LINKING AND ANCHORING TECHNIQUES IN TEST EQUATING USING THE RASCH MODEL

Guangzhong Luo, Anthony Seow &
Chew Lee Chin

Linking and Anchoring Techniques in Test Equating Using the Rasch Model

Guangzhong LUO, Anthony SEOW & CHEW Lee Chin
National Institute of Education
Nanyang Technological University
1, Nanyang Walk
Singapore 637616

Abstract

In analyzing computer assisted tests, the comparison between batches of examinees across different time frames is very important. In real testing situations, though great effort has been paid to construct parallel tests with the aim to ensure the fairness of the comparison between different cohorts using different tests, and overlap of test items in these different tests is often seen, an overall comparison is still not operational due to the existence of missing blocks in the data. In the traditional test theory (TTT), the standardization criterion is commonly used for the purpose of comparing the examinees' achievement. However, the comparison is unarguable only when the examinees are considered from the same population.

Using the Rasch model in the Item Response Theory (IRT), this paper presents the Linking and Anchoring techniques for comparing or equating tests. The advantages of these techniques are (1) the requirement that the examinees must be from same population is lifted and thus the comparison of examinee performance can be conducted across different academic levels; (2) the techniques are applicable even when the historic raw data are not available; and (3) the calibration of items using these techniques provides essential and comparable indexing reference for item banking.

Keywords

Test Equating, Linking and Anchoring Techniques, Rasch Model.

Introduction

One great challenge for practitioners of educational assessment is how direct comparisons of achievement can be made between cohorts of students across different years or different schools. Although traditional test theory (TTT) is still widely used in educational measurement, it has limited efficiency with regard to this measurement problem for when different tests are administered to different students and their performance on a test is expressed in terms of raw scores, student ability cannot be compared based on the total score obtained on the test. A point in case, is that of a student of lower ability who may have obtained a higher score on an easy test compared to another student of higher ability who may have obtained a lower score on a more difficult test. On the basis of the TTT, the z-score is used as a standardization criterion to overcome this problem; but a comparison using z-score is indisputable only when examinees considered are from the same population.

The development of a modern test theory and the rapid advances in computer technology over the last decade hold promise in meeting this challenge in educational measurement. One important characteristic of the modern test theory is its "*specific objectivity*" (Rasch, 1961). That is, a comparison of student achievement is independent of test-items used. Not only does this application of the theory allow for fairness in comparing test performances of students within a cohort who are administered a same test, but it also allows for comparison of students who are administered different tests or who are in different cohorts. This is possible because the theory allows for systematic missing blocks of data to be analysed but with the proviso that there is enough "linkage" between any two tests. In estimating person and item locations based on modern test theory, artificial origins are created on the scaling continuum, that is, the practice of data analyses applies the constraint that the locations of items involve a sum up to zero. This may be a potential limitation but it can be overcome by using the technique of anchoring analysis, which is operational under the modern test theory.

The National Institute of Education in Singapore has developed a comprehensive testing software known as "NIE Computerised English Language Test" or, NIECELT, that is capable of administering a test or a number of tests to any specified number of examinees at the same time or at different times (Hsui, Seow, & Chew, 1997; 1999). An issue challenging the use of NIECELT is the comparability of test scores when examinees take different test forms or when test forms are used interchangeably. Test equating techniques would be needed to determine the relationships between the scores obtained on the two test forms.

This paper presents the linking and anchoring techniques for equating different tests using the Rasch model. First, the algorithms operationalized in **RUMM** (Andrich, Sherridan & Luo, 1990-2001) are described. An example using a data set collected for an empirical study is used to illustrate the application of the techniques. Discussions on the use of test equating techniques and its attendant advantages are then provided.

Algorithms In RUMM

Linking Analysis. For a complete data set, the estimation procedure of RUMM is as follows:

Step 1 Calculate sufficient statistics for item parameters;

Step 2 Estimate item parameters (locations, units, skewnesses and kurtosises as well as derivative item thresholds) using pairwise estimation algorithm. This step does not involve person location parameters;

Step 3 Estimate person location parameters using the values for item parameters estimated in Step 2.

