Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 8, Number 23, November, 2003 ISSN=1531-7714

A Practical Use of Vertical Equating by Combining IRT Equating and Linear Equating

Shing On Leung,
University of Macau

It is a common practice to measure students' performance over a period of time, or to compare scores that assess different grades in a given time, or, even more, to analyze the trend of a subject in a particular grade over different years. For example, Russell (2000) used Expected Growth Size to estimate changes in test scores. These analyses quite often assume that scores have to be vertically equated. That is, scores for different students or at different times can be mapped in a common metric for comparison and analyses.

There are two types of test equating: horizontal and vertical equating (Kolen & Brennan 1995). Horizontal equating involves equating tests of different forms or at different times of a single grade; while vertical equating involves equating tests of different grades or levels. Usually, vertical equating is performed at a given time period, and more complications will be involved for equating tests of different time.

There are different studies on vertical equating for establishing developmental scales for grade equivalent and representation of growth (Schulz & Nicewander 1997; Williams, Pommerich & Thissen 1998 and Yen & Burket 1997). However, no common consensus is arrived and different results are obtained in different studies, particularly on whether variance will be increase or decrease when grade increases. In particular, there are serious off-level effects when vertically equates students with very different grades, e.g., grade one through six. The purpose of this paper is to introduce a simple use of Item Response Theory (IRT) equating with linear equating that can address this problem.

This study presents a practical, direct and simple method of vertical equating for the Chinese competence of primary school students in Hong Kong. It is a combination of three-parameter item response equating and linear equating. Item response equating is used to vertically equate persons in two adjacent grades with overlapping items (Harris & Hoover 1987). Then, students will have two different ability scores from two separate analyses. For example, grade two students will have one ability estimate from the analysis of grade one and two together, and another estimate from analyzing of grade two and three. Linear equating using simple least squared regression is then used to equate scores from different analyses (Kolen & Brennan 1995). Repeating uses of these procedures will result in a through metric from grade one to six. This can avoid the so-called off-level effect, which very often occurs when item response equating is used in equating abilities of very different levels (O'Brien & Tohn, 1984).

Data

Data is obtained through parts of pilot studies of a large scaled research on Chinese competence of primary students in 1999. This is a research based low stakes test and is conducted in classroom environment with the cooperation of the school administrators. Items are assigned to each level according to another study conducted by subject experts based on Chinese phrases used in textbooks in each grade. Theoretically, grade 2 students may not know items at level 4. However, apart from textbooks in formal school curriculum, there are other influence from family and society. It is the purpose of these studies to investigate the general Chinese competence in each grade.

An overlapping design in adjacent grades is employed to link up ability of students in adjacent grades. All items are multiple-choice items and are ready to be analyzed by usual software for Item Response Theory. An overlapping design is used with details in Table 1.

Table 1: Number of Items for Test Papers in Different Grades at Different Levels

Grade	LEVEL						Total	N: number of students
	1	2	3	4	5	6		
1	43	12					55	733
2	12	10	12	10			44	671

3			59	14			73	834
4			12	10	12	12	46	816
5					59	12	71	938
6						69	69	845
Total	43	12	59	14	59	69		4837
(Note: This is an overlapping design. For example, the 12 items used in grade 2 at level 1 are taken from those 43 items used in grade 1 at level 1. Hence, in the last row, the total number of items at each level is the maximum number used at each column, i.e., each level.)								

For papers in grade 1, 3, 5 and 6, the number of items at corresponding levels (i.e., level 1 for grade 1) is large in order to fulfill requirement in different content areas. For papers in grade 2 and 4, there is no such requirement and the only purpose is to provide as much overlapping as possible to link up information from grade 1 through 6, i.e., to develop a vertical metric among all grades. The sample sizes vary from over 600 to 900, depending on the grades, and are also reported in Table 1.

Method

One of the main purposes of the project is to establish a common metric from grade one to six on students' Chinese competence. Item response theory is used first as a flexible mean of equating test papers with overlapping items. However, there is off-level effect (or out-of-level effect) in putting students with very different ability level together (O'Brien & Tohn, 1984). The research team has tried to put all data together in one mega-step IRT equating, but results show that there are serious off-level effects. To alleviate this problem, a staged step-by-step procedure, which combines item response equating and linear equating, is described as follows.

To start with, students' abilities in two adjacent grades are equated with overlapping items and three-parameter item response model. Five analyses are denoted as P1P2 (for grade one & two), P2P3, P3P4, P4P5 and P5P6. There is one ability estimate for each student in each analysis. Hence, there will be two ability estimates for each student from grade two to five altogether, as described in Table 2.

