standard proceeded in the following fashion. Twenty-one Chapter I instructors, considered to be master teachers, were selected to participate in the study by the school district's Director of Chapter I services, the Chapter I mathematics coordinator and the school district's Supervisor of Mathematics. Of these 21 teachers, 11 were selected to participate in standard setting at the 4th and 5th grade levels and 10 were chosen to work on setting a standard for the 6th grade test.

Each teacher was furnished a copy of the appropriate grade level test and a recording form (see Appendix A for an example). The teachers were instructed to indicate on the recording form which test items an "average"

Chapter I student would answer correctly at least 50 percent of the time. Each teacher worked independently and was allotted one week to complete the task. After the results from the first round were tabulated, each teacher was furnished a copy of the group results in addition to their own initial judgments and instructed to study these results and make any changes in their first impressions which they felt were warranted. Once again each teacher worked independently and was given one week to complete the task. The results of the second round indicated those items which these teachers felt an average Chapter I participant would most likely pass. Following the tabulation of the results of the second round, the teacher judgments for each item on a given test were transformed into estimates of item logit difficulties for analysis (the BICAL program reports item difficulty estimates in logits, see Best Test Design, Wright and Stone, 1979), utilizing equation 1:

$$\text{Teacher judgment}^{1}\ \text{difficulty estimate for item } i = \log \left\{ \frac{\text{no of teachers not indicating pass for item } i + .5}{\text{no. of teachers indicating pass for item } i + .5} \right\} \quad (1)$$

In addition to the teacher judgment estimates, each of the tests were analyzed with BICAL and the corresponding student performance item difficulty estimates were obtained from the fourth through sixth grade student performances.

[1] Equation 1 is a modification of the item logit difficulty estimate appearing in Best Test Design. The addition of .5 in the numerator and denominator is to adjust for those items for which all teachers indicated pass or all indicated not pass. See Wonnacott and Wonnacott, Regression: A Second Course in Statistics, 1981, for a discussion.

CRITERION SCORES BASED UPON DIFFICULTY ESTIMATES

To derive a criterion score from these two independent estimates of item difficulties, the student performance difficulties estimated by the BICAL program were regressed on the teacher judgment difficulties. Because these sets of item difficulties are both estimates, ordinary least squares (OLS) regression on an error free independent variable is not appropriate. In this instance the OLS estimate of the slope coefficient is replaced by the ratio of the standard deviation of student performance difficulty estimates (S_d) divided by the standard deviation of teacher judgment difficulty estimates (S_t). Thus, in the regression equation $d = A + Bt$, $B = S_d/S_t$. The vertical intercept A, is the estimate of the criterion score and is defined in the usual manner; $A = \bar{d} - B\bar{t}$, where $\bar{d}$ is the mean of the student performance difficulty estimates and $\bar{t}$ is the mean of the teacher judgment difficulty estimates. An approximation for the standard deviation of criterion score A, is:

$$SD(A) = \sqrt{S_d^2/L + B^2(S_t^2/L)} \approx S_d\sqrt{2/L} \qquad (2)$$

where L is the number of test items, B is the constant calculated above, and $\bar{d}$, $\bar{t}$ are independent.

CRITERION SCORES BASED UPON ABILITY ESTIMATES

The preceeding section outlined the derivation of a criterion score based upon teacher and student difficulty estimates. A second approach to estimating a standard based on teacher estimates of student ability follows.

Each individual teacher's judgments on the items for each of the three grade level tests is totaled to yield a test score (1 = Pass item i; 0 = Fail item i). These total scores are then transformed into estimates of criterion abilities by the PROX technique described in Best Test Design (Wright and Stone, 1979):

$$b_j = H + G \log [r_j/(L - r_j)], \text{ teacher estimated ability criterion} \quad (3)$$

where r_j = number of items indicated "pass" by teacher j

L = total number of test items

$H = \sum_{i=1}^{L} d_i/L$, average item difficulty

$G = [1 + V/1.7^2]^{1/2}$, item variance expansion factor

$V = [\sum_{i=1}^{L} d_i^2 - LH^2]/(L - 1)$, item difficulty variance

Once each teacher's raw score is transformed into an estimate of criterion ability, these estimates are edited for outliers and averaged to yield a cut-off score for each grade level test. The standard deviation of the teacher estimates over the group of teachers serves as an indication of the coherence of the standard.

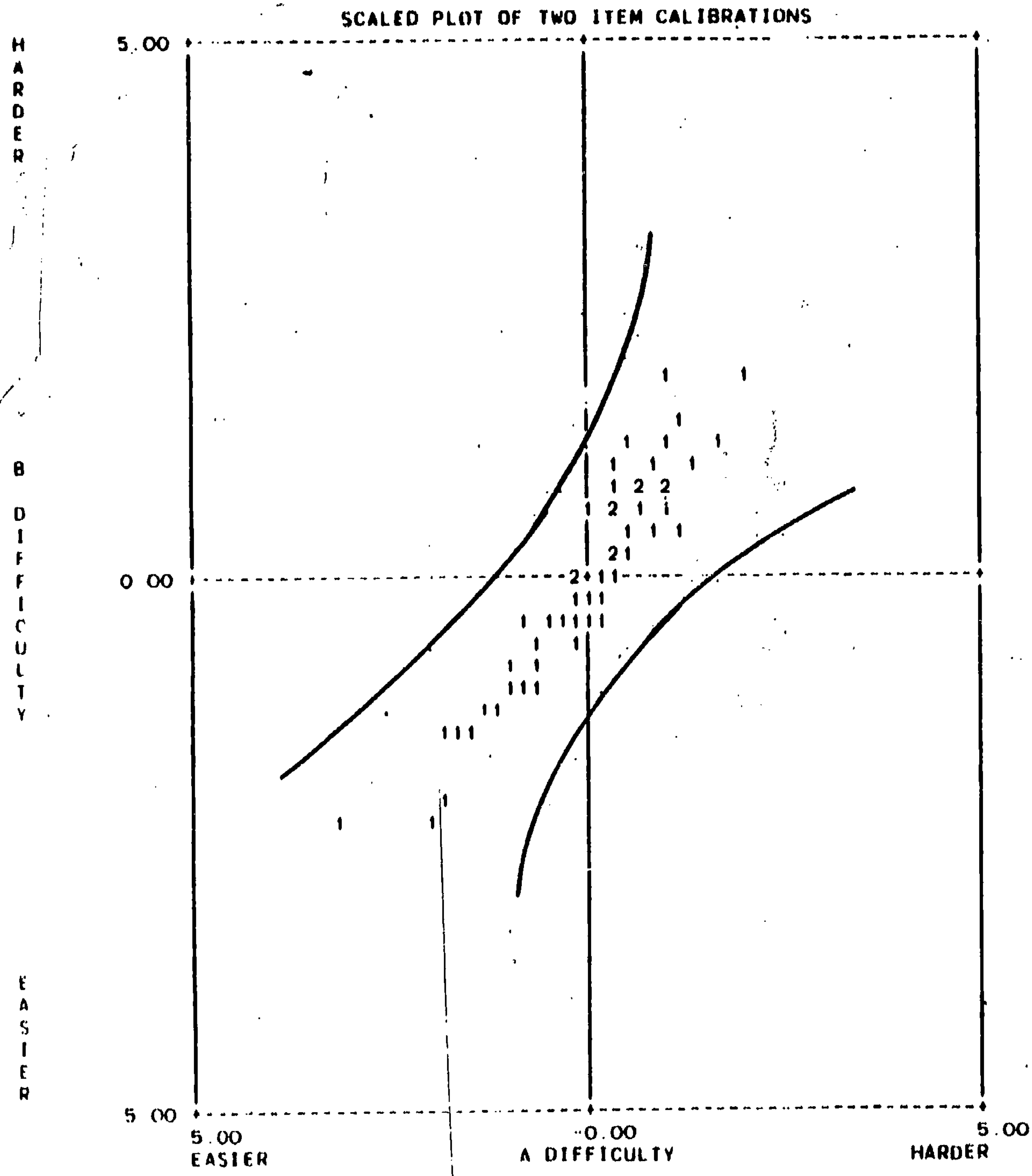
ABILITY MEASURES AND ITEM CALIBRATION INVARIANCE

Before applying the techniques outlined in the previous section a check of the invariance of item calibrations across the groups examined, within grades four through six, was conducted. Table 1 presents a description of the partitioned groups within grade level. For example, at 5th grade, 35 students participated in Chapter I mathematics remediation, in the 1981-82 school year, and on the basis of their spring 1982 test scores were served during the 1982-83 school year as 6th graders. One hundred twenty-two 5th graders were served during the 1981-82 school year, but on the basis of their fifth grade spring test scores were deemed ineligible for service as sixth graders the following school year. Finally, 50 students in Chapter I schools, not served as 5th graders were identified, on the basis of their fifth grade spring "82" test scores, and eventually served by Chapter I in the 1982-83 school year as 6th graders. A total of 207 students comprised the 5th grade group. Only those students for which complete test information existed across the grade levels were included in the study. This resulted in approximately 90 percent of the students being included in the study for each of the 3 grade levels.

Item difficulty calibration estimates for the entire group of students, at each grade level, were plotted against the difficulty estimates for the three partitioned groups to examine item difficulty invariance. Figure 1 presents the plot and the 99 percent confidence interval, of the difficulty estimates (5th grade test has 55 items) for the population of fifth grade students versus group one's (students served in both the 1981-82 and 1982-83 school years) item calibrations.

Figure 1

GRADE 5 ITEM CALIBRATIONS: COMPARISON OF GROUP 1 TO ALL STUDENTS   55 PAIRS



MEAN A= -0.00   SA= 1.00   R= 0.94   RMAX= 0.96
MEAN B= 0.00   SB= 1.00   PAIRS= 55

GROUP A: ITEM CALIBRATIONS FROM GRADE 5, GROUP 1
GROUP B: ITEM CALIBRATIONS FROM GRADE 5, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS

Figure 1 (cont'd)

GRADE 5 ITEM CALIBRATIONS: COMPARISON OF GROUP 1 TO ALL STUDENTS 55 PAIRS

NOTE. THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

SCALED POINTS BELOW ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
33	I133	1.18	0.45	0.73
28	I128	0.06	-0.50	0.55
29	I129	0.15	-0.39	0.55
30	I130	0.15	-0.26	0.41

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
37	I137	0.25	0.67	-0.42
44	I144	0.36	0.81	-0.45
49	I149	-0.05	0.63	-0.68
48	I148	0.36	1.04	-0.68
39	I139	1.01	1.78	-0.77
53	I153	0.48	1.25	-0.77

11

TABLE 1

Breakdown of Student Groups by Year of Chapter I Service

Grade Level	Group 1	Group 2	Group 3	All Students
4	In Chapter I for the 81-82 and 82-83 school years (N = 35)*	In Chapter I for the 81-82 school year and not the 82-83 school year (N = 102)	Not in Chapter I for the 81-82 school year but in for the 82-83 school year (N = 46)	N = 183
5	In Chapter I for the 81-82 and 82-83 school years (N = 35)	In Chapter I for the 81-82 school year and not the 82-83 school year (N = 122)	Not in Chapter I for the 81-82 school year but in for the 82-83 school year (N = 50)	N = 207
6	In Chapter I for the 81-82 and 82-83 school years (N = 34)	In Chapter I for the 81-82 school year and not the 82-83 school year (N = 86)	Not in Chapter I for the 81-82 school year but in for the 82-83 school year (N = 52)	N = 172

*Note: Numbers in parenthesis indicate the number of children falling in each of the three categories examined.