It is noted that as the pairwise algorithm is used, the calculation of the sufficient statistics for item parameters allows for missing data. Consequently, the estimation of item parameters are operational even when the data set involves *systematically* missing blocks which are the result of the test design rather than students' own "missing" responses. Therefore, when two or more cohorts using different tests with overlapping items between them and the responses for each test are available, the linking is relatively straightforward. The procedure is as follows:

Step 1 Merge data sets for different tests. For example Test A involves items 1, 2, 3 and 4; Test B involves items 3, 4, 5 and 6; Test C involves items 2, 6, 7 and 8; Students A1,A2, A3 and A4 take Test A, Students B1, B2, B3 and B4 take Test B and Students C1, C2, and C3 take Test C. Given these conditions the merged data set would be as those shown in Table 1.

Table 1 Merged Data Set For Tests A, B And C

Item Person	1	2	3	4	5	6	7	8
A1	▧	▧	▧	▧				
A2	▧	▧	▧	▧				
A3	▧	▧	▧	▧				
A4	▧	▧	▧	▧				
B1			▨	▨	▨	▨		
B2			▨	▨	▨	▨		
B3			▨	▨	▨	▨		
B4			▨	▨	▨	▨		
C1		▩				▩	▩	▩
C2		▩				▩	▩	▩
C3		▩				▩	▩	▩

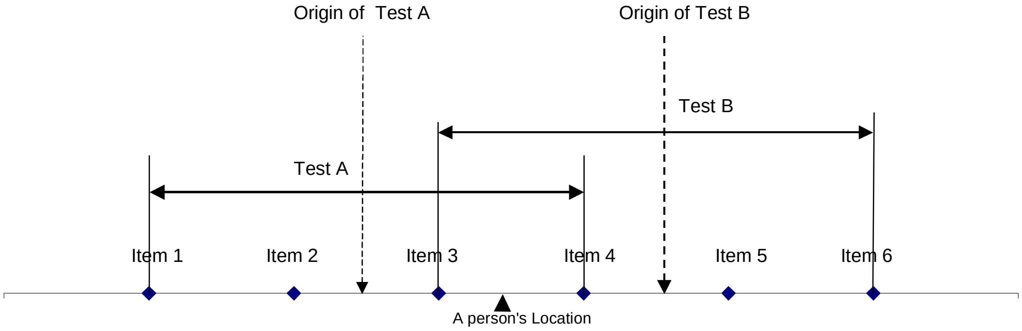
Step 2 Estimate item and person parameters using the merged data set. With the merged data set, all item and person parameters can be estimated in a single analysis. For detailed instructions for setting up a project and an analysis, see RUMM laboratory (2001).

In addition, RUMM provides a facility for comparing test forms separately based on the merged data set.

Anchoring Analysis. Linking analysis is the preferred procedure for test equating when all raw data are available. However, It is not uncommon in practice that the historic data to which the current test is to be equated are no longer available. Only the estimated values of item parameters based on the historic data are in the records and these parameters are often in the form of locations, units, skewnesses and kurtosises if the historic data were processed by RUMM, or in the form of item locations and thresholds if the data were processed by other programmes using IRT models.

As mentioned earlier, an advanced feature of the Rasch model is that a comparison of students' achievement is independent of the items used. However, the origin of the location continuum is arbitrary. Conventionally, to overcome the arbitrariness, a constraint on the sum of locations of items involved is set to be zero. Therefore, the origin of the location continuum depends on the "true" item locations in the test. With reference to the example illustrated in Table 1, items 1 and 2 are easier than items 5 and 6. If the data for Test A and Test B were processed separately, the origin of Test A would be smaller than that of Test B. Figure 1 shows the different locations of the origins of these two tests.

Figure 1. The Origins Of Test A And Test B



Therefore, in order to compare the estimation of person locations in Test B (the current test) to those in Test A (the historic test), it is necessary to adjust the origin of Test B. A typical way to do it is to anchor the origin of Test B back to the origin of Test A. This is termed ***Relative Anchoring Analysis*** in RUMM. The procedure is as follows:

Step 1 Identify the common items for the historic test and the current test. The number of common items is denoted as K;

Step 2 Retrieve the estimated values for the parameters of the common items. They can be in the form of locations, units, skewnesses and kurtosises or item thresholds;

Step 3 Calculate M_0, the mean of the retrieved locations for the common items;

Step 4 Estimate the parameters for all items in the current test with the constraint that the sum of the item locations is zero;

Step 5 Calculate M_1, the mean of the estimated locations for the common items based on the current test data. The difference between M_1 and M_0 is
$M= M_1- M_{0;}$

