JOURNAL OF APPLIED MEASUREMENT, *4*(1), 10–23

Rasch Simultaneous Vertical Equating for Measuring Reading Growth

Ong Kim Lee
National Institute of Education
Nanyang Technological University

The longitudinal measurement of ability growth requires that the measures that are taken at the various time points be obtained using the same yardstick. Tests given for this purpose, therefore, need to be equated. The popular practice still in use for the purpose of equating tests, is the grade-equivalent. This paper compares the observation of children's growth in reading using grade-equivalents with that using Rasch Simultaneous Vertical Equating procedure. It is found that grade equivalents differ much more between two different test forms compared to ability measures obtained using Rasch Simultaneous Vertical Equating. It is also found that the spread of students' grade-equivalents, increased over the years as they grow while the standard deviations of their Rasch measures remain relatively constant over the same period of time. Student responses to the ITBS Form 7 and the CPS90 and CPS 91 were used. A total of 5,623 students were tracked over eight years.

Request for reprints should be sent to Lee Ong Kim, National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore 637616, email: oklee@nie.edu.sg.

Introduction

An essential and important component of the teaching-learning process is "Assessment". Teachers need to know that students are progressing and progressing means that there is growth in the student's ability in the subject area concerned. Teachers commonly assess student progress through testing. What is important in the use of tests is the way test scores are interpreted. When a student obtains a raw score of 70 in a test this semester, and then obtains 65 in the following semester, does it mean he is retrogressive as implied by the teacher whose comment on the student's performance is "you have dropped slightly"? What if the second test is actually a more difficult one? A student who has grown in ability could still score lower on a harder test compared to his higher score on an easier test. His growth can only be shown if the tests have been equated and his ability measures obtained. For tests to be equated, the items will have to be calibrated on a linear scale of difficulty measures. The ability levels of students responding to tests consisting of items of known difficulty levels, can be estimated on a linear scale. His ability growth can then be determined from one time point to the next through testing.

It should be noted that raw scores are not linear and are therefore not measures. A raw score of five for example, on different parts of the raw score "scale" does not imply the same "amount" of the variable being measured. On a mathematics test for example, the five points between 40 and 45 for instance, is not the same amount of ability as the five points between 90 and 95 on the familiar zero to 100 "scale". The raw scores from a single test provide only the rank order of abilities of the persons who took the test, and they are not measures (Wright, 1992, 1993).

Teachers use test results to make many educational decisions. Students may be put into different learning groups, into remedial classes, into different streams and perhaps even be promoted or retained using their performances on tests as deciding criteria. Hence it is very important that these students are measured on a scale that allows for comparisons of performances to be made. Comparisons of performances on tests may be those of the same person across different time points or those of different persons at a given time point. The purpose of this paper is to demonstrate how several test forms and levels may be equated in a single run of the Rasch Analysis program, Winsteps (Linacre and Wright, 2000) for the purpose of making such comparisons through testing. Data that are available from three test forms from the Iowa Tests of Basic Skills (ITBS), namely, Form 7, CPS90 and CPS91 were used to demonstrate this equating process (see Lee, 1992; Lee and Wright, 1992). Tests for the same level from each of these forms, are parallel and may be used interchangeably by some school systems.

Equating Tests

A test may contain items that measure ability on a variable with only one dimension. The test items however, will vary in difficulty levels. When several items are put together, to make up a test form, the overall difficulty level of the test form is not necessarily the same as another test form that is constructed parallel to the first. A given number of correct responses on a harder test, would indicate a higher ability than the same number of correct responses on an easier test. A compensation factor is required to reflect this difference, which is Lord's requirement of "equity" (Lord, F. M., 1980). Rasch equating satisfies this equity requirement, which requires that a person's measure of ability should be independent of whether he takes a harder test or an easier test. This paper shows a quick and practical way of equating the Iowa Tests for Basic Skills (ITBS) tests through the use of the Winsteps program, which is a Rasch Model computer program (Linacre and Wright, 2000). An example of a control file used in the analysis is shown in Appendix A. Data from three forms of the ITBS used by the Chicago Public Schools Board were obtained for this equating. These are the Form 7, CPS90 and the CPS91 test forms. Level 7 of each of these forms (e.g. Form 7 Level 7 and CPS90 Level 7) are for Grade 1 students, Level

2 for Grade 2 and so on up to Level 14 which is for Grade 8 students. Data were also available for Level 6 of Form 7 with common persons with Form 7 Level 7 and CPS90 Level 7. These were therefore also added into the matrix for the equating. Item fit statistics are studied but not with the intention of dropping any item since all these forms of the ITBS are existing test forms already in use. The growth measures obtained through this Rasch equating are then contrasted with that of Grade Equivalents. The cohort of students tracked for this study, had taken at least one of these test forms at each grade level, at the various stages of their elementary school career.

Method

This simultaneous Rasch equating was used to equate twenty-five reading tests in all, nine levels of ITBS Form 7 Reading tests, and eight levels each of CPS90 and CPS91. The pattern of overlapping items for these three test forms is shown in Table 1. From this table we can see that items 19 through 44 of Form 7 Level 10 overlap the Level 9 test, that is, these 26 items are common in both levels 9 and 10. Similarly, 38 items are common between Level 11 and Level 10, and so on. Where there are no common items, such as for Levels 6, 7, and 8 of Form 7 and Levels 7 and 8 of CPS90, common persons are used in the response matrix. For common persons, a group of students were made to take both test forms administered one day apart. The SAS application software was used to set up the response matrix for the common items and common persons to overlap accordingly. The structure of the overlapping matrix is shown in Figure 1.

