DOCUMENT RESUME

ED 348 392 TM 018 815

AUTHOR Lee, Ong Kim; Wright, Benjamin D.
TITLE Mathematics and Reading Test Equating.
INSTITUTION Chicago Panel on Public School Policy and Finance, IL.; Chicago Public Schools, Ill.; Chicago Univ., IL. Center for School Improvement.
SPONS AGENCY Spencer Foundation, Chicago, Ill.
PUB DATE Apr 92
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Ability; Academic Achievement; Achievement Tests; Comparative Testing; *Educational Change; Elementary Secondary Education; *Equated Scores; *Mathematics Tests; *Reading Tests; *Scaling; Urban Schools
IDENTIFIERS Chicago Public Schools IL; *Iowa Tests of Basic Skills; Logits; Rasch Model; Reform Efforts; *Test Equivalence

ABSTRACT

As part of a larger project to assess changes in student learning resulting from school reform, this study equates levels 6 through 14 of the mathematics and reading comprehension components of Form 7 of the Iowa Tests of Basic Skills (ITBS) with levels 7 through 14 of the mathematics and reading comprehension components of the CPS90 (another version of the ITBS), using a Rasch analysis. The analysis results in the common calibration of all 1,031 mathematics items found in the 17 levels of the two test forms to define a mathematics variable and all 602 reading items to define a reading variable. Each item in each subject obtains a person free calibration (in logits) of its own level of difficulty on one common scale linking all items of that subject. The 17 levels of the two tests were successfully equated so that a person taking the CPS90 or Form 7 (or a combination of items from the forms targeted at his or her ability level) will obtain statistically equivalent measures of ability. Logit measures give a more accurate picture of student rate of growth than do grade equivalents, with rates of growth highest at the lower grades and decreasing in the higher grades. Four tables, 13 figures, and 6 references are included. An appendix lists the criterion definitions of variables. (SLD)

MATHEMATICS AND READING TEST EQUATING

Ong Kim Lee
University of Chicago

Benjamin D. Wright
University of Chicago

Paper presented at the annual meeting of the
American Educational Research Association
April 1992, San Francisco, CA.

Reading and Mathematics Equating Study[1]

Introduction

This study is part of a larger project intended to assess school change in student learning as a result of school reform. In order to do this we want to look at improvements in students' academic achievement over time. Current policy of the Chicago Public Schools is to change the form of the ITBS each year. Sufficient number of anomalies appeared after the first change in forms that prompted questions on adequacy of equating by grade equivalents, at least as applied to Chicago schools. Unless these test forms are equated, it is not possible to compare student performances from year to year to determine school change. The Easton, Dean, and Bryk paper (1991) points out that earlier studies (Frank and Seltzer, 1990) using longitudinal data bases had shown the inadequacy of the grade equivalent scores for determining growth. Schulz, Shen and Wright (1990), point out that the construction of the grade equivalent metric is such that students show an average annual gain of one grade equivalent irrespective of their actual changes in ability. The incorporation of time into grade equivalents removes the possibility of determining growth rates.

This study equates levels 6 through 14 of the Mathematics and Reading Comprehension components of the Iowa Tests of Basic Skills (ITBS Form 7) with levels 7 through 14 of the Mathematics and Reading Comprehension components of the CPS90 (another version of the ITBS), using Rasch analysis (Wright & Douglas, 1975, Wright, B.D., 1977, Wright & Stone, 1979). The analysis results in the common calibration of all 1031 mathematics items found in the 17 levels of the two test forms to define a math variable, and all 602 reading items to define a reading variable. Each item in each subject obtains a person-free calibration (in logits) of its own level of difficulty on the one common scale linking all items of that subject.

---

[1] This project is a collaboration between the Center for School Improvement under the directorship of Professor Anthony S. Bryk at the University of Chicago, The Chicago Panel on Public School Policy and Finance under the directorship of John Q. Easton, and the Chicago Public Schools, represented by Carole Perlman, and is supported by a grant from the Spencer Foundation to The Chicago Panel on Public School Policy and Finance.

We owe special thanks to Professor Anthony S. Bryk for his useful pointers in the course of the analysis and for his input and comments on the draft of this paper. We would also like to thank Paul Dean, John Q. Easton, Kenneth Frank, David Kerbow, Julia B. Smith and Arie van der Ploeg for their ideas and comments.

Design and Method

Test linking in this study was done with common persons and common items. The design is in Figure 1. Each arrow represents a group of persons taking a pair of tests. The initial design took into consideration the need to minimize the number of students involved in the study. Levels 10, 12 and 14 of Form 7 and Levels 10 and 12 of CPS90 were not administered because their items appear in levels 9, 11 and 13 of their respective forms. Level 14 of Form 7 shares 67% of its items with Level 13. Linking was strengthened by adding existing data[2] for Levels 10 and 12 of both Forms and Level 14 of Form 7. These data are from the regular student testing, from schools used in the study. Table 1 lists the number of items and number of students used in the analysis, for each of the test levels.