Table 2: Ability Estimate for Students from grade one to six in different analyses

Grade	Analysis				
	P1P2	P2P3	P3P4	P4P5	P5P6
1	√				
2	√	√			
3		√	√		
4			√	√	
5				√	√
6					√

Further, as each student only tested once, the estimate of his / her ability should be the same in two different analyses. To align two different scales in two different analyses, linear equating simple least squared regression of ability estimates from grade two to five is used. There are four linear equations for students in:

1. grade 2 (x from analysis P1P2; y from analysis P2P3),
2. grade 3 (x from analysis P2P3; y from analysis P3P4),
3. grade 4 (x from analysis P3P4; y from analysis P4P5), and

4. grade 5 (x from analysis P4P5; y from analysis P5P6).

From these four linear equations, a single metric can be formed as follows. The y in equations 1 will be substituted in x in equation 2 because they are from the same analysis, and similarly for y in equation 2 to x in equation 3, etc. In the end, a single metric will be formed. Regression here is used as a mean for interpolation rather than prediction.

Results

The first step of the analyses involves vertically equate ability scores of two grades using three-parameter item response theory together with overlapping items in two adjacent test papers in two grades. Table 3 shows the reliability coefficients from various analyses. In Table 3, analysis P1 refers to calibrating items use only grade 1 students, and similarly for P2 to P6. These analyses are used for individual calibration to assess the quality of equating, and will be described later. Most of the reliabilities in all analyses are very high with some of them are moderately high. These indicate that these scales are sufficient for further analysis.

Table 3: Reliabilities from various analyses

Analysis	N	Items	Reliability	Analysis	N	Items	Reliability
P1	736	55	0.900	P1P2	1525	77	0.840
P2	789	44	0.909	P2P3	1623	95	0.916
P3	834	73	0.907	P3P4	1674	97	0.889
P4	840	46	0.834	P4P5	1776	93	0.915
P5	936	71	0.842	P5P6	1775	128	0.717
P6	839	69	0.867				

Note: Total number of items in analysis P1P2 does not equal to the sum of P1 and P2 because of overlapping items, and similar to others.

Various statistics are computed for students' ability scores in Table 4. They all look roughly the same in different analyses and similar to the standard normal distribution (mean zero and standard deviation one), and this can be confirmed with frequency distribution curves. Since all these scale scores are obtained from different analyses, they cannot be compared directly unless we equate all scores into a common metric. Simple linear equating does this.

Table 4: Various Statistics of Students' Ability Scores in Different Analyses

Analysis	Mean	Median	Std. Deviation	First Quartile	Third Quartile
P1P2	-0.029	0.063	1.196	-0.647	0.797
P2P3	-0.039	0.159	1.211	-0.706	0.799
P3P4	-0.011	0.126	1.214	-0.625	0.769
P4P5	0.000	0.112	1.204	-0.631	0.786
P5P6	0.006	0.117	1.158	-0.600	0.734

For students in grade two, there will have two different scores from two analyses: P1P2 and P2P3. A simple linear equation is formed with x be the scale score from P1P2 and y from P2P3. This is performed by SPSS usual default of least square method. Similarly, three other equations are formed for in grade 3, 4 and 5, and results are shown in Table 5.

Table 5: Linear equations for Grade 2 to 5 Students

Grade	Analysis where x come from	Analysis where y come from	equation	Correlation
2	P1P2	P2P3	y = -0.790 + 1.117 * x	0.9713
3	P2P3	P3P4	y = -0.724 + 0.997 * x	0.9962
4	P3P4	P4P5	y = -0.790 + 1.118 * x	0.9736
5	P4P5	P5P6	y = -0.305 + 1.029 * x	0.9992

In a perfect situation, all correlation will be one. Since there are errors in scales, correlations are not one but high enough to show their predictive power. All slopes are roughly equal to one because the variances of scores are roughly the same, and one unit in one scale corresponds to also one unit to the other. It is worth to note that all intercepts are negative. Take grade two students as examples. Scores in analysis P2P3 is roughly 0.8 less than those in P1P2. This is because we are comparing them with grade three students (which have highly ability) in P2P3, but grade one students in P1P2. Obviously, they get lower scores in P2P3 than those in P1P2. Similar interpretation holds for other three equations. Therefore, when we move from grade one to six, there will be a shift from lower ability to higher ability.

