THE USE OF RASCH MEASUREMENT MODEL IN ENGLISH TESTING

Siti Eshah Mokshein*, Haliza Ishak and Hishamuddin Ahmad
Universiti Pendidikan Sultan Idris, Malaysia
*e-mail: eshah@fpm.upsi.edu.my

Abstract: This study aimed at determining the quality of the English Paper 1 (EP1) items of UPSR trial examination for six graders in terms of its reliability, validity and items characteristics. It also sought to determine the difficulty levels of 40 multiple-choice items consisting five constructs of vocabulary, language and social expression, grammar, cloze-comprehension and reading comprehension. A number of 525 primary schools students were randomly selected from 3876 students in Kuala Selangor, Malaysia. Using the Rasch measurement model, the validity evidences were shown through the results of Principle Component Analysis (PCA), fit statistics and item distractor analysis. The results from PCA analysis showed the absence of second dimension in the test, which met the assumption of modern testing theory. Fit statistics analyses have identified seven misfit items that are beyond the acceptable range (0.7 - 1.3 logit). Item distractor analysis has identified five problematic items whereby three of them are also misfit items. Summary statistics shows that the reliability indices of Cronbach's Alpha were greater than 0.80 and separation indices were greater than 2. This study would benefit teachers in improving existing assessment practice by spreading out the importance of item analysis in schools, particularly in language testing.

Keywords: ***item analysis, validity and reliability, Rasch measurement model***.

PENGGUNAAN MODEL PENGUKURAN RASCH DALAM TES BAHASA INGGRIS

Abstrak: Penelitian ini bertujuan untuk menentukan kualitas butir soal *English Paper 1* (EP1) dilihat dari segi kehandalan, validitas dan karakteristik butir pada ujian percobaan UPSR untuk siswa tahun enam. Hal ini juga untuk menentukan tingkat kesulitan dari 40 aitem pilihan ganda yang terdiri dari lima konstruksi kosakata yaitu, bahasa dan ekspresi sosial, tata bahasa, pemahaman *cloze* dan pemahaman membaca. Sejumlah 525 siswa sekolah dasar dipilih menggunakan metode proporsional *stratified random sampling* dari 3876 siswa di Kuala Selangor, Malaysia. Dengan menggunakan model pengukuran Rasch, bukti validitas ditunjukkan melalui hasil Analisis Komponen Utama (PCA), serta statistik fit dan analisis butir distraktor. Analisis PCA menunjukkan tidak adanya dimensi kedua dalam tes, yang memenuhi asumsi teori pengujian modern. Analisis statistik Fit telah mengidentifikasi tujuh aitem yang tidak sesuai dan berada di luar rentang yang dapat diterima (0,7 - 1,3 logit). Analisis item distraktor telah mengidentifikasi lima aitem bermasalah di mana tiga di antaranya juga aitem tidak sesuai. Kesimpulan dari analisis statistik menunjukkan bahwa indeks reliabilitas Cronbach's Alpha lebih besar dari 0,80 dan indeks pemisahan lebih besar dari 2. Penelitian ini akan bemanfaat pada guru dalam meningkatkan praktek penilaian yang ada dengan pentingnya analisis aitem di sekolah-sekolah, terutamanya dalam pengujian bahasa.

Kata Kunci: ***analisis aitem, validitas dan reliabilitas, model pengukuran Rasch***

INTRODUCTION

The mismatch of the current academic achievement in public examination with international assessment such as Trends in Mathematics and Science Study (TIMSS) and the Programme for International Pupils Assessment (PISA) result as reported in Malaysia Education Blueprint 2013-2015 has turned to be a vital issue in education assessment in Malaysia (Kementerian Pendidikan Malaysia, 2013). For English Paper 1 (EP1) for example, the quality of the items used in UPSR 2010 and 2011 was reported not up to the standard of international benchmark. Based on item analysis conducted by the Pearson Group, it was reported that the pupils have not been assessed with good quality items even they were developed by the highest authority in the Malaysia education assessment system, which is the Examination Syndicate.

The implementation of the educational transformation, which focuses on quality education system as stated in Malaysia Education Blueprint 2013 -2015, is timely so that Malaysian education can compete with the international standards. This is aligned with the views of Tavakol & Dennick (2013) which stated that the methods and the quality of assessment processes are as significant as the process of teaching and learning in any form of educational activities. Consequently, the assessment result would be meaningless if the considerations in determining the quality of the assessment have been omitted by the item developers during the item development process. Since it might violate the validity evidence, Martone & Sireci (2009) have emphasised that good development of the test items may pledge accurate assessment.

A good test consists of good quality of operational items that are capable to be an accurate indicator of pupils' knowledge, skills and abilities. Theoretically, this statement gives the impression that item building is easy but in reality, providing good quality items is not as easy as expected. To make testing as a highly valid and reliable measurement tool, the test items should be developed or written according to the standard set by the highest authority like Examination Syndicate. In Malaysia, the principles of writing quality items with HOTS features have been outlined by Lembaga Peperiksaan (2013). By the way, the common issues in testing are remain lingering around the process of item development especially in standardised testing. The ethical issues among the item developers can usually haunt the stakeholders regarding the use of language testing and its consequences result due to the test interpretation (Bachman, 2000). Hence, ethical issues involving item developers are given attention as it linked to the validity and reliability of the test. It was observed at the beginning of item development process until its transparent report (Bachman, 2000; Prapphal, 2008). This is consistent with the view of Wolf, Farnsworth, & Herman (2008) and Stobart (2001) who have emphasised that the item developers and the state education department should be responsible towards the compliance aspect and the validity of the carried out tests, and took appropriate actions, as a result of the test interpretation.