The various test levels were equated by setting up the response matrices from these test forms, into a single large matrix as if it was a single test. This means that we are equating the various test levels and forms through the analysis of responses set up as one test. This is different from the analysis of responses to sample items over all the grades

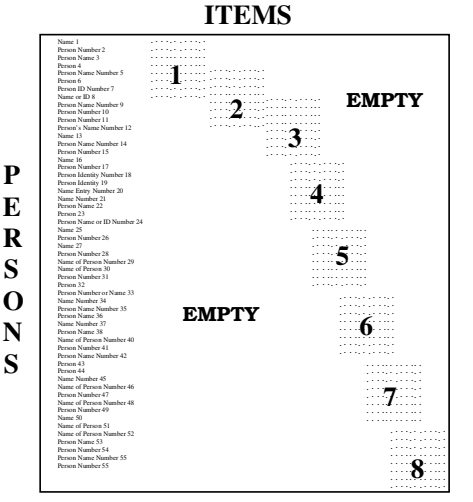

Figure 1. Structure of the Reading Data Matrix for Simultaneous Vertical Equating

Table 1

Overlapping items of Form 7, CPS90 and CPS91 Reading Tests.

Form 7			CPS90 and CPS91		
Level	No. of Items	Overlapping Items	Level	No. of Items	Overlapping Items
6	70	-			
7	66	None. Overlapping persons used	7	56	-
8	67	None. Overlapping persons used	8	61	None. Overlapping persons used
9	1-44 (44)	-	9	1-44 (44)	-
10	19-67 (49)	19-44 (26)	10	19-67 (49)	19-44 (26)
11	30-83 (54)	30-67 (38)	11	30-83 (54)	30-67 (38)
12	45-100 (56)	45-83 (38)	12	45-100 (56)	45-83 (38)
13	68-124 (57)	68-100 (33)	13	68-124 (57)	68-100 (33)
14	84-141 (58)	84-124 (41)	14	84-124 (58)	84-124 (41)

put together as one test and then implementing the test to a representative sample of students over the grades as is done in the determination of grade-equivalents. In order to equate all the test forms so that all the items are placed on a common scale, the individual response matrices cannot be divorced from each other in any part of the large matrix. It is necessary to link them to one another either by way of common persons or common items. This study employed simultaneously, common persons and common items linking in the structuring of a large response matrix for the Reading tests, where common persons over two tests were arranged in the same rows and common items were arranged in the same columns. Levels 6 through 8 of Form 7 and levels 7 and 8 of CPS90 and CPS91 do not have within form overlapping items and hence common persons were used. Also requiring common persons were all the "between-forms" linking. Levels 9 through 14 within each test form have common items and do not require common persons.

Before setting up the individual tests response strings into a single large data matrix, however, each data set has first to be cleaned as only reliable data may be used in the large matrix. In order to properly calibrate test items, students must have seriously attempted the items. Responses of students who "misbehave" in the test, such as making "multiple responses" and blind guessing as shown by fixed response patterns, for example, are not reliable data to be used for equating. Of course once the tests have been equated, the measures of all persons who have taken or who subsequently take any of these tests, including those who "misbehaved", can be estimated and reported. The data for each test level are from a wide range of student ability typical of that of the target population. The data obtained were already relatively "clean" as at the end of all stages of cleaning, less than four percent of the response strings were dropped. The data were cleaned in four stages as follows:

Stage 1: Students who fit the following criteria were dropped from the analysis:

(a) There were more than three multiple responses in the string;

(b) 50-70 percent of the string were light **and** there were more than one embedded omits;

(c) 80-100 percent of the string were light **and** there was at least one embedded omit.

The optical mark reader (OMR) was programmed to indicate a lightly shaded bubble where the intensity of the shading casts doubts as to the choice of response, as "light". The set of criteria used is to ensure that subjects used are those who are certain of their choice of answers. The number of cases fulfilling the above criteria and were thus dropped, was less than two percent.

Stage 2: The response strings were visually scanned on the computer screen. Response strings showing a series of successive zeroes and/or the same responses for 25% or greater of the total number of items, were dropped. These are likely to be the very slow test takers who may be able to answer some of the questions if given sufficient time. The fact that they were not able to reach those questions does not reflect the true difficulty levels of those items, making them appear to be more difficult than really are. The number dropped through this process was even smaller, at less than half a percent.

Stage 3: Response matrices from the test levels were each analyzed separately using the Rasch Analysis Computer Program. Persons with both infit and outfit mean squares greater than 2.5 were removed. An examination of their residuals show (a) a good number with well spread out residuals amongst the items with response strings showing a random pattern of "correct" responses for difficult items and "wrong" responses for easier items, which is quite likely due to blind or random guessing (see Smith, 1993), (b) several with fixed response patterns clearly indicating blind or random guessing, and a few with appreciable number of items towards the end of the test, not responded to. These are not considered reliable data. Less than one percent was removed this way. If only one of the fit statistics was greater than 2.5 and the other between 2.0 and 2.5, the item residuals were examined for possible test-taking misbehavior. The intention is that, if the misfit was due to only one or two

large residuals, the person should be retained, as it is quite natural to occasionally have an outlier or two. If the residuals are spread out through most of the items, the person was to be removed since most of his or her responses were unexpected. Under this category, none of the persons was removed as a result of the examination of residuals. The intention, in other words, was not to remove all misfitting persons as may be defined by a certain cut-off point of the infit and outfit mean squares, but rather to have a clean and reliable data set (free of misbehaviors) for the purpose of test equating.

