https://doi.org/10.15625/2615-8957/22210401

Test score equating and item anchoring for high stakes examination

Chieng Zouh Fong[1]✉, **Tong Yeah Chuen**[2]

[1] lalabai1108@gmail.com
University of Malaya
(Malaysia)

[2] tongyc@yahoo.com
University of Malaya
(Malaysia)

✉ Corresponding author

ABSTRACT: *Test equating becomes essential to safeguard the test fairness for sitting for the actual national examination. Thus, this paper describes a proposal to help teachers in Malaysia to ascertain the relative efficiency of test score equating methods in comparing students' high stakes examinations. The proposal addresses the practical implications of score equating by describing aspects of equating and item anchoring process which can be used by teachers. This study examined Principles of Accounting (PA) subject with Rasch measurement framework for dichotomous data analysis. A non-experimental quantitative research approach was adopted in which a set of equivalent test instrument were administered to two different groups of respondents comprising 429 students. Data collection was through stratified random sampling method and analysed using Winstep software. Results showed a good fit study by using Common Item Non-Equivalent Group Design (CINEG) also named as Non-equivalent Groups with Anchor Test (NEAT) design. Both test forms were reasonably predictable good fit of measurement. No single student's destiny should rely upon a single test paper (Wu et al., 2016). Hence, multiple sets of equivalent test papers should be developed by teachers in schools with the same standard as the actual exam papers. Subsequently, students will be more well prepared for the national examination and will be able to achieve desired grades.*

KEYWORDS: **Test equating, item anchoring, rasch model, test fairness.**

➔ Received 11/10/2022 ➔ Revised manuscript received 15/12/2022 ➔ Published 30/12/2022.

1. Introduction

Fairness and validity in education assessment reflects to knowledge and skills that are equally familiar and appropriate to all students and is as free as possible of cultural, ethnic, and gender stereotypes (Tierney, 2013). Concern about fairness and validity in education assessment have dated as early as 20th century. The early research done by Finklestein (1913, p.6) states that multiple choice questions test was hailed as a means of bypassing the "injustice" caused by teachers' inconsistent grading practices.

The analogy of defining the standard size code for a dragon fruit for a fair-trade principle have been led to the idea of education measurement in defining human cognitive measurement. In assessment, item developer building items with construct validity in mind. An experience item builder will always think of *"Does results reflect the content expectation?"* Results interpretation depends on types of content, knowledge, practice, and assessment. *"Do the interrelationships of dimensions measured by the test correlate with the construct of interest and test scores?"* The extent to which internal structure is consistent with the construct of domain become main topic for the scholars to study in the academic field (Messick, 1995).

Fairness in education can view as equal opportunities for learning and attending assessments. There were no bias against gender, disability, race, ethnicity, and socio-economic status. Issues about fairness in assessment have brought a debate in Malaysia's current education scenario. During the pandemic Covid outbreak, many school had been forced to temporally close for social distancing purposes. Children had been barred from sitting examination because of detected positive for Covid (McKinsey & Company, 2020). The interruption of schooling time for children is not only cause by pandemic, but the global warming has also contributed to the increase of flood frequency which affected some vicinity in Malaysia especially for the villages

near the riverbank. Because of this, schools also temporally closed, examination interrupted and issues of fairness in administering standardise high stake examination across the nation have become every teacher and parents' debate's topic. In this situation, test equating becomes essential to safeguard the test fairness (Kolen & Brennan, 2014) especially in high stake examinations. I would like to share how student's achievement can be compared in pandemic time. If students are giving different set of test form with different difficulty levels it will be not fair to say that the result from the different set of test form are on same examination standard. Here, raises the issue of how to prepare a parallel test form to compare students' cognitive abilities. If two different group of same cohorts of student are going to sit for an examination, we need an equivalent test form on a comparable scale. Test equating with item anchoring techniques can adjust the difficulty level of test form.

2. Literature review

There were variety of methods for transforming data on a comparable scale. They can be categorised follows:

2.1 Concurrent calibration

Concurrent calibration means several test instruments are analysed concurrently and the item parameter for anchor item are not known by the researcher until the time it is been calibrated. When two test forms are calibrated together in a single run of Winsteps, the value of the item parameter estimates display and reported on the comparable scale. This concurrent calibration method will generate only one set of anchor items. In which it needs no linkage for the scales (Cook & Eignor, 1991). This is the main different among concurrent calibration and separate calibration.

2.2 Separate calibration

The two instruments which distributed to two different respondent's group will be calibrated separately. In this case, the item parameter for each test form will be different and it all depend on the response of the respondent. Each instrument will have different value of anchor items estimate although both instruments are using same anchor items. Basically, anchor items play the role in linking both instruments. To equate the scales scores from the calculation of the difference value of means for that set of anchor items. Researcher study lots of linear equation method to transform the scales on an instrument to the other instrument. Example, the mean/sigma method by Marco et al. (1983), and the mean/mean method by Loyd and Hoover (1980).

In this study, the mean/mean method were utilized to find the Equating Coefficient. The correction term or known as equating coefficient is the difference of the mean for item difficulty parameters for anchor items which is used in DRM calibration by "mean/mean" linear equating method (Loyd & Hoover, 1980).

$$E_C = M(\sigma_X) - M(\sigma_Y) \qquad (1)$$

Equation 1 shows the correction term c, in which $M(\sigma_X)$ show the item difficulty of anchor item in old test form (TF-X) and $M(\sigma_Y)$ item difficulty of anchor item in new test form (TF-Y).