There is no doubt that the demand for high technical quality of the test items is high and tests should meet the intended statistical figures (Miller, Linn, & Grondlund, 2009). Consequently, the use of standardised testing, which is administered centrally has been disputed among educators related to the quality of the items used in this kind of test where the psychometric features of the test seldom be the basis of consideration (Martone & Sireci, 2009). Due to the demand of items with psychometric features, the researchers in education field have resorted to modern measurement models over classical due to its limitations (Hambleton & Jones, 1993). The application of Item Response Theory

(IRT) such Rasch measurement model in few studies from abroad in language testing have been carried out for validation purposes ((Baghaei & Amrahi, 2011; Choe, 2010; Lee-ellis, 2009). The studies have come into an agreement that this modern theory is worth enough to be applied in testing even the assumption of unidimensionality is quiet hard to be fulfil as stated by McNamara (1996).

Analysis of UPSR results highlighted imbalanced allocation of items according to cognitive levels, whereby 70% of the items were at the knowledge level (Kementerian Pendidikan Malaysia, 2013). It was uncalled for as it is highly expected that items for such standardised national test have conformed to the guidelines in item writing and have gone through the crucial steps in test development process (Lembaga Peperiksaan, 2013). However, the analysis of test items and documents related to the public examination are strictly confidential, and the access to these documents is strictly restricted.

In English testing, the validity issues that include the ethical issues in the test development process, nature of the test items and content validity have been highlighted in previous studies (Bachman, 2000; Martone & Sireci, 2009; McNamara, 1996; Wiliam, 2010). No denial that the process of item development requires double efforts, and contribution of great ideas to ensure that the built items have good psychometric features. There is a doubt whether the appointed item developers have gone through the test development process ethically (Bachman, 2000; Prapphal, 2008) based on the standards outlined by the Examination Syndicate (Lembaga Peperiksaan, 2013). As the demand for high technical quality of the test items with intended statistical figures is high (Miller, et al., 2009), the omission of any step in these guidelines is a great threat to test validity as it might affect the quality of the test items (McNamara, 1996). In addition, the development of multiple-choice item is quite challenging as it needs plenty of time and efforts (Hughes, 2008), especially in writing and selecting effective and plausible distractors (Stewart, 2014). Since the experience, knowledge and skill are not gifted to be good item developers, hands-on training for them is necessary (Chen, 2011; Downing, 2009). Hence, no matter how good are the people in the testing field, the quality of test items can still be questioned in terms of validity and reliability of the test items (Reich, 2013).

In the Malaysian education scenario, pupils need to sit for trial examination a few months before the actual examination takes place. Besides preparing the pupils for actual examination, this trial examination is believed to be the best predictor of actual performance in national examination. However, it is not always true. In the UPSR 2014 for instance, the results of the UPSR in School A was 66%, which was lower than the results of the actual examination (74%). Undoubtedly, the discrepancy in both trial and actual examination results indicated the lack of predictive validity element. The quality of EP1 items of the 2014 UPSR trial examination that was administered under State Education Department (SED) was questioned by English teachers, as the test specification table was not provided. Hence, teachers were not able to examine whether the intended difficulty level of the test items was based on the desired cognitive domain. It should be noted that the papers in this trial examination were set under the accountability of School Heads Council. The items were developed by a panel of selected experienced English teachers. Since the content validity of the test was unknown and was not accessible, item analysis should be conducted to provide empirical evidences to meet the demands of construct validity. Downing (2009) has stated that the quality of the test items are unknown until they have gone through try-outs and pilot testing where it can be proven by interpretation of statistical figures of the chosen measurement model. Thus, there is a need to conduct this study so that the empirical evidence of the quality items used in this test can be determined from the aspects of validity and reliability.

In Malaysia, few item analysis studies found the application of modern method for multiple-choice items in trial examination papers at primary and secondary level for various subjects such as Mathematics, Science and Islamic Studies (Kirfee, 2012). However, most of them were not published and could not be accessed by the public. To date, there is no application of modern measurement model for standardised achievement test at primary level in the country, which focuses on English Language. The only latest unpublished study found on language testing was by Rusilah Yusup (2012) who did item analysis using the Rasch model on Malaysian University English Test (MUET) for reading test at tertiary level. Thus, the application of modern testing theory using Rasch model in language testing at primary level, particularly in investigating the quality of EP1 items seems to be promising.

The purpose of this study was to determine the quality of the items in the UPSR Trial by providing empirical evidence of construct validity using the Rasch measurement model. Specifically, the study was intended to:

i) examine the extent to which the items in the test demonstrates evidence of validity,
ii) examine the extent to which the test demonstrates the evidence of reliability.
iii) evaluate the distribution patterns of items difficulty in relation to pupils' ability based on the item-person map.
iv) determine the difficulty levels of the items in the test based on the identified constructs.

METHODS

The population was made up of 37 national primary schools in Kuala Selangor district. A sample of 525 or 14% of the 3876 UPSR candidates was selected using the proportional stratified random sampling method. The instrument used in this study is English Paper 1 (EP1) of 2014 UPSR trial examination which was administered under the responsibility of SED of Selangor. In identifying the psychometric properties of EP1, the data of 40 dichotomous items from 525 UPSR candidates were analysed using WINSTEPS 3.68.2 software. The software of IBM SPSS version 21.0 was used to key in the data of 21000 responses together with the candidate's code. The diagnosis of test validity is determined by Principle Component Analysis (PCA), fit statistics (PTMEA Corr, MNSQ and Zstd) while test reliability is shown by item-person reliability and separation indices. The difficulty of EP1 items are displayed in item-person map that visualised the pattern of items' distribution on the same vertical scale of pupils' ability.

FINDINGS AND DISCUSSION

Findings

The validity of EP1 items is addressed in the first research question in this study. The validity evidences are evaluated based on PCA, fit statistics and item distractor analyses. PCA is one of the diagnosis by Rasch model to ensure that all items share the same dimension which capable to sense the ability of the instrument in measuring a uniformity of single dimensions with acceptable noise levels (Linacre, 2012). The analysis of PCA is presented in Table 1.