Stage 4: Each test level was individually analyzed and performances of common persons over two test forms were compared. The differences in their measures were standardized using the equation:

$$Z = \frac{[M_1A - M_2A]}{\sqrt{[E_1^2 + E_2^2]}}$$

where M_1A and M_2A are the person measures centered on their means (group's mean measure subtracted from each measure) for the two tests. E_1 and E_2 are the respective standard errors for the measures. Large values of Z would mean that the student performed extremely well on one test but not the other, or extremely poorly on one test but not the other. Response strings for students with standardized differences of greater than 2 or less than -2 were examined. Strings on the test for which the students performed poorly were dropped if they showed that they had "slept" on the test (did not respond to at least 25% of it). Students were also dropped if there was evidence of wild guessing as may be shown by string blocks of the same responses or by response-sets. They were excluded from that particular test level only. They were not dropped from the other test level, as their good performance on it is useful information for the analysis of that test level. An example of across-form common persons with large Z values is shown in Table 2. Less than one percent of the students were dropped through this process.

Persons with large Z values can also be viewed graphically using plots of person measures on one test against their measures on the second test, with the 95% control lines. Such plots show how large the number of outlying persons is and is a quick way of identifying the extreme persons. However, not all outliers must necessarily be dropped. We expect 5% of the sample to fall outside the 95% control lines. The locations of these 95% control lines are approximate since the standard errors (and the person measures for that matter) are all estimates. Hence only persons significantly far from these lines should be considered outliers and have their response strings examined.

After the data cleaning, the total number of different persons used in the study was 5,623 and the total number of different reading items over all the tests (Form 7, CPS90 and CPS91) was 1,118. The distribution of students by grade, including the overlapping counts for common persons, is shown in Table 3 with a total of 7,028 students. The total number of common persons between Form 7 and CPS90 and between CPS90 and

Table 2

Standardized Differences (z) between common-person measures for reading (154 records)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
STUID	SC1	M1	E1	INMS1	INST1	OMS1	OST1	SC2	M2	E2	INMS2	INST2	OMS2	OST2	M1A	M2A	P-Q	√E	Z=R/S
31286409	34	0.03	0.26	0.93	-0.93	0.90	-1.06	15	-1.22	0.31	1.08	0.55	1.25	1.19	0.01	-1.22	1.23	0.40	3.05
31934265	30	-0.24	0.26	1.08	0.98	1.06	0.62	14	-1.32	0.32	0.92	-0.43	0.89	-0.47	-0.26	-1.32	1.06	0.41	2.58
31034736	35	0.10	0.26	0.87	-1.79	0.84	-1.64	20	-0.79	0.28	1.02	0.18	1.12	0.82	0.08	-0.79	0.87	0.38	2.29
31735939	25	-0.58	0.27	0.92	-0.82	0.90	-0.78	12	-1.53	0.33	1.01	0.09	1.05	0.26	-0.60	-1.53	0.93	0.43	2.19
30985273	49	1.11	0.29	1.11	0.81	1.46	2.15	34	0.25	0.27	0.97	-0.32	0.94	-0.57	1.09	0.25	0.84	0.40	2.13
32087256	17	-1.20	0.29	1.28	1.83	1.38	1.71	25	-0.40	0.27	1.26	2.89	1.32	2.83	-1.22	-0.40	-0.82	0.40	-2.06
31988837	30	-0.24	0.26	0.97	-0.31	0.95	-0.47	38	0.55	0.28	0.83	-1.85	0.79	-1.86	-0.26	0.55	-0.81	0.38	-2.11
29806497	32	-0.10	0.26	0.90	-1.30	0.88	-1.27	40	0.71	0.28	0.92	-0.77	0.89	-0.78	-0.12	0.71	-0.83	0.38	-2.16
31933978	34	0.03	0.26	0.97	-0.41	0.95	-0.46	42	0.87	0.29	0.93	-0.57	0.85	-0.95	0.01	0.87	-0.86	0.39	-2.20
30460618	25	-0.58	0.27	1.03	0.32	1.02	0.17	35	0.33	0.27	1.03	0.41	1.00	0.04	-0.60	0.33	-0.93	0.38	-2.43
30869915	26	-0.51	0.26	0.94	-0.64	0.90	-0.78	36	0.40	0.27	0.94	-0.63	0.92	-0.78	-0.53	0.40	-0.93	0.37	-2.47
MEAN	33	0.02	0.28	1.01	0.10	1.03	0.15	30	-0.00	0.28	1.00	-0.02	1.01	-0.00	-0.00	-0.00	0.00	0.40	-0.03
STD	11	0.85	0.04	0.10	0.99	0.19	1.07	10	0.78	0.02	0.10	1.07	0.16	1.13	0.85	0.78	0.46	0.04	1.13

CPS90 LEVEL 8 IS HARDER THAN FORM 7 LEVEL 8 BY 0.016 LOGIT

N.B. $\sqrt{E} = \sqrt{(E_1^2 + E_2^2)}$;

SC: Raw Score; M: Measure; E: Standard Error of Measure; INMS: Infit Mean Square; INST: Infit Mean Square Standardized; OMS: Outfit Mean Square; OST: Outfit Mean Square Standardized; M1A and M2A: Measures Centered on their Means; Subscripts 1 and 2 refer to Test 1 (Form 7 Level 8) and Test 2 (CPS90 Level 8) respectively.

