Examining the Qualities of Anchor Items in Rasch Model Test Linking

Chieng Zouh Fong[1,a,*]

[1]Faculty of Education, University of Malaya, Kuala Lumpur, Malaysia

[a] chieng@moe.gov.my,

*corresponding author

I. Abstract

Study of invariance measurement in examining quality of anchor items in Rasch model Test Linking are essential in safeguard test fairness especially in high stake examinations. Rasch model provide sufficiency, invariance, and objectivity in resultant measures (Stenner, 2001). The primary objective of this study was to introduce a framework for schoolteachers, especially teachers in Malaysia to ascertain the relative efficiency of test linking methods in comparing students' high stake examination measures. This paper proposed the practical implications of score equating by describing aspects of the test linking and item anchoring process which can be used by teachers. This study examined the two Principles of Accounting Tests (ZPATx and ZPATy) within a Rasch measurement framework. Are anchor items measure in ZPATx and ZPATy invariance? We should investigate and monitor item parameters value and taking account the measurement error to prove the invariance property by using anchor items through concurrent and separate test equating analysis. Data collection was through stratified random sampling method to two different groups of respondents comprising 429 students each and resulting response data analysed using Winsteps software. Results showed a good fit by using the Non-Equivalent Groups with Anchor Test (NEAT) equating method. Calculation of the different of mean in logit for anchor items measure were used to place both tests form on same scale. Linear equating been chosen as the stronger model for this study for more accurate equating result as suggested by Linacre and Wright (1989). Equating interval scales as stated by Linacre and Wright, when plotted the measures common to the test must follow the identity line stochastically.

Keywords: Invariance measurement, Item Anchoring, Test Equating, and Test Fairness.

II. Introduction

Research in test administration are labelled as an observational study, in which an examinee receives a set of test items while test administrator observes examinee responses. Rasch model capable of recognizing every response and able to generate item parameter, and test taker ability parameter through real data base simulation. This unique relationship in test construction and administration have created wide platform for researchers to compare the probability of correctly answering items and draw interpretation of the effect of both item parameter and test takers abilities. The response of the item parameter is dependent upon students' ability. In this case, we are going to design an instrument with embedded anchor items to measure students' ability across multiple forms. What made the test form a parallel form? What evidence that test developer can contribute to testing program that the standardize exam can establish comparable score in an equivalent instrument? When can scores from different assessments be viewed as the same, essentially the same, pretty much the same, sort of the same, hardly the same, or only same for the name's sake only? There should be a strong evidence base in psychometric property to prove that both the test forms are equivalent. Therefore, capitalizing the equivalence of the two forms can be calculated directly from the value of anchor items measured in test linking (Kolen & Brennan, 2014).

Q. Zhang (ed.), *Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023)*,
Atlantis Highlights in Social Sciences, Education and Humanities 23,
https://doi.org/10.2991/978-94-6463-494-5_23

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" (Kolen & Brennan, 2014). Test Linking is essential in safeguarding test fairness especially in high stakes examination. The Rasch model provides sufficiency, invariance, and objectivity in resultant measures (Stenner, 2001). The primary objective of this study was to introduce a framework for schoolteachers, especially teachers in Malaysia to ascertain the relative efficiency of test linking methods in comparing students' high stake examination measures. This study examined the two Principles of Accounting Test (ZPATx and ZPATy within Rasch measurement Framework. Are anchor items measured in ZPATx and ZPATy invariance?

III. Literature Review

We begin with the theoretical and operational principles of item anchoring and test linking. Then discuss these approaches in the context of examining the qualities of anchor items in Rasch model test linking. Test equating is a particular statistical approach for developing equivalent scores for examinees of equal ability in standardised tests (Kolen & Brennan, 2014). Generally, equating adjusts the outcomes of two parallel tests of the same length so that the two forms have a comparable mean and standard deviation. Equating techniques and outcomes are different in classical and item response test theories. The main purpose of equating is to make a statistical transformation of test scores and adjusts for differences in difficulty among test forms. 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 (interchangeable) (Holland et al., 2006). Techniques other than equating, such as scaling and linking for developing equivalent scores are also envisaged under IRT.