Table 1. Dimensionality Map of EPI (in Eigenvalue Units)

	Eigenvalue	Empirical	Model
Total raw variance in observation	55.4	100.0%	100.0%
Raw variance explained by measure	15.4	27.8%	24.3%
Raw variance explained by persons	5.3	9.6%	8.4%
Raw variance explained by items	10.1	18.2%	15.9%
Total raw unexplained variance	40.0	72.2%	75.7%
Unexplained variance in 1st contrast	2.9	5.3%	

Table 2. Dimensionality Map by Construct (in Eigenvalue Units)

		Eigenvalue	Empirical	Model
A.	Vocabulary – Raw variance explained by measure	14.3	59.7%	59.8%
B.	Language and social expression – Raw variance explained by measure	19.8	79.8%	87.2%
C.	Grammar – Raw variance explained by measure	6.8	40.6%	41.3%
D.	Text completion – Raw variance explained by measure	3.6	41.7%	44.6%
E.	Reading & comprehension – Raw variance explained by measure	3.6	41.7%	44.6%

PCA analysis (Table 1) shows the 27.8% of raw variance was explained by the measurement model which exceeding the expected which is 24.3%. Nevertheless, 27.8% of the variance in the data based on the dimensions of Rasch measurement model is considered weak according to Rating Scale Instrument Quality Criteria by Fisher (2007). According to Rasch, in Azrilah, Saidfudin, & Azami (2013), the requirement of at least 20% instrument uniformity has been achieved, but the 40% minimum requirement of Rasch measurement model has not been met.

The raw variance explained by person of 9.6% shows that there is less variance in person ability as compared to 18.2% of item difficulty. This is due to the smaller value of standard deviation for persons (0.87) compared to the standard deviation for item (1.12). Unexplained variance the 1st contrast was 5%, with eigenvalue 2.9 (<3.0), indicating the absence of second dimension and the test is probably unidimensional (Linacre, 2012).

Analyses by construct however, have yielded greater values of variance explained by measure, which all were above the 40% of the Rasch requirement (Table 2). The values were 59.8%, 79.8%, 40.6%, and 41.7% respectively for vocabulary, language and social expression, grammar, text completion, and reading and comprehension.

Fit statistics is a summary of the discrepancies between what is observed with what is expected is intended to identify the misfitting items as predicted by the model. Item fit indices are examined through infit-outfit of mean square (MNSQ) and standardised form (Zstd) while diagnosis of item polarity through PTMEA Corr analysis is also discussed in investigating the linkage of EP1 items as a part of content and construct validity specifically. Table 3 shows the summary of seven misfit items in EP1.These items are considered misfit as their values are beyond the range of productive measurement which is between 0.7 and 1.3 for infit – outfit MNSQ and -2 to and +2 for z-std as proposed by (Bond & Fox, 2012). Items CE38, CC19, CC20, CA3, CC23 and CA6 are considered as underfit items as outfit MNSQ > 1.3 and z-std or t > 2.0 whereas CB15 is considered overfit where the outfit MNSQ index is < 0.7 and t < -2.

Table 3. Item Statistics of EPI Items: Misfit Order

Entry No	Total Score	Measure	Model S.E	Infit		Outfit		PT Measure	
				MNSQ	Zstd	MNSQ	Zstd	Corr.	Exp.
38	51	2.78	.15	1.19	1.6	2.42	6.0	-.12	.23
19	98	2.96	.12	1.24	3.4	2.03	7.4	-0.8	.30
20	129	1.57	.11	1.13	2.3	1.35	3.8	.15	.33
3	154	1.29	.10	1.20	4.2	1.46	5.8	.09	.35
23	154	-.34	.10	1.19	4.0	1.31	4.1	.12	.35
6	330	-1.35	.10	1.31	7.5	1.55	9.1	.00	.37
15	422		.12	.84	-.2.6	.69	-3.4	.50	.31

Chi-Square: 22477,71; d.f: 20436; p=.000

Misfit order statistics (Table 3) has identified seven misfit items, CE38, CC19, CC20, CC23, CA6, CA3 and CB15. As one of the misfit items, CC38 item is considered underfit even the responses pattern meet the criterion as the most difficult item in EP1 with +2.78 logit. The infit-outfit MNSQ values for this item is between the range of 1.19 ~ 2.14 which do not fit with MNSQ ideal value = 1. As the infit MNSQ > 1, there is a possibility that pupils with high ability did not succeed on this item due to carelessness. However, MNSQ values between 0.7 to 1.3 are acceptable for dichotomous items (Bond & Fox, 2012). This is reinforced by the value of outfit zstd, where *t* = +6.0 which is quite high and negative value of PTMEA Corr = -0.12. Other underfit items from grammar and vocabulary constructs are also behaving just as item CC38, based on the MNSQ and *t* values shown.

Item polarity is determined through PTMEA Corr that shows the direction or orientation of pupils' responses towards latent variable. Positive item polarity is gained if responses to the item are positively correlated with the latent variable. The analysis identified 37 items with positive polarity between the range of 0.09 – 0.61 which means high ability pupils succeeded on difficult items or low ability pupils succeeded on easy items. The other two items; CE38 and CC19 showed negative polarity, while item CA6 with a value of 0.00. Since CE38 is the most difficult item in EP1, further investigation based on other analysis will be done. Negative polarity indicates a weak correlation between the items in these three constructs; reading and comprehension, grammar and vocabulary with the latent variable or might be due to extremely tough item, miskeyed option or data entry errors (Linacre, 2012).

Distractor analysis has identified five items; CE38, CC19, CC24, CA10 and CA6 that have been marked with (*) in Table 4. According to Linacre (2012), the acceptance of items with distractors problem should meet these conditions. Items that have good fit values and the average measure of incorrect options are smaller than the average measure of correct options can be accepted and kept for further use. Nevertheless, the items that are misfitting and the average measure of the incorrect options are greater than the average measure of the correct option must be checked or removed.

In brief, CE38, CC19 and CA6 are also misfit items that do not fit the Rasch measurement model based the values of MNSQ, which are out of productive measurement. Therefore, the rejection will be directed towards the misfit items, no matter how the values of average measure were derived from the analysis. However, based on the above discussion, item CA10 that has good fit values needs to be excluded, as the average values of incorrect options are greater than the correct option. The only item which can be kept from five problematic items is CC24 as it has a good fit value and the average measure of each options is as desired.