Step 6 Adjust the origin of the location continuum for the current test by M. That is, for any item i in the current test with an estimated location δ_i, the adjusted location is

$$\tilde{\delta}_i = \delta_i - M \ .$$

After the adjustment, the mean of the estimated locations for all items in the current test is

$$\frac{\sum_{i=1}^{I} \tilde{\delta}_i}{I} = \frac{\sum_{i=1}^{M} (\delta_i - M)}{I} = \frac{\sum_{i=1}^{M} \delta_i - IM}{I} = \frac{0 - IM}{I} = -M \ ;$$

where I is the number of items in the current test. As the result, the mean of the estimated locations for the common items in the current test is:

$$\frac{\sum \tilde{\delta}_i}{K} = \frac{\sum (\delta_i - M)}{K} = \frac{\sum \delta_i - KM}{K} = \frac{KM_1 - KM}{K} = M_1 - (M_1 - M_0) = M_0 \, .$$

Step 7. Finally, the person locations are estimated using the item parameters estimated in Step 6.

Simulation studies using RUMM have shown that when both historic and current tests fit the Rasch model, adjusted locations for the common items in the current test are similar to their historic values. But when neither the historic nor the current test fits the Rasch model well, there would be a noticeable difference between the two sets of location estimation for the common items after the adjustment described in Step 6.

In practice, it is often that the data of a historic test are not available and that the parameters of common items cannot be altered as these have been reported earlier to an authority or the public. In this case, ***Absolute Anchoring Analysis*** is used to accommodate the problem. The procedure in RUMM is as follows:

Step 1 Identify the common items for the historic test and the current test. The number of common items is denoted as K;

Step 2 Retrieve the estimated parameters of the common items;

Step 3 Calculate the sufficient statistics for item parameters using data from the current test;

Step 4 Estimate all item parameters. After each estimation cycle, restore the location values for common items back to the values retrieved in Step 2. Repeat the estimation cycle until convergence is achieved;

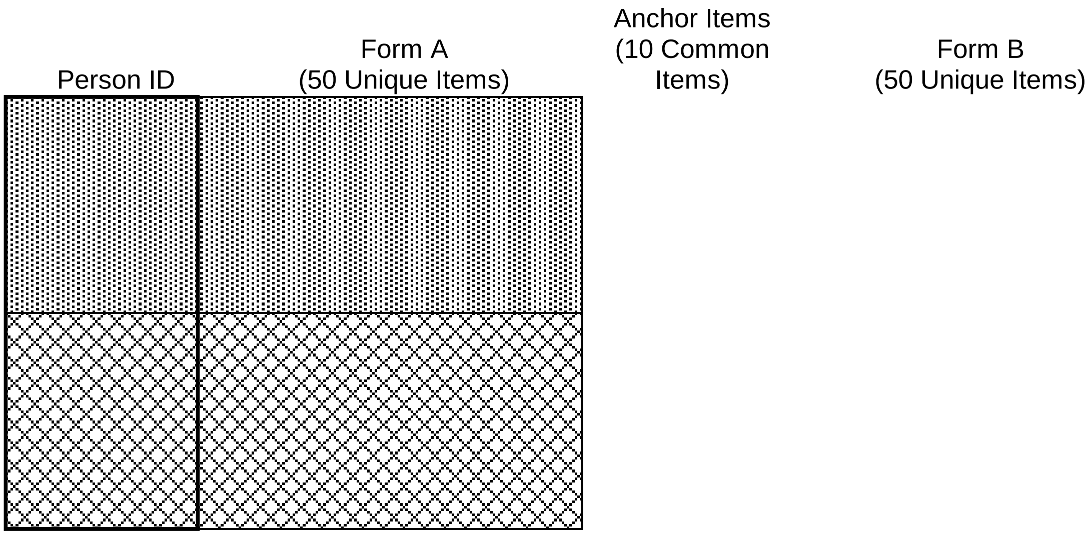
Step 5 Estimate person locations using the values of item parameters estimated in Step 4.

An Example On An Application Of RUMM

An example using a data set collected for an empirical study is now used to illustrate an application of RUMM to equate two parallel forms of a test.

Data Source. The data source came from a population of Singapore secondary students who took a 2-year biology course. Using a sampling framework based on school type (government/government-aided/independent schools) and the geographical location of schools, about 2000 students from 26 schools were sampled. During the data collection, the two biology test forms, named Form A and Form B, were spiralled and given out to students as they were ordered. This was to obtain two randomized but equivalent groups of students from each participating school. The test length of each form was 60 items with 50 unique items and 10 anchor items (i.e. common items to both test forms). Figure 2 shows the test design.