Related studies on equating designs can be found by Haebara (1980), Harris (1989), Cook and Eignor (1991), Hambleton et al. (1991), von Davier et al. (2004), and Kolen & Brennan (2014). The three most used equating designs were Random Group (RG) design, Single Group (SG) design, and Non-Equivalent Group with Anchor Test (NEAT). Current research project has adopted an innovative approach to equating test by using the Rasch model, sometimes, but erroneously, called the one parameter IRT model. The Rasch model was first developed by George Rasch in the 1960s, a Danish psychometrician who presented a theory and approach to the field of social science measurement with his probabilistic models for tests of intelligence and achievement (Rasch, 1993). This approach is now referred to as the Rasch measurement theory.

Rasch model provided the invariance measures of person estimates across two equivalent tests with concurrent validity. Rasch analysis used Winsteps which allows the user to input known values such as anchor items values, so that output values such as person values can be compared across different research contexts (Bond et al., 2021). I have the privilege to study common item linking with Prof. Bond. He has guided me how to plot the invariance plot. The learning process of utilising anchor items in test linking will be reported in this study. Hence, this study intends to proof anchor items invariance properties across different set of equivalent instruments.

IV. Research Purpose

The purpose of this research is to introduce a framework to examine qualities of anchor items in test linking. What means qualities anchor items? And why using Rasch model in test linking? The main research objective seeks to investigate the claim that Rasch measures should remain invariant in appropriate circumstances with eight selected anchor items across the two samples. The anchor item measure has the property to compare test taker ability across different set of instruments. Test linking is essential in safeguard test fairness especially in high stake examinations. Rasch model provided sufficiency, invariance, and objectivity in resultant measures (Stenner, 2001).

V. Conceptual Framework

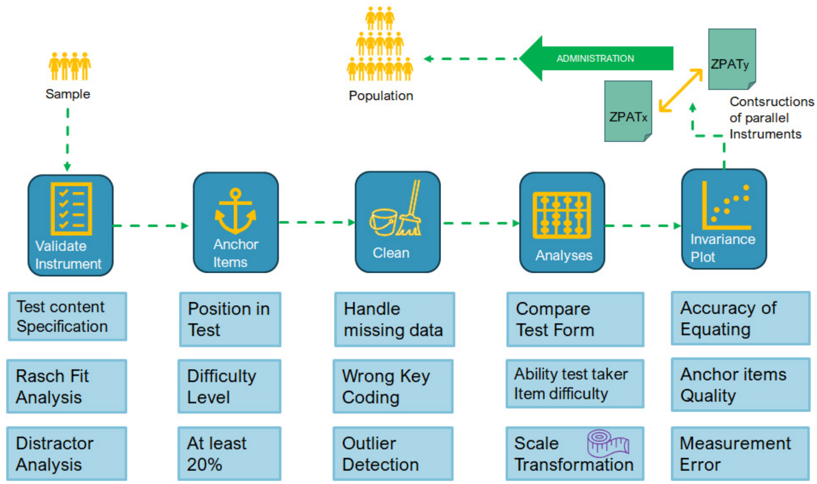


Figure 1 Conceptual framework for ZPATx Vs ZPATy.

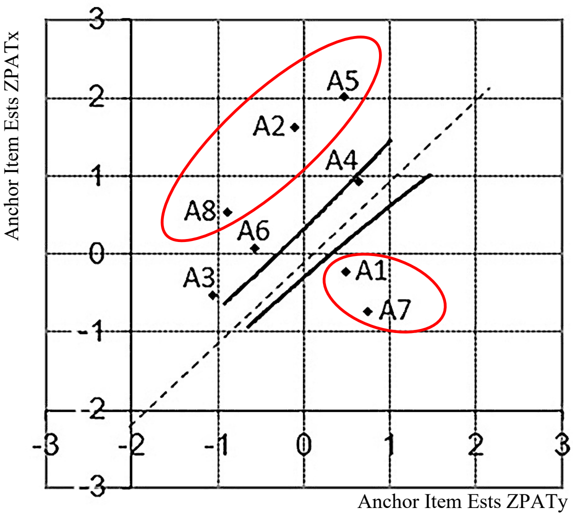
Figure 1 illustrates framework to examine qualities of anchor items in test linking. This study was a continuation of the previous work that I did for my dissertation (Chieng & Tong, 2022). ZPATx and ZPATy were designed for high school (age 16 to 17 years old) students' cognitive development for fundamental knowledge about Principle of Accounting. It was a high-stake examination's paper. The purpose of this research is to provide an essential guideline for the test equating process by transforming a new test form (ZPATy) to an earlier test form (ZPATx) to possess the same levels of item difficulty that match the examinees' ability. In general, the Principal of Accounting (PA) subject is the knowledge about bookkeeping rules and guidelines for reporting financial data. From year 2018 onward, Ministry of Education Malaysia had endorsed a dual certification collaboration with Pearson LCCI. For those who pass in Malaysia Certificate of Examination (MCE) Principle of Accounting (PA) exam, will be awarded a Level 2 in LCCI Accounting certificate (Book-Keeping & Accounting) (Pearson, 2018).