Overall item calibrations remained stable with a few exceptions. Those items which lie outside the confidence interval region do reveal, however, an interesting phenomenon, "instructional sensitivity." Referring to figure one, the horizontal axis represents the item calibrations of all fifth grade students, while the vertical axis represents the plot of item calibrations for fifth grade, group one students. Those items lying below the confidence control lines (#?8, 29, 30 and 33) are all division problems, with or without remainders. They are, since they lie below the confidence band, much easier for group one students than for the group as a whole. Conceivably, this group of students received the necessary amount of instruction to overcome their difficulties relative to division problems.

Those items lying outside and above the confidence band (#37, 39, 44, 48, 49 and 53) were more difficult for the group one students as compared to the total population. Three of the items (#44, 48, and 49) are measurement items, area of a rectangle, measure of an angle using a protractor and the average of three numbers; items 37 and 53 are decimal numeration problems; and item 39 is a geometry item. These items, represent a different and possibly higher order conceptual understanding as compared to the mechanics of division and less intructional time was probably devoted to these higher order concepts to insure that students know the rudements of calculating (Appendix B contains the complete results of the item invariance analyses). Overall the results of the item calibration analyses confirm stability of the item difficulty estimates for the tests and groups of students studied. The evidence of the stability of item calibration implies the stability of ability measures and subsequently permits the application of a standard setting procedure circumventing the concerns of traditional psychometric problems outlined in the introduction.

APPLICATION OF THE TECHNIQUES

An earlier section outlined two techniques which incorporate teacher judgments in the framework of an "objective" measurement process (Rasch Model), to establish a priori content-based test standards for the identification of students in need of remedial mathematics instruction. Since both techniques are mathematically equivalent in terms of the resulting cut-score (see interim report, March 1983), the results for setting criterion scores based upon ability estimates will be presented. Table 2 presents the resulting cut-score, in logits and their respective raw score values. After each teacher's raw score was converted to and ability estimate, outlined on pages seven and eight, outliers were removed by including only those estimates lying within plus or minus one standard deviation from the teacher's average ability estimate.

Figures 1, 2 and 3 present the results of mathematics testing at fourth grade in the spring of 1982. In each figure the total test raw score, frequency count, unconditional student ability estimate corresponding to a particular raw score, error estimate, and frequency distribution are displayed. In addition, the error bands centered on the teacner derived cut-score are depicted (refer to Table 2 for derived cut-score and error bands).

Figure 1 refers to the group of fourth grade Chapter I students (N = 35) who were served in 1981-82, tested in the spring, 1982, and eventually served in the 1982-83 academic year. Based upon the testing results relative to the error band, three students (8.6 percent) scored

TABLE 2

A Comparison of District Established and Model Derived Criterion Test Scores

Grade Level	District Criterion Spring 82	# of Test Items	Ability Method Cut-Score Logit	Ability Method Cut-Score Standard	Ability Method Cut-Score Equivalent Raw Score	Error Band Logits	Error Band Raw Score
4	34	50	.42	.29	29	(.13, .71)	(26, 32)
5	33	55	-.05	.24	27	(-.29, .19)	(24, 29)
6	35	60	.21	.12	32	(.09, .33)	(31, 34)

well above the cut-score and outside the error or retest band. Twenty-one students (60 percent) scored below the lower limit of the error band, while eleven students (31.4 percent) fell within the error band. The error band functions to identify those students who should probably be retested to verify their scores before a final determination is made to serve those individuals. Consequently, according to the results presented, three students scored high enough on the test to question their eligibility for further service.

Figure 2 presents the results of the group of fourth grade students (N = 102) who were served in the 1981-82 school year and were determined ineligible for service in the 1982-83 academic year. Based upon the tests scores relative to the estimated criterion score, 64 students (62.7 percent) scored above the upper limit of the retest (error) band; 24 students (23.5 percent) fell within the retest region and 14 individuals (13.7 percent) fell below the error band and probably should have been given further consideration relative to a second year of remediation.

TABLE 3

Summary of Eligibility Decisions Based Upon the Teachers Predetermined Cut-Score

Grade Four
N = 183

	Group I N = 35	Group II N = 102	Group III N = 46
Above Error Band	3*	64	7*
Within Error Band	11	24	21
Below Error Band	21	14*	18

Grade Five
N = 207

	Group I N = 35	Group II N = 122	Group III N = 50
Above Error Band	3*	83	10*
Within Error Band	13	27	20
Below Error Band	19	12*	20

Grade Six
N = 172

	Group I N = 34	Group II N = 86	Group III N = 52
Above Error Band	9*	42	20*
Within Error Band	10	9	14
Below Error Band	15	35*	18

*Indicates the number of students for which an improper determination as to eligibility was made based upon the teacher derived cut-score as compared to the district's existing standards. For example, for fourth grade, three students in Group I who were served in both the 1981-82 and 1982-83 school year scored well above the cut-score and outside the retest (error) band, and consequently should have not been declared eligible for service based upon their test results. The text provides a complete description of the table.

Finally, the results of the group of fourth grade students (N = 46) tested in the spring of 1982, and served for the first time in the 1982-83 school year are presented in Figure 3. Seven students (15.2 percent) fell outside and above the error band and possibly did not need further service; twenty-one (45.7 percent) fell within the retest band and eighteen (39.1 percent) fell outside and below the error band. Based upon the results of the entire fourth grade population examined here (N = 183), 24 students (13.1 percent) on the basis of their test score, relative to the error band, should or should not have been served relative to the group they were in. A total of 56 students (30.6 percent) fell inside the retest region and should have been retested to verify their scores. Of the total group 103 individuals (56.3 percent) were correctly identified to be served or not served. Appendix C contains the results for the fifth and sixth grade analysis.

Table 3 presents a summary of the results for each of the three grade levels. In fifth grade a total of 25 (12.1 percent) students were improperly identified relative to the group they resided in; 60 (28.9 percent) fell in the retest (error) band, and 122 (58.9 percent) were correctly identified based upon the teacher established criterion score. In sixth grade a total of 64 (37.2 percent) students were incorrectly identified relative to their group, 33 (19.2 percent) fell in the retest zone and 75 (43.6 percent) were properly identified. The reader should recall that the information presented in Table 1, reflects the service and no service conditions as they existed at the time of the study based upon the district's and not the teacher derived cut-score. Furthermore, the results discussed in this section and depicted by Figures 1, 2 and 3 and Appendix B represent a comparison of decisions based

Figure 1

GRADE 4 MATH, GROUP-1 STUDENTS

50 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION: X=1 PERSON
49	0	4.98	4.75	1.05	
48	0	4.06	3.95	0.77	
47	0	3.52	3.45	0.64	
46	0	3.12	3.08	0.57	
45	0	2.81	2.78	0.52	
44	0	2.55	2.53	0.48	
43	1	2.32	2.31	0.45	X
42	0	2.12	2.12	0.43	
41	0	1.94	1.94	0.41	
40	0	1.77	1.78	0.39	
39	0	1.62	1.63	0.38	
38	0	1.47	1.49	0.37	
37	2	1.34	1.35	0.36	XX
36	0	1.21	1.23	0.35	
35	0	1.08	1.10	0.35	
34	0	0.96	0.99	0.34	
33	0	0.85	0.87	0.34	
32	0	0.74	0.76	0.33	
31	1	0.63	0.65	0.33	X
30	0	0.52	0.54	0.33	
29	0	0.4	0.43	0.33	
28	3	0.31	0.33	0.32	XXX
27	5	0.20	0.22	0.32	XXXXX
26	2	0.10	0.12	0.32	XX
25	2	0.0	0.01	0.32	XX
24	5	-0.10	-0.09	0.32	XXXXX
23	2	-0.20	-0.20	0.33	XX
22	2	-0.31	-0.30	0.33	XX
21	2	-0.41	-0.41	0.33	XX
20	0	-0.52	-0.52	0.33	
19	2	-0.63	-0.63	0.34	XX
18	1	-0.74	-0.75	0.34	X
17	0	-0.85	-0.86	0.34	
16	3	-0.96	-0.98	0.35	XXX
15	1	-1.08	-1.11	0.35	X
14	1	-1.21	-1.24	0.36	X
13	0	-1.34	-1.37	0.37	
12	0	-1.47	-1.51	0.38	
11	0	-1.62	-1.65	0.39	
10	0	-1.77	-1.81	0.40	
9	0	-1.94	-1.97	0.41	
8	0	-2.12	-2.15	0.43	
7	0	-2.32	-2.35	0.45	
6	0	-2.55	-2.56	0.48	

Figure 2

GRADE 4 MATH, GROUP-2 STUDENTS

50 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
49	0	5.21	5.10	1.10	
48	1	4.26	4.22	0.82	X
47	1	3.69	3.66	0.69	X
46	3	3.27	3.24	0.60	XXX
45	4	2.94	2.91	0.54	XXXX
44	1	2.67	2.64	0.50	X
43	3	2.43	2.40	0.47	XXX
42	3	2.22	2.19	0.44	XXX
41	4	2.03	2.00	0.42	XXXX
40	6	1.86	1.83	0.41	XXXXXX
39	3	1.70	1.67	0.39	XXX
38	2	1.54	1.52	0.38	XX
37	8	1.40	1.38	0.37	XXXXXXXX
36	7	1.27	1.24	0.36	XXXXXXX
35	5	1.14	1.11	0.36	XXXXX
34	5	1.01	0.99	0.35	XXXXX
33	8	0.89	0.87	0.35	XXXXXXXX
32	6	0.77	0.75	0.34	XXXXXX
31	3	0.66	0.63	0.34	XXX
30	3	0.54	0.52	0.34	XXX
29	5	0.43	0.40	0.33	XXXXX
28	5	0.32	0.29	0.33	XXXXX
27	2	0.21	0.18	0.33	XX
26	1	0.11	0.07	0.33	X
25	1	0.0	-0.04	0.33	X
24	3	-0.11	-0.15	0.33	XXX
23	1	-0.21	-0.26	0.33	X
22	2	-0.32	-0.37	0.33	XX
21	1	-0.43	-0.48	0.34	X
20	3	-0.54	-0.59	0.34	XXX
19	0	-0.66	-0.71	0.34	
18	0	-0.77	-0.82	0.34	
17	0	-0.89	-0.94	0.35	
16	0	-1.01	-1.07	0.35	
15	0	-1.14	-1.19	0.36	
14	1	-1.27	-1.32	0.36	X
13	1	-1.40	-1.45	0.37	X
12	0	-1.54	-1.59	0.38	
11	0	-1.70	-1.74	0.39	
10	0	-1.86	-1.89	0.40	
9	0	-2.03	-2.05	0.41	
8	0	-2.22	-2.23	0.43	
7	0	-2.43	-2.42	0.45	
6	0	-2.67	-2.63	0.47	

