ITEM ANALYSIS OF ENGLISH PAPER 1 (EPI) OF 2014 UPSR TRIAL EXAMINATION USING RASCH MEASUREMENT MODEL

HALIZA BINTI ISHAK

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

DISERTATION SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF EDUCATION (EDUCATIONAL EVALUATION) (MASTER BY MIXED MODE)

FACULTY OF HUMAN DEVELOPMENT
UNIVERSITI PENDIDIKAN SULTAN IDRIS

2017

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

ABSTRACT

This study was aimed to identify the quality of English Paper 1 items of 2014 UPSR trial examination in terms of reliability, validity and items characteristics based on the analysis using Rasch measurement model. 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 by examining the distribution of items in the item-person map. A number of 525 primary school pupils in Kuala Selangor was selected using proportionate stratified random sampling method. The validity evidences are shown through the Principle Component Analysis (PCA), fit statistics and item distractor analysis. Even though the raw variance explained by measure in PCA analysis is rather weak, it achieves the minimum uniformity and it clearly shows the absent of second dimension in the test. The fit statistics analyses have shown seven misfit items that are beyond the acceptable range (0.7 - 1.3 logit) and two of them have negative PTMEA Corr values. Item distractor analysis has identified five problematic items whereby three of them are also misfit items. Summary statistics of items and persons show the reliability indices of Cronbach's Alpha are greater than 0.80 and separation indices greater than 2. Item-person map has demonstrated good distribution pattern where most of the items fall within the range of person ability. It is expected that this study will benefit teachers in improving their assessment practice. This would also help the establishment of the item banks with validated items and enhance teachers' ability in analysing and interpreting items accurately using modern test theory such as Rasch model. This study contributes to the field of educational measurement and evaluation in terms of expanding the use of modern test theory in language testing in Malaysia.

ANALISIS ITEM BAHASA INGGERIS KERTAS 1 (EP1) PEPERIKSAAN PERCUBAAN UPSR 2014 MENGGUNAKAN MODEL PENGUKURAN RASCH

ABSTRAK

Kajian ini bertujuan untuk mengenal pasti kualiti item-item Peperiksaan Percubaan UPSR Bahasa Inggeris Kertas 1 2014 daripada aspek kesahan, kebolehpercayaan dan ciri-ciri item berdasarkan analisis menggunakan model pengukuran Rasch. Kajian ini juga bertujuan untuk mengenal pasti sejauh mana tahap kesukaran 40 item aneka pilihan yang terdiri daripada 5 konstruk; *vocabulary, language and social expression, grammar, cloze-comprehension* dan *reading and comprehension* dengan meneliti *item-person map*. Seramai 525 murid sekolah rendah di Kuala Selangor telah dipilih menggunakan kaedah persampelan berstrata berkadaran. Eviden kesahan ditunjukkan melalui *Principle Component Analysis* (PCA), *fit statistics* dan analisis *item distractor*. Walaupun analisis PCA bagi *raw variance explained by measure* agak lemah, ia masih mencapai keseragaman minimum dan jelas menunjukkan ketidakhadiran dimensi kedua dalam set ujian tersebut. Analisis *fit statistics* menunjukkan tujuh item *misfit* yang berada di luar julat yang boleh diterima (0.7 - 1.3 logit) dan dua daripadanya mempunyai nilai *PTMEA Corr* negatif. Analisis *item distractor* mengenal pasti lima item bermasalah yang mana tiga daripadanya juga merupakan item *misfit. Summary statistics* bagi item dan individu menunjukkan nilai *Cronbach's Alpha* lebih besar daripada 0.80 dan indeks pengasingan lebih besar daripada 2. *Item-person map* juga menunjukkan pola taburan yang baik di mana kebanyakan item berada dalam julat kebolehan individu. Melalui kajian ini, diharapkan para pendidik akan mendapat faedah dalam menambah baik amalan pentaksiran mereka. Hal ini juga dapat membantu penubuhan bank item dengan item yang telah divalidasi, di samping meningkatkan kemahiran guru dalam menganalisis dan menginterpretasi item-item ujian dengan tepat menggunakan teori ujian moden seperti Rasch. Kajian ini menyumbang kepada bidang pengukuran dan penilaian pendidikan dari segi memperluaskan penggunaan teori ujian moden dalam pengujian bahasa di Malaysia.