Fischer et al. (2021) investigated the performance of linking Rasch-scaled test to a small item bank by examining four types of IRT linking methods. (1) fixed parameter calibration, (2) simultaneous calibration, (3) mean/mean linking and finally (4) weighted mean/mean linking. They suggested that the proportion of anchor points should be more than 20%. In the study they found out that among the four methods of test equating, the result of weighted mean/mean linking produced more accurate equating results. The imprecision of the difficulty of the anchor items increases the standard error (SE) of the test. This means that the mean and variance of the difficulty of the test instrument must closely match the ability distribution of the sample.

2.3 Statements of problems

Despite the best effort of item developers, there were no two tests form provides the same test instrument in terms of difficulty level and quality. Some examinees could be advantaged by assigned to the easier test forms while other examinees might be disadvantaged by assigned more difficult

forms. Equating can play an important role in providing a comparable score for multiple test forms. When equating is performed successfully, all the examinees can have equal grading system over multiple set of instruments administered to them. This brings out the issues that test equating become essential to safeguard the test fairness (Kolen & Brennan, 2014).

Thus, it is important in producing a technically-sound calibrated items for a common scale in assessment to replenish enough items in bank item so that it can cover all the well-defined content for test specification. Thus, students' achievement across Malaysia can be evaluated fairly in theoretically aspect. With this point of view, the accuracy of test equating had to be conducted so that a comparable scale can be built for multiple set of test instrument. Data from group of examinees with difference ability also can be evaluated.

2.4 Research questions

The questions here were how can items from difference instrument to be calibrated on a comparable scale? For the purpose to calibrate Multiple-choice of Questions on a comparable scale, there were some important aspects to go through such as (1) Does empirical data can fit Rasch model study? (2) Does the selected pair of anchor items functioning on a common scale using a linear transformation? (3) Does test form-X can be equated to test form-Y?

3. Methodology

This study mainly adopts a non-experimental quantitative research design to obtain the findings required for data analysis. The descriptive method was conducted for sampling analysis in which data was separated into two parts by gender and domicile interpretation and the outcome was displayed in Figure 1.

3.1. Population and sampling

Kolen and Brennan (2014) proposed two rules of thumb for choosing sample size for study in test equating. The first rule is based on the standard deviation unit while the second rule is based on comparison with the identity equating. Generally, the sample size as 400 is needed for IRT Rasch model equating (Kolen & Brennan, 2014). The sampling method utilised in the study was call stratified random sampling (Hayes, 2021). Figure 1 shows the comparison of respondent's gender and domicile in Malaysia.

There were 702 upper secondary schools which took the PA test. Out of 702 schools, there were 258 urban schools and 444 sub-urban schools. The ratio of urban to sub-urban school was approximately 1:2. The total population (N) was 67,893 respondents taking the PA test (Ministry of Education, 2021). Table 3.2 shows the population of secondary schools offering PA test since 2013-2019. A total of 702 schools offered PA subject and a total of 67,893 students for the whole population. Total number of samples equal to 858 students (n=858). Out of this figures, 438 male students and 420 female students. The samples of the study split in half. 429 candidates answered the first test form (TF-X), while the second test form (TF-Y) answered by another group of 429 candidates.

Kolen and Brennan proposed two rules of thumb for choosing sample size for study in test equating. The first rule is based on the standard deviation unit while the second rule is based on comparison with the identity equating. Generally, the sample size as 400 is needed for IRT Rasch model equating (Kolen & Brennan, 2014). The sampling method utilised in the study was call stratified random sampling (Hayes, 2021). Stratified random sampling (SRS) classified as a type of probability sampling, which divide in strata/groups within population, equal chance of selecting sample from the whole population by reduce human bias in the selection process and can provide high representative sampling. SRS allow us to make statistical conclusions that data collected will be valid (Hayes, 2021). The advantage of stratified random sampling approach apply in this study is to minimizing sample selection bias that best represents the entire population been divide into strata group follow demographic such as domiciles and gender. Here domiciles represent urban and sub-urban school. Gender represents male and female respondent.