This study mainly adopted a non-experimental quantitative research design to obtain the findings for data analysis. The data from a large representative sample was divided using the spiralling method. Spiralling refers to the way test booklets are assembled, packaged, delivered to testing sites, and distributed to respondents. Examinees were alternately assigned a different set of instruments (Kolen & Brennan, 2014). The population for this study consisted of government secondary school students. There are 702 schools in Malaysia which take the Principles of Accounting (PA) test, with the ratio of urban to sub-urban schools being approximately 1:2 (258 schools to 444 schools) (Malaysia Ministry of Education, 2021). The total population (N) was 67,893 students taking the Principle of Accounting Subject. However, the sample (n) selected was 858 students, which was split equally between two test forms, each consists of 429 students. Stratified random sampling (SRS) was utilised for this study by dividing the sample into strata groups according to demographic characteristics such as domicile and gender (Hayes, 2021).

We start with validation of the instrument by studying the test content specification table, Rasch fit analysis and distractor analysis. The requirement for test linking to become successful, both test instruments should measure the same construct, same context and have the same level of difficulty and reliability. The test instruments were designed to achieve the objectives of this study and be able to show a statistically significant result to generalise to the whole population. The first set of instruments was named ZPATx, and the second as ZPATy. Both instruments consisted of 40 multiple-choice questions (MCQs), among which 32 were unique items. Only eight items were common items labelled as anchor items (20% of total test length as suggested by Cook & Eignor, 1991). Anchor items were placed at almost the same position on each instrument.

Data for this study was first collected from the examinee's responses to ZPATx and ZPATy. Then, the responses were transferred to a Microsoft Excel sheet, and subsequently imported into Winsteps. The psychometric attributes, such as item and person reliability, separation index, mean and standard deviation were examined as well. Item fit statistics such as Pt-Measure Correlation, infit and outfit MNSQ, ZSTD, Standard Error (SEs), Item Characteristic Curve (ICC), and distractor analysis for each item were examined too. After identifying each item's function and fit, an item map was used to show how items were distributed in terms of their difficulty in comparison to respondent's ability.

Data cleaning is a very important process in quantitative research. Improperly cleansed or calibrated data can lead to big mistake in reporting results and create inaccurate test linking results. Data cleaning in this study involved spotting incorrect key answer for Multiple Choice Question (MCQ). I have experience of keying wrong answer key for my anchor items and end up with all my anchor items points scattered all over the invariance plot (Figure 3). It became the indicator spotted by Professor Bond that something must be very wrong with the analysis, and he asked me to reconsider the rescoring and check for my anchor item's answer key. After rescoring, I managed to spot the wrong answer key and produce a better invariance plot for my anchor items (see Figure 3). In this case, the invariance plot of Rasch model can assist test developers to build parallel test form with high accuracy and secure low measurement error. Examining the quality of anchor item can improve the efficiency of test linking. My mistake in data cleaning procedure demonstrate the benefit of Rasch model invariance plot for test linking study. I will show it in **VII(A)** Preliminary study and **VII(B)** final invariance plot.



*Ests = Estimates.

Figure 3 Anchor items Invariant Plot-ZPATx Vs ZPATy (Preliminary study).

VI. Results: Rasch Analysis for ZPATx and ZPATy

The statistical analysis starts with two key aspects for the variable of mean and standard deviation: (1) Item difficulty estimates and, (2) Person ability estimates. Other key parts of the analysis include (3) Precision of person and item estimates (SEs); (4) The fit of the items; (5) The fit of the persons; and (6) the separation and reliabilities index of the person and item estimates are reported in detail in Rasch's output tables. We utilised two methods of analysis in this study which is separate analysis and concurrent analysis results for ZPATx and ZPATy shown in Table 1.

Table 1 Summary Statistics of Person and Items for instruments ZPATx and ZPATy.