TABLE OF CONTENT

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

1PL	1Parameter Logistic
2PL	2 Parameter Logistic
3PL	3 Parameter Logistic
CTT	Classical Test Theory
DEO	District Education Office
EP1	English Paper 1
EP2	English Paper 2
EPRD	Education Planning and Research Development
GPS	*Gred Purata Sekolah*
HOTS	High Order Thinking Skills
IRT	Item Response Theory
KBSR	Kurikulum Bersepadu Sekolah Rendah
KPI	Key Performance Indicator
KSSR	Kurikulum Standard Sekolah Rendah
MGB	*Majlis Guru Besar*
MNSQ	Mean-Square
MOE	Ministry of Education
MUET	Malaysian University English Test
NKRA	National Key Results Areas
OBE	Outcome Based Education

PCA	Principle Component Analysis
PISA	Programme for International Pupils Assessment
PMR	*Pencapaian Menengah Rendah*
PP	Pentaksiran Pusat
PPPM	*Pelan Pembangunan Pendidikan Malaysia*
PTMEA Corr	Point Measure Correlation
SBA	School-Based Assessment
SED	State Education Department
SPM	*Sijil Peperiksaan Malaysia*
TIMSS	Trends in Mathematics and Science Study
TOV	Take- off Value
UPSR	*Ujian Pencapaian Sekolah Rendah*
ZSTD	Z-standardised

APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The vision and aspirations of Malaysia Education Blueprint (*Pelan Pembangunan Pendidikan Malaysia, PPPM*) 2013 – 2025, which was launched in 2013, focuses on improving the quality of the education system so that it is at the international quality standards. As Malaysia placed at one third of the lowest group of international assessments of Trends in Mathematics and Science Study (TIMSS) and the Programme for International Pupils Assessment (PISA), the improvement for future performance is necessary. Therefore, Malaysia needs to employ steps for improvements so that it can meet the international standards, specifically in the field of Mathematics, Science and English. This statistic is rather alarming although the pupils' achievement results in national examinations have showed an increase through their School Average Grade

(*Gred Purata Sekolah, GPS*). The yearly pass percentage among Malaysia pupils Primary School Achievement Test (*Ujian Pencapaian Sekolah Rendah, UPSR*), Lower Secondary Assessment (*Penilaian Menengah Rendah, PMR*) and Malaysia Certificate Examination (*Sijil Peperiksaan Malaysia, SPM*) for 2000 -2011 was almost constant each year (*Kementerian Pendidikan Malaysia*, 2013).

Consequently, as one of the important achievement test for primary level, the credibility of the items used for the UPSR examination should be within the international standard should not be questioned by any parties. Nevertheless, the quality of the items used in UPSR was dubious, as shown by the findings reported in the Malaysia Education Blueprint 2013-2025. Based on item analysis conducted for English Paper 1 (EP1) items in the 2010 and 2011 UPSR examination by the Pearson Group, it was reported that the pupils have not been assessed with good quality items. In this regard, there was an imbalanced allocation of items according to cognitive domains, where 70% of the items only measure pupils only at the knowledge level only. It was quite surprising as EP1 is a high stake test for primary school. Since our pupils are examined through these items in public examination, there is a query on how are they going to excel in international assessment.

Thus, to be at the same level with the international standard, pupils need to be exposed to items that really measure their cognitive skills, even at the beginning of schooling session. Without a doubt, the improvement stages of pupils' learning and development are necessary in moulding and producing 21st century pupils who acquire knowledge as well as the skills required. This improvement is in line with the

requirement of international benchmarks, and further step should be done in the assessment framework.

Assessment of learning which focused on formative and summative assessments at school level is one of the best platform to equip pupils with Higher Order Thinking Skill (HOTS) and to prepare them for international assessments. Consequently, items with balanced difficulty, inclusive of the HOTS domains and represents the whole curriculum must be developed (Lembaga Peperiksaan, 2013). In addition, to develop test items with HOTS features, the steps in item development that constitute a good test should not be neglected.

Due to the high awareness among educators in testing and measurement field these days, the hunt for quality items has turned item analysis to be one of the most important components in assessment practice, as part of item development process. Hence, identifying the technical quality of the developed items as an empirical evidence of the test validity is a crucial stage for item developers (Haladyna, 2004). In this regard, high validity values and good reliability indices are among the criteria that constitute a good test, and as a result, in any forms of test, valid and reliable items are capable to accurately measure the ability or knowledge of the tested field (Arshad Abd. Samad, 2010).

1.2 Problem Statement

The mismatch of the current academic achievement in public examination with international assessment result as reported in Malaysia Education Blueprint 2013-2025 has turned to be a vital issue in education assessment (Kementerian Pendidikan Malaysia, 2013) in Malaysia. In this light, the quality of the items used in EP1 for 2011 and 2010 UPSR examination was not to the standard of international benchmark (Kementerian Pendidikan Malaysia, 2013), even they were developed by the highest authority in the Malaysia education assessment system, which is the Examination Syndicate. It is highly expected that the chosen operational items for such as standardised test have conformed to the guidelines in item writing and have gone through the crucial steps in test development process systematically as provided by Pusat Perkembangan Kurikulum (2001) and being revised recently by 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 very restricted.

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 primary school was 66%, which was slightly lower than the result of the actual examination, which was 74%. Undoubtedly, the discrepancy in both trial and actual examination results indicated the lack of predictive

validity element. Due to the trial examination's vital role towards actual achievement, the quality of the items used in the examination should not be compromised.

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 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, Linn, & Grondlund, 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).

The quality of EP1 items of the 2014 UPSR trial examination that was administered under State Education Department (SED) was questionable 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 or is known as *Majlis Guru Besar* (MGB). 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.