Table 3

Number of Persons for Each Grade Level with Double-Counting for Common Persons

Grade	Test Level	Form 7	CPS90	CPS91	Total
K	6	238	-	-	238
1	7	406	363	225	994
2	8	516	504	512	1532
3	9	290	413	412	1115
4	10	227	226	234	687
5	11	177	209	124	510
6	12	236	228	216	680
7	13	329	151	113	593
8	14	239	175	265	679
				Total	7028

Table 4

Item Infit and Outfit Mean Squares for Each Test Level

	Infit Mean Square					Outfit Mean Square				
Test	Min	Max	Mean	S.D.	Z-Std	Min	Max	Mean	S.D.	Z-Std
Form 7 L6	0.85	1.45	1.00	0.12	-0.1	0.77	1.66	0.99	0.27	-0.1
Form 7 L7	0.83	1.44	1.00	0.12	-0.1	0.71	1.71	1.03	0.25	0.0
Form 7 L8	0.79	1.65	1.00	0.16	-0.2	0.72	1.96	1.00	0.25	-0.1
Form 7 L9	0.83	1.31	1.01	0.11	0.0	0.72	1.84	1.01	0.21	0.0
Form 7 L10	0.82	1.39	0.99	0.14	-0.4	0.74	1.66	1.03	0.23	0.0
Form 7 L11	0.80	1.32	1.00	0.12	-0.1	0.70	2.32	1.04	0.27	0.1
Form 7 L12	0.79	1.98	1.02	0.18	0.0	0.71	2.51	1.03	0.26	0.0
Form 7 L13	0.54	1.20	0.97	0.12	-0.5	0.40	1.59	0.98	0.19	-0.3
Form 7 L14	0.66	1.17	1.00	0.10	-0.1	0.49	1.35	1.01	0.15	0.0
CPS90 L7	0.85	1.33	1.00	0.12	-0.1	0.74	1.44	1.00	0.17	0.0
CPS90 L8	0.82	1.34	1.01	0.14	0.0	0.80	1.57	1.01	0.20	0.1
CPS90 L9	0.87	1.30	1.00	0.10	-0.3	0.86	1.40	1.02	0.14	-0.1
CPS90 L10	0.84	1.46	1.02	0.14	-0.1	0.78	1.63	1.04	0.21	0.0
CPS90 L11	0.72	1.37	0.98	0.16	-0.4	0.58	1.64	0.98	0.23	-0.3
CPS90 L12	0.62	1.65	1.00	0.15	-0.3	0.51	1.68	1.02	0.26	-0.1
CPS90 L13	0.67	1.30	0.98	0.12	-0.3	0.56	1.58	0.99	0.18	-0.2
CPS90 L14	0.74	1.32	0.97	0.13	-0.5	0.65	1.50	0.97	0.20	-0.4
CPS91 L7	0.80	1.49	1.00	0.14	-0.1	0.68	2.04	1.00	0.27	-0.1
CPS91 L8	0.75	1.37	1.00	0.16	-0.3	0.62	2.96	1.01	0.35	-0.1
CPS91 L9	0.84	1.32	1.00	0.12	-0.2	0.65	2.33	1.01	0.28	-0.0
CPS91 L10	0.79	1.38	1.01	0.16	-0.1	0.70	1.48	1.02	0.26	-0.2
CPS91 L11	0.75	1.29	1.00	0.12	-0.2	0.70	1.47	1.02	0.21	-0.0
CPS91 L12	0.75	1.29	0.99	0.11	-0.3	0.66	1.48	1.00	0.16	-0.1
CPS91 L13	0.81	1.23	1.00	0.11	-0.1	0.72	1.37	1.00	0.17	0.0
CPS91 L14	0.91	1.17	1.00	0.06	-0.1	0.87	1.27	1.00	0.10	-0.1

CPS91 is 1,405. Hence there were a total of 5,623 (7028–1405) *different* students used in the equating.

Item Fit Statistics

For each independent Rasch analysis of a test level, values of the minimum, maximum, mean, standard deviation, and mean standardized infit and outfit mean squares for items are shown in Table 4. For any given test, the items share a common dimension that we may define as "reading" or "math", etc. Invariably, each item may contain several dimensions, each one deviating somewhat from the shared dimension across items. We would like, therefore, to see that the shared dimension to be the main overriding dimension so that the variations within items only

contribute to the "random noise" in dimension. Each of the test levels used has both infit and outfit mean squares very close to 1 and a small standard deviation of less than 0.3. Items with infit and outfit mean squares of between 0.4 to 1.6 still contribute meaningfully to measurement. In Table 4, the infit and outfit mean squares that are outside this range are for only one or two items as observed for each test level analysis. For each of them, however, the response strings and residuals were examined to determine if the low or large fit statistics could bring any real harm to the equating. It was found (for example for the items with large outfit mean squares of 2.32, 2.51, 2.04 and 2.96 of Form 7 Level 11, Form 7 Level 12, CPS91 Level 7 and CPS91 Level 8 respectively), that each of the large mean squares was contributed by only about six to eight students who had responded wrongly or correctly in an unexpected manner. Wright and Linacre (1994) explains that there are no "hard and fast rules" as to what range of infit and outfit mean squares that can be considered tolerable. A certain mix of test items while still useful to the purpose of the test, may be seen as off-target in a fit analysis. We can conclude that the items on all these ITBS test levels do share a dimension that can be defined as "reading" ability. These ITBS tests are "tests in use". The purpose of the fit analysis is not for dropping items that would tantamount to proposing a change in the ITBS test Forms, but more to ensure that the equating is reasonably successful in order that we can make a fair comparison between the use of Rasch person measures and that of grade equivalents in the measurement of growth.