Summary Statistics of ZPATx & ZPATy	Separate Analysis				Concurrent Analysis	
	ZPATx		ZPATy		ZPATx & ZPATy	
	Person	Item	Person	Item	Person	Item
N	429	40	429	40	858	72
Measure (logit)						
Mean	-0.48	0.00	-0.46	0.00	-0.52	0.00
SD(Standard Deviation)	0.74	0.81	0.69	1.13	0.78	1.05
SEs (Standard Error)	0.04	0.13	0.03	0.18	0.03	0.13
Outfit Mean-square (MNSQ)						
Mean	1.03	1.03	1.02	1.02	1.03	1.02
SD (Standard Deviation)	0.26	0.19	0.33	0.17	0.32	0.19
Separation	1.73	6.97	1.46	9.31	1.76	8.99
Reliability	0.75	0.98	0.68	0.99	0.75	0.99

*Item and person outliers were removed.

A. Overview of Results in Separate Analysis

By default, the mean of item estimates (measures) is located at 0 logits (Bond et al., 2021). We see the results summary statistic for ZPATx and ZPATy in Table 1, have mean index as 0 logits and standard deviation for item estimates for ZPATx as 0.81 and ZPATy as 1.13. The reliability of the item difficulty estimates for ZPATx is 0.98 and ZPATy is 0.99 on a 0-to-1 scale. The high item reliabilities with close to 1 indicating a wide difficulty range as explained by Wright & Master (1982). Reliability is calculated as the number of standard errors of spread among the items. Both ZPATx and ZPATy have item reliability index close to 1, this indicates ZPATx can be equated to ZPATy.

B. Overview of Results in Concurrent Analysis

We can see that the eight anchor items manage to secure concurrent analysis results as close as separate analysis results. Summary statistics for concurrent analysis for ZPATx and ZPATy in Table 1 shows mean item measure as 0 logit and person's mean as -0.52. Standard deviation for item measure 1.05 and person measure 0.78. Item separation 8.99 and person separation 1.76. Item reliability 0.99 and person reliability 0.75. The measures of mean, standard deviation, and separation index in concurrent analysis did not differ much if compared to separate analysis. Data from each subsample (ZPATx = 459; for ZPATy = 459) were analysed concurrently and contribute to total sample as n = 858.

To do the concurrent analysis, we need to "stake" together 72 items in one analysis and share 8 pairs of anchor items (shaded with yellow and green colour) as shown in Figure 2. In which 32 unique items measure (and Standard Errors = SE) from ZPATx, 8 pairs of anchor items (and SEs) for ZPATx and ZPATy, and 32 unique items (and SEs) from ZPATy were imported into an Excel spreadsheet. The eight pairs of anchor items in concurrent analysis were named as Ai1, Ai2, Ai3, Ai4, Ai5, Ai6, Ai7, and Ai8 (see Table 3) which have been plotted in the invariance plot and have identical measure for each pair. The value of anchor items measures for ZPATx and ZPATy are assumed to be identical in this concurrent analysis. We can draw a diagonal line 45° (slope=1) through the point of calibration mean of ZPATx and ZPATy (0.0 logits) required for invariance (and the 95% control lines (two black solid lines) are based on the SEs for each of the item pairs). Qualified anchor items are now calibrated based on the whole data set; the results yield a quality control line as a dotted line show in Figure 3 and Figure 4. This dotted line is not a regression line but the Rasch-modelled relationship requirement for invariance control lines.