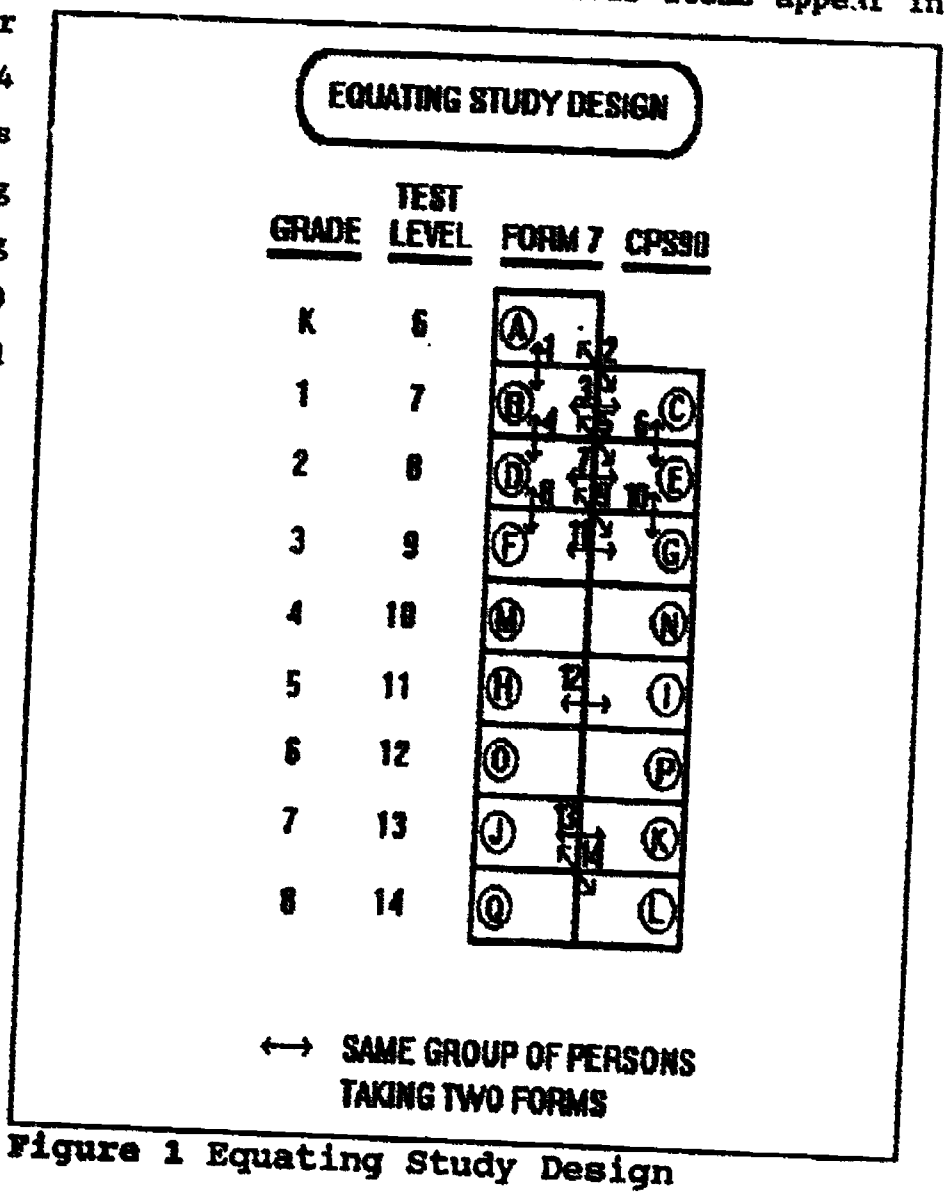


Figure 1 Equating Study Design

The Calibration Matrices

The data were cleaned in four stages: (1) Only response strings marked valid as defined by standard Chicago

---

[2] We are grateful to Paul Dean for extracting response strings for Levels 10, 12, and 14 from his files to provide additional data for the linking.

Table 1 Number of Items and Persons by Test Level

Test	MATHEMATICS				READING			
	Form 7		CPS90		Form 7		CPS90	
	Items	Persons	Items	Persons	Items	Persons	Items	Persons
1. Level 6	33	265	-	-	70	238	-	-
2. Level 7	81	459	82	365	66	466	56	383
3. Level 8	88	502	96	550	67	566	61	544
4. Level 9	90	282	86	380	44	299	44	453
5. Level 10	99	196	95	205	49	227	49	236
6. Level 11	109	156	101	157	54	177	54	209
7. Level 12	114	196	109	198	56	236	56	238
8. Level 13	117	370	113	178	57	329	57	151
9. Level 14	121	200	117	191	58	239	58	175

Public Schools' procedures[3], were included[4]; (2) Response strings showing series of zeroes and/or same responses for 25% or greater of the total number of items, were dropped; (3) Misfitting persons on Rasch estimates were removed; and (4) Persons with large standardized differences in performances on their pair of tests were removed. About 12% of data were lost through cleaning. After data cleaning, the item response strings were linked into one giant calibration matrix such that strings for a person taking two tests are aligned into the same row and responses to a given item fall into the same column. This is diagrammed in Figure 2.

Tests are arranged from the lowest test levels of Form 7 and CPS90 to the highest. This results in a Mathematics calibration matrix with 1031 different items taken by 2995 different persons, and a Reading calibration matrix with 602 different items taken by 3159 persons.

Notice that these calibration matrices are only 15 percent filled with data. Nevertheless, reliable equating was accomplished from Grade 1 through Grade 8. Rasch equating does not need complete data to calibrate items successfully onto a common scale or to obtain good estimates of person measures.

---

[3] Test strings were flagged when they failed evaluation under one or more of the following criteria: (1) More than 3 multiples; (2) 50-70% lite and > 1 embedded omits; (3) 80-100% lite and > 0 embedded omits.

[4] We would like to thank the Chicago Public Schools for doing the first stage of cleaning by flagging invalid response strings.

Each matrix was Rasch-analyzed in a one-step equating procedure and all tests were placed on a common logit scale. Items calibrations in difficulty logits, the log odds of an item provoking failure from a person with ability equal to the scale zero. We now have a bank of 1031 Mathematics items and another bank of 602 Reading items. Fit statistics do not suggest the existence of dimensions other than Mathematics and Reading in these two tests.