Table 4. Summary of Item Distractor Analysis

Item	Data Code	Score Value	Data Count	Average Measure	S.E Mean	Outfit MNSQ	PTMEA Corr.
CE38	4	0	58	.00	.12	.9	-.11
	2	0	229	.18	.06	1.0	-.09
	1	0	187	.56	.06	1.3	.24
	3	1	51	-.06*	.12	2.6	-.12
CC19	1	0	27	-.56	.13	.4	-.22
	3	0	164	.12	.06	.9	-.12
	4	0	236	.53	.05	1.4	.27
	2	1	98	.13*	.10	2.2	-.08
CA10	2	0	98	-.33	.07	.7	-.33
	1	0	69	.04	.11	1.2	-.11
	4	0	109	.58	.08	1.8	.18
	.	0	1	.78	-	1.6	.03
	3	1	248	.44*	.05	1.2	.18
CC24	1	0	28	-.49	.15	.6	-.21
	2	0	133	.01	.07	1.0	-.17
	3	0	149	.24	.07	1.3	-.03
	.	0	1	.92	-	1.8	.03
	4	1	214	.56*	.05	1.1	.27
CA6	1	0	34	-.79	.13	.6	-.32
	4	0	14	-.68	.19	.6	-.18
	3	0	147	.61	.07	2.1	.24
	2	1	330	.27*	.04	1.2	.00

Table 5. Summary Statistics of 525 Persons

		Measure	Infit MNSQ	Zstd	Outfit MNSQ	Zstd
Person	Mean	.27	.99	-.1	1.04	.1
	S.D	.87	0.20	1.2	.45	1.3
	Reliability	.80				

Cronbach Alpha (KR-20) Person raw Score Reliability = .82

The reliability of EP1 items are demonstrated based on the second research question where the evidence of reliability was shown through item-person reliability and item-person separation indices. Summary Statistics output provides the analysis result for person-item reliability and person-item separation indices as displayed in the following tables. Table 5 shows a summary of the statistics on persons where the person separation for 525 pupils indicates two strata (2.09) of pupils' abilities as measured in the constructs of EP1. Linacre (2012) has proposed the minimum value for individual strata is two but if the separation > 2, the test items have performed better in distinguishing the strata of pupils' ability. The value of person separation for this test is accepted. The person reliability index, which is 0.82, is equivalent to the interpretation of Cronbach's Alpha in CTT. The high value signifies the wide range of person measure or EP1 contains adequate number of items (Bond & Fox, 2012).

The indices of items reliability and separation in Table 6 show that the order of items difficulty is consistent and reproducible with other sample. Rasch measurement model (1PL) assumes the same discrimination of 1.00 for all items. Item separation of 10.15 indicates that EP1 test items have been separated into 10 levels of difficulty or 10 groups of item strata at 2 S.E. The indices indicate that these items have been distributed well in which the location of the items on the logit scale has high reliability. The index of items separation which is > 5 and items reliability of 0.99 are considered excellent (Linacre, 2012). To conclude the summary statistics, high item reliability of EP1 shows the wide range of items and adequate items and sample were used in the test.

In Figure 1, 525 pupils and 40 EP1 items are located on the map based on the estimation of their ability and item difficulty on a single measurement continuum ranging from the easiest to the most difficult and the highest to the lowest ability.

Table 6. Summary Statistics of 40 Measured Items

		Measure	Infit MNSQ	Zstd	Outfit MNSQ	Zstd
Item	Mean	0.00	.99	-.13	1.04	.0
	S.D	1.12	.13	3.1	.34	3.6
	Reliability	0.99				
	Separation	10.15				