Figure 2. Test Design Using Common Test-Items Anchors



Data Analyses. Descriptive statistics of the test data were first determined. For each test form, the mean and standard deviation of the test scores were computed. In addition, the Cronbach alpha coefficient was calculated as a way of assessing the internal consistency of the test items.

Estimations of item and person locations were performed using the RUMM (Andrich, Sherridan & Luo, 1990-2001) statistical software. This software modeled the empirical data based on the one-parameter logistic model (Rasch) using the pairwise likelihood estimation (PLE) procedure to estimate both the item and the person locations. Concurrent calibration of the items on the two parallel test forms was first conducted. The anchor-item test design implemented in the study provided the linking of the test forms to place parameter estimates on the same scale. Estimates of item locations were also obtained from separate calibrations of the parallel forms. Finally, analyses were performed on the basis of relative anchoring and of absolute anchoring. The following are the results of the analyses.

Descriptive statistics. Descriptive statistics of the raw scores for the two test forms are shown in Table 2. One notable result is that the means and standard deviations of the two test forms differ slightly. Form B shows a higher mean score than Form A (47.01 versus 44.62) but a lower standard deviation (7.08 versus 7.61). The Cronbach alpha reliability coefficients of the two forms ranged from .83 to .84. These results indicate good test reliabilities.

Table 2 Descriptive Statistics Of Raw Scores For The Test Forms

	Form A	Form B
No of items	60	60
No. of persons	975	942
Mean score	44.62	47.01
Standard deviation	7.61	7.08
Cronbach alpha	.84	.83

Linking Analysis. Summary statistics based on the Rasch model for the two test forms are shown in Table 3. As the scale was centred on item location for all analyses, the mean of item locations for each test is zero. Comparing the two parallel forms, Form A shows greater dispersion of item locations around the mean (1.01 versus .87). Both forms had good test reliabilities. Results show a mean of 1.51 (.84) for the person locations on the overall test. Compared to Form A. Form B had a higher mean (1.64 versus 1.40) for person locations.

Table 3 Summary Statistics Of The Estimated Item And Person Locations

	Overall Test	Form A	Form B
Item Estimates			
No. of Items	110	60	60
Mean	.00	.00	.00
SD	.94	1.01	.87
Person Estimates			
No. of persons	1916	974	941
Mean	1.51	1.40	1.64
SD	.84	.86	.82
Reliability of test	.81	.82	.79

Table 4 Location Estimates Of Anchor Items

Item No.	Overall Test	Form A	Form B
3	-0.66	-0.74	-0.56
7	-1.22	-1.39	-1.03
11	-0.11	-0.23	0.03
16	0.46	0.45	0.49
30	0.43	0.37	0.51
39	1.23	1.21	1.27
42	1.24	1.16	1.34
47	-0.53	-0.61	-0.43
54	-0.36	-0.55	-0.14
58	-1.65	-1.67	-1.61
Mean	-1.17	-2.00	-0.13
SD	0.96	0.99	0.95

Table 4 shows the location estimates of the anchor items obtained from a concurrent analysis of the overall test data, and those from separate analyses of the data for parallel forms. For the overall test, the mean location of the anchor items is -1.17. Compared to Form A, Form B had a smaller mean location (-0.13 versus -2.00). A plot of the location estimates of the anchor items on the two test forms is shown in Figure 3. It is notable, from a visual inspection of the plots, that the locations of anchor items in general had higher values in Form B than in Form A.