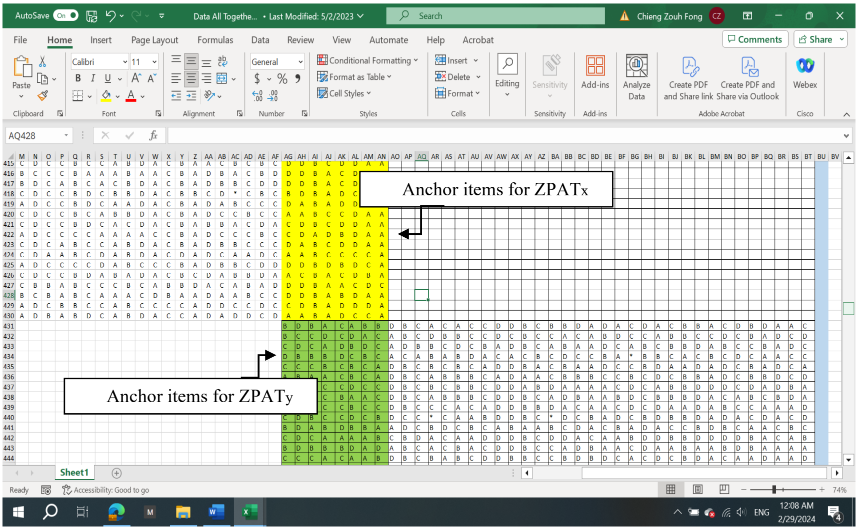


Figure 2 Concurrent analysis for ZPATx and ZPATy.

VII. Implementation and results of the anchoring process

Anchor item values should remain invariance. “Does instrument ZPATy measures the same construct as instrument ZPATx?” This is one of the important arguments to help item developers to use invariance principle to build a parallel test form by using anchor items measure to secure the item difficulty parameters in each test. We use the linking procedure from Rasch modelling to see whether anchor items in ZPATx and ZPATy will remain invariance. The relative difficulty of the two tests would be identical (within error). The intercept of that diagonal line on the y-axis and x-axis is the adjustment that must be made for ZPATy estimates to bring them to the same origin as the ZPATx estimates as shown in Figure 4 at the end part of this article.

A. Information for Anchor items (Preliminary Study)

Table 2 Preliminary Invariance Plot for ZPATx and ZPATy.

Separate Analysis ZPATx			Separate Analysis ZPATy		
Anchor Item	Difficulty Estimate	Error Estimate	Anchor Item	Difficulty Estimate	Error Estimate
A1(X1)	0.48	0.11	A1(Y5)	-0.06	0.11
A2(X3)	-0.11	0.10	A2(Y4)	1.81	0.17
A3(X6)	-1.05	0.10	A3(Y8)	-0.34	0.10
A4(X14)	0.64	0.12	A4(Y15)	1.10	0.13
A5(X25)	0.46	0.11	A5(Y23)	2.20	0.20
A6(X27)	-0.56	0.10	A6(Y28)	0.26	0.11
A7(X29)	0.75	0.12	A7(Y29)	-0.55	0.10
A8(X36)	-0.89	0.10	A8(Y37)	0.71	0.12

In preliminary study, each set of anchor items has their own special value of item difficulty measure as shown in Table 2. The measure estimate will change depending on who we are administer to the ZPATx and ZPATy. If the sample of respondent differ very much in their ability level, the measure estimates for the set of anchor items will differ. When the anchor items estimate is differed, the others unique items estimate also will change by corelation to the anchor items measure estimate. The test results can then be analysed with the anchor items measured and become a scale for measuring purposes. This supports the claim of Rasch model as "An acceptable method of scaling must result in a derived scale which is independent of the original scale and of the original group tested" (Loevinger, 1947, p.46).

Precision in measurement are derives from some good information that we have on the model that we measure (Bond et al., 2021). Application of the model provides diagnostic information and display on invariant plot. In preliminary study we spotted that 5 out of 8 chosen anchor items are placed out of the boundary of the 95% control lines as shown in Figure 3. The findings demonstrate that anchor items A1, A2, A5, A7 and A8 are misfit (In red circle). This study identifies that these problematic anchor items need a further investigation. Discrepancy between the anchor values and the values that would have been estimated from the current data can be reported as displacement. If the two tests were both measuring the same ability and were of the same mean difficulty, at least 95% of the plotted points would lie within the control lines, and the diagonal line representing the modelled relationship between the performances would pass through the origin point of the graph (Bond et al., 2021). This invariance plot indicated something must be very wrong with the analysis and triggered me to consider rescoring and replotting of invariance plot. There was no more shortcut in data analysis. I need to work diligently and systematically to get all the answer key keying correctly and do the data cleaning properly. Re-analysed and replotted the invariance plot. We will see a more convincing results in the final plot at next session (Figure 4).

B. Final Invariance Plot

Figure 5 shows separate analysis for anchor item estimates ZPATx and ZPATy in black dots with labels as A1, A2, A3, A4, A5, A6, A7 and A8. The position of anchor item in ZPATx name as A1(X1). A1 is anchor item number 1. And X1 means that anchor item was placed in number one of instrument ZPATx. Meanwhile A1(Y5) have the position of item number 5 in instrument ZPATy.