Figure 3

GRADE 4 MATH, GROUP-3 STUDENTS

50 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
49	0	5.06	4.80	1.05	
48	0	4.14	4.01	0.77	
47	0	3.58	3.51	0.64	
46	0	3.18	3.13	0.57	
45	0	2.86	2.83	0.52	
44	0	2.59	2.58	0.48	
43	1	2.36	2.35	0.46	X
42	0	2.16	2.16	0.43	
41	0	1.97	1.98	0.41	
40	0	1.80	1.81	0.40	
39	1	1.65	1.66	0.39	X
38	0	1.50	1.51	0.37	
37	0	1.36	1.37	0.37	
36	0	1.23	1.24	0.36	
35	2	1.10	1.12	0.35	XX
34	0	0.98	1.00	0.35	
33	3	0.86	0.88	0.34	XXX
32	4	0.75	0.76	0.34	XXXX
31	1	0.64	0.65	0.33	X
30	2	0.53	0.54	0.33	XX
29	3	0.42	0.43	0.33	XXX
28	0	0.31	0.33	0.33	
27	6	0.21	0.22	0.33	XXXXXX
26	5	0.10	0.11	0.33	XXXXX
25	3	0.0	0.01	0.33	XXX
24	1	-0.10	-0.10	0.33	X
23	3	-0.21	-0.21	0.33	XXX
22	1	-0.31	-0.31	0.33	X
21	2	-0.42	-0.42	0.33	XX
20	0	-0.53	-0.53	0.33	
19	2	-0.64	-0.64	0.34	XX
18	2	-0.75	-0.76	0.34	XX
17	1	-0.86	-0.87	0.34	X
16	0	-0.98	-0.99	0.35	
15	0	-1.10	-1.12	0.35	
14	0	-1.23	-1.25	0.36	
13	1	-1.36	-1.38	0.37	X
12	2	-1.50	-1.52	0.38	XX
11	0	-1.65	-1.66	0.39	
10	0	-1.80	-1.82	0.40	
9	0	-1.97	-1.99	0.42	
8	0	-2.16	-2.17	0.43	
7	0	-2.36	-2.36	0.46	
6	0	-2.59	-2.59	0.48	

upon the teacher derived standard versus the existing conditions based upon the district's standard.

It is clear from the results presented that the greatest error associated with a determination of service or no service is concentrated at the sixth grade level. The reasons for this situation are not entirely apparent, but one reasonable speculation can be advanced relative to the information presented in Table 3. Of the three groups studied within the three grade levels, Group 1 has the smallest proportion of students served in two successive years (1981-82/1982-83) relative to Group 2 students served in 1981-82, or Group 3 students served for the first time in 1982-83. For those fourth grade students falling in Group 2, and served only in 1981-82, 14 (13.7 percent) of the students, on the basis of the a priori standard, should have been served as fifth graders, and were not. The number could potentially increase based upon the retesting of those students in Group 2 (24 or 23.5 percent) falling in the error (retest) band. Consequently, for those fourth graders served in 1981-82 and not served as fifth graders, the lack of continuing remediation when the results so indicate, can potentially compound their deficiences at the succeeding grade level (in this case fifth grade) and eventually result in those students being selected for service at the sixth grade after two full years of being behind! Subsequently, the lack of adhering to a predetermined "fixed" standard (along with suitable accompanying anecdotal information) as opposed to a "floating" standard can result in an in-and-out migration pattern within the Chapter II program, vis-a-vis successive years of continuous remediation, adversely affecting a student's future earning experiences.

REFERENCES

Bejar, I. Biased Assessment of Program Impact Due to Psychometric Artifacts. Psychological Bulletin, 1980, 87(3), 513-524.

Bunda, M. A. and Sanders, J. R. (Eds.). Practices and Problems in Competency-Based Measurement. National Council on Measurement in Education, 1979.

Divgi, D. R. The Logic of Standard Setting: Some Issues and Questions. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY, 1982.

Echternacht, G. (Ed.). Measurement Aspects of Title I Evaluations. Jossey-Bass, Inc., Publishers, 1980.

Jaeger, R. M. Measurement Consequences of Selected Standard-Setting Models. Presented in a symposium on Measurement Issues Related to Performance Standards on Competency-based Education, National Council on Measurement in Education, San Francisco, CA, April 1976.

Popham, W. J. Setting Performance Standards. Los Angeles, CA: Instructional Objectives Exchange, 1978.

Rascn, G. Probability Models for Some Intelligence and Attainment Tests. Copenhagen: Denmark Paedagogiske Institute, 1960 (reprinted by the University of Chicago Press, 1980).

Shepard, L. A. Setting Standards and Living With Them. Paper presented at the National Council on Measurement in Education, San Francisco, 1976.

Simpson, N. and Smith, A. Achievement Gain: Different Estimates from Different Tests. A paper presented at the joint convention of the Iowa Educational Research and Evaluation Association and the Midwest Educational Research Association, Des Moines, IA, November 1981.

Wargo, M. and Green, D. (Eds.). Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation. CTB/ McGraw-Hill, 1977.

Warm, T. A. A Primer of Item Response Theory. Technical Report 940279. National Technical Information Service. Springfield, VI, December 1978.

Wright, B. Sample-free Test Calibration and Person Measurement. Invitational Conference on Testing Problems. ETC, Princeton, NJ, October 28, 1967.

Wright, B. and Stone, M. Best Test Design. University of Chicago, MESA Press, 1979.

Zieky, M. and Livingston, S. A Manual of Methods for Setting Standards on Criterion-Referenced Tests of Basic Skills. Educational Testing Service, Princeton, NJ, 1977.

APPENDIX A

Example Teacher Recording Form

Name ____________________

Judges Recording Form

4th Grade Math

Addition Strand

Stop at item # ________ Group Decision ____________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				

Subtraction Strand

Stop at item # ________ Group Decision ____________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
9.				
10.				
11.				
12.				
13.				
14.				
15.				
16.				

4th Grade Math Continued

Multiplication Strand

Stop at item # __________ Group Decision ______________________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
17.				
18.				
19.				
20.				
21.				
22.				
23.				
24.				
25.				
26.				

Division Strand

Stop at item # __________ Group Decision ______________________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
27.				
28.				
29.				
30.				
31.				
32.				
33.				

4th Grade Math Continued

Numeration Strand

Stop at item # ________ Group Decision ____________________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
34.				
35.				
36.				
37.				
38.				

Geometry Strand

Stop at item # ________ Group Decision ____________________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
39.				
40.				

4th Grade Math Continued

Fractions Strand

Stop at item # ________ Group Decision ____________________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
41.				
42.				
43.				
44.				

Measurement Strand

Stop at item # ________ Group Decision ____________________

Item	Your Decision P F	Group Decision	Incorrect Option Most Often Chosen A B C D E	Group Decision
45.				
46.				
47.				
48.				
49.				
50.				

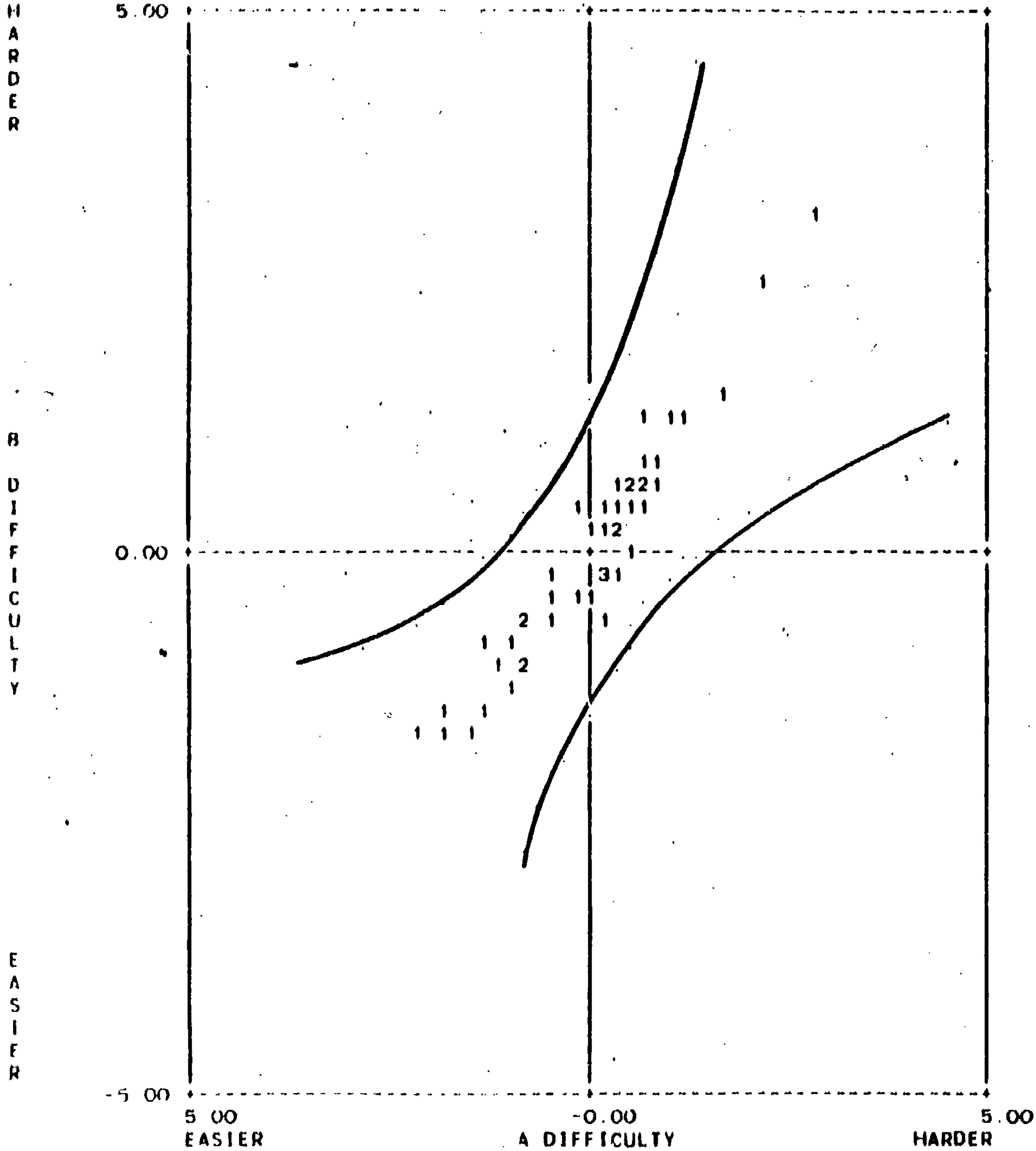
APPENDIX B

Fourth, Fifth and Sixth Grade
Item Invariance Plots

(Note: see Figure 1, page 10, for invariance plot for fifth grade, group 1 vs. all fifth graders.)