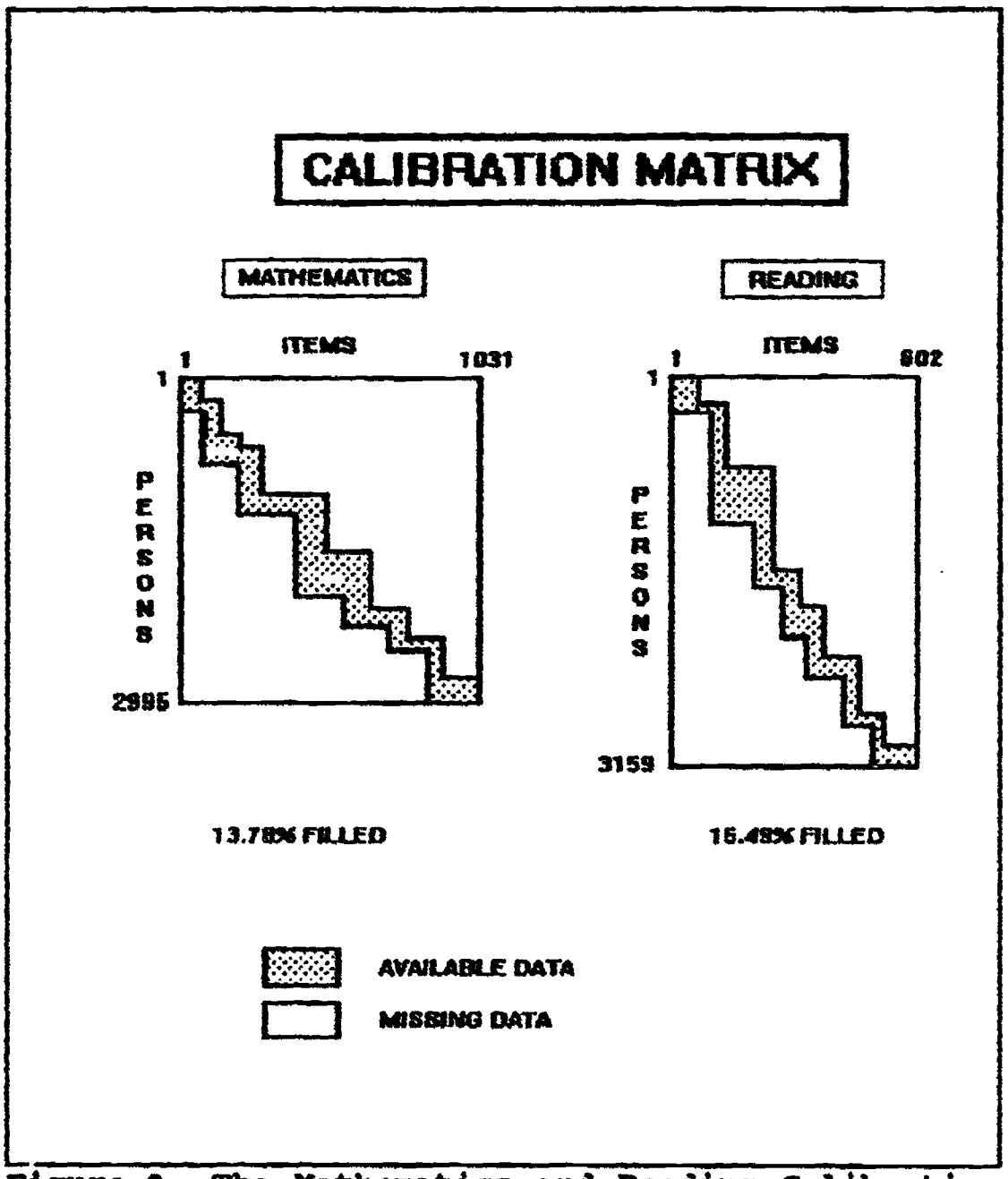


Figure 2 The Mathematics and Reading Calibration Matrices

Determining Person Measures Using the Mathematics and Reading Banks

(a) When response strings are available

Persons responding to any of the ITBS test levels equated here, can have their abilities estimated from their responses by running a Rasch analysis on their responses while anchoring the item difficulties on their bank values. Any set of items can be selected from these banks to form a test targeted on a given group of persons, and person abilities estimated in the same way. A realistic standard error for each measure can be estimated inflated for observed person misfit. This is because Rasch estimates are based on perfect fit and the standard errors for misfitting persons tend to be underestimated.

(b) When response strings are not available

In longitudinal studies where tests were implemented years ago, response strings are no longer available. The student measures therefore cannot be determined from an analysis of their responses. An indirect method based on their recorded grade-equivalents (GE's) must be used. The method is to regress the direct person measures for each test level from the equating study, on their GE's for that test level. The person measures used were those of the individual test analyses of uncleaned data, with item difficulties for this step anchored on their bank values. The regression coefficients can then be used to predict student ability measures from the GE's they obtained in their earlier tests.

Standard errors for these measures must also be estimated. Again regression analysis was used. This time the dependent variables were the standard errors (inflated for misfit) of the measures from the direct analyses of uncleaned data.

Mean Item Difficulty of Form 7 and CPS90

Tables 2(a) and 2(b) show the mean item difficulty for each test level. The last columns of Tables 2(a) and 2(b) show the differences between the mean logit measures of CPS90 and Form 7. It is clear that CPS90 is slightly harder than than Form 7 at most test levels. Mean test difficulties were plotted against

Table 2(a) Mean Item Difficulty for Reading

ITBS Test Level	Grade	Form 7 (F7)			CPS90 (C9)			Logit Diff. Bet. Forms (C9-F7)
		Number of Items	Item Difficulty		Number of Items	Item Difficulty		
			Mean	S.D.		Mean	S.D.	
6	K	70	-2.24	0.99	-	-	-	-
7	1	66	-2.15	0.61	56	-1.86	0.62	0.314
8	2	67	-1.10	0.75	61	-1.00	0.68	0.091
9	3	44	-0.17	0.89	44	0.37	0.72	0.542
10	4	49	0.96	0.85	49	1.12	0.68	0.166
11	5	54	1.47	0.87	54	1.51	0.73	0.041
12	6	50	1.97	0.76	56	2.07	0.82	0.086
13	7	57	2.80	1.06	57	2.69	0.81	-0.113
14	8	58	3.40	0.90	58	3.30	0.84	-0.096

Table 2(b) Mean Item Difficulty for Mathematics

ITBS Test Level	Grade	Form 7 (F7)			CPS90 (C9)			Logit Diff. Bet. Forms (C9-F7)
		Number of Items	Item Difficulty		Number of Items	Item Difficulty		
			Mean	S.D.		Mean	S.D.	
6	K	33	-3.51	1.33	-	-	-	-
7	1	81	-2.84	1.12	82	-2.82	1.17	0.02
8	2	88	-1.77	1.36	96	-1.48	1.36	0.29
9	3	90	-1.07	1.02	86	-1.06	1.29	0.01
10	4	99	-0.09	1.11	95	0.07	1.20	0.18
11	5	109	1.05	0.88	101	0.98	1.14	0.09
12	6	114	1.90	0.95	109	1.87	1.02	-0.03
13	7	117	2.51	0.97	113	2.73	0.95	0.22
14	8	121	3.07	0.98	117	3.22	0.95	0.15

grade and shown in Figures 3 and 4 for Reading and Mathematics respectively. The difference in mean difficulties between CPS90 and Form 7 for Level 9 (Grade 3) of the Reading test is at 0.54 and that for Level 7 (Grade 1) is at 0.31 logits. For Mathematics, the largest differences in mean difficulties between CPS90 and Form 7 are at Levels 8 (Grade 2), 10 (Grade 4), 13 (Grade 7) and 14 (Grade 8) with 0.29 logits, 0.18 logits, 0.22 logits and 0.15 logits respectively. To show these differences more clearly, they were plotted against grade and shown in Figures 5 and 6.

Notice from Tables 2(a) and 2(b) that the standard deviations of item calibrations for Reading increase with grade, that is the Reading items become more spread out in difficulty. The item calibration standard deviations for Mathematics decrease with grade, that is, the items are closer together in difficulty level at the higher test levels. This requires further investigation as to why it is so.

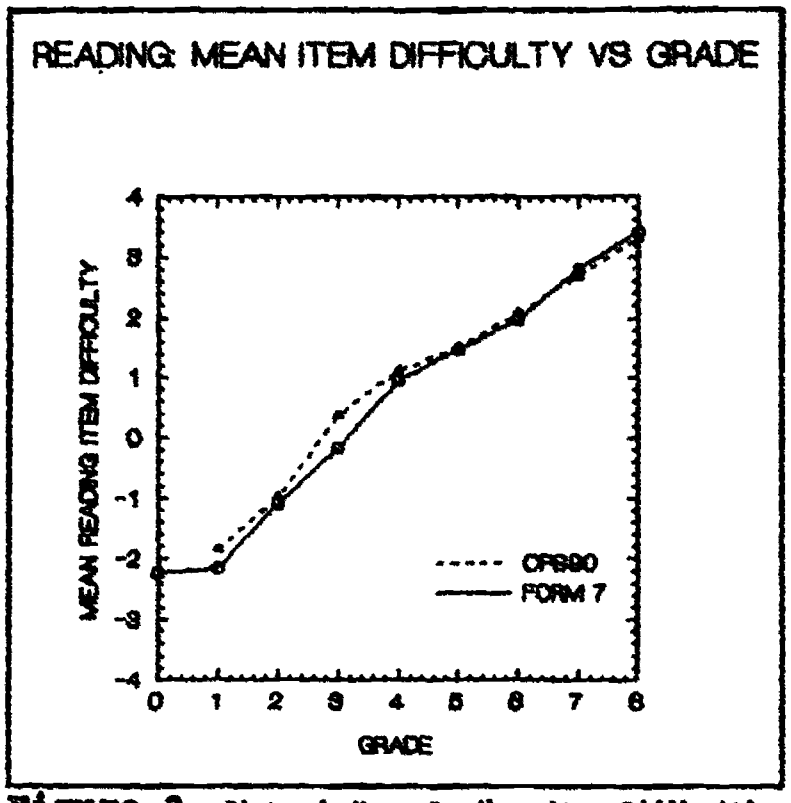


Figure 3 Plot of Mean Reading Item Difficulties against Grade for Form 7 (Levels 6 through 14) and CPS90 (Levels 7 through 14).