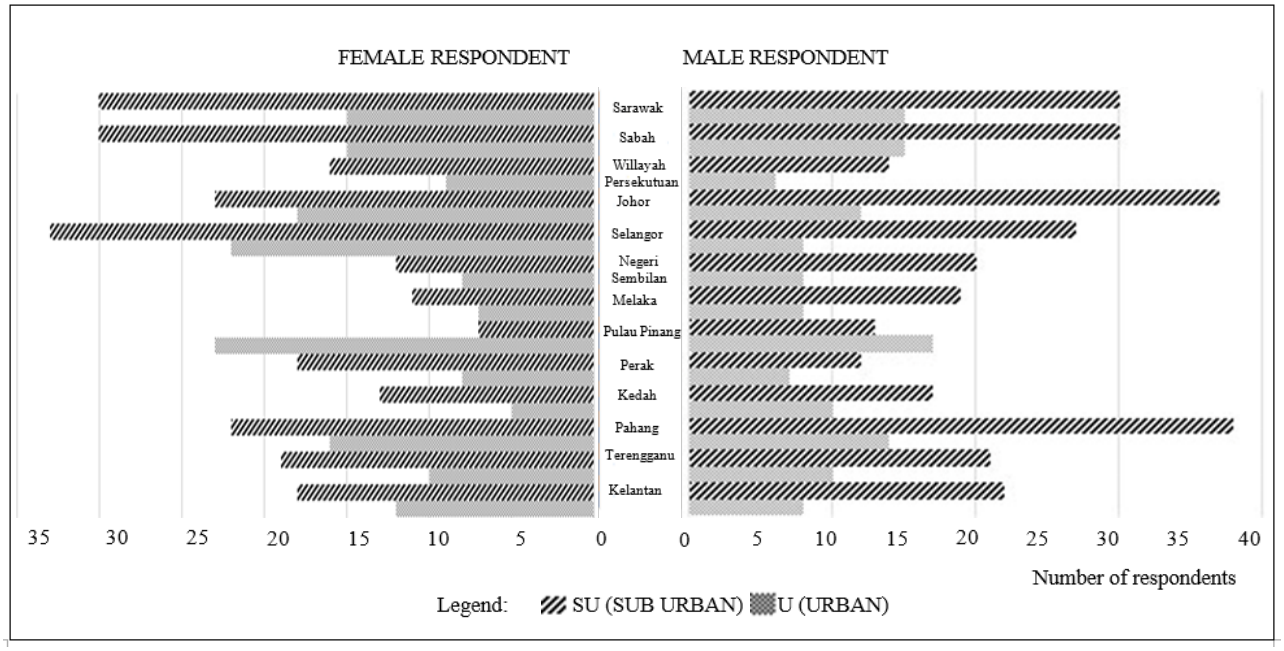


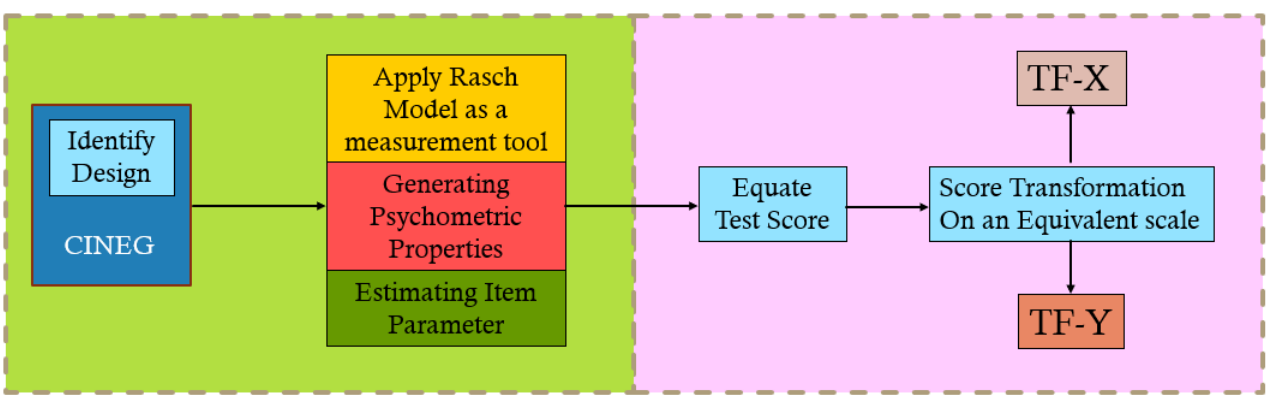
Figure 1. Comparison of respondent's gender and domicile in Malaysia.

3.2. Conceptual framework

Specific test designs and statistical procedures are implemented in the conceptual framework for placing different test form on a common scale as shown in Figure 2. Use Rasch model as a tool to generate psychometric properties such as person and item parameter (define Unidimensionality, Reliability, Validity, Discrimination index, person ability and item difficulty). To equate test score on an equivalent scale, we need at least two test form. One is the existing test form name as TF-X and another one is the new test form name as TF-Y. Equating test score been carry out by designing the calibration design and score transformation use anchor items. Anchor items included in both tests (Internal anchor items) so that can be used to link the two test and create an equivalent scale.

3.3. Identify the test design

Data from this study was conducted in a quantitative research method. The data from a large representative sample divided in spiralling method. Spiralling refers to the way test booklets are assembled, packaged, delivered to testing sites, and distributed to respondents. In this spiralled test administration, examinees are randomly assigned with different set of



Source: Kolen and Brennan, 2014

Figure 2. Conceptual Framework of the current study.

instruments which are administered to them in a different form alternatively. Each examinee will only get one test form. According to Kolen and Brennan (2014), the spiralled method of distribution typically leads to a parallel and comparable test form.

The method that was utilised in this study is called Common Item Non-Equivalent Group Design (CINEG) also name as (NEAT) design. This method been chosen because of the format of the instrument is 40 MCQ question. The test forms were built and develop with internal anchor items so that become a parallel test instrument in which the test scores are comparable. The composition and arrangement position of internal anchor items must be representative in test blueprint to the full test and was embedded in both forms at same position and specification as shown in Table 1.

Table 1. Test blueprint for Principles of Accounting

Topic	Content	Low Level Item Knowledge/Comprehension	Medium Level Item Application/Analysis	High Level Item Synthesis/Evaluation
1	Introduction to Principles of Accounting	2, 4	7	
2	Accounts Classification	5, 8, 9		
3	Business Document as A Source of Information	10	11	
4	General Journal as First Record of Information	(1)	13	
5	General Ledger	15, 16, 17		22
6	Balance Sheets		18, 23	
7	Financial Statements	19		
8	Cash Accounting	20, 21		
9	Accounting for Studies	26		24
10	Partnership		(3)	28
11	Limited Company	30	(6) – 31	
12	Club and Associations	(14) – 33		32
13	Incomplete Record	(25)	34	35
14	Introduction to Management Accounting	(27) – (29)		
15	Information for Decision Making	12	37, 39	40 – (36) ,38
	Total Item (40)	21	11	8
	Percentage (Ratio)	52.5 % (5)	27.5 % (3)	20 % (2)

* Red circles indicate the anchor items. Example shown here is for test TF-X.

