Submission ID = 3968

Test Score Equating in Chinese Language Classroom Assessment

Chieng Zouh Fong, Dr. Rahimah Adam and Dr. Shahrir Jamaluddin
Examinations Syndicate, Examinations Syndicate and University Malaya)
chieng@moe.gov.my, rahimah.adam@moe.gov.my, jamaluddinshahrir@um.edu.my

Equating test scores is an important issue in large scale assessment. Almost all standardized tests have several forms which vary in difficulty. Two parallel forms of a test are considered "equated" for a single group of examinees when the standard deviations and the means of the two test forms are equal (Baghaei, 2010). Scores obtained from different tests are often added up by teacher to get students' total score. When scores are treated in this manner, they are assumed to be interchangeable or comparable, even when they are not. The urge to use raw scores to make comparisons compel teachers to add up students' scores from different tests and divide by total number of tests in to get the relative test performance and students' achievement in the class. This act of adding up raw scores may lead to misinterpretation of marks because each assessment tool is crafted for a specific purpose and may not have the same mean and standard deviation. Therefore, for any comparison to be made over students' achievement in tests, their test scores should be standardized through an appropriate method. However, teachers in schools usually lack the knowledge to do test equating in classroom assessments. Thus, this paper describes a proposal to help teachers to ascertain the relative efficiency of test score equating methods in the comparison of students' continuous classroom assessment measures of Chinese Language test in primary Chinese schools in Kuala Lumpur, Malaysia. The proposal addresses the practical implications of score equating by describing aspects of equating and practices associated with the equating process which is going to be implemented by the teachers. It is hoped that by applying test equating in classroom assessments, teachers are able to make better judgement of students' performance.

Keywords: Test equating, Item Response Theory, Test fairness.

Introduction

Malaysia provides free education for all Malaysians aged six (6) up to 19 years old through preschool, primary, secondary, post-secondary and tertiary education. The public primary schools consist of the national and the national-type or vernacular primary schools. The medium of instruction in the national school is the Malay language (national language) while Chinese and Tamil languages are the medium of instructions in the Chinese and Tamil national-type schools respectively.

Primary School Achievement Test, also known as Ujian Pencapaian Sekolah Rendah (commonly abbreviated as UPSR), is a standardised examination taken by students in Malaysia at the end of their sixth year in primary school before they leave for secondary school. It is prepared and examined by the Examinations Syndicate, Ministry of Education Malaysia. UPSR is just a check point of students' abilities in literacy, numeracy and reasoning skills after six years of primary education. These constructs are assessed through the subjects of languages (Malay, English, Chines and Tamil), Science and Mathematics.

Problem Statement

In line with the Malaysia Education Blueprint 2013-2025, the issue of teaching to the test has often translated into debates over whether the UPSR examination should be abolished. Summative national examinations should not in themselves have any negative impact on students. Therefore, school-based assessment and high stake examination systems must maintain equal level of item difficulty. If not, there might be a fluctuation in the passing rate.

Although the passing rate was affected by item difficulty, the ability of the examinees in a year also may be another factor of influence to the passing rate. School districts seek to become more "outcome oriented," they will need to invest in better testing and reporting systems in order to know whether they are making genuine progress towards equality of educational opportunity for examinees. Better testing and reporting systems are needed for charting academic productivity. The general problem to be considered in this research is how to support schools in making performance judgement based on their assessment data by introducing test equating to monitor students' achievement and plan for future teaching and learning.

Research Purposes and Questions

The purposes of this study are (1) to investigate the impact of item-type multidimensionality on Chinese language test equating results, (2) to explore psychometric properties of Rasch model for Chinese language test, and (3) to evaluate an equating function. Specifically, this study addresses the following questions with regards to Chinese language test:

(1) How does item-type multidimensionality influence equating result?
(2) What are the psychometric properties of Rasch Model for Chinese language test?
(3) When equating scores on dichotomous data, how to evaluate an equating function?

What is Test Equating?