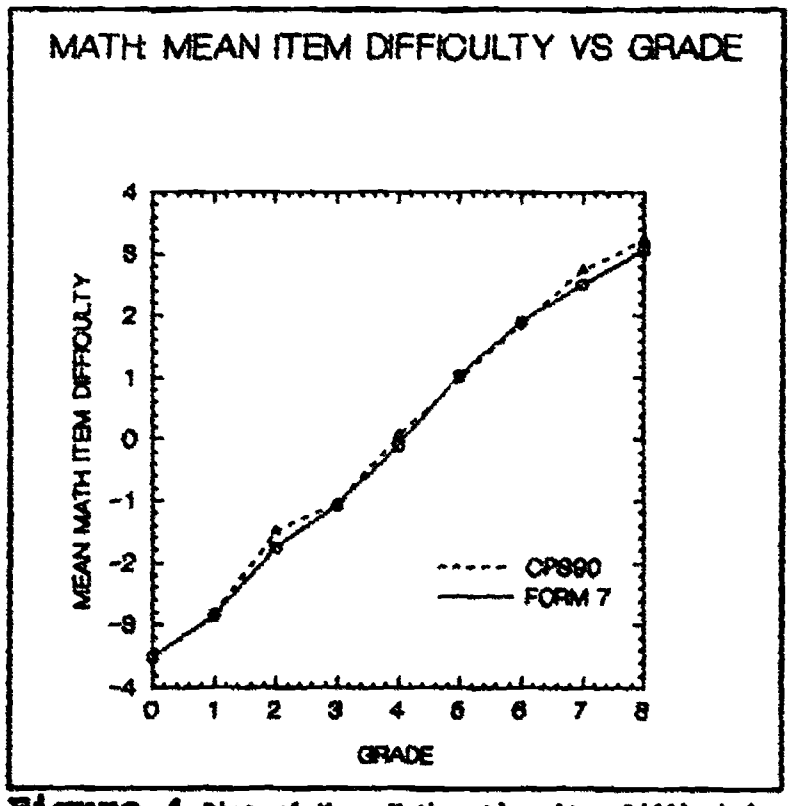


Figure 4 Plot of Mean Mathematics Item Difficulties against Grade for Form 7 (Levels 6 through 14) and CPS90 Levels (7 through 14).

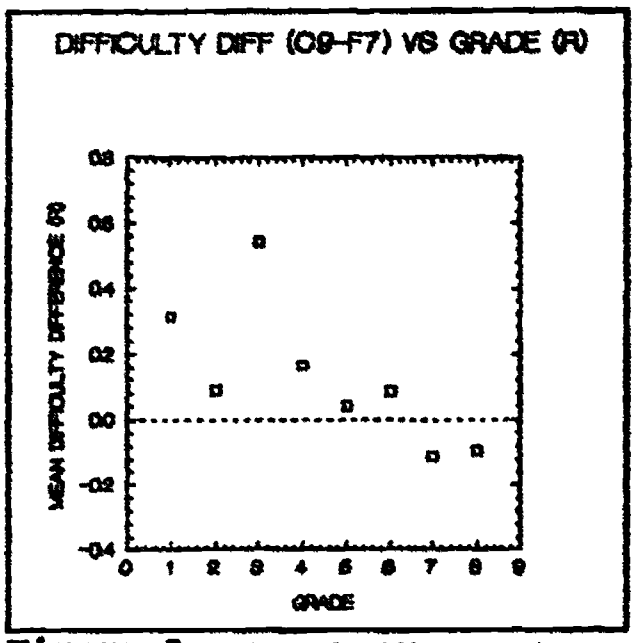


Figure 5 Plot of Differences between Reading Mean Difficulties of CPS90 and Form 7 against Grade.

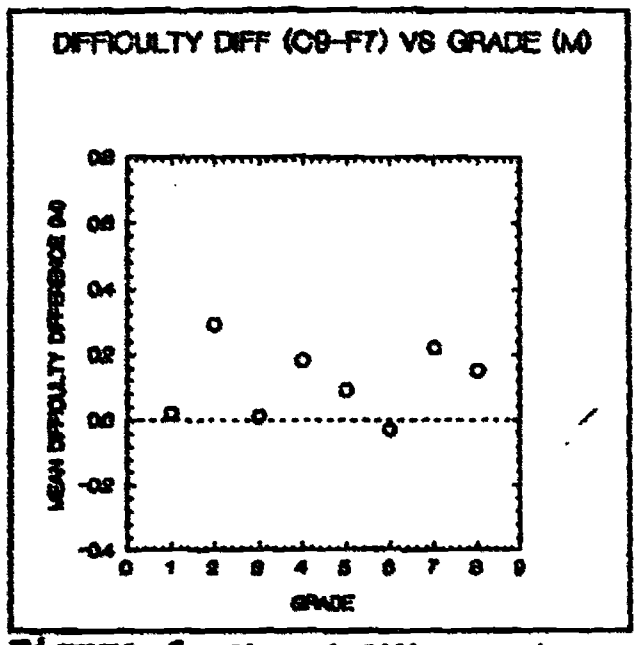


Figure 6 Plot of Differences between Mathematics Mean Difficulties of CPS90 and Form 7 against Grade.

Mean Measures and Mean Grade Equivalents of Common Persons taking Pairs of Tests

For the common persons taking a CPS90 and a Form 7 test at the same levels, (arrows 3, 7, 11, 12, and 13 in Figure 1), the mean measures and grade equivalents were calculated. Results are shown in Tables 3 and 4 for Reading and Mathematics.

Table 3 Mean Measures, Mean Grade Equivalents, and Standard Deviations of Common Persons Between Form 7 and CPS90 Reading Tests.