GRADE 4 ITEM CALIBRATIONS: COMPARISON OF GROUP 1 TO ALL STUDENTS 50 PAIRS

SCALED PLOT OF TWO ITEM CALIBRATIONS



MEAN A= -0.00 SA= 1.00 R= 0.96 RMAX= 0.97
MEAN B= 0.00 SB= 1.00 PAIRS= 50

GROUP A: ITEM CALIBRATIONS FROM GRADE 4, GROUP 1
GROUP B: ITEM CALIBRATIONS FROM GRADE 4, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS
CREATED BY L.H. LUDLOW: MESA, THE UNIVERSITY OF CHICAGO

GRADE 4 ITEM CALIBRATIONS: COMPARISON OF GROUP 1 TO ALL STUDENTS 50 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

SCALED POINTS BELOW ERROR BAND

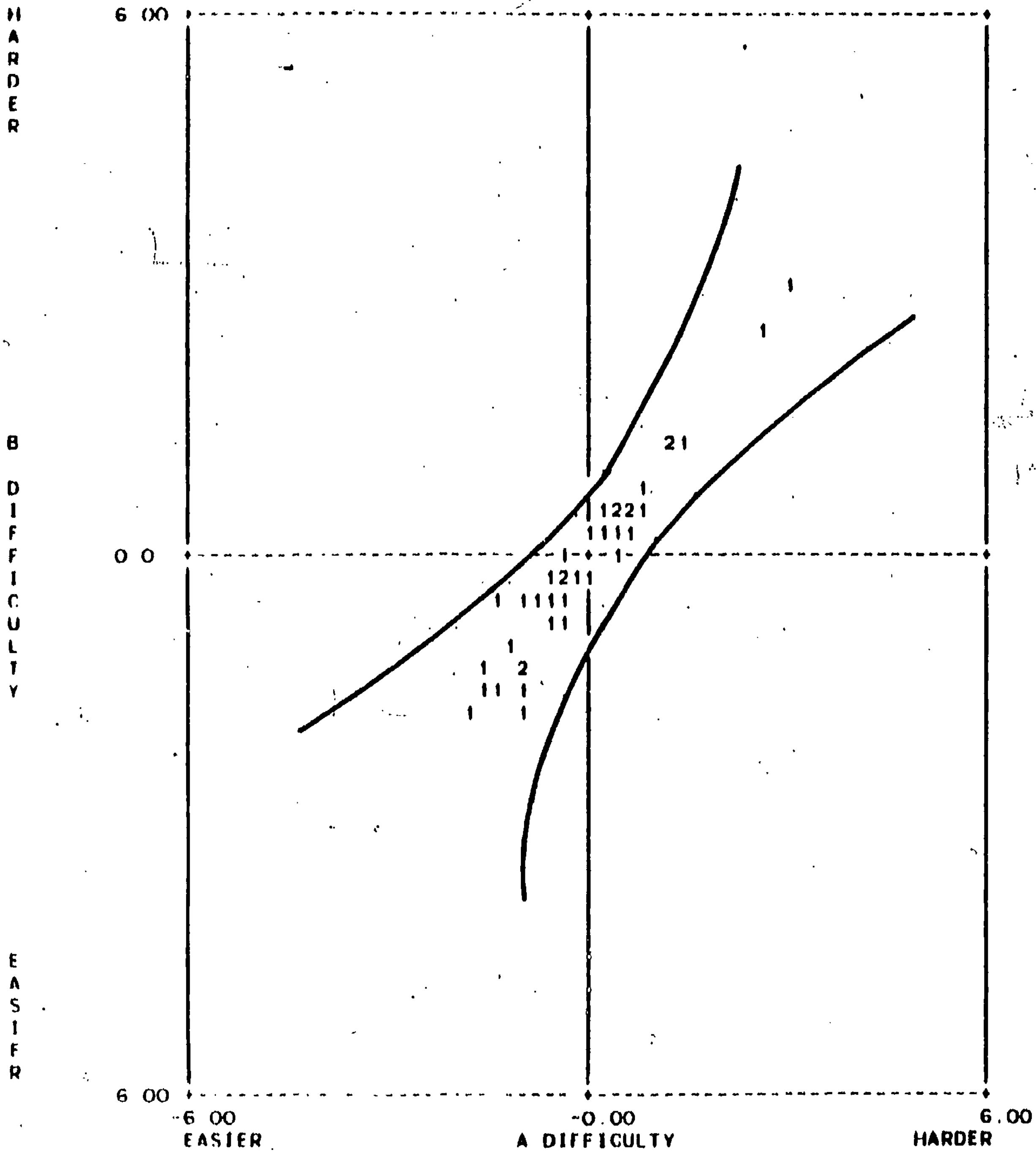
ID	NAME	X AXIS	Y AXIS	DIFF
21	IT21	0.19	-0.54	0.73
12	IT12	0.38	-0.20	0.58
10	IT10	0.48	-0.01	0.49
28	IT28	-0.00	-0.49	0.49
11	IT11	0.09	-0.30	0.39

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
39	IT39	-0.10	0.34	-0.44
26	IT26	0.70	1.25	-0.55

GRADE 4 ITEM CALIBRATIONS: COMPARISON OF GROUP 2 TO ALL STUDENTS 50 PAIRS

SCALED PLOT OF TWO ITEM CALIBRATIONS



MEAN A= -0.00 SA= 1.00 R= 0.97 RMAX= 0.98
MEAN B= 0.00 SB= 1.00 PAIRS= 50

GROUP A: ITEM CALIBRATIONS FROM GRADE 4, GROUP 2
GROUP B: ITEM CALIBRATIONS FROM GRADE 4, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS
CREATED BY L.H. LUDLOW: MESA, THE UNIVERSITY OF CHICAGO

GRADE 4 ITEM CALIBRATIONS: COMPARISON OF GROUP 2 TO ALL STUDENTS 50 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

SCALED POINTS BELOW ERROR BAND

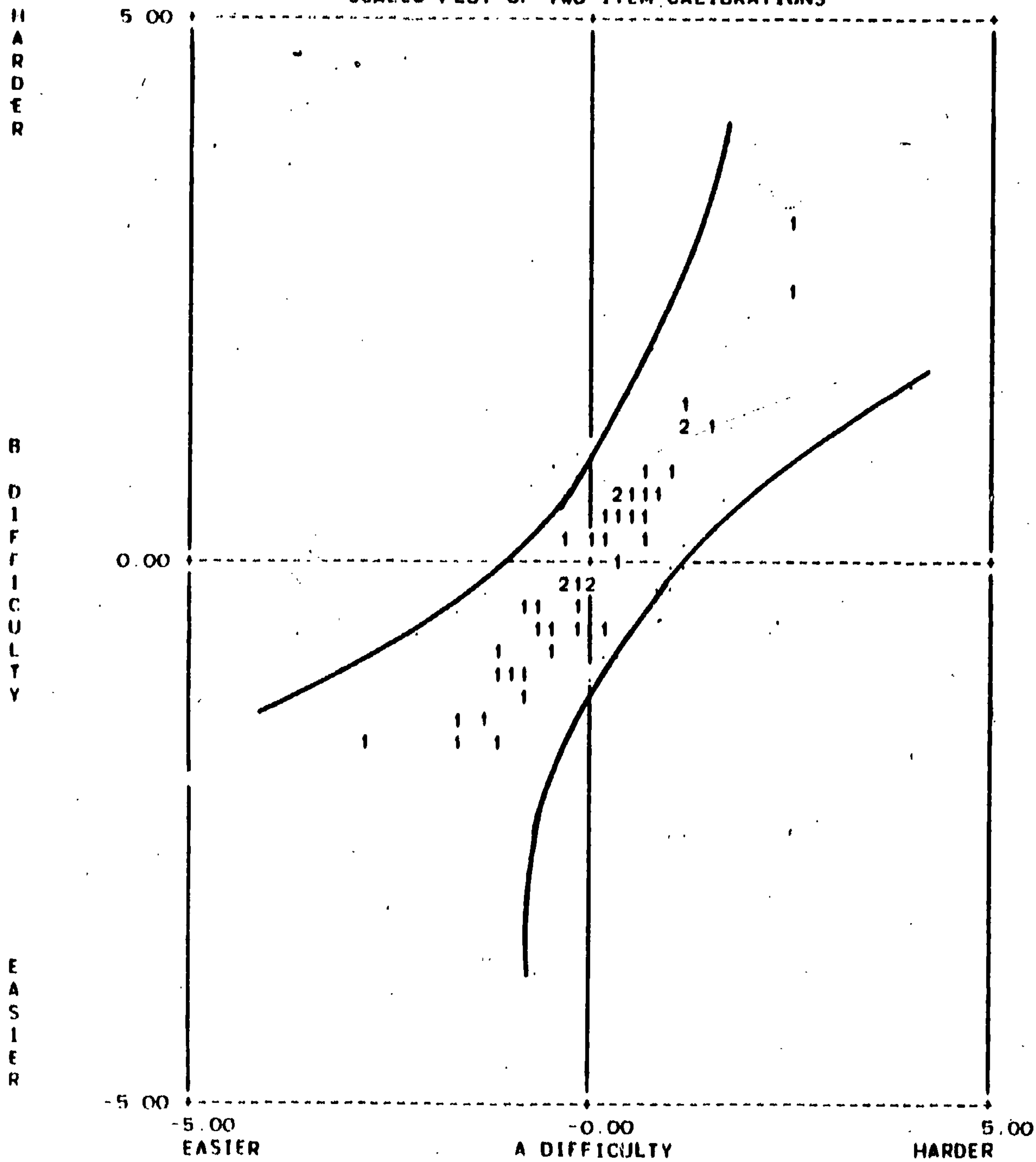
ID	NAME	X AXIS	Y AXIS	DIFF
3	IT03	-0.97	-1.64	0.66
1	IT01	-1.06	1.53	0.46
5	IT05	-0.37	-0.78	0.41
49	IT49	-0.03	-0 33	0.30
50	IT50	0.52	0.28	0.24
6	IT06	-0.37	-0.58	0.22
40	IT40	0.33	0.11	0.21
37	IT37	0.56	0.38	0.18
47	IT47	0.68	0.52	0.16