Test equating is a statistical procedure to establish the relationships between scores from two or more tests, or simply to place two or more tests on a common scale as stated by Hambleton & Swaminathan (1985). Other terms for it are such as linking (Vale,1986), calibration (Wright, 1968), and Scaling (Hambleton, Swaminathan, & Roger, 1991). Kolen & Brennan (2014) believed that a procedure can be called "equating" only if it is used strictly to equate two testing forms with the same content, and other related procedures should be called "scaling" or "linking".

The goal of equating is to produce a linkage between scores on two test forms such that the scores from each test form can be used as if they had come from the same test. Strong requirements must be put on the blueprints for the two tests and on the method used for linking scores in order to establish an effective equating. There are five requirements that are widely viewed as necessary for a linking to be an equating (Holland & Dorans, 2006):

(1) The Equal Construct Requirement: The two tests should both measure the same construct (latent trait, skill, ability).
(2) The Equal Reliability Requirement: The two tests should have the same level of reliability.
(3) The Symmetry Requirement: The equating transformation for mapping the scores of Y to those of X should be the inverse of the equating transformation for mapping the scores of X to those of Y.
(4) The Equity Requirement: It should be a matter of difference to an examinee as to which of two tests the examinee actually takes.
(5) The population Invariance Requirement: The equating function used to link the scores of X and Y should be the same regardless of the choice of (sub) population from which it is derived.

How Do We Equate Test?

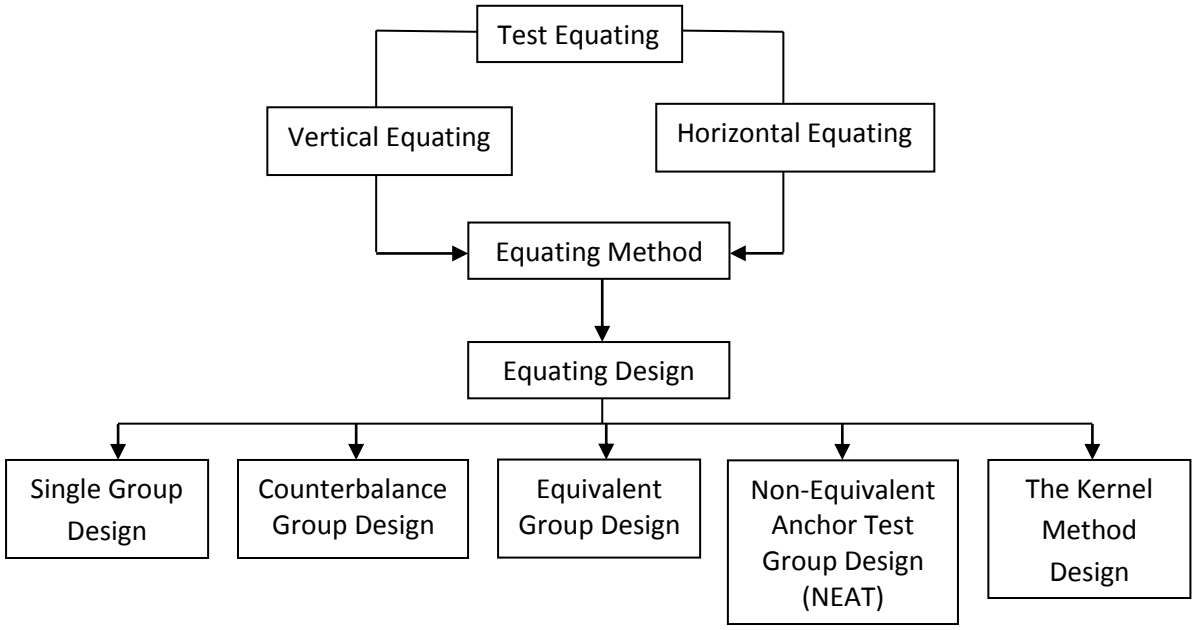

Figure 1: Theoretical Framework of Test Equating
Source: Holland & Dorans (2006)

Test equating, like other aspects of test development, starts with data collection. Although the names of these test-equating procedures are different, they are generally classified into two categories: Horizontal and vertical equating (Holland & Dorans, 2006) as shown in Figure 1. Vertical equating involves equating test of different grades or levels. It allows comparison to be made between students at different levels and also comparison of their growth over time. Vertical equating is also called across-grade-scaling. Horizontal equating on the other hand involves equating test of different forms or at different time of a single grade of level. It is also called within-grade-scaling. Horizontal equating places students' scores on two tests at the same level, for the same content and for the same population so that their scores can be directly compared (Holland & Dorans, 2006).