ITBS Test Level	Grade	Number of Common Persons	Person Measures					Grade Equivalents				
			Form 7		CPS90		Logit Diff Bet Forms (C9-F7)	Form 7		CPS90		GE Diff Bet Forms (C9-F7)
			Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.	
6	K	-	-	-	-	-	-	-	-	-	-	-
7	1	120	-1.80	1.27	-1.85	0.66	-0.05	1.67	0.85	1.25	0.51	-0.42
8	2	154	-1.08	0.86	-1.00	0.78	0.08	2.09	0.76	1.77	0.69	-0.32
9	3	160	0.03	1.01	0.00	0.97	-0.03	3.63	1.05	2.95	0.99	-0.68
10	4	-	-	-	-	-	-	-	-	-	-	-
11	5	175	1.47	1.00	1.50	1.00	0.03	5.51	1.40	5.13	1.38	-0.38
12	6	-	-	-	-	-	-	-	-	-	-	-
13	7	144	2.56	0.83	2.49	0.80	-0.07	6.93	1.54	7.14	1.55	0.21
14	8	-	-	-	-	-	-	-	-	-	-	-

Table 4 Mean Measures, Mean Grade Equivalents, and Standard Deviations of Common Persons Between Form 7 and CPS90 Mathematics Tests.

ITBS Test Level	Grade	Number of Common Persons	Person Measures					Grade Equivalents				
			Form 7		CPS90		Logit Diff Bet Forms (C9-F7)	Form 7		CPS90		GE Diff Bet Forms (C9-F7)
			Mean	S.D.	Mean	S.D.		Mean	S.D.	Mean	S.D.	
6	K	-	-	-	-	-	-	-	-	-	-	-
7	1	90	-2.42	0.86	-2.48	0.78	-0.06	1.64	0.56	1.41	0.48	-0.23
8	2	116	-1.06	0.96	-1.05	0.94	0.01	2.86	0.78	2.72	0.79	-0.14
9	3	118	-0.40	1.05	-0.33	1.25	0.07	3.82	0.86	3.41	0.92	-0.41
10	4	-	-	-	-	-	-	-	-	-	-	-
11	5	150	1.64	0.81	1.68	0.87	0.04	6.17	0.86	5.72	0.91	-0.45
12	6	-	-	-	-	-	-	-	-	-	-	-
13	7	173	2.73	0.94	2.73	0.85	0.00	7.81	1.33	7.25	1.24	-0.56
14	8	-	-	-	-	-	-	-	-	-	-	-

Since the same persons took both tests, their matched mean measures on the two tests should be statistically equivalent. It is shown graphically by plotting the mean measures against grade in Figure 7 (for Reading). The same was done for grade equivalents in Figure 8. Similar plots are shown for Mathematics in Figures 9 and 10. Note that the matched mean GE's for the same persons are not the same over the two test forms they took, for both the Reading and Mathematics. Students obtain higher grade equivalents from Form 7 for both Reading and Mathematics, except for Grade 7 Reading. This shows a bias in grade-equivalent equating of the ITBS, that is, GE's produced by the two forms are not directly comparable. The GE plots are not even the straight lines we expect from GE scoring. For Grade 7 Reading, the mean Rasch logit measure shows that CPS90 is slightly harder than Form 7. In grade equivalents, however, the same students appear to have done better on CPS90. This apparent contradiction suggests the possibility that the norm group used for the CPS90 Grade 7 Reading could have been a less able group compared to the norm group for Form 7. Hence the same group of students in the equating study when seen in terms of GE appear to have performed better on the CPS90 than on the Form 7 Grade 7 Reading. When compared in logit measures for common persons, Form 7 and CPS90 differences for all the grade levels are very close to zero as expected. This is shown in Figures 11 and 12.

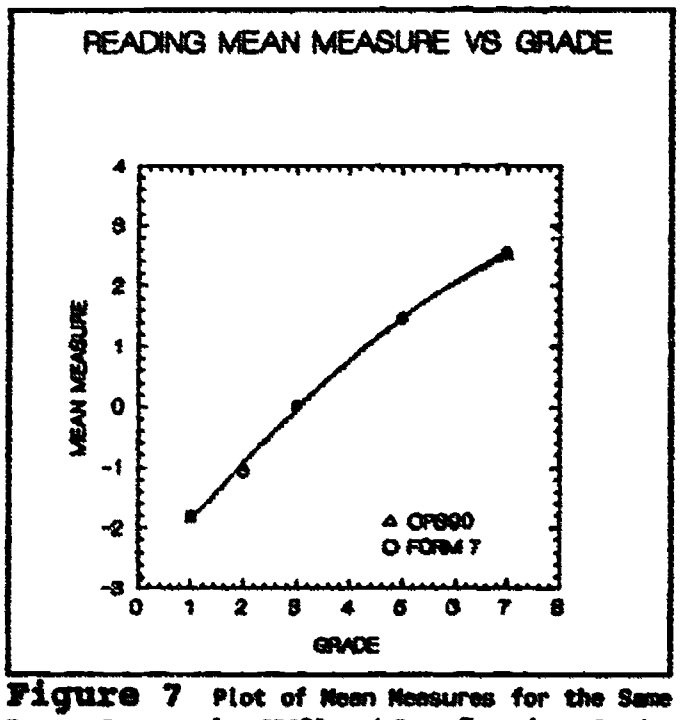


Figure 7 Plot of Mean Measures for the Same Person Groups, for CPS90 and Form 7 against Grade, for Reading.

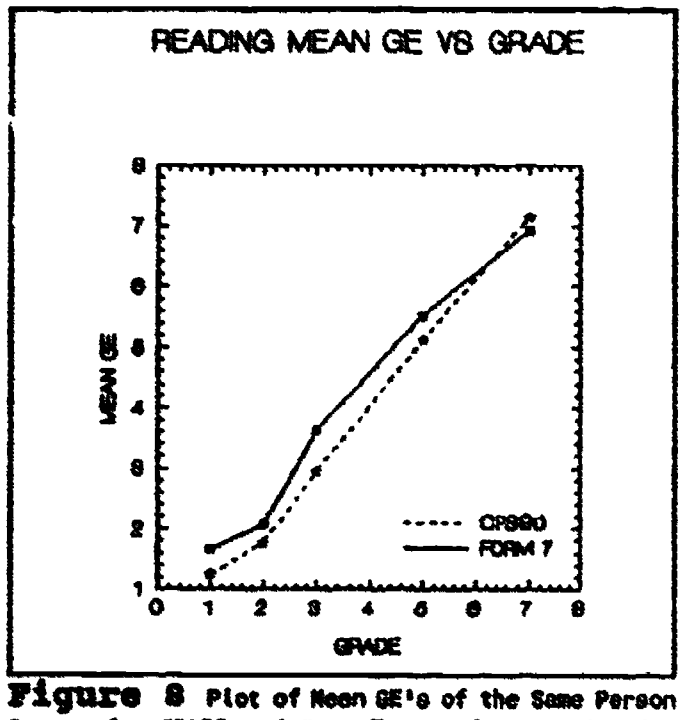


Figure 8 Plot of Mean GE's of the Same Person Groups for CPS90 and Form 7, against Grade, for Reading.

