CONSTRUCTION AND VALIDATION OF THE TESL FOUNDATION ENTRANCE TEST IN A PUBLIC UNIVERSITY

GEETHANJALI A/P NARAYANAN

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

CONSTRUCTION AND VALIDATION OF THE TESL FOUNDATION ENTRANCE TEST IN A PUBLIC UNIVERSITY

GEETHANJALI A/P NARAYANAN

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2018

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: GEETHANJALI A/P NARAYANAN

Matric No: PHA 070019

Name of Degree: DOCTOR OF PHILOSOPHY

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

CONSTRUCTION AND VALIDATION OF THE TESL FOUNDATION ENTRANCE TEST IN A PUBLIC UNIVERSITY

Field of Study: TEACHER EDUCATION

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be

subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date:

Subscribed and solemnly declared before,

Witness's Signature

Date:

Name:

Designation:

ABSTRACT

In the 2015 Global Monitoring Report, there is a large disparity in the pupil-teacher ratios across the globe. However, to address this issue, the quality of the recruitment of teachers should not be brushed off. Teachers must be well versed in their subject matter to make them authoritative and credible. The development of a highly qualified and committed teaching force is one of the main concerns of the global educational reform. As such the aim of this study was to construct and validate an English Language entrance test (ATET) for the TESL foundation programme in a public university, so that the graduating students from this programme will be highly sought after at the TESL degree level, making them more knowledgeable and authoritative in their field. The ATET was developed and validated over 4 years, 2010-2013, to determine the items have psychometric properties. The construction and validation of ATET employed the framework that maps the Rasch Model and Kane's Validity Argument. Winsteps was used to analyse the bulk of the quantitative data which showed fit and order validity. The final outcome, after three sequentially improvised versions, was ATETv3 with 60 MCQ items, testing Reading, Grammar and Writing skills and one essay question (open-ended). These items range in difficulty between -3.24 and 4.38 logits, with Person Reliability of 0.86 and Item Reliability of 0.98 and Cronbach Alpha, 0.8. The essay component was analysed using Facets as well as EduG and the D-Study (Decision Study) suggested one rater (an experienced one) is sufficient, with a G-coefficient of 0.91. Meanwhile this study also established Predictive (r=0.85) and Construct (r=0.78) validity using SPSS. On the whole the cut-off point to be accepted into the programme was set at -1 logit. The paper-pencil ATET, which was originally meant to select appropriate candidates for the TESL foundation programme, is currently adopted as a digitized diagnostic test, conducted online in a public university.

PEMBINAAN DAN PENGESAHAN UJIAN KEMASUKAN ASASI TESL DI SEBUAH UNIVERSITI AWAM

ABSTRAK

Dalam laporan pemantauan 2015 *Global Monitoring Report*, nisbah murid-guru adalah sangat besar di seluruh dunia. Namun dalan keghairahan mengambil lebih ramai guru, kita tidak harus mengabaikan kualiti bakal guru. Seorang guru perlu mengetahui perkara yang mereka mengajar dengan baik supaya mereka mempunyai kredibiliti yang tinggi. Salah satu daripada perkara yang penting dalam reformasi pendidikan global ialah untuk mendapat satu pakatan tenaga pengajar yang menpunyai kelulusan yang tinggi serta komited. Dengan sedemikian, matlamat kajian ini adalah untuk membina dan mevalidasi satu ujian kemasukan (*entrance test*) Bahasa Inggeris (ATET) untuk program Asasi TESL di sebuah universiti umum supaya membolehkan graduan program ini menjadi pilihan di peringkat ijazah sarjana muda TESL kelak dan seterusnya menjadikan mereka lebih berpengetahuan dan berwibawa. ATET telah dibangunkan dalam masa 4 tahun, 2010-2013, dengan perincian psikometrik. Ini dilakukan berdasarkan *Rasch Model* dan *Kane's Validity Argument*. Winsteps telah digunakan untuk menganalisa data yang telah menunjukkan *fit dan order validity*. Hasil daripada analisis ini, selepas tiga versi yang melalui proses penambahbaikan, ialah ATETv3 dengan 60 item berbentuk aneka pilihan yang menguji kemahiran Pemahaman, Penulisan dan Tatabahasa Bahasa Inggeris dan satu soalan esei yang berjenis soalan terbuka. Item-item ini adalah dalam tahap kesukaran antara -3.24 dan 4.38 logit, kebolehpercayaan orang= 0.86, kebolehpercayaan item=0.98 serta Cronbach Alpha=0.8. Komponen esei dianalisis dengan Facets bersama EduG dan kajian kesimpulan (Decision-study) mengesyorkan seorang pemeriksa esei yang berpengalaman, dengan koeffisien G= 0.91. Kajian ini juga telah meneliti keesahan dari segi ramalan (r=0.85) dan konstruk (r=0.78) dengan menggunakan SPSS. Akhirnya, kata putus (*cut-off point*) untuk calon yang boleh diterima dalam program ditetapkan pada logit -1. Walaupun ujian kertas-pensil ATET dibina untuk pengambilan pelajar ke dalam program Asasi TESL, ia kini telah diadaptasikan sebagai ujian diagnostik yang telah dimuatnaik dalam laman web dan diuji secara atas talian di sebuah universiti umum.