3.4 Method of data analysis

This study used Dichotomous Rasch Model (DRM) to analyse the items in Principles of Accounting instrument. The data first collected through the answer sheet from respondents' response. Then, transfer the response to Microsoft Excel and imported it into Winsteps 3.75 program. The psychometric attributes such as item and person reliability, separation index, mean and standard deviation were examined as well. Item fit statistic such as Pt-Measure Correlation, infit and outfit MNSQ, ZSTD, Standard Error, ICC curve and distractor analysis for each item were examined. After identifying each item functioning and fit study, item map (Wright map) was displayed to show how the items distributed in levels of item difficulty in compare to respondent's ability. To link two instruments on a comparable scale. We use the anchor item to do the test equating procedure. For the process to equate the two tests, the separate calibration for test TF-X and TF-Y were carry out. Each test form was analyses separately to determine the value of the item parameter such as item difficulty and person ability. After both test form had been analyses separately, the calculation of the different value of mean in logit for anchor item were used to place both tests form on a same scale as what explained by the "mean/mean" linear equating method (Loyd & Hoover, 1980).

4. Results

Test equating using Item Response Theory (IRT) usually involves a three-step process (Kolen & Brennan, 2014). First, item parameters for DRM are calculated by using Winsteps program (Linacre, 2018). Second, the parameter estimates are brought to a common scale using a linear transformation. Third, the numerically correct scores on the new test form are converted to a scale by using the numerically correct scale on the old form. The procedure of stepwise equating of test scores is explained in detail in this section with reference to the teaching modules on IRT presented by Harris & J. Kolen, (1990), and the IRT equating modules of Cook and Paterson (2015).

4.1. Item parameter estimated by using winsteps

Item parameters were assessed separately for each test form using Winsteps. The summary statistic of person and items for TF-X and TF-Y are shown in Table 2. Subjects administered TF-X had standard deviation of 0.81 for TF-X (Item parameter). The subjects administered TF-Y had standard deviation of 1.12 (Item Parameter).

Table 2. Summary Statistics of Person and Items for TF-X and TF-Y

	TF-X		TF-Y	
	Person	**Item**	**Person**	**Item**
N	429	40	429	40
Measure (logit)				
Mean	-0.48	0.00	-0.52	0.00
SD (Standard Deviation)	0.74	0.81	0.67	1.12
SE (Standard Error)	0.04	0.13	0.04	0.18
Outfit Mean-square (MNSQ)				
Mean	1.03	1.03	1.03	1.03
SD (Standard Deviation)	0.26	0.19	0.39	0.24
Separation	1.73	6.97	1.71	9.09
Reliability	0.75	0.98	0.74	0.99
Cronbach's Alpha	0.77		0.77	
Total raw variance in observation		50.0 (100%)		54.9 (100%)
Raw unexplained variance (Total)	40.0 (79.9%)		40.0 (79.9%)	
Raw variance explained by measures	10.0 (20.1%)		14.9 (27.1%)	
Unexplained variance in 1st contrast	2.5 (4.9%)		2.1 (3.7%)	

Thus, comparison of the standard deviation shows that the group of examinees had slightly different in the range of ability levels. In most cases, changes of the value of standard deviation in item parameters known as item parameter drift (IPD). It happens over times when concept of invariant no more holds with the cause of changes in curriculum or differential item functioning (DIF) in items (Mislevy & Bock, 1990). Item parameters were assessed separately for each test form using Winsteps. The parameter estimates for CINEG design are shown in Table 3.

The Rasch Model is obtained from the general formula for the 3PL model, as shown in Equation 2. Knowing the value of the parameters *a*, *b*, and *c* for each item, we can identify the item used. Value of discrimination = *a*. Value of item's difficulty = *b*. And value of guessing = *c*. The higher the discrimination value of *a* (e.g: *a* = 1.7), the steeper the curve of ICC (item characteristic curve), the more item discriminates between subjects. The value of *b* and *c* parameter for an item must be in the range of 0 to 1. Typically, the *c* parameter for an item ranges from 0 to the probability that a test item will answer a task correctly by random guessing, with a calculation of 1 divided by the number of options (Kolen & Brennan, 2014).

Dichotomous Rasch Model (DRM) requires that all item to be equally discriminative. Therefore, in this study, the value of *a* is set to 1. While *c* is set to 0 because Rasch Model does not allow guessing. Rasch model considered becomes a 1PL model (Wu et al., (2016).

$$p = P(X = 1) = c + \frac{(1-c)}{1+\exp(a(\theta - b))} \quad (2)$$

Table 3. Item Parameter Estimates for CINEG Design

		TF-X		TF-Y
BIL.	**Item ID**	**b_x**	**Item ID**	**b_y**
1	XQ2	-0.15		0.09
2	XQ4	1.44		0.86
3	XQ5	-0.32		0.00
4	XQ7	-0.26		0.03
5	XQ8	-1.08		-0.37
6	XQ9	-0.18		0.07
7	XQ10	-0.77		-0.22
8	XQ11	1.08		0.69
9	XQ12	0.72		0.51
10	XQ13	0.97		0.63
11	XQ15	0.24		0.28
12	XQ16	-0.51		-0.09
13	XQ17	0.13		0.22
14	XQ18	-0.76		-0.21
15	XQ19	2.02		1.14
16	XQ20	-0.45		-0.06