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
15	IT15	0.05	0.25	-0.20
14	IT14	0.19	0.40	-0.21
11	IT11	-0 63	0.30	0.33
9	IT09	-0.82	-0.49	0.33
10	IT10	-0.42	-0.01	0.40
20	IT20	-1.06	-0.54	-0.52
21	IT21	-1.37	-0 54	-0.83

GRADE 4 ITEM CALIBRATIONS: COMPARISON OF GROUP 3 TO ALL STUDENTS 50 PAIRS

SCALED PLOT OF TWO ITEM CALIBRATIONS



MEAN A= -0.00 SA= 1.00 R= 0.95 RMAX= 0.97
MEAN B= 0.00 SB= 1.00 PAIRS= 50

GROUP A: ITEM CALIBRATIONS FROM GRADE 4, GROUP 3
GROUP B: ITEM CALIBRATIONS FROM GRADE 4, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS

GRADE 4 ITEM CALIBRATIONS: COMPARISON OF GROUP 3 TO ALL STUDENTS 50 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

SCALED POINTS BELOW ERROR BAND

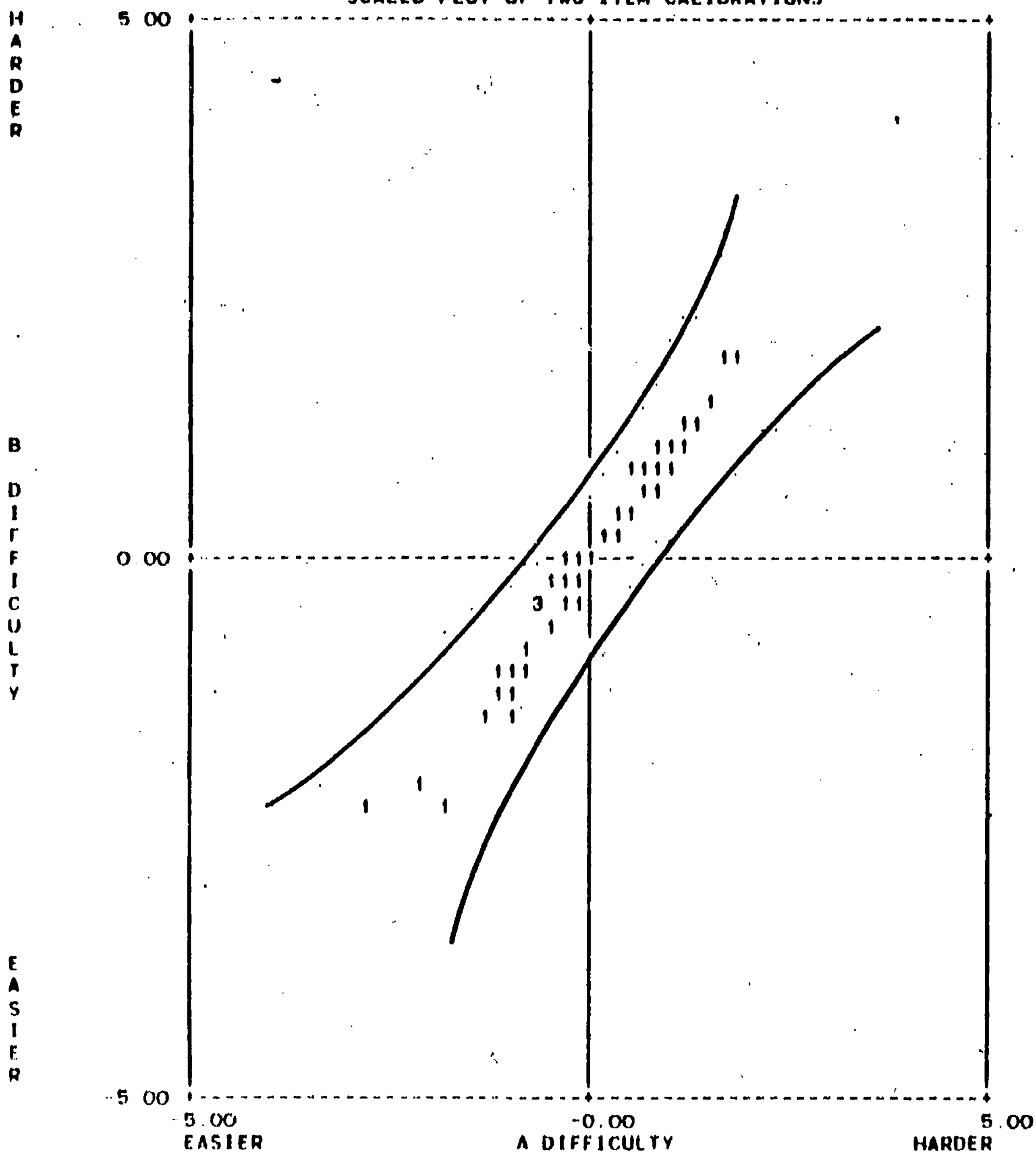
ID	NAME	X AXIS	Y AXIS	DIFF
20	IT20	0.11	-0.54	0.65
10	IT10	0.38	-0.01	0.40
21	IT21	-0.17	-0.54	0.37
15	IT15	0.60	0.25	0.36
9	IT09	-0.17	-0.49	0.32
11	IT11	-0.03	-0.30	0.27

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
37	IT37	0.11	0.38	-0.27
28	IT28	-0.90	-0.49	-0.41
40	IT40	-0.32	0.11	-0.43

GRADE 5 ITEM CALIBRATIONS: COMPARISON OF GROUP 2 TO ALL STUDENTS 55 PAIRS

SCALED PLOT OF TWO ITEM CALIBRATIONS



MEAN A= 0.00 SA= 1.00 R= 0.99 · RMAX= 0.99
MEAN B= 0.00 SB= 1.00 PAIRS= 55

GROUP A: ITEM CALIBRATIONS FROM GRADE 5, GROUP 2
GROUP B: ITEM CALIBRATIONS FROM GRADE 5, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS
CREATED BY L.H. LUDLOW: MESA, THE UNIVERSITY OF CHICAGO

GRADE 5 ITEM CALIBRATIONS: COMPARISON OF GROUP 2 TO ALL STUDENTS 55 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS OO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

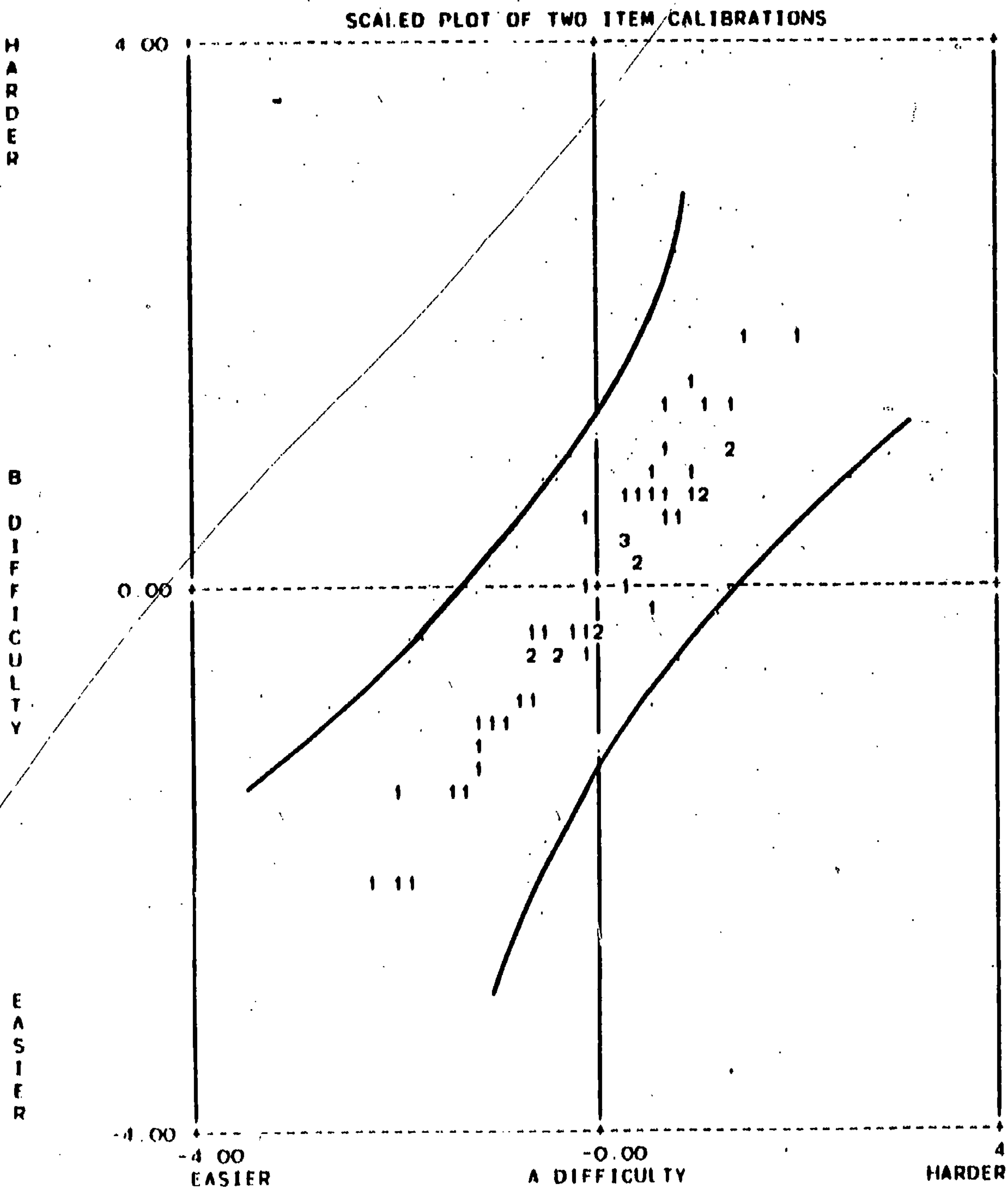
SCALED POINTS BELOW ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
3	IT03	-1.08	-1.53	0.45
38	IT38	-0.25	-0.44	0.19
49	IT49	0.79	0.63	0.16

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
23	IT23	-0.06	0.10	-0.16
7	IT07	-0.28	-0.08	-0.19
8	IT08	0.54	0.73	-0.19
17	IT17	0.82	1.02	-0.19
29	IT29	-0.60	-0.39	-0.21

GRADE 5 ITEM CALIBRATIONS: COMPARISON OF GROUP 3 TO ALL STUDENTS 55 PAIRS



MEAN A= -0.00 SA= 1.00 R= 0.96 RMAX= 0.97
MEAN B= 0.00 SB= 1.00 PAIRS= 55

GROUP A: ITEM CALIBRATIONS FROM GRADE 5, GROUP 3
GROUP B: ITEM CALIBRATIONS FROM GRADE 5, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS
CREATED BY L.H. LUDLOW: MESA, THE UNIVERSITY OF CHICAGO