With these four equations, we can form a single metric for students' competence in Chinese from grade one to six. Further, the scale, denoted by Y1, is re-scaled linearly T-score with mean 50 and standard deviation 10 for easy interpretation. Different statistics in this single scale from grade one to six are shown in Table 6.

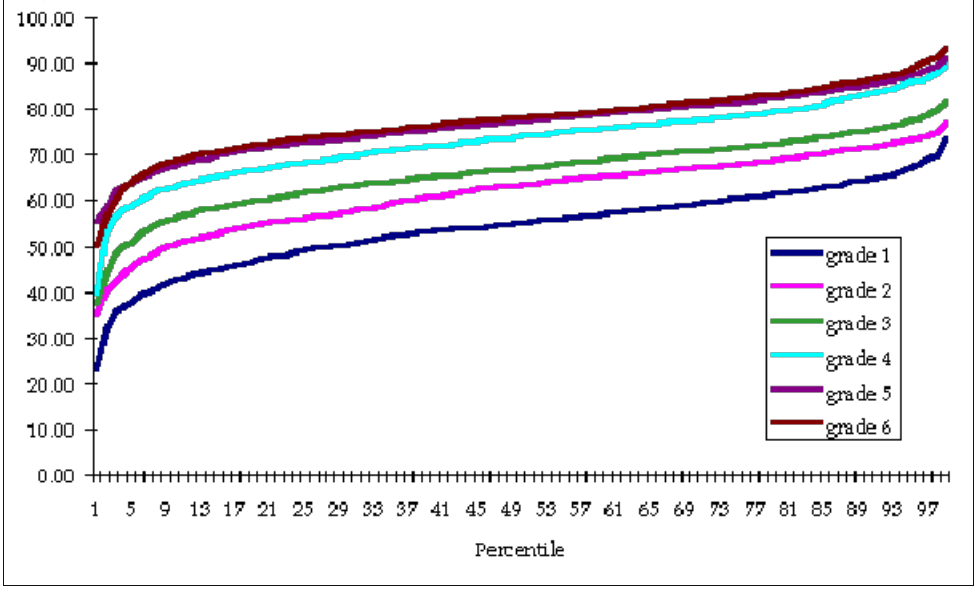
Table 6: Various Statistics of Students' Ability Scores in the Single Scale (Y1) from Grade One to Six

Grade	Mean	Median	Std. Deviation	First Quartile	Third Quartile	Inter-quartile Range
1	37.71	38.54	8.14	33.61	43.07	9.46
2	44.08	45.60	7.78	39.39	49.33	9.95
3	47.76	48.52	7.13	44.19	52.45	8.26
4	53.71	54.46	7.28	49.76	58.26	8.50
5	56.73	57.30	6.07	53.29	60.62	7.33
6	57.35	57.92	6.70	54.20	61.50	7.29

From Table 6, as expected, both mean and median increase with grades, indicating increases in Chinese competence. Further, the marginal increases in competence diminish as grade increases and students getting older: the difference between grade one and two is the biggest while between grade five and six is the smallest. This confirms the views that students learn faster when they are younger. Regarding the changes in variation in competence through grades, there is no consistent pattern both in terms of standard deviation and inter-quartile range. This is probably consistent with ***real*** situations as no consensus is found in literature. Please note that results in this study are obtained not from simulation but real pilot. Other kinds of comparison can also be made, e.g., what is the proportion of students in grade two below the mean of grade one, and similar.

Further, from Figure 1, when we look at the cumulative frequency curve in each grade, the distribution is nearly the same as indicated by six parallel curves for six grades. There is an upward shift to adjacent higher grades. And, the shift gradually diminishes as grade increases. This means that the improvement at lower grades is higher than those in the upper. It happens at all ability level within each grade, as the shift is roughly the same for each value in the ability scale.

Figure 1: Cumulative Frequency Curve from Grade 1 to 6



Comparison with other method and quality of equating

To compare this equating method with others, another single metric (Y2) is formed by fixed parameter IRT equating. Items in grade 1 are first calibrated. Then, by fixing overlapping items at the values in grade 1, items in grade 2 are also calibrated, and similarly for grade 3 to 6. Finally, Y2 is re-scaled to T-score with mean 50 and standard deviation 10. To assess the quality of equating, Y1 and Y2 are correlated with scales from individual calibration in each grade, i.e., use only samples in each grade. Table 7 shows these correlations, together with mean and standard deviation of Y1 and Y2 in each grade.