Results and Findings

For each test level of Form 7, CPS90 and CPS91, the item calibrations were extracted from the analysis and the mean values calculated. Also the mean person measures were calculated for the group of persons taking each level. In addition to these mean calibrations and person measures obtained from the Rasch analysis, each person's Grade Equivalent and Percentile Rank were also obtained from the ITBS manuals. Table 5 shows the mean measures, mean grade equivalents and their respective standard deviations for common persons for Form 7 and CPS90 while Table 6 shows the corresponding values for CPS90 and CPS91. Table 7 shows the mean percentile ranks and their standard deviations for common persons for Form 7 and CPS90 while the corresponding values for CPS90 and CPS91 are shown in Table 8.

A series of graphs were plotted of the mean measures, mean grade equivalents and mean percentile ranks for these common persons, against grade. Form 7 Level 7 is for Grade 1, Level 8 is for Grade 2, etc. Similarly, for CPS90 and CPS91, Level 7 is for Grade 1, Level 8 is for Grade 2, etc. Another series of graphs were also plotted of the respective standard deviations against grade level. These graphs are shown in Figure 2 through Figure 13.

Table 5

Mean Measures, Mean Grade Equivalents and Standard Deviations of Common Persons for Form 7 and CPS90

ITBS Test Level	Grade	No. of Common Persons	Person Measure					Grade Equivalents				
			Form 7		CPS90		Diff Bet. Forms	Form 7		CPS90		Diff Bet. Forms
			Mean	S.D.	Mean	S.D		Mean	S.D.	Mean	S.D.	
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

Table 6

Mean Measures, Mean Grade Equivalents and Standard Deviations of Common Persons for CPS90 and CPS91

ITBS Test Level	Grade	No. of Common Persons	Person Measure					Grade Equivalents				
			Form 7		CPS90		Diff Bet. Forms	Form 7		CPS90		Diff Bet. Forms
			Mean	S.D.	Mean	S.D		Mean	S.D.	Mean	S.D.	
7	1	73	-2.17	0.40	-2.13	0.54	0.04	1.00	0.33	0.99	0.45	-0.01
8	2	77	-0.93	0.92	-0.97	0.79	-0.04	1.86	0.82	1.80	0.66	-0.06
9	3	130	0.13	0.82	0.22	0.90	0.09	3.08	0.84	3.22	0.84	0.14
10	4	-	-	-	-	-	-	-	-	-	-	-
11	5	119	1.11	0.84	1.13	0.86	0.02	4.61	1.19	4.52	1.20	-0.09
12	6	-	-	-	-	-	-	-	-	-	-	-
13	7	112	2.23	0.71	2.24	0.86	0.01	6.51	1.49	6.39	1.61	-0.12
14	8	141	2.86	0.80	2.28	0.71	0.02	6.93	2.18	7.23	1.82	0.30

Table 7

Mean Percentile Ranks and Standard Deviations of Common Persons for Form 7 and CPS90

ITBS Test Level	Grade	No. of Common Persons	Person Persentile Ranks				Difference Between Forms
			Form 7		CPS90		
			Mean	S.D	Mean	S.D.	
7	1	120	40.25	24.61	25.41	18.12	-14.84
8	2	154	29.04	21.97	20.09	19.45	-8.95
9	3	160	44.86	24.58	28.25	22.30	-16.61
10	4	-	-	-	-	-	-
11	5	175	43.81	25.04	37.54	24.85	-6.27
12	6	-	-	-	-	-	-
13	7	144	37.90	21.72	39.26	23.23	1.36

Table 8

Mean Percentile Ranks and Standard Deviations of Common Persons for CPS90 and CPS91

ITBS Test Level	Grade	No. of Common Persons	Person Persentile Ranks				Difference Between Forms
			Form 7		CPS90		
			Mean	S.D	Mean	S.D.	
7	1	122	14.55	10.59	14.72	13.64	0.17
8	2	91	20.54	22.16	42.41	29.76	21.87
9	3	131	30.61	20.55	34.37	20.76	3.76
10	4	-	-	-	-	-	-
11	5	121	27.57	22.05	26.31	21.69	-1.26
12	6	-	-	-	-	-	-
13	7	114	31.14	21.69	30.11	23.28	-1.03
14	8	144	33.71	22.60	33.47	20.22	-0.24

The data matrix shown in Figure 1 is only about 15% filled with data but Rasch Equating treats the empty spaces as missing data. This means that each person is considered to have taken a test of length equal to the number of items he attempted.

The large data matrix for Form 7 and CPS90 was analyzed so that each of the items is calibrated and the person-measures obtained. Then the large matrix of CPS90 and CPS91 was analyzed anchoring the item difficulties of CPS90 on the values obtained from the first analysis.