17	XQ21	-1.59		-0.61
18	XQ22	-0.71		-0.19
19	XQ23	0.01		0.16
20	XQ24	1.66		0.97
21	XQ26	-1.02		-0.34
22	XQ28	-0.06		0.13
23	XQ30	-0.39		-0.03
24	XQ31	-0.65		-0.16
25	XQ32	-1.00		-0.33
26	XQ33	-0.69		-0.18
27	XQ34	0.57		0.44
28	XQ35	0.37		0.34
29	XQ37	0.25		0.28
30	XQ38	-0.14		0.09
31	XQ39	0.37		0.34
32	XQ40	1.18		0.73
33	ANX1(XQ1)	0.48	ANY1(YQ5)	0.06
34	ANX2(XQ3)	-0.11	ANY2(YQ4)	0.12
35	ANX3(XQ6)	-1.05	ANY3(YQ8)	-0.23
36	ANX4(XQ14)	0.64	ANY4(YQ15)	0.39
37	ANX5(XQ25)	0.46	ANY5(YQ23)	0.52
38	ANX6(XQ27)	-0.56	ANY6(YQ28)	0.57
39	ANX7(XQ29)	0.75	ANY7(YQ29)	-0.43
40	ANX8(XQ36)	-0.89	ANY8(YQ37)	0.14
41			YQ1	-0.42
42			YQ2	-2.22
43			YQ3	-1.60
44			YQ6	0.01
45			YQ7	-1.71
46			YQ9	-1.43
47			YQ10	1.06
48			YQ11	0.04
49			YQ12	1.43
50			YQ13	-1.71
51			YQ14	-2.16
52			YQ16	1.93
53			YQ17	0.49
54			YQ18	1.20
55			YQ19	-0.16
56			YQ20	0.12
57			YQ21	-0.05
58			YQ22	1.90
59			YQ24	-0.17
60			YQ25	1.31
61			YQ26	0.93
62			YQ27	1.16
63			YQ30	-1.07
64			YQ31	1.20
65			YQ32	-1.77
66			YQ33	2.35
67			YQ34	-0.84
68			YQ35	-0.99
69			YQ36	-0.11
70			YQ38	-0.25
71			YQ39	0.12
72			YQ40	0.29

Notes. For TF-X (a_x=1.00, c_x=0.00), and TF-Y (a_y=1.00, c_y=0.00).

item in TF-X and TF-Y. Winsteps version 3.73 (Linacre, 2018) was utilized to calculate approximately Rasch item parameters. Anchor item in TF-X is given the name as ANX1, ANX2, ANX3, ANX4, ANX5, ANX6, ANX7 and ANX8. While anchor item in TF-Y is given name as ANY1, ANY2, ANY3, ANY4, ANY5, ANY6, ANY7 and ANY8.

The arrangement of anchor items in both test forms must match the content descriptive of the test specification and psychometric properties of the entire test (Cook & Peterson, 2015). In this study, the arrangement of anchor items was indicated by their position in the test forms with numbering in parentheses. Example: ANX1(XQ1) means 1st anchor item located in question number one in TF-X. And ANY2(YQ4) means 2nd anchor

4.2. Distribution of anchor items

Table 4 shows the distribution of anchor item located in question number four in TF-Y. Eight anchor items chosen for test form TF-X and eight anchor items chosen for test form TF-Y. Both set of anchor items is the same item in term of wording, content, and test specification.

The value for item difficulty parameter for TF-X () was calculated separately from test Form TF-Y () by using separate calibration running on Winsteps program. For anchor item in TF-X the highest estimated value was 0.75 (item ANX7(XQ29)) and the lowest estimated value are -1.05 (item ANX3(XQ6)). For anchor item in TF-Y the highest estimated value was 0.57 (item ANY6(YQ28)) and the lowest estimated value was -0.43 (ANY7(YQ29)).

The Difference between the value of anchor item for TF-X and anchor item for TF-Y (-)

Table 4. Distribution of Anchor Items in TF-X and TF-Y

No.	Ancho Item Text Form X	Item Difficulties Parameter (b_x)	Anchor Item Text Form Y	Item Difficulties Parameter (b_y)	Difference between (b_x-b_y)
1	ANX1(XQ1)	0.48	ANY1(YQ5)	0.06	0.42
2	ANX2(XQ3)	-0.11	ANY2(YQ4)	0.12	-0.23
3	ANX3(XQ6)	-1.05	ANY3(YQ8)	-0.23	-0.82
4	ANX4(XQ14)	0.64	ANY4(YQ15)	0.39	0.25
5	ANX5(XQ25)	0.46	ANY5(YQ23)	0.52	-0.06
6	ANX6(XQ27)	-0.56	ANY6(YQ28)	0.57	-1.13
7	ANX7(XQ29)	0.75	ANY7(YQ29)	-0.43	1.18
8	ANX8(XQ36)	-0.89	ANY8(YQ37)	0.14	-1.03

show the lowest for the pair item XQ25 and YQ23 is (-0.06). While the others three pair of anchor item exceed logit of 1. That is the pair of items XQ27 and YQ28 (-1.13), item XQ29 and YQ29 (1.18), and item XQ36 and YQ37 (-1.03). Overall, anchor item's parameter drift of 0.2 logit is unproblematic. Somehow, drift of 0.5 logits will affect the reduction of equating accuracy over time. But drift of 1.0 logits indicated the anchor item is not functioning well and mainly cause by DIF (Kopp & Jones, 2020).

4.2. Scale transformation for TF-X and TF-Y

Harris (1989) has provided a discussion on how to do the scale transformation for two different test forms on a common scale by using Rasch Model. The method is based on a linear transformation. In Rasch Model, the term of the ICC is a function of (θ - *b*). Just the origin of the person ability parameter (*b*) and item difficulty are indeterminate. In this study, the mean of θ is set to zero because the data are a standard normal distribution. Now suppose that TF-X is equated into TF-Y by a linear transformation in Equation 2, following the linear equation method called "mean/mean" introduced by Loyd and Hoover (1980), which is used in Dichotomous Rasch Model.