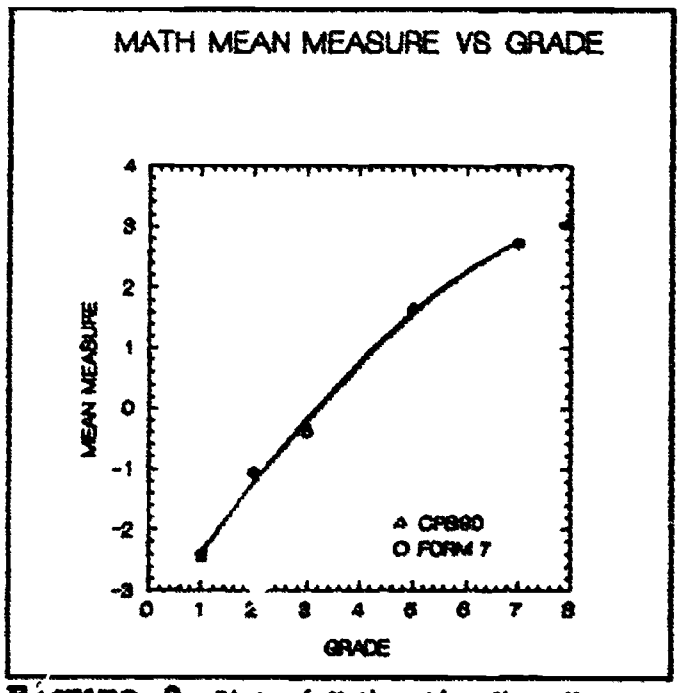


Figure 9 Plot of Mathematics Mean Measures for the Same Person Groups, for CPS90 and Form 7 against Grade.

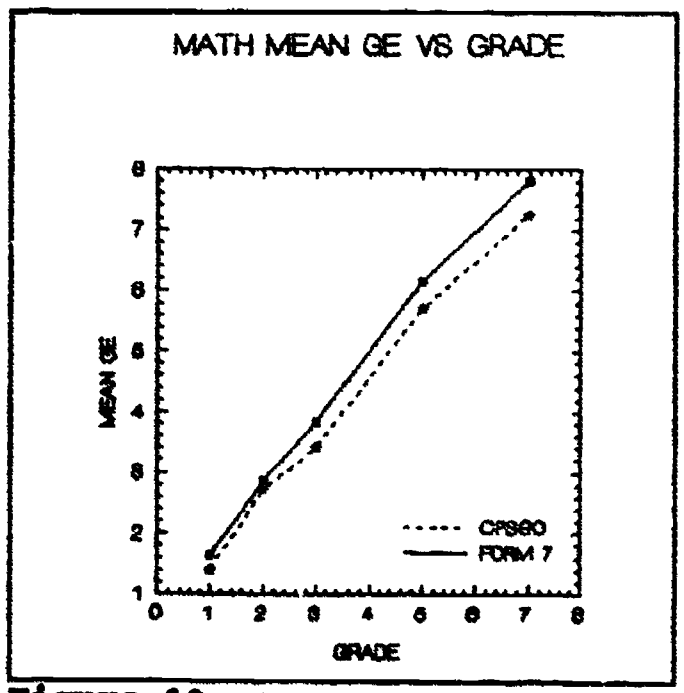
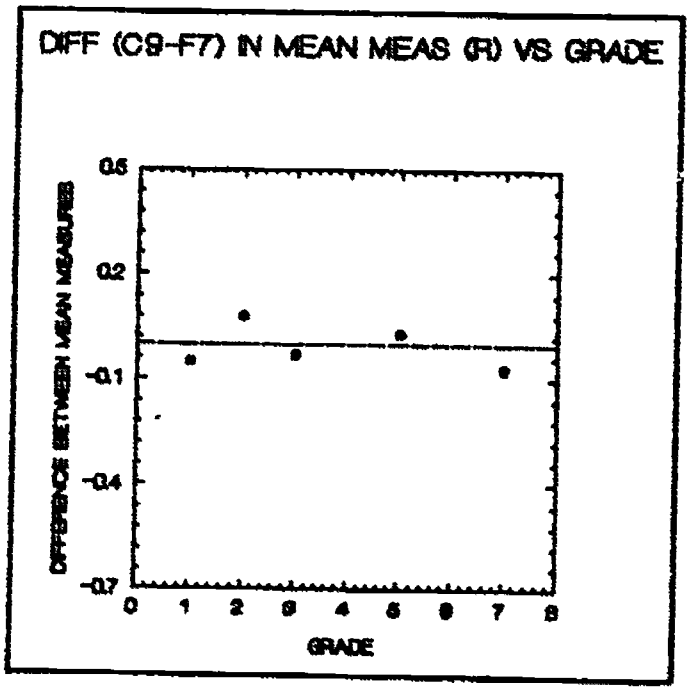


Figure 10 Plot of the Mean GE's for the Same Person Groups for CPS90 and Form 7, against Grade.



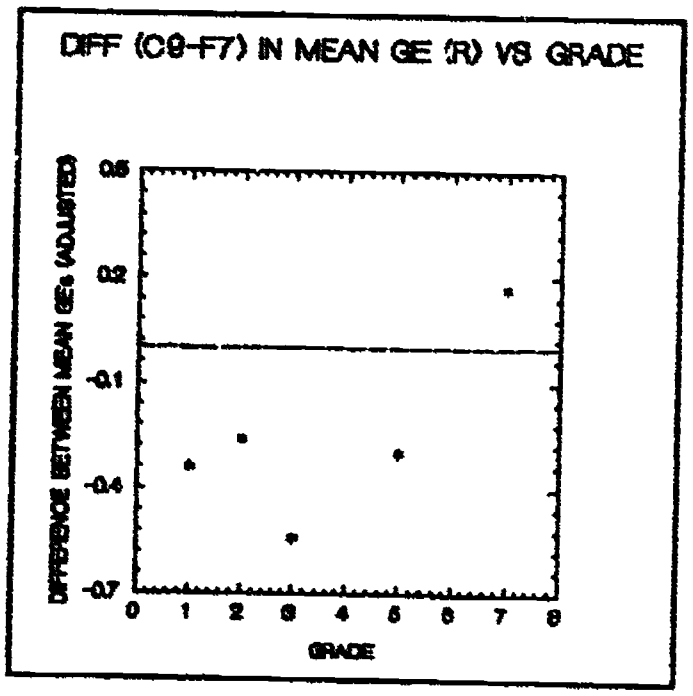
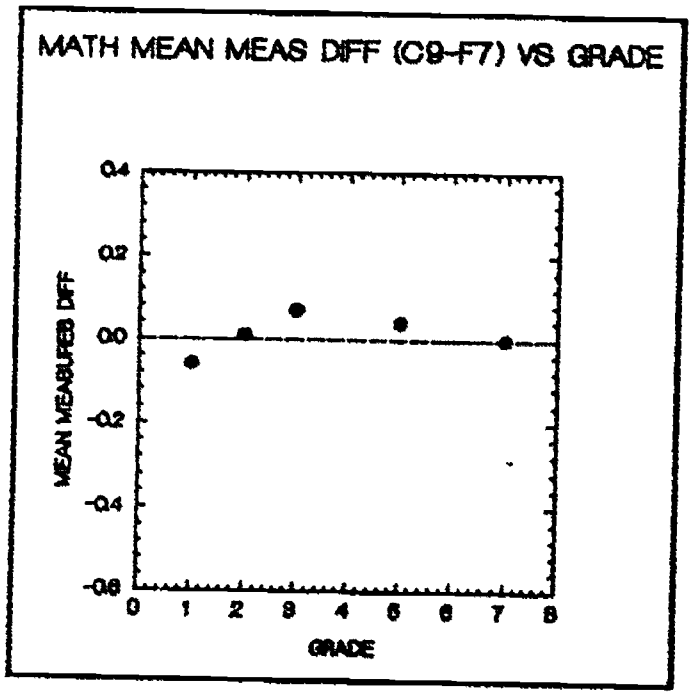