Chi-Square: 22477.71 with 20436 d.f. p= .0000

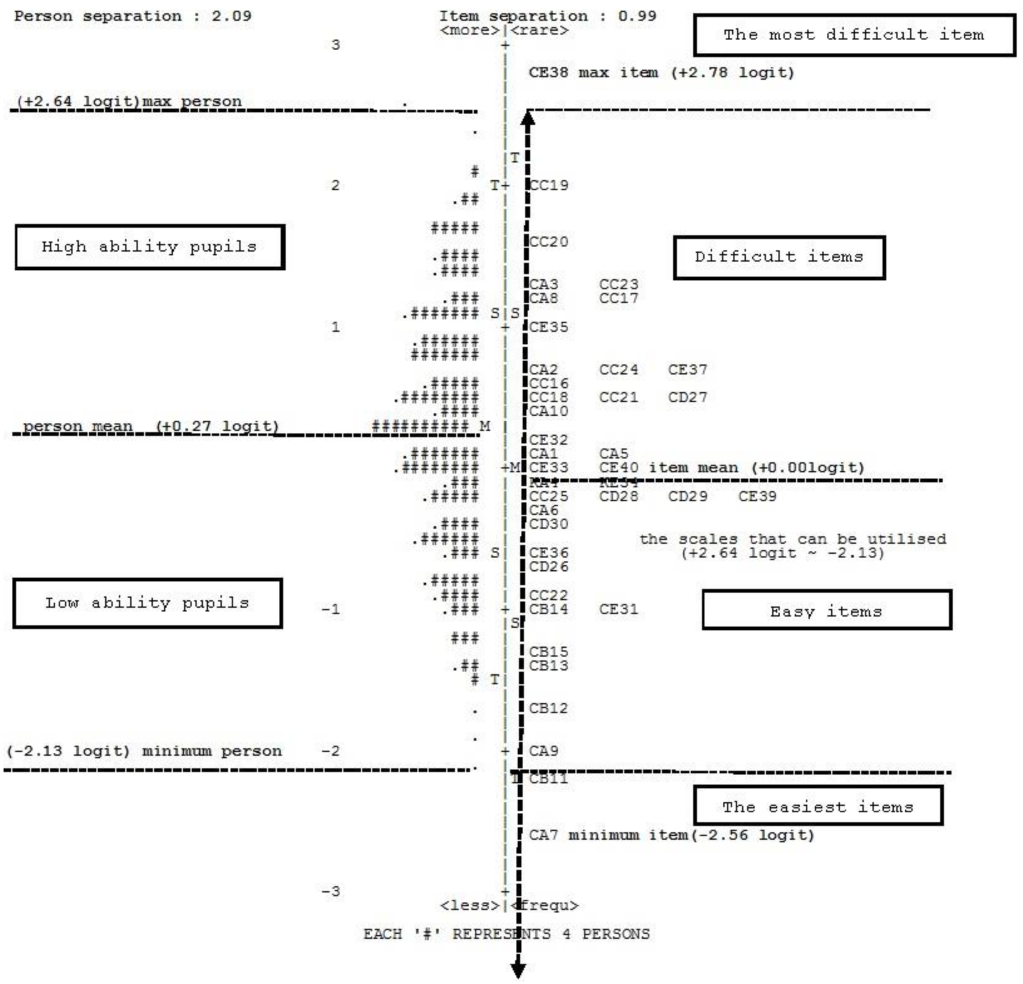

Figure 1. Item-person Map

This map clearly illustrates the distribution of the test items, which was dispersed in the range of 2.78 to 2.56 logit. The development of ruler scale is really helpful to identify the levels of item difficulty. EP1 items have been categorised into four categories after the formation of item-person ruler scale; 1A: the easiest items, 1B: easy items, 2A: difficult items and 2B: the most difficult items. Though, these 40 multiple-choice items only measure two persons separation which are high and low ability pupils ranging from -2.13 to 2.64 logit with persons' mean = 00:27 logit.

Based on the ruler scale illustrated in item-person map, the difficulty level of items in correspond with pupils' ability can be identified easily. Table 7 shows the analysis summary of difficulty levels of items by constructs. The analysis of measure order shows that the items are almost distributed equally. It is based on the derived difficulty levels, which are in the range of -2.56 to 2.78 logit.

This means that the EP1 items spread in a wide band of 5.34 logits. For Construct A – vocabulary, it ranges between -2.56 to 1.29. It is between the range of -2.20 to -1.04 for Construct B – language and social expression. For Construct C – Grammar, it ranges between -0.92 to 1.96. It ranges between 0.46 to 0.75 and -.98 to 3.76 for Construct D – cloze-comprehension and Construct E – reading and comprehension respectively. It indicated that the widest band of item difficulties distribution was for Construct A – vocabulary (3.85 logits).

Table 8 summarises the problematic items, which have been mentioned a few times to answer the research questions of this study. It will ease us to seek their correlatedness in making fair judgement towards these items. The worst item with major statistical problems is CE38. Instead of possessing negative value of PTMEA Corr, this misfit item also has distractor problem. As the most difficult item in the test, the logit measure is above the maximum person where only a few pupils capable to succeed on this item by chance. CC19 item also has the same problematic features as CE38 excluding the logit measure of 1.96 which is still within the range of high ability pupils (0.27 ~ 2.64 logit).

Table 7. Items Difficulty Level by Constructs

Category of item difficulty level	Constructs of EPI					No of item
	CA	CB	CC	CD	CE	
The most difficult item (+2.78 logit)					38	
No of item/construct	-	-	-	-	1	1
Difficult items (+1.96 ~ + 0.01 logit)	3 , 8 2 , 10 1 , 5	-	19 , 20 23 , 17 24 , 16 18, 21	27	35 , 37 32 , 33 40	
No of item/construct	6		8	1	5	20
Item mean: 0.00 logit						
Easy items (-0.04 ~ -2.01 logit)	4 , 6 9	14 , 15 13 , 12	25 , 22	26 , 28 29 , 30	34 , 39 36 , 31	
No of item/construct	3	4	2	4	4	17
The easiest items (-2.20 ~ -2.56 logit)	7	11	-	-	-	
No of item/construct	1	1	-	-	-	2
Sum Total	10	5	10	5	10	40

Note: CA – Vocabulary, CB = Language and Social Expression, CC = Grammar,
CD = Cloze-comprehension and CE = Reading and Comprehension

Table 8. Problematic Items in EPI

Item	PTMEA Corr	Misfit		Item Distractor	Item-person map		Remarks
	-ve / 0.00	Underfit (>1.3)	Overfit (<0.7)		(>2.64)	(<-2.13)	
CE38	/	/		/	/		////
CC19	/	/		/			///
CC20		/					/
CC23		/					/
CA6	/	/		/			///
CA3		/					/
CB15			/				/
CA10				/			/
CC24				/			/
CA7							/
CB11							/

Note: '/' indicates type of problem in EPI

Similarly, CA6 item, which is underfit, and having problem with distractor, the zero index of PTMEA Corr also needs further discussion. Another four misfit items; CC20, CC23 and CA3 are considered underfit items while CB14 is overfit. As the only overfit item in this study, CB15 item has no problem regarding other statistical issues, which is similar to CA10 and CC24 where both are having problem with distractors only. Nevertheless, two fit items; CA7 and CB11 are also considered as problematic items since they have been identified as the easiest items further down the logit of minimum person.

Discussion

The major finding of this study is the issue of unidimensionality in EP1. As achievement test in language testing usually encompasses of various skills, knowledge, processes, and different strategies, the measurement of unidimensionality does not require the performance of items as a psychological process. McNamara (1996) has emphasised that the items just need to work in unison to form a latent pattern in the data matrix. Although scholars debated that this assumption is not very appropriate for language testing data, still, there is an agreement by some researchers to this assumption. Unidimensional assumption has been met by several studies on language proficiency test like "Korean C-Test" and "Vocabulary Size Test" (Beglar, 2010; Choe, 2010; Lee-ellis, 2009). The failure of unidimensional diagnosis to reach 40% requirement of measurement model might be due to the misfit items of CE38 and CC19 which did not meet the Guttman pattern that emphasises on "success on all easy items and failure on all difficult items" (Linacre, 2012). On the other hand, the result was contradicted the Guttman scalogram of responses as discussed earlier. It can be seen that more pupils with low ability succeeded on the most difficult items compared to pupils with high ability. For this particular study however, unidimensionality was not really an issue as the variance explained by measure increased greatly (40.6% -79.8%) when separate analyses were performed by construct, indicating that the constructs were actually quite different from one another.