4.3. Equating test scores

If the parameter estimates for TF-Y were set to the same scale as TF-X. The ability estimate achieved for an examinee will be the equivalent within measurement error irrespective of which test form it takes. Thus, when examinees can be told the ability estimates are their test scores, the process of test equating is complete. This situation had proved the assumption of Equity property which had been proposed by Lord (1982).

$$E_C = M(\sigma_X) - M(\sigma_Y) \qquad (1)$$

Equating Constant is the difference of the mean for item difficulty parameters estimate for anchor items. is the mean for anchor items difficulty in TF-X while is the mean for anchor items difficulty in TF-Y.

$$M(\sigma_X) = \frac{0.48+(-0.11)+(-1.05)+0.64+0.46+(-0.56)+0.75+(-0.89)}{8}$$

$$= -0.035$$

$$M(\sigma_Y) = \frac{0.06+(0.12)+(-0.23)+0.23+0.39+(0.52)+0.57+(-0.43)+(0.14)}{8}$$

$$= 0.1425$$

$$E_C = -0.035 - 0.1425 = -0.1775$$

To fit all item's difficulty parameter approximates from the first measurement to the scale of the second measurement, just add together the equating constant of (-0.1775) to each item difficulty parameter. The same value of the constant would be added up to the ability estimates from the second measurement to bring them to the same scale as the first measurement. In Dichotomous Rasch Model, only the origin of the ability measured (item difficulty) is indeterminate, so the correlation between measures developed from two different 1PL measured transformations differs only by the constant (Lyod & Hoover, 1980).

Figure 3. shows test equating results for TF-X and TF-Y on an equal scale using the CINEG design. TF-X is represented by the blue line on the graph. TF-Y is represented by an orange line on the graph. The ogive shows the relationship between the two test forms.

The slope of the graph's estimates represents the thresholds for an individual's ability compared to the scaled score. Comparative to the mean of the thresholds, the two test forms were identical with a difference of only 0.1775. TF-Y is slightly more difficult for individuals in the upper group, but also slightly easier for individuals in the lower group. This conclude that instrument TF-X is parallel to TF-Y and quite identical in term of psychometric properties concern. Both instruments only different by 0.1775 measure of difficulty. The test form difficulty bias was small and can be negligible, anchor item drift up to 0.5

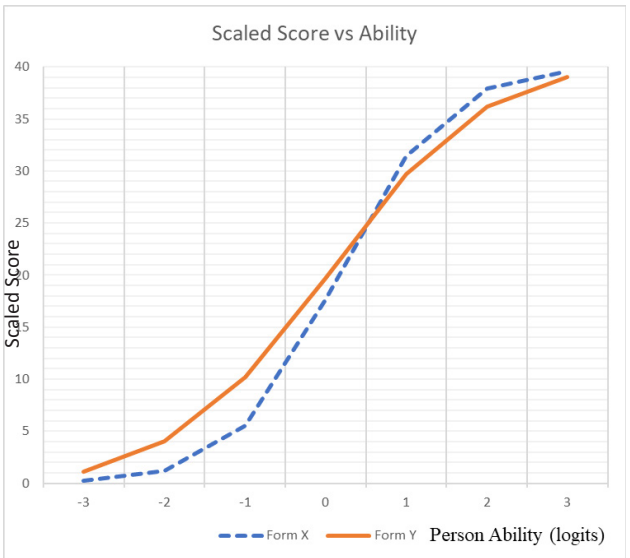


Figure 3. Scaled Score Vs Person Ability for TF-X and TF-Y

Table 5. Data for Test Equating TF-X and TF-Y generated from Microsoft Excel

Pi						
-3	-2	-1	0	1	2	3
0.0437	0.1104	0.2523	0.4784	0.7137	0.8714	0.9485
0.0206	0.0541	0.1346	0.2971	0.5347	0.7575	0.8946
0.0473	0.1188	0.2682	0.4991	0.7303	0.8804	0.9524
0.0460	0.1158	0.2625	0.4918	0.7246	0.8773	0.9511
0.0670	0.1634	0.3467	0.5906	0.7968	0.9142	0.9666
0.0443	0.1119	0.2551	0.4821	0.7167	0.8730	0.9492
0.0582	0.1438	0.3134	0.5537	0.7713	0.9016	0.9614
0.0245	0.0638	0.1564	0.3350	0.5780	0.7883	0.9101
0.0290	0.0751	0.1809	0.3751	0.6201	0.8160	0.9234
0.0258	0.0671	0.1636	0.3471	0.5910	0.7971	0.9144
0.0364	0.0931	0.2182	0.4313	0.6734	0.8486	0.9384
0.0516	0.1289	0.2868	0.5222	0.7482	0.8898	0.9564
0.0383	0.0977	0.2274	0.4445	0.6851	0.8553	0.9414
0.0579	0.1432	0.3123	0.5525	0.7704	0.9012	0.9612
0.0156	0.0414	0.1049	0.2417	0.4642	0.7019	0.8649
0.0502	0.1256	0.2808	0.5149	0.7426	0.8869	0.9552
0.0843	0.2002	0.4049	0.6491	0.8341	0.9318	0.9738
0.0566	0.1402	0.3071	0.5465	0.7661	0.8990	0.9603
0.0405	0.1030	0.2379	0.4590	0.6975	0.8624	0.9446
0.0186	0.0489	0.1226	0.2753	0.5080	0.7373	0.8841
0.0652	0.1594	0.3401	0.5835	0.7921	0.9119	0.9657
0.0419	0.1062	0.2441	0.4675	0.7047	0.8664	0.9463
0.0488	0.1224	0.2750	0.5076	0.7370	0.8840	0.9539
0.0551	0.1367	0.3009	0.5392	0.7608	0.8963	0.9592
0.0646	0.1581	0.3380	0.5812	0.7904	0.9111	0.9654
0.0561	0.1390	0.3050	0.5440	0.7643	0.8981	0.9599
0.0312	0.0804	0.1920	0.3924	0.6371	0.8268	0.9284
0.0342	0.0879	0.2076	0.4159	0.6593	0.8403	0.9346
0.0362	0.0927	0.2173	0.4301	0.6723	0.8480	0.9381
0.0435	0.1099	0.2514	0.4772	0.7127	0.8709	0.9483
0.0342	0.0879	0.2076	0.4159	0.6593	0.8403	0.9346
0.0233	0.0610	0.1500	0.3243	0.5660	0.7800	0.9060
0.0448	0.1130	0.2573	0.4850	0.7191	0.8744	0.9498
0.0423	0.1072	0.2460	0.4700	0.7068	0.8676	0.9468
0.0590	0.1455	0.3165	0.5572	0.7738	0.9029	0.9619
0.0326	0.0839	0.1994	0.4037	0.6479	0.8334	0.9315
0.0287	0.0745	0.1795	0.3729	0.6177	0.8146	0.9227
0.0274	0.0711	0.1722	0.3612	0.6059	0.8069	0.9191
0.0711	0.1722	0.3612	0.6059	0.8069	0.9191	0.9686