After rescoring the answer key for ZPATy, we have new results for all 8 anchor items as shown in Table 3. After investigating the eight pairs of anchor items estimate separate analysis (Green colour circle), we get new results for anchor items. The results of the anchor items pair were plotted in the invariance plot as shown in Figure 4. All eight pairs of anchor items in separate analysis were fall inside the 95% control lines. The eight pairs of anchor items in concurrent analysis were named as Ai1, Ai2, Ai3, Ai4, Ai5, Ai6, Ai7, and Ai8 in blue dots which were plotted in invariance plot and have identical measure for each pair.

Table 3 Anchor items Invariance Plot- ZPATx and ZPATy.

Separate Analysis ZPATx			Separate Analysis ZPATy			Concurrent Analysis ZPATx & ZPATy		
Anchor Item	Difficulty Estimate	Error Estimate	Anchor Item	Difficulty Estimate	Error Estimate	Anchor Item	Difficulty Estimate	Error Estimate
A1(X1)	0.48	0.11	A1(Y5)	0.45	0.11	Ai1	0.01	0.09
A2(X3)	-0.11	0.1	A2(Y4)	-0.14	0.1	Ai2	-1.02	0.07
A3(X6)	-1.05	0.1	A3(Y8)	-1.19	0.11	Ai3	-1.59	0.07
A4(X14)	0.64	0.12	A4(Y15)	0.74	0.12	Ai4	-0.14	0.08
A5(X25)	0.46	0.11	A5(Y23)	0.47	0.11	Ai5	0.09	0.09
A6(X27)	-0.56	0.1	A6(Y28)	-0.61	0.1	Ai6	-1.09	0.07
A7(X29)	0.75	0.12	A7(Y29)	0.86	0.12	Ai7	-0.88	0.07
A8(X36)	-0.89	0.1	A8(Y37)	-0.95	0.1	Ai8	-1.08	0.07

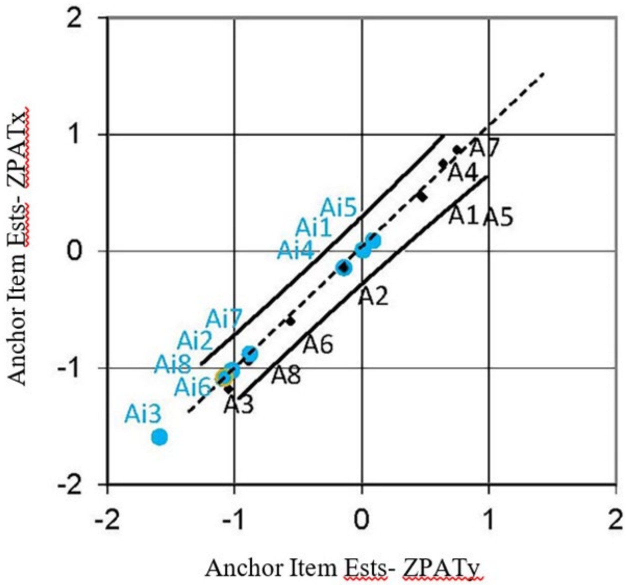


Figure 4 Information for Anchor items after rekeying correctly with correct answer code.

The value of anchor items measures for ZPATx and ZPATy are assumed to be identical in this concurrent analysis. Qualified anchor items are now calibrated based on the whole data set; results yield a quality control line as a dotted line show in Figure 4. This dotted line is not a regression line but the Rasch-modelled relationship requirement for invariance control lines. The 95% of control lines were set with the value of Standard Error (SEs) for anchor items measured. The higher the number of SEs, the more dispersed, the control lines are (Bond et al., 2021). In this case, by observing the standard error produce by each anchor item either in concurrent or separate analysis, all of them are uniformly distributed.

We must consider measurement error when doing correlation statistics. The more information, the smaller the error and higher unattenuated corelation value should be. Therefore, plotting of Rasch modelled estimated of person ability or item difficulty can be more useful when interpreted in the context of the precision (SEs) or measurement error of these estimates (Bond et al., 2021). The finding demonstrate that anchor items estimate information is beneficial for test score equating and provide useful practical guidelines for item developers.