Six items in test were diagnosed as misfit-underfit with the outfit MNSQ values > 1.3 and one of the item was misfit-overfit where the value is < 0.7. Similarly, the values of zstd for misfit items, which are beyond the scale specify that the responses made by pupils are unexpected by the model. The *t* value >

+2.0 indicates that there are data which are beyond the expectation while t < -2.0 signifies that the data are too easy to be expected or there might be other dimensions that limit the pupils' response patterns. There are three problematic items; CE38 and CC19 that have negative PTMEA Corr while one item; CA6 with zero index. This indicates that these three items have low correlation indices in reading and comprehension, grammar and vocabulary constructs, and cannot distinguish different types of ability among pupils (Linacre, 2012). Besides, these three items will not contributed to the measurement of respondents (Rahayah, Omar & Sharif, 2010). Since EPI items could provide empirical evidence, there were no miskeyed options or data entry errors detected as suggested. The item might be too tough for the pupils to comprehend their understanding and the stem of item should be precise and not ambiguous (Hammouri & Sabah, 2010).

The high values of MNSQ > 1.3 logit found in this test signify that there are unexpected responses possibly due to poorly developed items such as the use of vague words, more than one answer for that item and ambiguous stem of item (McNamara, 1996). Any abandonment steps in the item development process by item developers are the starting point to the occurrence of misfit items. In addition, it is also an indication that misfit items measure different constructs and caused by chance or randomness (Pae, 2012). These scenarios lead to a negative value of item polarity, which give negative impact to the quality of the item. An overview of the extent to which a pattern of responses meets the normal expectation can be seen through the value of zstd as well as helping the researcher to investigate the unforeseen and unexpected pattern of responses (Azrilah et al., 2013).

In testing and measurement field, any misfit items found in the test are due to the unforeseen and haphazard pattern of responses made by pupils. Item CE38, CC19, CC20, CA3, CC23 and CA6 are misfit (underfit) as the logit measures are > 1.30 logit. The *t* value for each item is also > +2. These underfit items indicate that the items were not properly built or else they measure different construct in the same set of construct (Bond & Fox, 2012). In other words, these items are not homogeneous with other items in a measured scale and should be reviewed to identify the possible causes of the occurrence of misfit item. High value of MNSQ suggested that the items were too erratic and had high possibility to falsify the data and vitiate the measurement system (Pae, 2012). Besides, this scene has created a bigger treat to validity of the test items as the pupils' responses were afar of item developers' expectation (Hammouri & Sabah, 2010).

The only overfit item found in this study is item CB15 with the outfit MNSQ value < 0.7 and *t* value < -2. It shows that the pupils' responses towards this item are foreseeable due to the small variations as expected by the model. Yet, it does not provide a lot of information related to the ability of pupils. It would not harm the measurement interpretation as it only signifies that the pupils who have higher abilities than the level of items difficulty are more likely to respond correctly more often than the expectation of Rasch measurement model (Athanasou & Lamprianou, 2002). On the other hand, bear in mind that overfit item is also considered redundant or dependent on other items in the test (McNamara, 1996). Thus, it violates one of the guidelines in item development process and leads to low quality of item, which gives significant impact for test validity.

In fact, the misfit items might be influenced by problematic item distractors that affect the quality of the test items. Between, it provides information related to the real root cause of the weak item by reflecting the existing teaching and learning process among teachers (Hammouri & Sabah, 2010). In addition, Koizumi, Sakai, Ido, Ota, Hayama, Sato &

Nemoto (2011) have stressed out the importance of good and quality distractors in providing information for error pattern profiles development. Since both stem and distractor constitute an item in the test, it may reflect of how the test items are constructed. Therefore, it involves the responsibility of the entire item developers to get involved from the planning stage until after the test administration. Hughes (2008) has identified some flaws distractor for items that are problematic whereas there is more than one correct answer options, there are no right answers, there are clues to the correct answer option and the option of ineffective responses. Effective distractors should be able to attract a good number of pupils to select but if there is no choice or very little choice made, the selection of distractors for these items are useless (Anderson, Clapham & Wall, 1995). The distractors that do not function well are not able to differentiate between pupils of high and low ability pupils. This option can be removed or fixed to improve the quality of items where it can reduce the time allotted to answer this item and the length of the repeated reading as well as setting aside the potential sources of confusion towards the item (Rodriguez, Kettler, & Elliott, 2014).

The test consists of 40 multiple-choice items, which represents five substantive constructs in language testing, shows good and acceptable indices of reliability of items and individual as well. The values > 0.80 indicate that the items used in the test are consistent. This means that if the items are reproducible to other group of pupils who have the same ability, the probability location of the items on the scale of measurement is high. This signifies that the items in EP1 are credible and measure what should be measured.

The mapping of item difficulty – pupils' ability revealed normal distribution where the distribution pattern of items difficulty and persons' ability are well scattered in the range of 2.78 to 2.56 logit and 2.64 to 2.13 logit respectively. Majority of pupils' measure were in the range of item difficulty though the mean ability of the pupils of 0.27 logit, which was slightly higher than the item mean of 0.00 logit. Above all, with the logit of 2.78, item CE38 is also an item that exceeds the maximum level of 2.64 logit on pupils' ability. It indicates that the level of items difficulty is higher than pupils' ability. Indeed, as emphasised by previous scholars and studies in testing fields the misfit value indicated by this item, shows the weaknesses of item development process. By focusing to the issues in testing in previous chapter, the quality of item writer in developing test items and lack of assessment practice among teachers have turned to be a great threat to validity (Asim, 2013; Downing, 2009) or else the item is measuring different constructs.