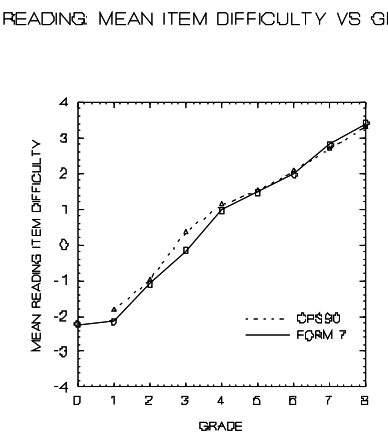

Figure 2. Mean Reading Item Difficulty vs Grade for CPS90 and Form 7

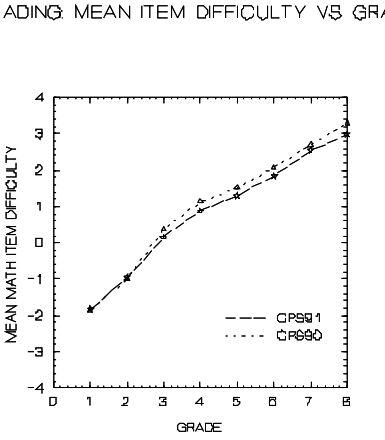

Figure 3. Mean Reading Item Difficulty vs Grade for CPS91 and CPS90

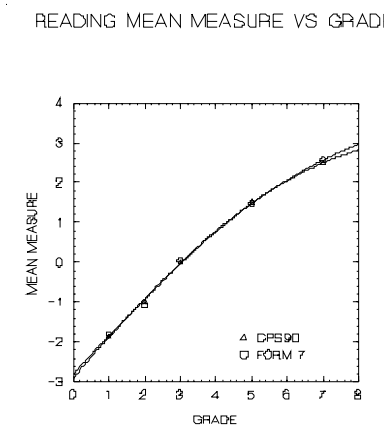

Figure 4. CPS90 and Form 7 Common Persons' Mean Measures

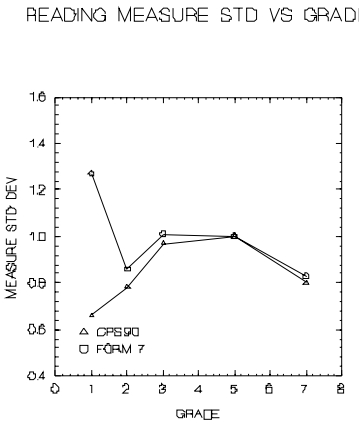

Figure 5. The Standard Deviations of Common Persons' Reading Measures for CPS90 and Form 7

This puts all the items of Form 7, CPS90 and CPS91 on to the same scale.

Figure 2 is a plot of the mean item difficulty of each test level by grade for CPS90 and Form 7 while Figure 3 is a similar plot for CPS90 and CPS91. These plots show that Form 7, CPS90 and CPS91 Reading tests difficulty levels are not the same, with CPS90 being a little harder than both Form 7 and CPS91 for most of the test levels. It is because these tests are of different difficulty levels that their item calibrations have to be known. When these item difficulties are known quantities that can be distributed along a line of measure, the tests are said to have been equated. Person measures obtained from tests with known difficulty levels, will be the same (within limits of measurement error) independent of which set of items is used to construct the test for them. This is clearly shown in Figure 4 where the mean measures for the groups of people who took the tests from grades 1 through 7, fall on the same curve for both test forms. Similarly, the mean measures fall on the same line for CPS90 and CPS91 as shown in Figure 10.

In contrast, the mean grade equivalents of the same groups of persons who took the tests (CPS90 and Form 7) are markedly different for all grades (Grades 1 through 7) as shown in Fig-

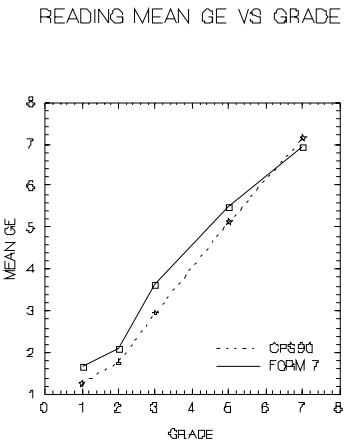

Figure 6. Common Persons' Mean Grade Equivalents vs Grade for CPS90 and Form 7

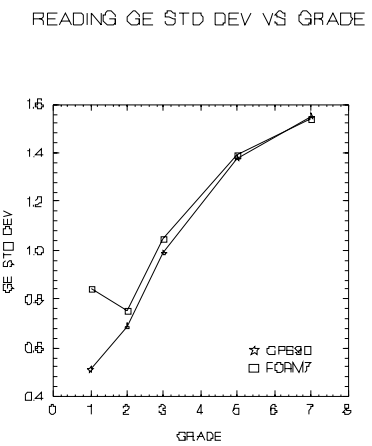

Figure 7. Standard Deviations of Common Persons' Reading Grade Equivalents vs Grade for CPS90 and form 7

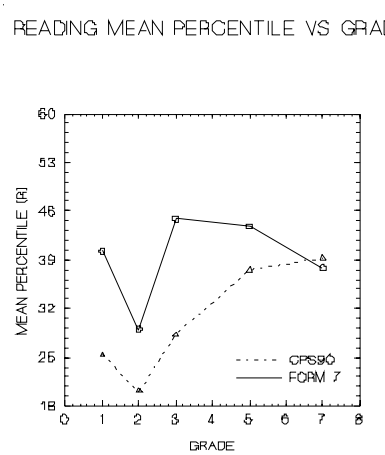

Figure 8. Reading Mean Percentiles of Common Persons' against Grade for CPS90 and Form 7