GRADE 5 ITEM CALIBRATIONS: COMPARISON OF GROUP 3 TO ALL STUDENTS 55 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

SCALED POINTS BELOW ERROR BAND

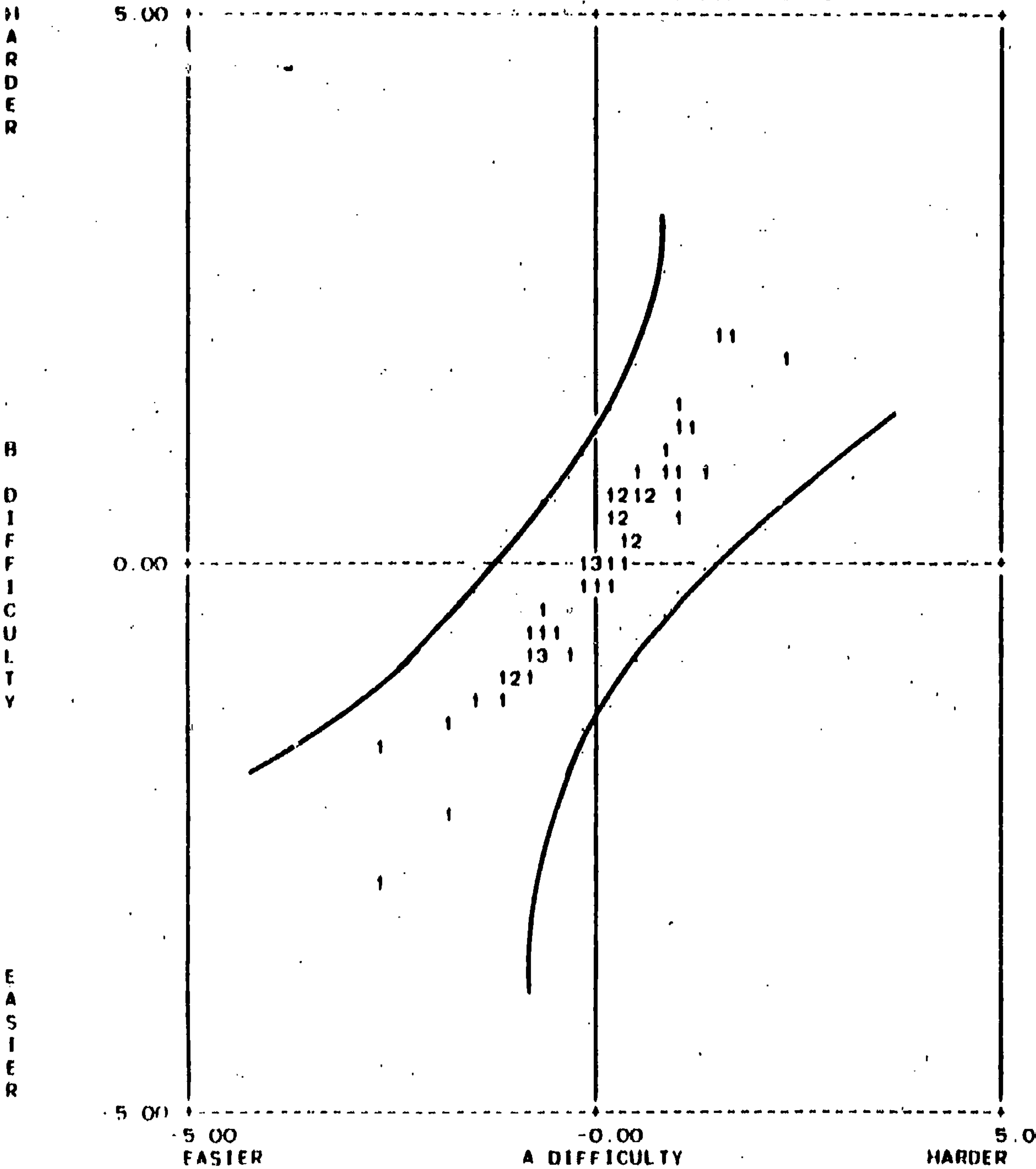
ID	NAME	X AXIS	Y AXIS	DIFF
7	IT07	0.59	-0.08	0.67

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
25	IT25	0.23	0.75	-0.52
45	IT45	0 68	1.30	-0.62
33	IT33	-0 19	0.45	-0.64

COMPARISON OF 'GROUP 1' TO 'ALL STUDENTS' ITEM CALIBRATIONS, GRADE 6 60 PAIRS

SCALED PLOT OF TWO ITEM CALIBRATIONS



MEAN A= 0.00 SA= 1.00 R= 0.96 RMAX= 0.96
MEAN B= -0.00 SB= 1.00 PAIRS= 60

GROUP A: ITEM CALIBRATIONS FROM GRADE 6, GROUP 1
GROUP B: ITEM CALIBRATIONS FROM GRADE 6, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS
CREATED BY L.H. LUDLOW: MESA, THE UNIVERSITY OF CHICAGO

COMPARISON OF 'GROUP 1' TO 'ALL STUDENTS' ITEM CALIBRATIONS. GRADE 6 60 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WI '. NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

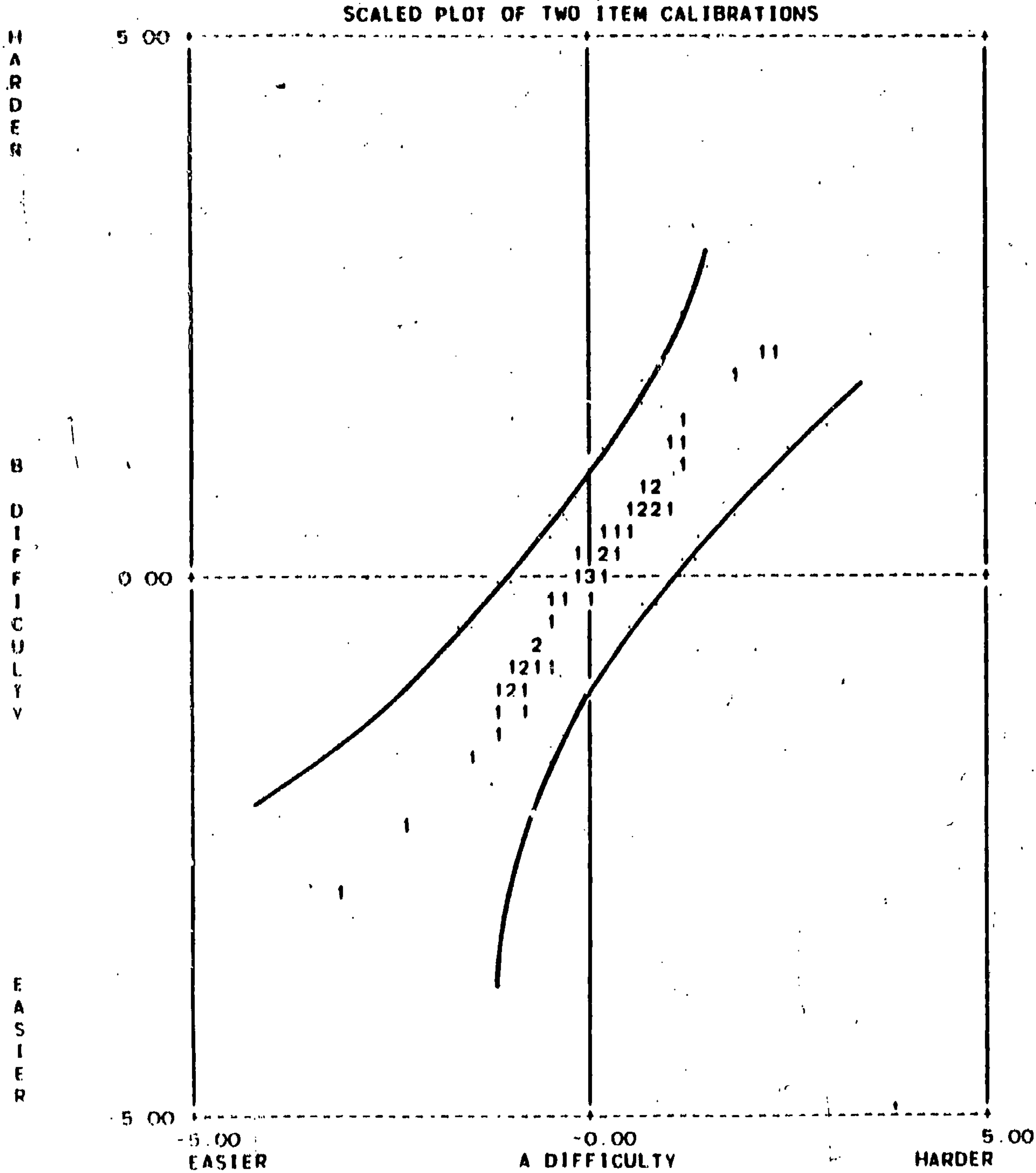
SCALED POINTS BELOW ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
47	IT47	0.92	0.44	0.48
13	IT13	-0.32	-0.76	0.44
7	IT07	0.58	0.15	0.43

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
44	IT44	0.28	0.68	-0.40
34	IT34	0.18	0.60	-0.42

GRADE 6 ITEM CALIBRATIONS: COMPARISON OF GROUP 2 TO ALL STUDENTS 60 PAIRS



MEAN A= 0.00 SA= 1.00 R= 0.98 RMAX= 0.98
MEAN B= -0.00 SB= 1.00 PAIRS= 60

GROUP A: ITEM CALIBRATIONS FROM GRADE 6, GROUP 2
GROUP B: ITEM CALIBRATIONS FROM GRADE 6, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS
CREATED BY L.H. LUDLOW: MESA, THE UNIVERSITY OF CHICAGO

GRADE 6 ITEM CALIBRATIONS: COMPARISON OF GROUP 2 TO ALL STUDENTS 60 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