Figure 11 Plots of Differences between Common Persons' (a) Mean Measures and (b) Mean GE's against Grade for Reading.



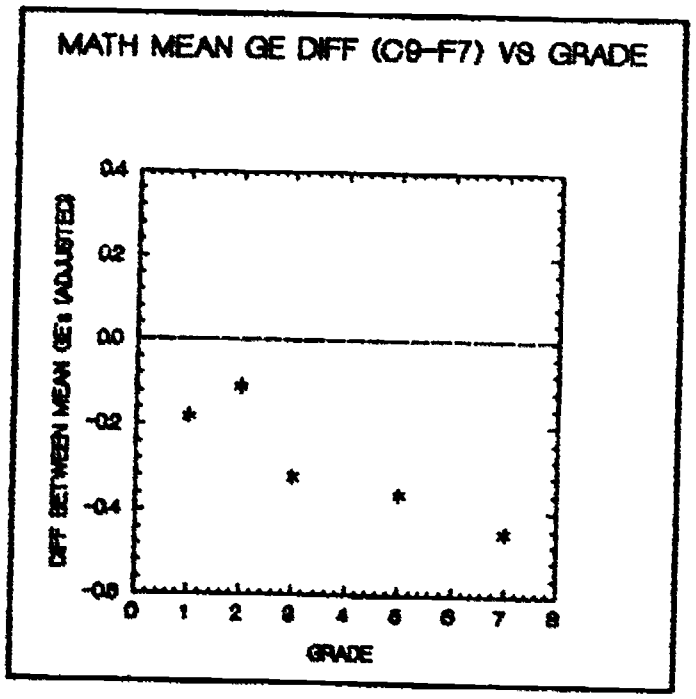
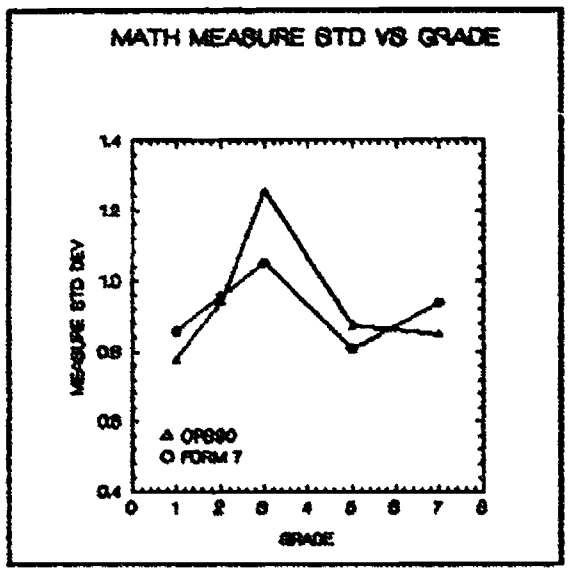


Figure 12 Plots of Differences between Common Persons' (a) Mean Measures and (b) Mean GE's against Grade for Mathematics.

The differences in mean GE's, adjusted to the logit scale using the average exchange of 0.8 logits per grade so that the vertical scales are all comparable, were also plotted against grade in Figures 11 and 12. Here the differences in GE's between the two test forms are much larger than zero.

Standard Deviations of Measures and Grade Equivalents

From Tables 3 and 4 we see that for Reading and Mathematics, the standard deviations of GE's increase with grade while those of measures do not. The spread of students in logit measures does not change much from grade to grade. The increasing standard deviations of the grade equivalents give the misleading impression that student spread increases, that they get further apart. Figures 13(a) and 13(b) plot standard deviation against grade. Note the relative constancy of the logit standard deviations and the systematic increase of the GE's standard deviations across the grades. The illusion of increasing spread



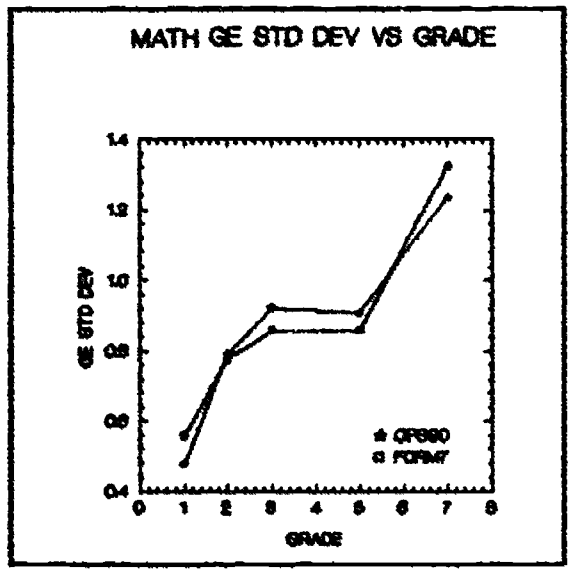
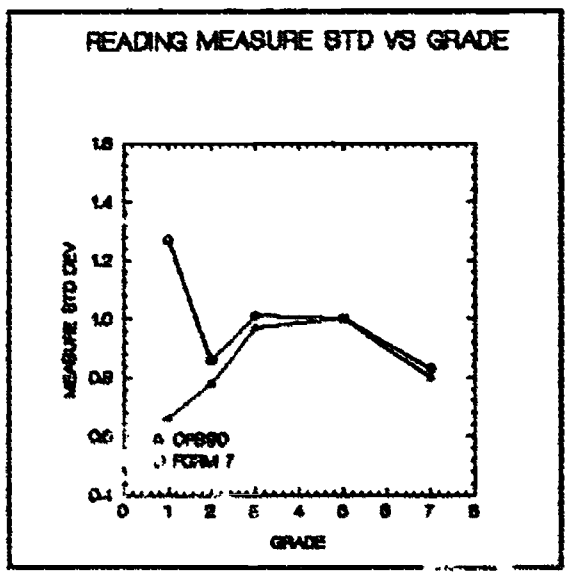


Figure 13(a) Plots of Standard Deviations of (i) Measures (ii) Grade Equivalents against Grade for Mathematics.