0.0415	0.1053	0.2423	0.4651	0.7027	0.8653	0.9458
0.0704	0.1708	0.3589	0.6035	0.8053	0.9183	0.9683
0.3143	0.5548	0.7721	0.9020	0.9616	0.9855	0.9946
0.1978	0.4013	0.6457	0.8320	0.9309	0.9734	0.9900
0.0470	0.1182	0.2670	0.4975	0.7291	0.8797	0.9521
0.2159	0.4280	0.6704	0.8468	0.9376	0.9761	0.9911
0.1722	0.3612	0.6059	0.8069	0.9191	0.9686	0.9882
0.0170	0.0448	0.1130	0.2573	0.4850	0.7191	0.8744
0.0457	0.1151	0.2611	0.4900	0.7231	0.8765	0.9507
0.0118	0.0314	0.0809	0.1931	0.3941	0.6388	0.8278
0.2159	0.4280	0.6704	0.8468	0.9376	0.9761	0.9911
0.3015	0.5399	0.7613	0.8966	0.9593	0.9846	0.9943
0.0072	0.0193	0.0507	0.1268	0.2829	0.5175	0.7446
0.0296	0.0766	0.1839	0.3799	0.6248	0.8191	0.9248
0.0148	0.0392	0.0998	0.2315	0.4502	0.6900	0.8581
0.0552	0.1371	0.3015	0.5399	0.7613	0.8966	0.9593
0.0423	0.1072	0.2460	0.4700	0.7068	0.8676	0.9468
0.0497	0.1246	0.2789	0.5125	0.7408	0.8859	0.9548
0.0074	0.0198	0.0522	0.1301	0.2891	0.5250	0.7503
0.0557	0.1382	0.3036	0.5424	0.7631	0.8975	0.9597
0.0133	0.0352	0.0903	0.2125	0.4231	0.6660	0.8442
0.0193	0.0507	0.1268	0.2829	0.5175	0.7446	0.8880
0.0154	0.0407	0.1034	0.2387	0.4601	0.6985	0.8629
0.1268	0.2829	0.5175	0.7446	0.8880	0.9556	0.9832
0.0148	0.0392	0.0998	0.2315	0.4502	0.6900	0.8581
0.2262	0.4428	0.6835	0.8545	0.9410	0.9775	0.9916
0.0047	0.0127	0.0339	0.0871	0.2059	0.4134	0.6570
0.1034	0.2387	0.4601	0.6985	0.8629	0.9448	0.9790
0.1182	0.2670	0.4975	0.7291	0.8797	0.9521	0.9818
0.0527	0.1312	0.2911	0.5275	0.7521	0.8919	0.9573
0.0601	0.1480	0.3208	0.5622	0.7773	0.9047	0.9627
0.0423	0.1072	0.2460	0.4700	0.7068	0.8676	0.9468
0.0359	0.0920	0.2159	0.4280	0.6704	0.8468	0.9376

1.7380	4.3607	9.8401	18.4868	27.7038	34.2495	37.6391
3.0515	6.6163	12.3843	19.8936	27.5177	33.4336	37.0178
-3	-2	-1	0	1	2	3

logits as acceptable (Kopp & Jones, 2020). Table 5. Shows the data for test equating TF-X to TF-Y generated from Microsoft Excel.

5. Discussion

Based on the results and findings in this study, the data generated from the 858 respondents gives a significant result on item fit study for Rasch model. Both instruments (TF-X and TF-Y) for Principles of Accounting test incorporated a reasonable fit index for item validity and reliability study. The findings shows that all the items in the instruments are unidimensional. Examining the fit of data can be viewed as a quality control of the data. Schoolteachers can apply this Rasch model approach for test equating use in items writing in schools. A few outliers in the data set may be negligible, but a few outlier items raise serious questions about test administration, data entry accuracy, and latent trait definition (Linacre, 2021).

Result of the finding suggested that by adding the equating constant can generate a parallel test form on a comparable scale. The test form difficulty bias was small. Means value for test form TF-X was reduced by the value of 0.1775 if compared to test form TF-Y. Kopp and Jones (2020) in their previous study had suggested that anchor item drift up to 0.5 logits as acceptable. Meanwhile, Linacre (2021) suggested that anchor item drift greater than 0.6 logits were flagged as not functioning well.

Although, CINEG test equating design seem very promising, yet it has limitation also. Fischer et al., (2021) in their study of four test equating method (1) Equating because of Fixed Parameter, (2) Equating by simultaneous calibration, (3) Equating by mean/mean linking and lastly the (4) weighted mean/mean linking. They suggested the weighted mean/mean method get more accurate test equating result with condition that the value of mean and variance for anchor item's difficulty parameter must match the ability distribution of the sample. The imprecision estimated of anchor item parameter will contribute to the increment of measurement standard error.