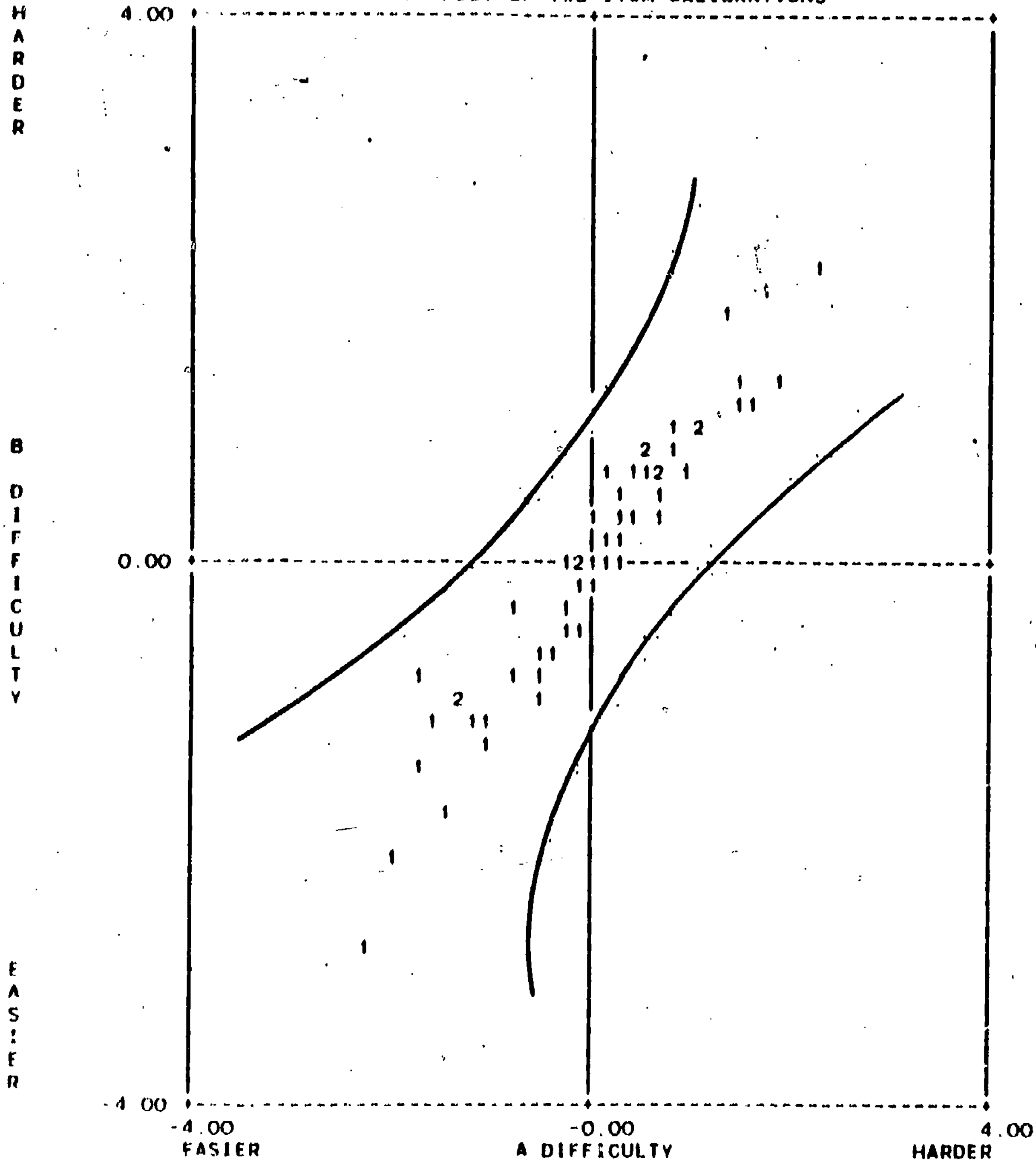
SCALED POINTS BELOW ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
29	IT29	-0.56	-0.91	0.35

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
37	IT37	-0.50	-0.22	-0.28
7	IT07	-0.16	0.15	-0.31
40	IT40	0.59	0.92	-0.33

SCALED PLOT OF TWO ITEM CALIBRATIONS



MEAN A= 0.00 SA= 1.00 R= 0.96 RMAX= 0.97
MEAN B= -0.00 SB= 1.00 PAIRS= 60

GROUP A: ITEM CALIBRATIONS FROM GRADE 6, GROUP 3
GROUP B: ITEM CALIBRATIONS FROM GRADE 6, ALL STUDENTS
THE CONFIDENCE INTERVAL REPRESENTS 4. STANDARD ERRORS
CREATED BY L.H. LUDLOW: MESA, THE UNIVERSITY OF CHICAGO

GRADE 6 ITEM CALIBRATIONS: COMPARISON OF GROUP 3 TO ALL STUDENTS 60 PAIRS

NOTE: THE DOTS REPRESENTING THE STANDARD ERROR BANDS ARE ENTERED INTO THE GRAPH LOCATIONS AFTER THE NUMERALS. THIS MEANS THAT PAIRS FALLING AT THE SAME LOCATION AS DO DOTS ARE NOT SHOWN AND THE TOTAL NUMBER OF OCCURENCES THAT COULD BE COUNTED IN THE PLOT WILL NOT EQUAL THE NUMBER OF PAIRS ACTUALLY PLOTTED. THIS CHOICE WAS MADE AS AN AID IN HIGHLIGHTING WHERE THE BANDS LIE AND WHICH PAIRS CLEARLY LIE OUTSIDE THE BAND. PAIRS FALLING ON OR OUTSIDE THE BANDS ARE LISTED IN THE FOLLOWING TABLES

SCALED POINTS BELOW ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
59	IT59	-0.13	-0.57	0.43
17	IT17	-0.59	-1.00	0.41

SCALED POINTS ABOVE ERROR BAND

ID	NAME	X AXIS	Y AXIS	DIFF
32	IT32	0.36	0.68	-0.32
48	IT48	1.30	1.79	-0.49
39	IT39	0.15	0.66	-0.51
31	IT31	-0.86	-0.29	-0.57
29	IT29	-1.77	-0.91	-0.86

APPENDIX C

Fifth and Sixth Ability Estimate Distributions
Relative to Teacher Determined Retest Band

(Note: see pages 15, 16 and 17 for fourth grade results.)

GRADE 5 MATH. GROUP-1 STUDENTS

55 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
54	0	4.99	4.57	1.01	
53	0	4.10	3.84	0.73	
52	0	3.57	3.39	0.60	
51	0	3.18	3.06	0.53	
50	0	2.88	2.80	0.49	
49	0	2.62	2.58	0.45	
48	0	2.41	2.38	0.43	
47	0	2.21	2.21	0.41	
46	0	2.04	2.05	0.39	
45	0	1.88	1.90	0.38	
44	0	1.73	1.77	0.36	
43	0	1.60	1.64	0.35	
42	0	1.47	1.52	0.35	
41	0	1.34	1.40	0.34	
40	0	1.23	1.29	0.33	
39	0	1.11	1.18	0.33	
38	0	1.01	1.07	0.32	
17	0	0.90	0.97	0.32	
36	1	0.80	0.87	0.32	X
35	0	0.70	0.77	0.31	
34	0	0.60	0.67	0.31	
33	0	0.51	0.58	0.31	
32	0	0.41	0.48	0.31	
31	2	0.32	0.39	0.31	XX
30	0	0.23	0.30	0.31	
29	1	0.14	0.20	0.31	X
28	2	0.05	0.11	0.31	XX
27	2	-0.05	0.02	0.31	XX
26	3	-0.14	-0.08	0.31	XXX
25	4	-0.23	-0.17	0.31	XXXX
24	1	-0.32	-0.27	0.31	X
23	7	-0.41	-0.37	0.31	XXXXXXX
22	0	-0.51	-0.47	0.32	
21	4	-0.60	-0.57	0.32	XXXX
20	1	-0.70	-0.67	0.32	X
19	1	-0.80	-0.77	0.33	X
18	2	-0.90	-0.88	0.33	XX
17	0	-1.01	-0.99	0.34	
16	1	-1.11	-1.11	0.34	X
15	0	-1.23	-1.22	0.35	
14	1	-1.34	-1.35	0.35	X
13	1	-1.47	-1.48	0.36	X
12	0	-1.60	-1.61	0.37	
11	0	-1.73	-1.76	0.38	

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
10	1	-1.88	-1.91	0.40	X
9	0	-2.04	-2.07	0.41	
8	0	-2.21	-2.25	0.43	
7	0	-2.41	-2.45	0.4[illegible]	
6	0	-2.63	-2.67	0.48	
5	0	-2.88	-2.92	0.52	
4	0	-3.18	-3.22	0.56	
3	0	-3.57	-3.59	0.65	
2	0	-4.10	-4.09	0.76	
1	0	-4.99	-4.89	1.07	

GRADE 5 MATH, GROUP-2 STUDENTS

55 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
54	0	5.12	4.68	1.01	
53	1	4.21	3.94	0.74	X
52	0	3.66	3.49	0.60	
51	0	3.27	3.16	0.54	
50	0	2.96	2.89	0.49	
49	0	2.70	2.66	0.46	
48	3	2.47	2.46	0.43	XXX
47	1	2.27	2.29	0.41	X
46	2	2.09	2.12	0.40	XX
45	2	1.93	1.97	0.38	XX
44	0	1.78	1.83	0.37	
43	4	1.64	1.70	0.36	XXXX
42	1	1.51	1.57	0.35	X
41	7	1.38	1.45	0.35	XXXXXXX
40	3	1.26	1.33	0.34	XXX
39	9	1.14	1.21	0.33	XXXXXXXXX
38	3	1.03	1.10	0.33	XXX
37	10	0.92	1.00	0.33	XXXXXXXXXX
36	6	0.82	0.89	0.32	XXXXXX
35	2	0.72	0.79	0.32	XX
34	11	0.62	0.69	0.32	XXXXXXXXXXX
33	6	0.52	0.58	0.32	XXXXXX
32	3	0.42	0.48	0.32	XXX
31	4	0.33	0.39	0.31	XXXX
30	5	0.23	0.29	0.31	XXXXX
29	7	0.14	0.19	0.31	XXXXXXX
28	6	0.05	0.09	0.31	XXXXXX
27	2	-0.05	-0.01	0.31	XX
26	3	-0.14	-0.11	0.31	XXX
25	4	-0.23	-0.21	0.32	XXXX
24	5	-0.33	-0.31	0.32	XXXXX
23	5	-0.42	-0.41	0.32	XXXXX
22	0	-0.52	-0.51	0.32	
21	1	-0.62	-0.61	0.32	X
20	1	-0.72	-0.72	0.33	X
19	0	-0.82	-0.82	0.33	
18	1	-0.92	-0.93	0.33	X
17	0	-1.03	-1.05	0.34	
16	1	-1.14	-1.16	0.34	X
15	2	-1.26	-1.28	0.35	XX
14	1	-1.38	-1.41	0.36	X
13	0	-1.51	-1.54	0.36	
12	0	-1.64	-1.67	0.37	
11	0	-1.78	-1.81	0.38	