In practice, three data collection designs are commonly used for test equating purpose, that is single group (SG) design, equivalent group (EG) and Non-equivalent Anchor Test (NEAT) design which are illustrated in Tables 1, 2 and 3.

Table 1: The Single Group (SG) Design

Population	Sample	X	Y
P	1	@	@

Table 2: The Equivalent Group (EG) Design

Population	Sample	X	Y
P	1	@	
P	2		@

Table 3: The Non-equivalent Group with Anchor Test (NEAT) Design

Population	Sample	X	A	Y
P	1	@	@	
Q	2		@	@

Note: @ indicates examinees in sample for a given row take tests indicated in a given column; Lack of @ indicates score data were not collected for that combination of row and column.

Source: Dorans, Moses and Eignor (2010).

The SG design shown in Table 1 controls for any possibility of differential examinee proficiency by having the same examinees take both tests. It has several major uses in the practice of scaling and equating. The advantage of this design is that the measurement error is relatively small. The disadvantage of this design is fatigue and practice effects. To avoid the fatigue and practice effects, some sort of spiralling process should be applied (Kolen & Brennan, 2014).

In equivalent group design (EG) as shown in Table 2, two tests to be equated are administered to two equivalent groups of examinees. The groups may be chosen randomly, which is why this design is sometimes also called the random group design. The advantage of this design is that problems related to single-group design such as fatigue and practice effects can be eliminated. Furthermore, testing time is minimized, and testing can be completed in a single administration. The disadvantage of this design is that unknown degree of bias is introduced in equating process because groups are often not exactly the same in their ability distributions. To control sample-related bias, larger samples generally are required for this design (Kolen & Brennan, 2014).

In NEAT design shown in Table 3, there are two populations, P and Q with a sample of examinees from P taking test X and a sample from Q taking test Y. In addition, both samples take an anchor test, A. By following the terminology of von Davier et al. (2004a), this design is known as NEAT design. Kolen and Brennan (2014) and others have referred this as common-item non-equivalent design or simply the common item or anchor test design. This design is extremely useful when measuring growth in which two groups are known to be not equivalent or when it is impossible to administer more than one test due to test security or other practical concerns. This design is also necessary when developing an item bank, in which testing items are cumulated into a common scale. Petersen et al. (1989) proved strong statistical assumptions are required to remove the confound effects of group and test differences, and quite often, statistical procedures can provide only limited adjustments.

Kernel Equating (KE) is a powerful, modern, and unified approach to test equating. It is based on a flexible family of equipercentile-like equating functions and contains the linear equating function as a special case. Any equipercentile equating method has five steps parts.

They are (1) pre-smoothing; (2) estimation of the score-probabilities on the target population; (3) continuization; (4) computing and diagnosing the equating function; and (5) computing the standard error of equating and related accuracy measures. KE brings these steps together in an organized whole rather than treating them as disparate problems (A von Davier, 2004).

Method

Participants

The sample of this study comprised 200 primary school Year 6 students from two Chinese national-type schools in Kuala Lumpur, Malaysia. Several Chinese language teachers were requested to administer the tests during their teaching and learning sessions.

Instrumentation

Two parallel Chinese Language comprehension tests, Form X and Form Y, were employed in this study. Each of the two forms contained 20 multiple-choice items and 4 common items. Items 2, 4, 7-13 and 15-20 in Test Form X and items 21-36 in Test Form Y were unique items. Only items 1, 3, 6 and 14 were set as anchor items or classified as common items. The anchor item number was determined to be 20% for both conditions suggested by the research done by Angoff (1971).