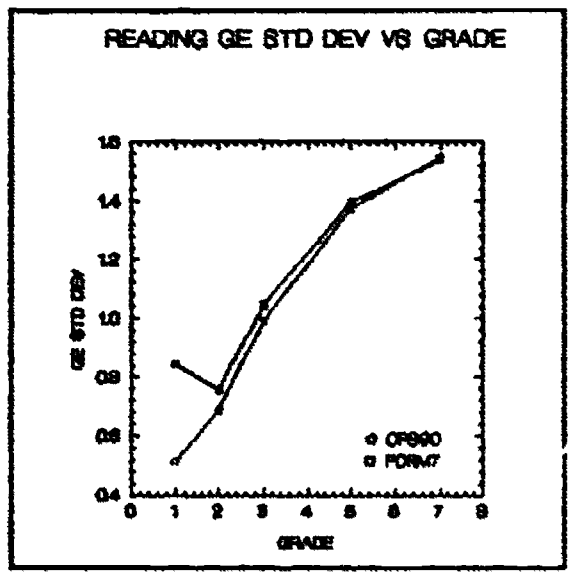


Figure 13(b) Plots of Standard Deviations of (i) Measures and (ii) Grade Equivalents against Grade for Reading.

produced by GE standard deviations could easily be misunderstood to prove that schooling increases the differences among students. The logit measure plots show that this is clearly not so.

Criterion Definition of Variable

Appendix A is an example of a criterion definition of the variable called Mathematics Computation. D on the vertical axes is a linear transformation of item calibrations. D = 26 + 5*(item difficulty). The vertical axis on the right shows the locations of the mean student ability at each grade.

Such item maps can readily be constructed once items have been calibrated, which an item bank of this kind enables. The math items increase in complexity as the difficulty level increases. This is useful to teachers. Students' measures are directly comparable to item difficulty calibrations. Reference to an item map such as this, enables a teacher to determine what a student has or has not mastered, where the student is in his mathematics education, and to plan his lessons accordingly.

Conclusion

The 17 levels of the ITBS Mathematics and Reading tests used in this study have been successfully equated and are each on a common scale of item difficulty from K to 8. A person taking either CPS90 or Form 7 (or any combination of items from these two test forms targeted at his ability level) will obtain statistically equivalent measures of his ability.

In the grade-equivalent metric, the difficulty of the test depends on the ability level of the norming sample. A student's grade-equivalent depends on which test form he takes. As a result it is impossible to compare student abilities by studying the grade equivalents. Students scoring lower grade-equivalents on a given test may be thought to be less able, when the test may actually be harder or the norming sample more able. Similarly, students scoring higher grade-equivalents may not necessarily be of higher ability since the test form may in

fact be easier or the norming sample less able. Using grade-equivalents results in misleading interpretations of student performance. These have serious policy implications. Teachers may recommend remedial programs for students who do not actually need them. Students may be thought to have acquired the desired level of competency when they have not. Funds may be channelled to the wrong programs for the wrong students.

Students' rates of growth will never be shown by grade equivalents. Every year they are forced to have one unit of grade-equivalent higher. A plot of GE growth against grade is forced close to a straight line giving the false impression that the rate of growth is uniform at all ages. With logit measures, however, rates of growth are shown to be highest at the lower grades, and to decrease in the higher grades.

References

Easton, J. Q., Dean, P., & Bryk, A. S. (1991). School change and its correlates: Studying Chicago Public Elementary Schools after school reform. Paper presented at the annual meeting of the American Educational Research Association.

Frank, K. & Seltzer, M. (1990). Modeling growth in reading achievement. Paper presented at the annual meeting of the American Educational Research Association.

Schulz, E. M., Shen, L., & Wright, B. D. (1990). An equal-interval scale for studying reading growth. Paper presented at the annual meeting of the American Educational Research Association.

Wright, B. D. (1977). Misunderstanding the Rasch Model. Journal of Educational Measurement, 14(3).

Wright, B. D. & Douglas, G. A. (1975). Better procedures for sample-free item analysis (Research Memorandum No. 20). Chicago: MESA Psychometric Laboratory.

Wright, B. D. & Stone, M. H. (1979). Best test design: Rasch Measurement. Chicago: MESA Press.

Appendix A

CRITERION DEFINITION OF VARIABLE

D	ADDITION	SUBTRACTION	MULTIPLICATION	DIVISION	D	
SINGLE DIGIT SIMPLE OPERATIONS (ADDITION AND SUBTRACTION)						
1	0 + 9 =				1	
5	"4 + 3 ="	9 - 4 =			5	
10	3 + 43 = ▼	26 - 12 = ▼			10	
MULTIDIGIT SIMPLE OPERATIONS / MULTIPLICATION	765 + 204 = ▼	789 - 121 = ▼	3 x 8 =		15	Grade 1
	14 + 30 + 45 =	285 - 118 =	8 x 1000 =		20	Grade 2
▲ 25	35+26+40+34=▼	8862 - 8341 = ▼	423 x 2 =	16 / 4	25	Grade 3
DIVISION / MULTIPLICATION	8.32 + 2.79 = ▼	3/8 - 4/8 1527 - 1483	80 x 300 = 7803 x 6 = ▼	89 / 7 =	30	Grade 4
SIMPLE OPERATIONS WITH DECIMALS AND FRACTIONS / DIVISION WITH DECIMALS AND FRACTIONS / MULTIPLICATION WITH DECIMALS AND FRACTIONS / MULTIPLICATION WITH FRACTIONS / MULTIPLICATION WITH REMAINDERS / COMBINED FRACTIONS WITH UNLIKE DENOMINATORS / DIVISION WITH DECIMALS / DIVISION WITH FRACTIONS / MULTIPLYING MIXED NUMBERS	1.873 + 0.042 = ▼	7920 - 4.76 = ▼ 7.83 - 4.76 = ▼	608 x 26 = ▼ 52 x 37 = ▼ 2/3 X 4/5 =	3600/60 = L 390/26 = L	35	Grade 5 Grade 6
	3/5 + 1/10=?	1/2 - 1/3 = 5 3/7 - 3 1/3 =	.68 x .5 = ▼ 703 x .36 = ▼ 614 x .02 = ▼	191/18 = L 544/27 = L 1.44/0.36 = L	40	Grade 7 Grade 8
	.26 + .48 + .72 = ▼ 1/4 + 2/3 + 5/6 = ▼	6 7/8 - 2 5/8 =	1/6 x 5 = ▼ 4/7 x 3/8 =	4/7 / 1/6 = 24 / 3/8 =	45	
			20% of 180= What % is 90 of 360? 2 1/2 x 7/8 =	5/12 / 3 = 20 / 1 1/2 = 15,168/300 = L	50	
INTRODUCTION TO FRACTIONS						

N.B. D is a linear transformation of difficulty: D = 26 + 5*(difficulty).