6. Conclusions

By introducing tailor made instrument with item anchoring and Common Item Non-equivalent Group test equating approach, the author strives to improve assessment system in Malaysia, and prevent grade inflation. The main objective of this research is to safeguard fairness in assessment by developing a justified equivalent test form with comparable standard for the item developer, examination body, students, schools, and parents. Consequently, this study managed to ascertain the applicability of test equating method in building an equivalent test instrument by using eight internal anchor items unto an equivalent scale.

6. Contributions

The results of this study would seem very useful for many stakeholders for improving their professional careers. The item developers can follow the guidelines in doing test equating to generate multiple equivalents test form by using anchor items. The test item can be carefully tailored made following the level of item difficulties to match the student abilities to ensure that it is representative of student's responses and performance. The item developers can made sure that item analysis is conducted in a proper way that the test content is accurate, relevant, and significant.

The finding of the study can be used by examination officers to reflect on current practices, more importantly to accumulate more quality items in item bank and conducting item's field test to ensure statistical evidence for standardised testing across the nation. Thus, it brings a quantum leap in developing computer adaptive testing to replenish bank items. Examination body can gain more confidence in maintaining exam's standard for administer multiple equivalent sets of exam instruments within different time frame as its instrument is equated with comparable scale score. Test equating can safeguard fairness in assessment and avoid grade inflation in Malaysia's education system.

Item anchoring and test equating have great impact on social science research especially in education assessment study. The Rasch Model framework provide guidelines in item response theory (IRT) for schoolteachers in creating Principles Accounting test instrument by studying item fit statistic such as test reliability and

validity. These techniques allow schoolteachers to develop test item with different levels of difficulty to assess students.

Student's performance in school can be measure by using vertical and horizontal test equating across different cohort and grade of students. Students and parents become more confidence in school assessment reporting system. In which, test reports created from schools become more transparent and trustworthy because schoolteachers have more knowledge in analysing test data. Every single test score represents a student's future. Sharing test score data is an effective strategy that schools can use to engage families and communities for student latent skill improvement and monitoring.

The findings of this study contribute to the knowledge in the relevant literature especially in item response theory and psychometric field. The education assessment has changed over time to suit the needs and demands of society. For example, the abolishment of Form Three assessment PT3 (*Pentaksiran Tingkatan* 3) in Malaysia from year 2022 onwards and replacing it with classroom-based assessment (PBD-*Pentaksiran Bilik Darjah*) in setting up items bank for schools (Rajaendram, 2022). These PBD consists of formative and summative assessments would be carried out every year. In this case, the applicability of test equating by using item anchoring will be beneficial.

References

Cook, L. L., & Peterson, N. S. (2015). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*(3), 225–244.

Cook L. L. & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Finklestein, I. E. (1913). *The marking system in theory and practice*. Baltimore, MD: Warwick & York, Inc.

Fischer, L., Rohm, T., Carstensen, C. H., & Gnambs, T. (2021). Linking of rasch-scaled tests: Consequences of limited item pools and model misfit. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.633896

Harris, D. (1989). Comparison of 1-, 2-, and 3-Parameter IRT models. *Educational Measurement: Issues and Practice*, *8*(1), 35–41. https://doi.org/ 10.1111/j.1745-3992.1989.tb00313.x

Harris, D., & Kolen, J. M. (1990). A Comparison of two equipercentile equating methods for common item equating. In *Educational and Psychological Measurement - EDUC PSYCHOL MEAS* (Vol. 50). https://doi.org/10.1177/0013164490501006

Kopp, J. P., & Jones, A. T. (2020). Impact of item parameter drift on rasch scale stability in small samples over multiple administrations. *Https://Doi.Org/10.1080/08957347.2019.1674303*, *33*(1), 24–33. https://doi.org/10.1080/08957347.2019.1674303

Hayes, N. (2021). Doing psychological research. In *Open University Press* (2nd ed.). Open University Press. https://books.google.com.my/books

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices: Third edition. *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. (2021). *Winsteps rasch measurement computer program user's guide*. Winsteps.Com.

Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, *1*(3), 165–174.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the dichotomous rasch model. *Educational Measurement*, *17*(Fall 1980), 179–193.

Marco, G. L., Petersen, N. S., & Elizabeth, E. S. (1983). A test of the adequacy of curvilinear score equating models. In *New Horizons in Testing* (pp. 147–177). Elsevier. https://doi.org/10.1016/B978-0-12-742780-5.50018-4

McKinsey & Company. (2020). *How Covid-19 has pusched companies over the technology tipping point*. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/how-covid-19%20has-pushed-companies-over-the%20technology-tipping-point-and-transformed-business-forever

Messick, S. (1995). Validity of Psychological Assessment. Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. American Psychologist, 5D, 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Ministry of Education. (2021). Quick facts 2021: Malaysia educational statistics. *Educational Planning and Research Division*.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, India. https://doi.org/10.1177%2F014662169702140006

Rebecca Rajaendram. (2022, June 2). PT3 exam abolished, says education minister | The Star. *The STAR*. https://www.thestar.com.my/news/nation/2022/06/02/pt3-exam-abolished-says-education-minister

Tierney, R. D. (2013). Fairness in classroom assessment. In J. H. McMillan (Ed.), *SAGE Handbook of Research on Classroom Assessment* (pp. 125-144). Thousand Oaks, CA: SAGE Publications.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational measurement for applied researchers. *Educational Measurement for Applied Researchers*. https://doi.org/10.1007/978-981-10-3302-5