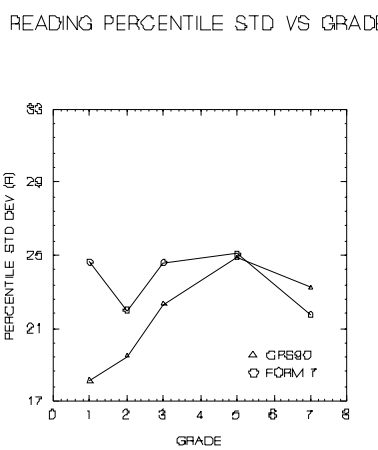

Figure 9. Standard Deviations of Percentiles against Grade of Common Persons' for CPS90 and Form 7

ure 6. This shows that equating using grade equivalents is not adequate. Figure 12 shows the same case for CPS90 and CPS91 although the differences are not as large.

Percentile ranks are norm-referenced values. As different test levels can be normed on different norm groups, the mean percentile ranks for the groups of persons at the different grade levels, fluctuate almost randomly as shown in Figure 8 (CPS90 and Form 7) and Figure 14 (CPS90 and CPS91). This does not make it possible to measure growth using percentile ranks.

Figure 5 shows the standard deviations of the mean measures by grade for CPS90 and Form 7 while Figure 11 shows the same plot for CPS90 and CPS91. Figure 7 is the plot for the standard deviations of the grade equivalents by grade for CPS90 and Form 7 while Figure 13 shows the same plot for CPS90 and CPS91. The corresponding plots of standard deviations of percentile ranks against grade are shown in Figures 9 and 15. The standard deviations of grade equivalents show that the students grow further apart as they develop their reading skills while the stan-

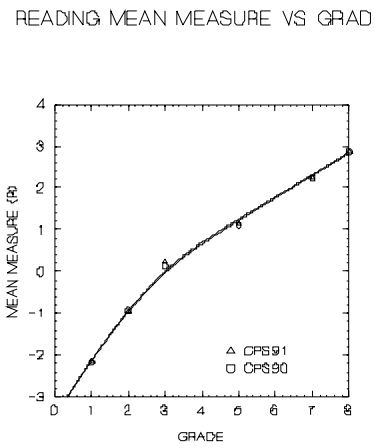

Figure 10. Reading Mean Measures of Common Persons against Grade for CPS90 and CPS91

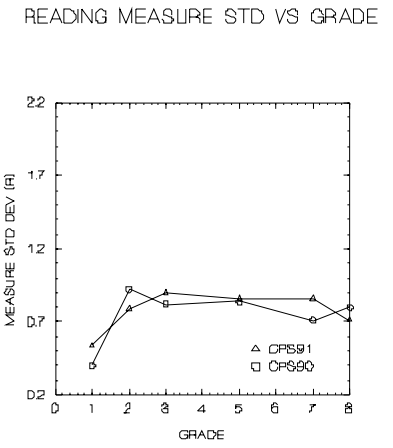

Figure 11. Standard Deviations of Measures of Common Persons against Grade for CPS90 and CPS91

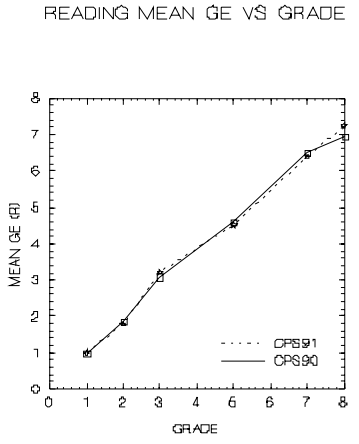

Figure 12. Reading Mean Grade Equivalents of Common Persons against Grade for CPS90 and CPS91

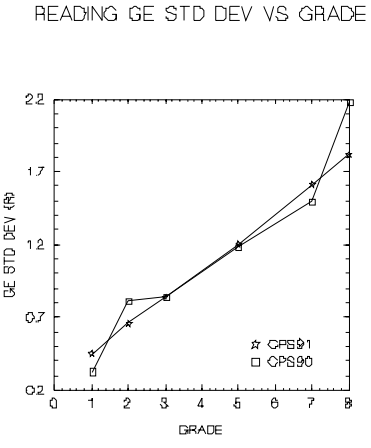

Figure 13. Standard Deviations of Grade Equivalents of Common Persons Against Grade for CPS90 and PS91

dard deviations of the percentile ranks and the Rasch measures do not show any definite pattern but rather random in nature. The standard deviations of the percentile ranks and Rasch measures show that the students can have different growth rates and that at any time point, the spread of reading abilities within the group is somewhat random. This is shown in Figures 5 and 11 where the overall standard deviation is more or less within a margin of 0.4 (from 0.6 to 1.0 in Figure 5). This observation of increasing standard deviation of grade equivalents is consistent with the explanation given by Berk (1981) in his article lamenting the problems with grade equivalents. He stated the following:

> Due to the relative stability of the average reading level in the population after junior high school and to the extrapolation of grade equivalents at the upper grade levels from the earlier grades, the growth curve between age