Procedure for Data Analysis

In this study, non-equivalent anchor test equating design (NEAT) was used. Form X and Form Y were randomly distributed among the 200 student participants of year 6 (12 years old). Form X was taken by 100 students and Form Y was taken by another 100 students from the schools mentioned above. Figure 1 shows part of the data setup for common item equating. Students 1 to 100 took Form X and students 101 to 200 took Form Y. As Figure 2 shows the 4 anchor items (Item no. 1,3, 6 & 14) in Form X, which become the anchor items of Form Y (Item no. 4, 5, 8 & 15). These anchor items taken by both groups and produced the linkage among the two datasets. The rest of the items in the two forms were unique items.

```
DDDCBBCBAABCDADBCCDAUUUUUUUUUUUUUUUUU                92
DCBABBDABCBDDBACBBDAUUUUUUUUUUUUUUUUU                93
BBCABCABAACAABDACABCUUUUUUUUUUUUUUUUU                94
DBCCBBCBDACBDCCDADCCUUUUUUUUUUUUUUUUU                95
BCACCBCBABADCBACACCBUUUUUUUUUUUUUUUUU                96
DADCBBCBDDDCDCCDCDCCUUUUUUUUUUUUUUUUU                97
DABCBBDBCDCADBABACCAUUUUUUUUUUUUUUUUU                98
BBCBDCBCBCBCCACADABBUUUUUUUUUUUUUUUUU                99
BBDDBDCACAABDDABAADAUUUUUUUUUUUUUUUUU               100
DUDUUBUUUUUUUAUUUUUUDBCACACCDDBCBBDA 101
CUDUUCUUUUUUUDUUUUUUABCDBBCCDCBCCACA 102
DUCUUCUUUUUUUAUUUUUUADBBCDCBADBCABAA 103
BUBUUBUUUUUUUCUUUUUUACABABDACACBCDCC 104
CUCUUCUUUUUUUCUUUUUUDBCBCBCADDBACBAA 105
BUAUUAUUUUUUUAUUUUUUDBCABBBCADAACBBB 106
CUCUUCUUUUUUUBUUUUUUDBCBCBBCDDABDAAA 107
```

Figure 2: Data setup for common item equating in WINSTEPS

NEAT equating design was used to place the items and persons from the two tests on the same scale so that the comparison of the abilities of the persons who had taken the two

different test forms could become possible. In NEAT equating design, after setting up the data in the fashion displayed in Figure 2, the entire dataset is calibrated in a single analysis. The anchor items take care of the difference in the difficulty of the two forms and bring the item and person estimates onto the same scale. Thus, the procedure allows the comparison of the difficulty estimates of the items in the two test forms and the ability estimates of the persons who have taken the two forms on a common scale. To analyse the data, one parameter (1PL) IRT model or Rasch model as implemented in WINSTEPS (Linacre, 2012) version 3.75.0 was chosen as shown in Figure 2. Before running the analysis, the quality of the anchor items should be checked. The difficulty estimates of the anchor items in two separate analyses should not be very different from each other; otherwise they cannot be used as anchor items (Baghaei, 2010). Items that fall outside the parallel quality control lines should be dropped from the analysis.

Results

First, item separations, person separations, and reliabilities were computed for Form X and Form Y. Whereas Table 4 shows the results of item separation and item reliability of the two forms, Table 5 depicts the results of person separation and reliability indices for the two forms.