The location of CB11 and CA7 with the logit values of -2.56 and -2.20 for each item at the bottom end of the map, are below the minimum level of item difficulty of -2.13 logit. Since both are items are not measuring any pupils' ability, they should not be on the test and need to be discarded or revised for future use. However, a study on language proficiency by Lee-ellis (2009) has suggested to retain these kind of items due to the objective of the test which is developed to assess the broad range of pupils' ability. Besides, the location of redundant items are clearly visualised in the map indicating that they have the same logit measure in the same construct.

EP1 items were divided into 4 levels of difficulty; the most difficult item, difficult item, easy item and the easiest item whereas the persons' ability falls into two categories of high and low ability. The division of 52% and 44% of pupils in these categories are visualised in the map while another 4% might be scattered beyond the map. The difficult items fall in the range of 1.95 to 0.01 logit. There are 20 items from the constructs of vocabulary, grammar, cloze-comprehension, and reading and comprehension have been identified. Since there are 80% of items measuring grammar construct fall in this category, it indicates that the pupils have difficulty to succeed on these kind of items. It implies low acquisition of grammar knowledge among pupils in Kuala Selangor district. Without a doubt, the logit measures of pupils' ability below the item mean are expected not capable to succeed on these items and vice versa. Similarly, to the acquisition of vocabulary knowledge and, reading and comprehension among pupils that can be considered as weak due to the number of items identified in this category.

All 17 easy items from all constructs in EP1 fall in the range of -0.04 to -2.01 logit. Obviously, 80% items from the constructs of language and social expression, and cloze-comprehension fall into this category. The last difficulty levels, which are the easiest items, are shown by CA7 and CB11 in vocabulary and, language and social expression construct respectively. With the range of -2.20 to -2.56 logit, both items are below the minimum level of pupils' ability.

IMPLICATIONS AND RECOMMENDATIONS

The implications of the study might benefit the testing and measurement field by providing useful insight for teachers who directly involved in educational assessment. Measuring the quality of the items used in standardised test, such as UPSR trial examination is necessary so that better prediction can be made for the upcoming real test. The strengths and weaknesses of constructs acquisition by pupils can be identified through this kind of analysis. For sure, there might be a bit gap between the expected and the observed result but at least final preparation can be done for those pupils.

The innovations in measurement methods from classical to modern measurements provide an opportunity for teachers to challenge their existing skills in analysing the test items. Besides that, they need to equip themselves to cope with teaching and learning of 21st century classroom. The findings in this study have proven, that the application of Rasch measurement model (1PL) is capable to provide meaningful descriptive output which can be interpreted in details as well as providing psychometric information for each item in the constructs of language testing.

The study has identified the quality of test items in EP1 based on the shown values using the Rasch measurement model. The misfit items, which are beyond or below the measurement range, and problematic items have also been identified. The consideration whether to keep, improve or remove those items, depends on the item developers. In this study, no such items will be discarded, as EP1 is a standardised test set by the state. On the other hand, the gained information may help us to write better items for summative or formative test in school. Investigating those problematic items, leads us to the source of occurrence of misfit items which due to the weaknesses of item development process as well as unexpected responses made by pupils.

The visualisation of item-person map eases the interpretation of the study whether the test items have been well dispersed which corresponded to pupils' ability. The items that meet the model's expectations will be stored in the item bank. The selection of the items is calibrated thoroughly via item development process for future use is something that is truly worthwhile for teachers despite the effort, time and expertise needed.

Literacy assessment practices among teachers should be aligned with the expected skills needed in developing good quality of the test items. Besides measuring the pupils' ability according to the intended cognitive domains accurately, the enhancement of skills among teachers may happen simultaneously. The teachers themselves have to be very conversant in their field of expert, so that the developed items measure the intended constructs as outlined in the specification table. In addition, the skills in item analysis skills tend to be very crucial since the teachers' readiness to practice is getting low due to lack of exposure to use it.

Even most teachers notice the significance of item analysis in item development process but they are incapable to proceed, as it is quite time consuming and more analytical skills needed. Their skills might limit the classical measurement method rather than modern method such as IRT as the Examination Syndicate itself does not emphasise on this matter. The test specification table might be referred as a part of content validity but the need of statistical analysis is given the highest priority in any testing form in providing empirical evidence to identify the quality of the test items for validation purpose.

Limitations of classical method have resorted most of the researchers to modern measurement, which is more robust in providing better analysis output from psychometrics' aspects. Therefore, the pupils' ability and item difficulty should be given lots of attention in developing test items to avoid inaccurate evaluation of assessment that likely to happen if the pupils were tested with unverified items. Based on the findings, the use of modern testing theory could give a very spacious opportunity for teachers to explore the skills of item analysis, which is very vital in the field of testing. Thus, several recommendations are listed as follow:

i. Proceed with quantitative analysis of dichotomous items by comparing the validity and reliability of two standardised tests for trial examination from any states that is ranked the best in public examinations achievement.
ii. Conducting a study of item analysis for dichotomous items to identify Differential Item Functioning (DIF) items in the standardised test prepared by state educational department.
iii. Conducting a qualitative research for partial-credit items to identify the degree of inter-rater reliability among the examiners in the standardised test.
iv. Implement further study inclusive of three Parameter Logistics; item difficulty (1PL), item discrimination (2PL) and guessing (3PL) using modern measurement model which is Item Response Theory (IRT)

CONCLUSION

Based on overall findings and discussions, there is no doubt that the validity of any forms of test is an issue that cannot be compromised. The development of the test items at school level, district, state or national level cannot be underestimated as it represents validity evidence of the test. The failure or omission of not performing any of these procedures does not only affect the pupils on short-term basis but also to national education assessment system in the long run as well. It should be noted that the validity of a test does not only focus on content validity but also inclusive of direct and indirect consequences of the use of the test scores.

This study has successfully concluded the discussion of findings using the Rasch measurement model (1PL) to identify the extent of the validity of test items of EP1 through diagnosis of unidimensionality, item fit and distractor analysis. Criticism and the weaknesses of the problematic items have been highlighted in details. The reliability indices indicate that the test items of EP1 are reproducible to any samples who have the same abilities. The distribution pattern of item difficulty and pupils' ability shows a good matching of both although there are few items beyond the range of pupils' ability. The identified categories of item difficulty levels based on the assessed constructs may facilitate the researcher to enlighten the stakeholders to proceed with improvements in item development process. The constructs, which is quite tough for the pupils to succeed, also have been identified so that better planning can be made in teaching and learning strategies.