Figure 3. Plot Of Location Estimates Of Anchor Items

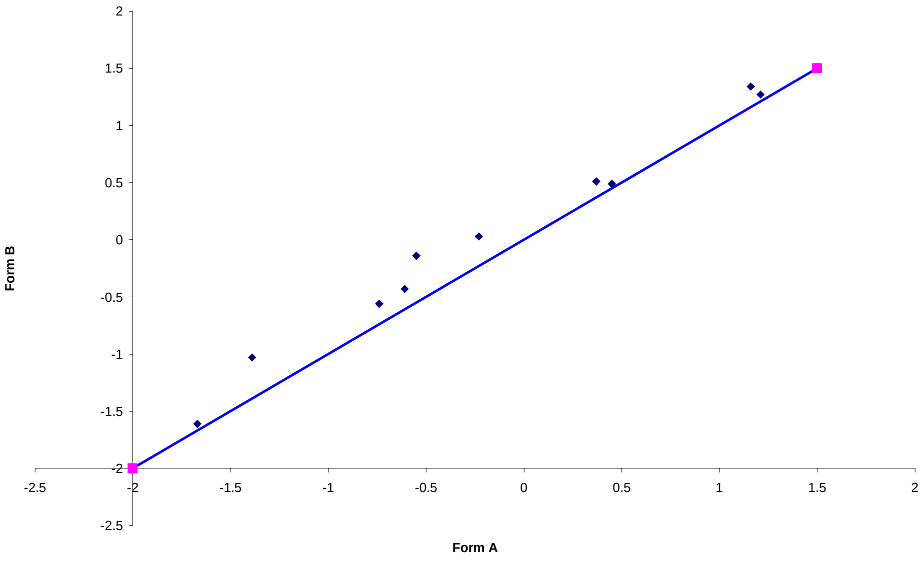
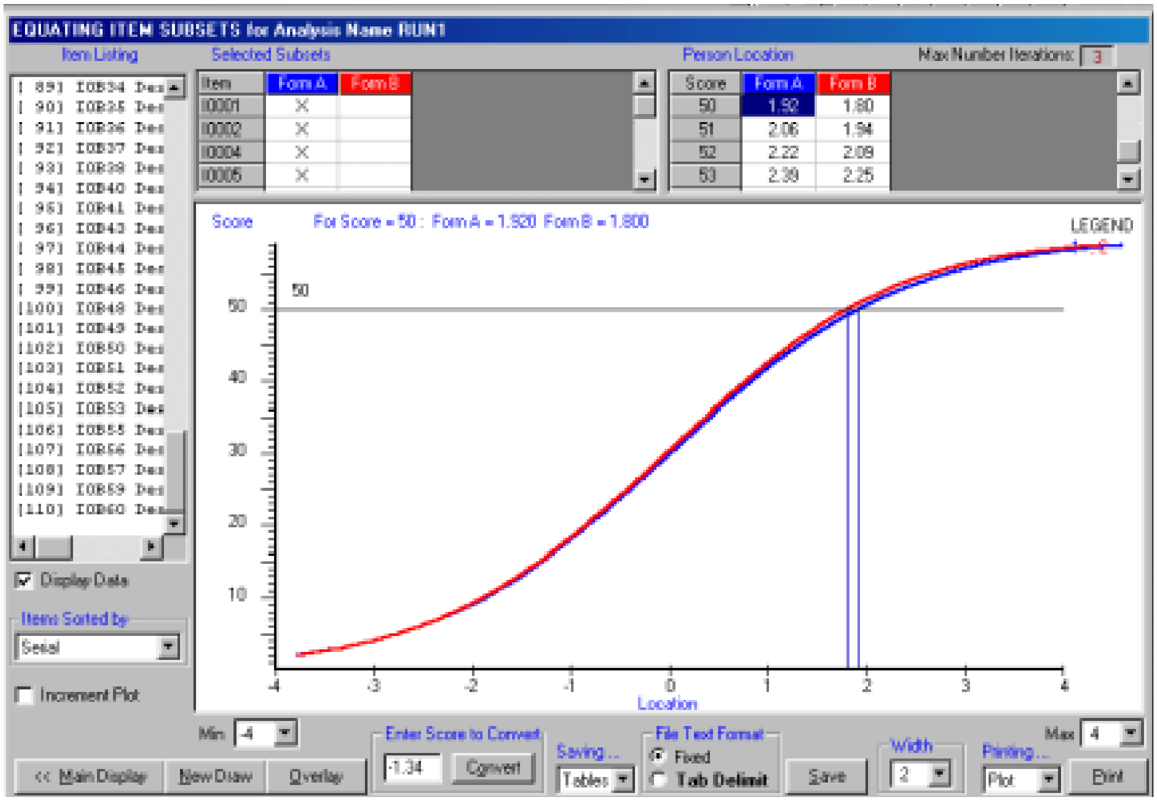


Figure 4. Results Of Linking Analysis Obtained With RUMM


EQUATING ITEM SUBSETS for Analysis Name RUN1

Item Listing

```
[ 89] IOB34 Des
[ 90] IOB35 Des
[ 91] IOB36 Des
[ 92] IOB37 Des
[ 93] IOB38 Des
[ 94] IOB40 Des
[ 95] IOB41 Des
[ 96] IOB43 Des
[ 97] IOB44 Des
[ 98] IOB45 Des
[ 99] IOB46 Des
[100] IOB48 Des
[101] IOB49 Des
[102] IOB50 Des
[103] IOB51 Des
[104] IOB52 Des
[105] IOB53 Des
[106] IOB55 Des
[107] IOB56 Des
[108] IOB57 Des
[109] IOB59 Des
[110] IOB60 Des
```