Table 4: Item Separation and reliability indices of Forms X and Forms Y

	Number of item, N	**Item Separation**	**Reliability of Item**
Form X	100	2.15	0.82
Form Y	100	5.04	0.96
Combined Analysis	200	4.15	0.95

Table 5: Person Separation and reliability indices of Forms X and Forms Y

	Number of item, N	**Person Separation**	**Reliability of Person**
Form X	100	0.89	0.44
Form Y	100	0.67	0.31
Combined Analysis	200	0.87	0.43

As demonstrated in Tables 4 and 5, Form X has a person reliability of 0.44 and an item reliability of 0.82. Figure 3 shows the WINSTEP diagnosis report of Data for Form X. The report shows that Root Mean Squared Error (RMSE) for the items is 0.24 and for the person is 0.55. RMSE is the square root of the average of squared standard errors of measurement for all items and persons. The small values here show that the measurement has been precise. The data showed good fit to the Rasch model with only two items (item 3 & 19) having infit mean square values outside the acceptable range of 0.7-1.3 (Bond & Fox, 2007).

data_X.xlsx

PERSON	100 INPUT		100 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	6.4	20.0	-.88	.54	1.00	.0	1.01	.1
S.D.	2.7	.1	.74	.10	.12	.6	.22	.7
REAL RMSE	.55	TRUE SD	.49	SEPARATION	.89	PERSON RELIABILITY		.44
ITEM	20 INPUT		20 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	32.1	99.9	.00	.24	1.00	-.1	1.01	.0
S.D.	10.7	.2	.57	.03	.10	1.0	.16	1.1
REAL RMSE	.24	TRUE SD	.52	SEPARATION	2.15	ITEM RELIABILITY		.82

Figure 3: WINSTEP diagnosis report for Form X.

Moreover, according to Table 4 and Table 5, Form Y has a person reliability of 0.31 and item reliability 0.96. Figure 4 shows the Root Mean Square Error (RMSE) for items is 0.28 and for the person is 0.60. This form also fitted the Rasch model well. There were also two items (item 4 & 18) having infit mean square values outside the 0.7-1.3 boundary. For the common items RMSE analysis, three out of four anchor items showed good fit in both analyses.

data_Y.xlsx

PERSON	100 INPUT		100 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	7.8	20.0	-.67	.60	1.00	.0	1.01	.0
S.D.	2.3	.0	.72	.07	.30	1.1	.60	.9
REAL RMSE	.60	TRUE SD	.40	SEPARATION	.67	PERSON RELIABILITY		.31
ITEM	20 INPUT		20 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	39.0	100.0	.00	.27	1.00	.0	1.01	.1
S.D.	25.4	.0	1.45	.07	.08	.7	.17	.9
REAL RMSE	.28	TRUE SD	1.42	SEPARATION	5.04	ITEM RELIABILITY		.96

Figure 4: WINSTEP diagnosis report for Form Y.

Table 5 indicates that the combined analysis, when the two forms are linked by means of the four anchor items, yield a person reliability of 0.43 and Table 4 shows an item reliability of 0.95 in combined analysis. Figure 5 shows the combined analysis of Form X and Form Y with linkage of four anchors items. The RMSE for the items is 0.25 and for the persons is 0.57; only two out of 36 items were misfits. The all four anchor items have good fit indices and cover a wide range of difficulty -2.25 to 2.64 with mean of 40.4 and a standard deviation of 21.4. In combined analysis, the items which were used as anchor items all had acceptable fit indices and spanned over the difficulty continuum as shown in Figure 6. The curved lines are the approximate 95% two-sided confidence bands for the items difficulty invariance. Acceptable fit, person and item reliability and separation indices, and small RMSEs in the combined analysis indicate that the equating procedure was successful.

join data_XY.xlsx

PERSON	200 INPUT		200 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	7.3	20.0	-.69	.56	1.00	.0	1.00	.0
S.D.	2.8	.1	.75	.08	.21	.8	.36	.8
REAL RMSE	.57	TRUE SD	.49	SEPARATION	.87	PERSON RELIABILITY		.43
ITEM	36 INPUT		36 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	40.4	111.1	.00	.24	1.00	.0	1.00	.0
S.D.	21.4	31.4	1.07	.06	.09	.9	.15	1.0
REAL RMSE	.25	TRUE SD	1.04	SEPARATION	4.15	ITEM RELIABILITY		.95

Figure 5: WINSTEP diagnosis report for combine analysis of Form X and Form Y.