C. Guidelines in Selecting Qualities Anchor Items in Test Linking

If you want to have eight good anchor items as suggested by Cook & Eignor (1991), which is at least 20% of total test length (40 questions); we need to have 12 or more items offer up in both version of instruments. Then we can pick the eight anchor items which remain close to invariant after administration. Anchor items should extend across the whole range of the variable, if possible, by (1) include enough test items from the reference form. (2) choose a set of test items that resembles the full test in construct and context. (3) Don't include any test items that have been changed. (4) Try to avoid breaking up an item set. (5) Put each anchor item in approximately the same position in the new form as it was in the reference form. (6) Include test items that represent the full range of difficulty. (7) Don't use test items at the end of test as anchor items unless the time limit is very generous. (8) Choose anchor items that correlate well with the total score.

VIII. Conclusion

The findings demonstrate that the information of anchor items estimate is beneficial for test score equating to align ZPATx and ZPATy on a same scale. For this purpose, eight pairs of anchor items have been chosen to link both forms. The study identified scenarios where in concurrent analysis, pairs of anchor items have identical value and become quality control line for us to check measures invariance claim by Rasch model. While in separate analysis, equating results shows that anchor items parameter drifted less than 0.5 and had secured the Rasch invariance measurement. All the eight pairs of anchor items fall in the boundary of 95% confidence line and maintains their invariance as shown in Figure 4. Anchor items parameter that drifted more than 0.5 were possibly caused by DIF effect and need further study (Kopp and Jones, 2020).

The main purpose of doing test equating is to give every examinee a fair assessment by preparing a parallel test form that is uniquely created for them. Test equating can safeguard test fairness. If we want to set up an examination, let it be fair for everyone. This study had introduced a Rasch framework for schoolteachers, especially teachers in Malaysia to ascertain the relative efficiency of test score equating methods in comparing students' high stake examination measures. "Research" is more about the learning process than the end results. We never know that the mistake we made can be the learning pathway for others too.

Acknowledgements

Trevor G Bond: for his knowledge in technical support (data analysis).

References

[1] Bond, T.G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model: Fundamental measurement in the human sciences* (4th ed.). New York, NY: Routledge.

[2] Chieng, Z.F., & Tong, Y.C. (2022). Test Score Equating and Item Anchoring for High Stakes Examination. *Journal of Educational Sciences*, *18*(4), 1-11. http://dpi.org/10.15625/2615-8957/22210401

[3] 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

[4] Haebara, T. (1980). Equating Logistic Ability Scales by A Weighted Least Squares Method. *Japanese Psychological Research*, *22*(3), 144-149. http://dpi.org/10.4992/psycholres1954.22.144

[5] 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.17453992. 1989. tb00313.x

[6] Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*, *30*(1), 84-87. Newbury Park, CA: Sage.

[7] Hayes, N. (2021). Doing Psychological Research. In Open University Press (2nd ed.). Open University Press. https://books.google.com.my/books

[8] Holland, P. W., Dorans, N. J. & Petersen, N. S. (2006). Equating Test Scores. Handbook of Statistics, 16(1), 169–203. https://doi.org/ 10.1016/S0169-7161(06)26006-1

[9] Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking: Methods and practices: Third edition. New York: Springer. 1–566. https://doi.org/10.1007/978-1-4939-0317-7

[10] Kopp, J. P., & Jones, A. T. (2020). Impact of Item Parameter Drift on Rasch Scale Stability in Small Samples Over Multiple Administrations. *Applied Measurement in Education*, *33*(1), 24–33 https://Doi.Org/10.1080/08957347.2019.1674303

[11] Linacre, J. M. (2021). Winsteps Rasch Measurement Computer Program User's Guide. Winsteps.Com

[12] Lovinger, J. (1947). A Systematic Approach to The Construction and Evaluation of Tests of Ability. *Psychological Monographs*, *61*(4).

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

[14] Pearson. (2018). *SPM Principles of Accounting Dual Certification-LCCI Asia*. https://www.lcciasia.com/spm-principles-of -accounting-dual-certification/

[15] Rasch, G. (1993). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: MESA Press.

[16] Stenner, J. (2001). The Necessity of Construct Theory. *Rasch Measurement Transactions*, *15*(1), 804-805.

[17] von Davier, A. A., & Han, N. (2004). Population Invariance and Linear Equating for The Non-equivalent Groups Design. *ETS Research Report Series*, *2004*(2), i–21. https://doi.org/10.1002/j.2333-8504.2004.tb01974.x

[18] Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.