To conclude, the use of modern testing theory with Rasch measurement model application is helpful in overcoming the arguments of validity in language test especially in achievement test for primary level. For further research, the utilisation of qualitative or quantitative method is recommended for dichotomous and partial-credit data.

REFERENCES

Anderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Asim, A. E. (2013). A diagnostic study of pre-service teachers' competency in multiple-choice item development. Research in Education, *89*(1), 13–23. doi: 10.7227/RIE.89.1.2

Athanasou, J. A., & Lamprianou, I. (2002). *A teacher's guide to assessment*. Sydney: Social Science Press.

Azrilah, A. A., Saidfudin, M. M., & Azami, Z. (2013). *Asas model pengukuran rasch: Pembentukan skala dan struktur pengukuran*. Bangi: UKM Holdings Sdn. Bhd.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing*, *17*(1), 1–42. doi: 10.1177/026553220001700101.

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice english vocabulary test with the rasch model. *Journal of Language Teaching and Research*, *2*(5), 1052–1060. doi: 10.4304/jltr.2.5.1052-1060.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101-118. doi:10.1177/0265532209340194

Bond, T. G., & Fox, C. M. (2012). *Applying the Rasch Model: Fundamental Measurement in the human Sciences* (Second Ed.). New York: Routledge.

Chen, J. (2011). Language assessment : Its development and future-an interview with Lyle F. Bachman. *Language Assessment Quarterly, 8*(3), 277-290. doi: 10.1080/15434303.2011.591464

Choe, M. (2010). A Rasch analysis of a nationwide english placement test in Korea. *English Teaching*, *65*(2), 3-23. 10.15858/engtea.65.2.201006.3

Downing, S. M. (2009). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–24). New York: Routledge.

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transaction, 21*(1), 1095.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues*, *12*(3), 38–47. doi: 10.1111/j.1745-3992.1993.tb00543.x

Hammouri, H., & Sabah, S. A. (2010). Analysis and assessment of the Jordan National Test for Controlling the Quality of Science Instruction (NTCQSI): A Rasch measurement perspective. *Educational Research and Evaluation: An International Journal on Theory and Practice,* *16*(6), 451-470. doi: 10.1080/09243453.2010.550469

Hughes, A. (2008). *Testing for language teachers*. Cambridge: Cambridge University Press.

Kementerian Pendidikan Malaysia. (2013). *Malaysia Education Blueprint 2013 - 2025*: *(Preschool to Post-Secondary Education)*

Kirfee, H. B. (2012). *Analisis item aneka pilihan peperiksaan percubaan SPM Pendidikan Islam (1223/1) berdasarkan Model Rasch*. University Malaya.

Koizumi, R., Sakai, H., Ido, T., Ota, H., Hayama, M., Sato, M., & Nemoto, A. (2011). Development and validation of a diagnostic grammar test for Japanese learners of English. *Language Assessment Quarterly, 8(1), 53–72.* doi: 10.1080/15434303.2010.536868

Lee-ellis, S. (2009). The development and validation of a Korean C-Test using Rasch analysis. *Language Testing, 26*(2), 245–274. doi: 10.1177/0265532208101007

Lembaga Peperiksaan. (2013). *Pentaksiran KBAT*. Putrajaya: Kementerian Pendidikan Malaysia.

Linacre, J. M. (2012). *Winsteps help for Rasch analysis*.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*(4), 1332–1361. doi: 10.3102/0034654309341375.

McNamara, T. F. (1996). *Measuring second language performance*. England: Addison Wesley Longman Limited.

Miller, M. D., Linn, R. L., & Grondlund, N. E. (2009). *Measurement and assessment in teaching* (10th Ed.). New Jersey: Pearson Education, Inc.

Pae, H. K. (2012). A psychometric measurement model for adult english language learners : Pearson Test of English Academic. *Educational Research and Evaluation: An International Journal on Theory and Practice, 18*(3), 211-229 doi: 10.1080/13803611.2011.650921.

Prapphal, K. (2008). Issues and trends in language testing and assessment in Thailand. *Language Testing*, *25*(1), 127–143. doi:10.1177/0265532207083748.

Rahayah, S., Omar, B., & Sharif, S. (2010). Validity and reliability multiple intelligent item using rasch measurement model, *9*(0), 729–733. doi: 10.1016/j.sbspro.2010.12.225.

Reich, G. A. (2013). Imperfect models, imperfect conclusions : An exploratory study of multiple-choice tests and historical knowledge. *The Journal of Social Studies Research*, *37*(1), 3–16. doi: 10.1016/j.jssr.2012.12.004

Rodriguez, M. C., Kettler, R. J., & Elliott, S. (2014). Distractor functioning in modified items for test accessibility. *SAGE Open, 4*(4), 1-10. doi: 10.1177/2158244014553586

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST ? Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271-282. doi: 10.1080/15434303.2014.922977

Stobart, G. (2001). The validity of national curriculum assessment. *British Journal of Educational Studies, 49(1), 26–39*. doi: 10.1111/1467-8527.t01-1-00161

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: an illustrative guide: AMEE guide no. 72. *Medical Teacher*, *35*(1), e838–e848. doi: 10.3109/0142159X.2012.737488

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist*, *45*(2), 107–122. doi: 10.1080/00461521003703060.

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing english language learners' language proficiency validity issues in assessing english language learners' language proficiency. *Educational Assessment*, *13*(2-3), 80-107. doi: 10.1080/10627190802394222

Yusup, R. B. (2012). *Item evaluation of the reading test of the Malaysian University English Test (MUET)*. Masters by Coursework & Shorter thesis, Melbourne Graduate School of Education, The University of Melbourne. Retrieved from http://hdl.handle.net/11343/37608