Table 7: Mean, standard deviation and correlation with individual calibration of Y1 and Y2 in each grade

Grade	Mean		Std Deviation		Correlation with individual grades	
	Y2	Y1	Y2	Y1	Y2	Y1
1	36.61	37.71	9.76	8.14	1.0000	0.9748
2	45.28	44.08	6.80	7.78	0.9946	0.9889
3	48.30	47.76	7.02	7.13	0.9867	0.9968
4	53.75	53.71	7.27	7.28	0.9163	0.9600
5	56.41	56.73	4.62	6.07	1.0000	0.9997
6	56.97	57.35	7.14	6.70	0.8857	0.9999
Overall	50.00	50.00	10.00	10.00	na	na-

Note: The label 'na' stands for not applicable, since it is meaningless to combine scale values from individual calibrations in each grade.

The means between Y1 and Y2 are very similar in each grade, indicating that both scales show a development in Chinese competence through grades. Further, the standard deviations are also similar, indicating that there is no clear trend of increases or decreases in variations through grades. In grade 1, the correlation between Y2 and individual calibration is 1 by definition, since it is the scale to start with. There seems a trend of decreases in correlations through grades in Y2, except in grade 5. On the other hand, the corresponding correlations in Y1 are fairly constant. Some correlations in Y1 are higher than Y2, with some are lower, with all very near or equal to 1, indicating that the quality in both equating methods is good. However, relatively low correlations of Y2 in grade 4 and 6 give us a warning.

The above results show that Y1 and Y2 are similar. Y1 is based on equating persons from different analysis, while Y2 is based on equating items in different calibrations. Since, usually, the number of persons far exceed number of items, there is reason to suspect that Y1 may perform a little bit more stable. If number of bad items is large, especially if large number of them exists in the overlapping items, there are possible danger is using Y2. Further, we start from grade 1 in forming the scale Y2 in fixed parameter equating, the low correlation in grade 6 lead us to suspect that there may be possible off-level effects if we are projecting too far away. However, having said all these, there is further research needed in this area.

Conclusion

This study demonstrates a practical use of vertical equating to obtain a metric of six grades. The method proposed is a simple staged procedure which recursively combines item response equating and linear equating. Item response equating is used in vertically equate two adjacent grades with overlapping items. However, there may be possible off-level effect if we tried to vertically equate six grades. Therefore, linear equating is used as a complement. Though this study uses Chinese as the subject, the method proposed here is applicable to other areas.

Finally, it is noted that vertical equating here is a method for numerical comparison among grades to overcome off-level effects. Whether it is meaningful to do so depends on other factors, e.g., content domain, curriculum and instructions conducted, etc. The end users have to consider these factors when interpreting scores from vertical equating.

Reference

Harris, Deborah J. & Hoover, H.D. (1987). An Application of the Three-Parameter IRT Model to Vertical Equating. *Applied Psychological Measurement,* 11(2), 151-159.

Kolen, M. J. & Brennan, R. L. (1995). Test equating: Methods and Practices. New York: Springer-Verlag.

O'Brien, Michael L. & Tohn, Diane (1984). Applying and Evaluating Rasch Vertical Equating Procedures for Out-of-Level Testing. Paper presented at the Annual Meeting of the Eastern Educational Research Association (West Palm Beach, FL, February 10, 1984).

Russell, Michael (2000). Using Expected Growth Size Estimates to Summarize Test Score Changes. *Practical Assessment, Research & Evaluation*, 7(6). Available online: http://pareonline.net/getvn.asp?v=7&n=6.

Schulz, E. Matthew & Nicewander, W. Alan (1997). Grade Equivalent and IRT Representations of Growth. *Journal of Educational Measurement*, 34 (4), 315-331.

Williams, Valerie S.L., Pommerich, Mary & Thissen, David (1998). A Comparison of Developmental Scales Based on Thurstone Methods and Item Response Theory. *Journal of Educational Measurement*, 35(2), 93-107.

Yen, Wendy M. & Burket, George R. (1997). Comparison of Item Response Theory and Thurstone Methods of Vertical Scaling. *Journal of Educational Measurement*, 34(4), 293-313.

Woldbeck, Tanya (1998). Basic Concepts in Modern Methods of Test Equating. Paper presented at the Annual Meeting of the Southwest Psychological Association (New Orleans, LA, April 1998).

Descriptors: equating; IRT; Item Response Theory; Test Equivalence ; Calibration; Equated Scores

Citation: Leung, Shing On (2003). A practical use of vertical equating by combining IRT equating and linear equating. *Practical Assessment, Research & Evaluation*, 8(23). Available online: http://PAREonline.net/getvn.asp?v=8&n=23.