: GRADE 5 MATH, GROUP-3 STUDENTS

55 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
54	0	4.71	4.44	1.01	
53	0	3.87	3.71	0.73	
52	0	3.37	3.27	0.60	
51	0	3.01	2.95	0.53	
50	0	2.72	2.69	0.48	
49	0	2.48	2.47	0.45	
48	0	2.27	2.28	0.42	
47	0	2.09	2.11	0.40	
46	1	1.93	1.95	0.38	X
45	1	1.78	1.81	0.37	X
44	0	1.64	1.68	0.36	
43	0	1.51	1.55	0.35	
42	0	1.39	1.43	0.34	
41	0	1.27	1.32	0.33	
40	0	1.16	1.21	0.33	
39	0	1.05	1.11	0.32	
38	1	0.95	1.00	0.32	X
37	1	0.85	0.91	0.31	X
36	0	0.75	0.81	0.31	
35	0	0.66	0.71	0.31	
34	0	0.57	0.62	0.31	
33	0	0.48	0.53	0.30	
32	2	0.39	0.44	0.30	XX
31	2	0.30	0.34	0.30	XX
30	2	0.22	0.25	0.30	XX
29	4	0.13	0.16	0.30	XXXX
28	4	0.04	0.07	0.30	XXXX
27	3	-0.04	-0.02	0.30	XXX
26	2	-0.13	-0.11	0.30	XX
25	3	-0.22	-0.20	0.30	XXX
24	4	-0.30	-0.29	0.30	XXXX
23	6	-0.39	-0.38	0.31	XXXXXX
22	3	-0.48	-0.47	0.31	XXX
21	1	-0.57	-0.57	0.31	X
20	0	-0.66	-0.67	0.31	
19	2	-0.75	-0.77	0.32	XX
18	1	-0.85	-0.87	0.32	X
17	0	-0.95	-0.97	0.32	
16	0	-1.05	-1.08	0.33	
15	4	-1.16	-1.19	0.34	XXXX
14	2	-1.27	-1.30	0.34	XX
13	0	-1.39	-1.42	0.35	
12	0	-1.51	-1.55	0.36	
11	0	-1.64	-1.68	0.37	

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
10	0	-1.78	-1.82	0.38	
9	0	-1.93	-1.98	0.40	
8	0	-2.09	-2.14	0.41	
7	0	-2.27	-2.32	0.43	
6	0	-2.48	-2.52	0.46	
5	0	-2.72	-2.75	0.49	
4	1	-3.01	-3.02	0.54	X
3	0	-3.37	-3.35	0.62	
2	0	-3.87	-3.80	0.73	
1	0	-4.71	-4.55	1.03	

GRADE 6 MATH. GROUP-1 STUDENTS

60 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
59	0	5.19	4.76	1.02	
58	0	4.28	4.02	0.74	
57	0	3.75	3.57	0.61	
56	0	3.36	3.23	0.54	
55	0	3.05	2.97	0.49	
54	0	2.79	2.74	0.46	
53	0	2.58	2.54	0.43	
52	0	2.38	2.37	0.41	
51	0	2.21	2.21	0.39	
50	0	2.05	2.06	0.38	
49	0	1.90	1.93	0.36	
48	0	1.76	1.80	0.35	
47	0	1.63	1.68	0.34	
46	0	1.51	1.56	0.34	
45	1	1.40	1.45	0.33	X
44	1	1.29	1.34	0.32	X
43	0	1.18	1.24	0.32	
42	2	1.08	1.14	0.31	XX
41	0	0.98	1.04	0.31	
40	0	0.88	0.95	0.31	
39	1	0.79	0.85	0.30	X
38	0	0.70	0.76	0.30	
37	3	0.60	0.67	0.30	XXX
36	1	0.52	0.58	0.30	X
35	0	0.43	0.49	0.30	
34	2	0.34	0.41	0.30	XX
33	3	0.26	0.32	0.30	XXX
32	3	0.17	0.23	0.30	XXX
31	2	0.08	0.14	0.30	XX
30	2	0.0	0.06	0.30	XX
29	1	-0.08	-0.03	0.30	X
28	0	-0.17	-0.12	0.30	
27	1	-0.26	-0.21	0.30	X
26	1	-0.34	-0.30	0.30	X
25	2	-0.43	-0.39	0.30	XX
24	4	-0.52	-0.48	0.30	XXXX
23	0	-0.60	-0.57	0.31	
22	0	-0.70	-0.67	0.31	
21	1	-0.79	-0.76	0.31	X
20	1	-0.88	-0.86	0.32	X
19	0	-0.98	-0.96	0.32	
18	1	-1.08	-1.07	0.33	X
17	0	-1.18	-1.18	0.33	
16	0	-1.29	-1.29	0.34	

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
15	0	-1.40	-1.40	0.34	
14	0	-1.51	-1.52	0.35	
13	0	-1.63	-1.65	0.36	
12	0	-1.76	-1.78	0.37	
11	0	-1.90	-1.93	0.38	
10	0	-2.05	-2.08	0.40	
9	1	-2.21	-2.24	0.41	X
8	0	-2.38	-2.42	0.43	
7	0	-2.58	-2.62	0.46	
6	0	-2.79	-2.84	0.48	
5	0	-3.05	-3.09	0.52	
4	0	-3.36	-3.39	0.57	
3	0	-3.75	-3.77	0.65	
2	0	-4.28	-4.26	0.76	
1	0	-5.19	-5.05	1.06	

GRADE 6 MATH, GROUP-2 STUDENTS

60 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
59	0	4.98	4.69	1.02	
58	0	4.12	3.95	0.74	
57	0	3.60	3.49	0.61	
56	0	3.23	3.16	0.54	
55	1	2.93	2.89	0.49	X
54	0	2.69	2.67	0.46	
53	2	2.47	2.47	0.43	XX
52	0	2.29	2.29	0.41	
51	1	2.12	2.14	0.39	X
50	1	1.97	1.99	0.37	X
49	0	1.83	1.85	0.36	
48	3	1.69	1.73	0.35	XXX
47	2	1.57	1.60	0.34	XX
46	1	1.45	1.49	0.33	X
45	0	1.34	1.38	0.33	
44	2	1.24	1.27	0.32	XX
43	3	1.13	1.17	0.32	XXX
42	3	1.04	1.07	0.31	XXX
41	4	0.94	0.98	0.31	XXXX
40	1	0.85	0.88	0.30	X
39	3	0.76	0.79	0.30	XXX
38	5	0.67	0.70	0.30	XXXXX
37	1	0.58	0.61	0.30	X
36	4	0.50	0.53	0.29	XXXX
35	5	0.41	0.44	0.29	XXXXX
34	2	0.33	0.36	0.29	XX
33	5	0.25	0.27	0.29	XXXXX
32	1	0.16	0.19	0.29	X
31	1	0.08	0.10	0.29	X
30	3	0.0	0.02	0.29	XXX
29	0	-0.08	-0.07	0.29	
28	5	-0.16	-0.15	0.29	XXXXX
27	4	-0.25	-0.23	0.29	XXXX
26	1	-0.33	-0.32	0.29	X
25	2	-0.41	-0.41	0.29	XX
24	3	-0.50	-0.49	0.30	XXX
23	2	-0.58	-0.58	0.30	XX
22	3	-0.67	-0.67	0.30	XXX
21	4	-0.76	-0.76	0.30	XXXX
20	1	-0.85	-0.85	0.31	X
19	1	-0.94	-0.95	0.31	X
18	0	-1.04	-1.05	0.31	
17	1	-1.13	-1.15	0.32	X
16	1	-1.24	-1.25	0.32	X

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
15	0	-1.34	-1.36	0.33	
14	1	-1.45	-1.47	0.34	X
13	2	-1.57	-1.59	0.35	XX
12	1	-1.69	-1.71	0.36	X
11	0	-1.83	-1.84	0.37	
10	0	-1.97	-1.98	0.38	
9	0	-2.12	-2.13	0.40	
8	0	-2.29	-2.30	0.42	
7	0	-2.47	-2.48	0.44	
6	0	-2.69	-2.69	0.47	
5	0	-2.93	-2.93	0.51	
4	0	-3.23	-3.21	0.55	
3	0	-3.60	-3.57	0.64	
2	0	-4.12	-4.05	0.75	
1	0	-4.98	-4.84	1.06	

GRADE 6 MATH, GROUP-3 STUDENTS

60 ITEMS MAX. OF 2 CATEGORIES

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
59	0	4.99	4.70	1.01	
58	0	4.12	3.96	0.74	
57	0	3.61	3.51	0.60	
56	0	3.23	3.18	0.54	
55	0	2.94	2.92	0.49	
54	0	2.69	2.69	0.46	
53	0	2.48	2.50	0.43	
52	0	2.29	2.32	0.41	
51	0	2.12	2.16	0.39	
50	0	1.97	2.01	0.38	
49	1	1.83	1.88	0.36	X
48	0	1.70	1.75	0.35	
47	0	1.57	1.62	0.34	
46	2	1.46	1.51	0.34	XX
45	1	1.35	1.40	0.33	X
44	2	1.24	1.29	0.32	XX
43	0	1.14	1.19	0.32	
42	2	1.04	1.09	0.31	XX
41	1	0.94	0.99	0.31	X
40	0	0.85	0.89	0.31	
39	1	0.76	0.80	0.30	X
38	3	0.67	0.71	0.30	XXX
37	1	0.58	0.62	0.30	X
36	2	0.50	0.53	0.30	XX
35	4	0.41	0.44	0.30	XXXX
34	2	0.33	0.36	0.29	XX
33	4	0.25	0.27	0.29	XXXX
32	1	0.16	0.18	0.29	X
31	7	0.08	0.10	0.29	XXXXXXX
30	1	0.0	0.01	0.29	X
29	2	-0.08	-0.07	0.29	XX
28	3	-0.16	-0.16	0.29	XXX
27	4	-0.25	-0.25	0.29	XXXX
26	3	-0.33	-0.33	0.30	XXX
25	1	-0.41	-0.42	0.30	X
24	0	-0.50	-0.51	0.30	
23	0	-0.58	-0.60	0.30	
22	1	-0.67	-0.69	0.30	X
21	0	-0.76	-0.78	0.31	
20	1	-0.85	-0.88	0.31	X
19	1	-0.94	-0.98	0.31	X
18	0	-1.04	-1.08	0.32	
17	0	-1.14	-1.18	0.32	
16	0	-1.24	-1.28	0.33	

SCALE SCORE EQUIVALENCE TABLE

RAW SCORE	COUNT	PROX POSITION	UCON POSITION	UCON ERROR	PERSON FREQUENCY DISTRIBUTION:X=1 PERSON
15	0	-1.35	-1.39	0.33	
14	0	-1.46	-1.51	0.34	
13	0	-1.57	[illegible]	0.35	
12	1	-1.70	-1.75	0.36	X
11	0	-1.83	-1.88	0.37	
10	0	-1.97	-2.02	0.38	
9	0	-2.12	-2.17	0.39	
8	0	-2.29	-2.34	0.41	
7	0	-2.48	-2.52	0.43	
6	0	-2.69	-2.72	0.46	
5	0	-2.94	-2.94	0.49	
4	0	-3.23	-3.21	0.54	
3	0	-3.61	-3.55	0.62	
2	0	-4.12	-4.00	0.73	
1	0	-4.99	-4.74	1.03	