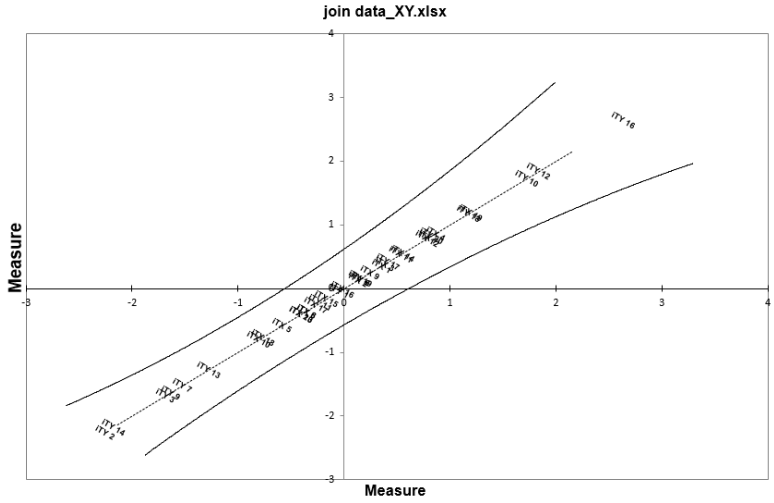

Figure 6: Item difficulty invariance.

In Figure 7, the Rasch dimension explains 47.3% of the total raw variance in the observations data. Meanwhile raw unexplained variance is 76.2%. The first contrast in the residuals explains 4.3% of the variance-somewhat greater than around 4% that would be observed in data like these simulated to fit the Rasch model. In these data, the variance explained by the items, 14.9% is only three times the variance explained by the first contrast 4.3%, so there is a noticeable secondary dimension in the items. The eigenvalue of the first contrast is 2.0% - this indicates that it has the strength of about 2 items out of 36 items. An eigenvalue of 2 is the smaller amount that could be considered a "dimension". For dimensionality analysis, we are concerned about the "Variance explained by the first constrast in the residuals". If this is big, then there is a second dimension at work. Infit and Outfit statistics are too local (one item or one person at a time) to detect multidimensionality productively. They are too much influenced by accidents in the data (e.g., guessing, response sets).