Selected Subsets

Item	Form A	Form B
I0001	X	
I0002	X	
I0004	X	
I0005	X	

Person Location

Max Number Iterations: 3

Score	Form A	Form B
50	1.92	1.80
51	2.06	1.94
52	2.22	2.09
53	2.38	2.25

For Score = 50 : Form A = 1.920 Form B = 1.800

☑ Display Data

Items Sorted by: Serial

☐ Increment Plot

Min: -4

Max: 4

Enter Score to Convert: -1.34

Saving ...: Tables

File Text Format: ● Fixed ○ Tab Delimit

Width: 2

Printing ...: Plot

Figure 4 shows the results of the linking analysis for Form A and Form B. To illustrate the obtained score-ability conversion, a student with a total score of 50 would obtain a location of 1.30 on Form B but only .92 on Form A.

Anchoring Analysis. In the anchoring analyses, Form A was used as the historic test and Form B as the current test. Table 5 shows the results of the locations of items on Form B obtained with and without anchoring. The first 10 items are the anchor items, and the next 10 items are unique items of Form B. It is noted that the values of the locations of anchor items obtained from the relative anchoring analysis are similar to those values of the historic test, that is, Form A (see values of anchor items under the “absolute anchoring” column).

Table 5 Results Of The Locations Of Items On Form B Obtained With And Without Anchoring

Item	No Anchoring	Relative Anchoring	Absolute Anchoring
I0003	-0.556	-0.742	-0.742
I0007	-1.032	-1.219	-1.391
I0011	0.032	-0.154	-0.234
I0016	0.490	0.304	0.453
I0030	0.507	0.321	0.372
I0039	1.271	1.085	1.210
I0042	1.337	1.151	1.158
I0047	-0.429	-0.615	-0.607
I0054	-0.141	-0.328	-0.550
I0058	-1.613	-1.800	-1.666
I0B01	-0.192	-0.378	-0.377
I0B02	-1.128	-1.315	-1.315
I0B04	-1.280	-1.466	-1.467
I0B05	-1.484	-1.671	-1.671
I0B06	0.293	0.107	0.109
I0B08	-0.115	-0.301	-0.300
I0B09	0.869	0.682	0.685
I0B10	-0.806	-0.993	-0.993
I0B12	-0.146	-0.332	-0.331
I0B13	0.812	0.625	0.628

Conclusion and Discussion

When all raw test data can be merged together to form a single data matrix (with missing blocks), the preferred procedure for test equating is linking analysis as described in this paper. Anchoring analysis, on the other hand, provides a solution for test equating when the data of a historic test are no longer available. In this case, both relative and absolute anchoring analyses could be conducted. A significant difference between the results obtained from these two anchoring analyses may imply that the data of either the historic or current test, or both, does not fit the Rasch model well. This provides an opportunity for diagnosing the model-data fit. It is noted, however, that when the merged data do not have enough linkage between the tests, then some items in the tests may be detected as extreme items, and consequently these items cannot be estimated.

When students are administered different items in computer-assisted assessments, it is critical that there is a fair comparison of test scores. The techniques described in this paper can be programmed into a CAA system so that test equating can be conducted within the system. The development of NIECELT has now reached a development stage of integrating these test equating techniques.

In sum, the use of a Rasch model in linking and anchoring techniques for comparing or equating tests has several advantages. One advantage is that the requirement that examinees must be from same population is now lifted, thus allowing for comparisons of examinees' test performance across different cohorts or academic levels. Another advantage is that the test equating techniques are applicable even when the historic raw data are not available. Lastly, there is also the advantage for calibrating of items using these techniques to provide for essential and comparable indexing reference in item banking.

References

Andrich, D., Sheridan, B. and Luo, G. (1990-2001). *RUMM: Rasch Unidimensional Measurement Model*. RUMM laboratory. Australia.

Hsui, V., Seow, A. & Chew, L. C. (1997). *Implications of IT for English language testing* in the proceedings of the Conference on information technology in English language learning, Singapore.

Hsui, V., Seow, A. & Chew, L. C. (1999). *Developing computerized language proficiency tests at NIE* in Margit Wass (ed.), Enhancing Learning: Challenge of Integrating Thinking and Information Technology into the Curriculum Vol. II (pp. 566-571). Singapore: Educational Research Association.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology* in J. Neyman (ed.) proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. IV, 321-334. Berkeley CA: University of California Press.

RUMM laboratory (2001). *RUMM 2010: getting started*. RUMM laboratory. Australia.