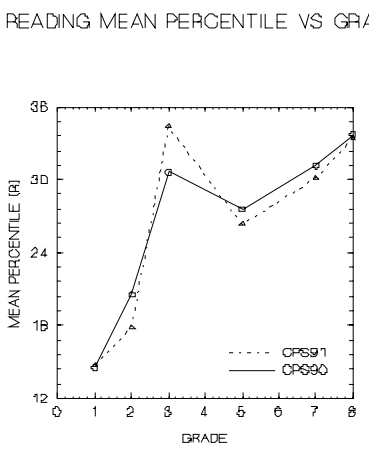

Figure 14. Reading Mean Percentiles for Common Persons against Grade for CPS90 and CPS91

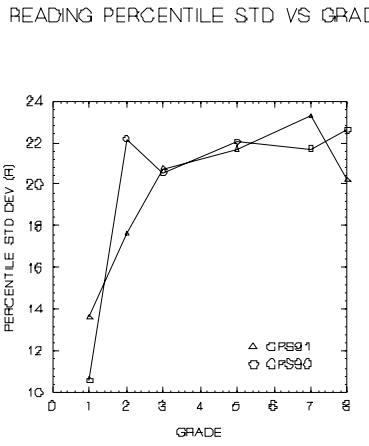

Figure 15. Reading Percentiles Standard Deviations of Common Persons against Grade for CPS90 and CPS91

> and achievement flattens out in the upper grades. This means that grade equivalents at these grade level become ludicrous with large fluctuations resulting from a raw score difference of only 2 or 3 points on a 100 item test. (p.136)

Reynolds (1981) provided a similar explanation stating that grade equivalents exaggerate small differences between individuals and that they are difficult to interpret for the higher grades because of their extrapolations from the lower grades.

Schulz, Perlman, Rice and Wright had also used Rasch Simultaneous Equating in their study of vertical equating of reading tests but the differences between that study and this, are that Schulz, et al. (1) covered grades 3 through 8; (b) equated a different set of tests, namely 24 Criterion Referenced Tests on reading, and (3) used MFORMS and MICROSCALE, older versions of the present Rasch model computer program. This study also differs from the investigation of the grade equivalent metric done by Frank and Seltzer (1990) in that (1) it covers the whole of the elementary school grades (1 to 8) and hence requires equating of the 8 levels of three test forms namely Form 7, CPS90 and CPS91 and (2) comparisons are also made between percentile ranks and Rasch measures.

Conclusion

Simultaneous vertical equating using Rasch analysis has the great advantage of running the calibration analysis on only a single matrix which therefore makes the estimates of item calibrations and person measures on a large amount of data thereby reducing the standard errors of measurement for each of the estimates. Calibration of a single test whose item difficulty levels are then used as an anchor for the equating on to the next test, would result in cumulative errors of measurement. Tests that are equated using the Rasch Simultaneous Vertical Equating, result in person measures that are the same, independent of which of the tests they take. The person's grade equivalent, however, does depend on which test form he takes. Percentile ranks on the other hand, may not have been calibrated using the same norm groups of the various test levels. When norm groups differ, a person moving up the grades may not show a systematic growth in percentile rank as he is now seen relative to his new group of peers which is different from the group in the earlier grade.

Teacher-made tests for school-based assessments can be similarly equated. With items that are calibrated, teachers can better select items that are more appropriate for their target groups. As

the known item calibrations can be used for anchoring, student abilities can be better estimated by running a Rasch analysis on each test given to students, irrespective of the test length for each student, or which items are given to them. Item-free student measures are then being put to good use in the classroom.

References

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical linear models*. London: Sage.

Berk, R. A. (1981). What's wrong with using grade-equivalent scores to identify LD children? *Academic Therapy*, *17*(2). Pp 133-140.

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

Lee, O. K., and Wright, B. D. (1992). *Mathematics and reading test equating*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lee, O. K. (1992). Measuring mathematics and reading growth. Unpublished doctoral dissertation, University of Chicago, Chicago, IL.

Linacre, J. M., and Wright, B. D. (2000). *A user's guide to WINSTEPS: Rasch model computer program*. Chicago: MESA Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum.

Reynolds, C. R. (1981). The fallacy of "two years below grade level for age" as a diagnostic criterion for reading disorders. *Journal of School Psychology*, *19*(4). Pp 350-358.

Smith, R. M. (1993). Guessing and the Rasch Model. *Rasch measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association*, *6*(4), pp 262-263.

Schulz, E. M., Perlman, C., Rice, W.K., and Wright, B. D. (1992) Vertically equating reading tests: An example from Chicago Public Schools. In Wilson, M. (Ed). *Objective Measurement: Theory into practice* (pp. 138-156). Norwood, New Jersey: Alex Publishing.

Wright, B. D. (1992). Scores are not measures. *Rasch measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association*, *6*(1), p. 208.

Wright, B. D. (1993). Thinking raw scores. *Rasch measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association*, *7*(2), p. 299.

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association*, *8*(3), p. 370.

Appendix A

Example of a Winsteps Control File

```
&INST
TITLE= C9L7 (R): UNCLEANED DATA
DATA=c9l7r.dat
MUCON=0
NAME1=3
NAMLEN=9
ITEM1=12
NI=56
XWIDE=1
CODES='ABCD12349  `
NEWSCR='111100000'
CATEGS=10
IFILE=c9l7r.itm
PFILE=c9l7r.per
TABLES=10100010001011001101110000
&END
Q1
Q2
Q3
.
.
.
Q54
Q55
Q56
END NAMES
```