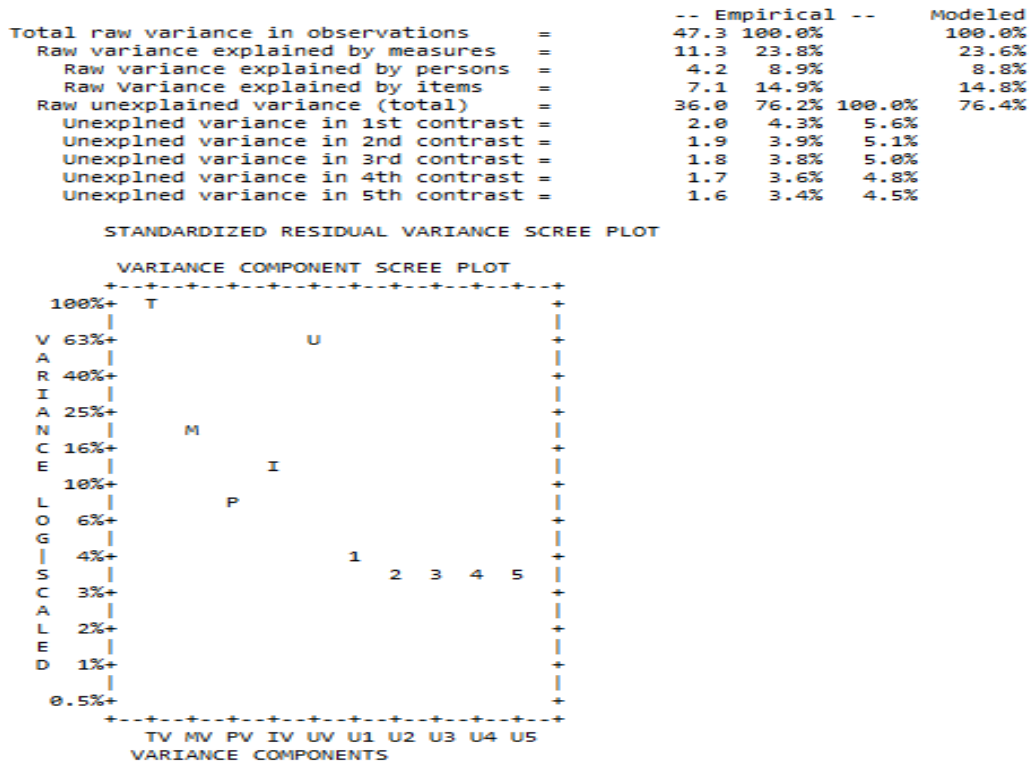

Figure 7: Standardized Residual Variance (in Eigenvalue units)

Discussion and Conclusion

The main reasons for developing standardised tests are to measure examinees' abilities objectively and fairly. Scores from standardised tests are often used to make important decisions about individuals' lives. With a wrong decision, an examinee may be excluded from the academic programme or the practice of their favourite profession. Furthermore, important decisions about education policies and curricula are made on the basis of standardised tests. Due to the importance of the results of standardised tests, every effort should be made to provide a fair measurement of the abilities of interest. The lack of equating and reporting raw scores across numerous forms which are used in multiple runs of an assessment over years and comparing examinees' raw scores with cut-point score which is a raw score may result in nonstandard measurement and unfair evaluation of examinees' skills (Cook, Eignore, 1991). In order to ensure validity and fairness, schools have to maintain the same standards from year to year. The standardization of students' continuous school based assessment score should be done through test score equating. This will allow for the comparison of scores and test forms. Developing a fair assessment instruments for schools is very important as to prepare students for the standardized examination. Therefore, schools should try to provide opportunities to teachers to further develop knowledge and skills in psychometrics so that they can develop good quality items to be saved in item bank.

References:

Baghaei, P. (2010). *Test Score Equating and Fairness in Language Assessment*. JELS, Vol.1, Spr (November), 113–128.

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Mahwah, NJ:LEA.

Cook, L.L., & Eignor, D. R. (1991). *IRT Equating Methods. Educational measurement: Issues and Practices*, 10, 191-199.

Hambleton, R.K., Swaminathan, H., (1985). *Item Response Theory: Principles and Applications. Boston*, MA: Kluwer Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Hewbury Park, CA: Sage.

Holland, P. W., & Dorans, N. J. (2006). *Linking and Equating*. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 189-220). Westport, CT: Praeger.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices: Third Edition*, (January 2006), 1–566. https://doi.org/10.1007/978-1-4939-0317-7

Linacre, J. M. (2012). *WINSTEPS, MINISTEP: Rasch-Model Computer Programs* [Program Manual version 3.75.0] Chicago, IL: Winsteps.com.

Ministry of Education. (2015). *Executive Summary Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education).* http://www.moe.gov.my/cms/upload_files/articlefile/2013/articlefile_file_003114.pdf

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). *Scaling, Norming, and Equating*. *In Educational Measurement, 3rd Ed*. (pp. 221-262). American Council on Education.

Vale, C.D. (1986). *Linking Item Parameters onto A Common Scale*. Applied Psychological Measurement, 10, 333-344.

von Davier, A. (2010). *Equating and Scaling. International Encyclopedia of Education*. https://doi.org/10.1016/B978-0-08-044894-7.00261-X

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). *The Chain and Post Stratification Methods for Observed-score Equating: Their relationship to population invariance*. Journal of Educational Measurement, 41, 15-32.

Von Davier, A. A., Holland, P.W., Thayer, D. T. (2004). *The Kernel Method of Test Equating*. New York, NY: Springer-Verlag.

Wright, B. D. (1968). *Sample free test calibration and Person Measurement*. Proceeding of the 1967 Invitational Conference on Testing Problems (pp. 85-101). Princeton, NJ: Educational Testing Service.