The researchers in education field have resorted to modern measurement models over classical due to its limitations (Hambleton & Jones, 1993). 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 (Hafizah Kirfee, 2012; Norsilah Ismail, 2011; Syakima Ilyana Ibrahim, 2012; Yee, 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. The application of modern testing theory using Rasch measurement model seems to be a good move in language testing and measurement field to investigate the quality of EP1 items psychometrically. Hence, there is a rationale to come out with item analysis of EP1 items in UPSR standardised trial examination study at primary level in Selangor.

1.3 Research Objectives

The objectives of this study are based on the statistical analyses of Winsteps® 3.68.2 by Rasch measurement model. Specifically,

(i) to examine the extent to which the EP1 items of 2014 UPSR trial examination demonstrates evidence of validity based on:

a) Principle Component Analysis (PCA)

b) Item Fit Statistics in terms of

- point measure correlation (PTMEA Corr) / item polarity
- mean-square (MNSQ)
- Z-standardised (Zstd)

c) Item Distractor

(ii) to examine the extent to which the EP1 items of 2014 UPSR trial examination demonstrates the evidence of reliability based on:

a) reliability indices

b) separation indices

(iii) to evaluate the distribution patterns of items difficulty to pupils' ability based on item-person map.

(iv) to determine the extent of EP1 items difficulty level of 2014 UPSR trial examination based on the identified constructs in the test.

1.4 Research Questions

Based on the mentioned objectives, the following research questions were formulated for this study.

(i) To what extent does the EP1 items of 2014 UPSR trial examination demonstrate evidence of validity based on:

a) Principle Component Analysis (PCA)?

b) Item Fit Statistics in terms of

- point measure correlation (PTMEA Corr) / item polarity?
- mean-square (MNSQ)?
- Z-standardised (Zstd)?

c) Item Distractor?

(ii) To what extent does the EP1 items of 2014 UPSR trial examination demonstrates the evidence of reliability based on:

a) reliability indices?

b) separation indices?

(iii) How is the distribution pattern of items difficulty to pupils' ability based on item-person map?

(iv) To what extent does the level of EP1 items difficulty of 2014 UPSR trial examination based on the identified constructs in the test?

1.5 Significance of the Study

In brief, this study will significantly affect these three groups; teachers, pupils and schools as well. First, this study will provide guidelines for item developers on how to construct high quality test items that fulfil evidence of validity. Any omission of the important steps in test development process by them might be a threat to the validity. It also provides valuable exposure to new and experienced teachers in determining the quality of the test items using modern measurement model, such as the Rasch 1 Parameter Logistic (1PL). Besides that, this study will assist teachers especially in language field, which lack mathematics background to easily interpret the output derived from the analysis.

The change in current practice of item analysis, from classical to modern measurement method, should be done, not only on dichotomous items, but is also applicable to polytomous items. As an empirical analysis, it facilitates the teachers in the development of high quality items with psychometric features. Only items with fit values will be kept while the misfit items will be revised or discarded.

Furthermore, the establishment of items bank will be a reality. It will ease the teachers in selecting the items according to the intended difficulty levels with the formation of items bank, which contain the validated items at the school level. Consequently, the pupils would likely be assessed by items according to their ability, and in return, the items that are well targeted to pupils' ability will yield accurate assessment results. Besides teachers, pupils will actually know their acquisition level of particular subject.

In brief, accurate and precise reporting on pupils' achievement would benefit teachers in enlightening the current school assessment practice such as testing, measurement and evaluation and improving the teaching and learning process. Besides nurturing the importance of item analysis among teachers and enhancing their skills in writing good quality items, school absolutely would gain positive impacts due to this improvement.

1.6 Limitations of the Study

This study focused on 2014 UPSR candidates from National School *(Sekolah Kebangsaan or SK)* in Kuala Selangor district with no involvement of candidates from National-type Schools *(Sekolah Jenis Kebangsaan or SJK)* of Tamil and Chinese schools. Therefore, the findings of this study cannot be generalised to other schools other than SK type.

The application of modern measurement model in this study, rather than classical, was limited to the statistical properties offered by Rasch measurement model using Winsteps® 3.68.2. On the other hand, the analysis using Item Response Theory (IRT) model, such as the 1 Parameter Logistic (1PL), 2 Parameter Logistic (2PL) or 3 Parameter Logistic (3PL), was not applicable in this study, even though discussions of this model have been highlighted in the literature review to show comparison between the two.

1.7 Research Conceptual Framework

This study aimed to measure the quality of EP1 items in the identified constructs of 2014 UPSR trial examination. Forty multiple-choice items, consisting of five substantive components in language testing including vocabulary, language and social expression, grammar, text completion and reading comprehension were analysed to answer the research questions of this study.

These items were analysed using Winsteps® 3.68.2 of Rasch measurement model to determine the psychometric properties of the items in terms of validity and reliability. This analysis would also help to identify the distribution pattern of item difficulty and pupils' ability and determine the extent of EP1 item difficulty level. Rasch's property offers significant diagnosis to determine the technical quality of EP1.