ACKNOWLEDGEMENTS

Namo Amituofo. It has been an extremely long journey that I had embarked since 2007. Many a times, I was at the verge of giving up. Trials and tribulations were my hurdles.

Nevertheless with the guidance of my supervisor, Dr Shahrir Jamaluddin, who always reassured me that my ability to see the 'bigger picture' was my plus point. His patience and motivation has kept me going for this long.

I'm also indebted to Prof. Lyle Bachman, who advised me via email on my methodology and conceptual framework. He mooted the idea of considering to Kane's Argument-based Validity with the endless reading list and pages off his manuscript of the book he was writing was a great lead for my dissertation.

The other instrumental person behind my Conceptual Framework is Dr. Vahid Aryadoust of National University of Singapore (NUS) who gave me a personal insight of his mapping of Kane's view of validity against Rasch's analysis.

My gratitude goes to the entire PROMS (Pacific Rim Objective Measurement Society) dignitaries who gave constructive feedback at the symposiums that I had presented and initiated me to Winsteps as well as the trainers of MPA (Malaysian Psychometric Association) who guided me with Facets.

I'm also grateful to Universiti Teknologi MARA for having granted me a scholarship to pursue this programme and for giving me all the support in completing my thesis. Thanks to the Asasi TESL students who took the tests and provided me with such rich data to be analysed.

My PhD would not have been possible if it was not for the support of my family and friends. A special thanks to my family, my husband, Baskaran Rajagopal, my son, Rohiit, my daughter, Nilasha, my mother, Subashini, my sister, Lingeswari and my late mother-in-law, Sarojini. This thesis is especially dedicated to my late father, Narayanan, who saw me off on this journey, and am sure is rejoicing in heaven upon the completion of my PHd. Thanks everyone!

TABLE OF CONTENTS

CONTENTS **Page**

Chapter 2: Literature Review

Chapter 3: Methodology

Chapter 4: Results

Chapter 5: Discussion and Conclusion

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Asasi	:	in Malay, Foundation
ATET	:	Asasi TESL Entrance Test
EduG	:	Software to perform generalizability analysis
FACETS	:	Software to perform Multi Faceted Rasch Model analysis
ICC	:	Item characteristic curve
IELTS	:	International English Language Testing System
IRT	:	Item Response Theory
MCQ	:	Multiple choice question
MUET	:	Malaysian University English Test
SAT	:	Scholastic Aptitude Test
SPM	:	in Malay, Malaysian Certificate of Education
SPSS	:	Statistical Package for the Social Sciences (software)
TESL	:	Teaching English as a Second Language
Winsteps	:	Software to perform Rasch Analysis

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

In the landscape of education, the meadows are filled with a wide array of knowledge, ranging from useful to harmful information. It takes a teacher to permeate his/her presence to teach the children of what is good and bad in those meadows. At home, parents are the teachers. In schools and educational institutions, the instructors are the teachers. Teachers are a vital part of everyone's lives, be it formal or informal. A child who goes to school is taught by qualified teachers while a child who does not attend school is taught by the society around him/her. The experience gained from this learning process is synonymous with education. As Nelson Mandela befittingly said "Education is the most powerful weapon which you can use to change the world" (Mandela).

The 1948 Universal Declaration of Human Rights manifested that everyone has the right to education, but many have been deprived. To materialise this dream, United Nations in a joint venture with UNDP, UNESCO, UNICEF and World Bank organised the World Conference on Education for All (1990) in Jomtien, Thailand, where the World Declaration on Education for All and the Framework for Action to Meet the Basic Learning Needs were embraced (Inter-Agency Commission, 1990). These documents prescribed universal route to primary education and reduction in illiteracy by 2000. As the targets were not met as planned, another World Education Forum was organised in 2000, giving birth to the Dakar Framework for Action where education for all by 2015 is an obligation of national governments. The focus of this framework include "… early childhood development, girls' education, literacy,

education in emergencies, HIV/AIDS and health issues and the role of information and communication technologies in education" (UNESCO, Final Report of World Education Forum 2000, 2000). The sequel to this was the World Education Forum (2015), formulating the Incheon Declaration, "Education 2030: Towards Inclusive and equitable quality education and lifelong learning for all". This declaration is a renewal of what was set in the previous frameworks, fine-tuned towards enhancing the quality of education, transforming teaching and learning through technology, solidifying early childhood care and education, ensuring ability to read and write within a lifelong learning perspective, adopting new trends in tertiary studies, ensuring fairness and gender impartiality in education, supporting education for peaceful and supportive societies, tapping into inclusive quality education and strengthening education and crisis (UNESCO, World Education Forum 2015 Final Report, 2015). Based on these frameworks, it is very clear that education has been given serious consideration and a substantial amount of money and time has been endowed.

However, the provision of education and its peripherals alone do not ensure a person becomes educated. It is how much of these learning experiences that are sustained over time and applied in the most favourable way not only to mankind, but also to Mother Earth encompasses an educated person. The United Nations Earth Summit in Rio de Janeiro (1992) enlisted 40 chapters of Agenda 21, the Rio Declaration on Environment and Development and one of the many things addressed in this declaration is education for sustainability (Chapter 36) (United Nations Sustainable Development Agenda 21, 1992). A decade later, the Rio Summit achievements were reviewed at the 2002 World Summit on Sustainable Development (WSSD) in Johannesburg. It was here that the realization that education is not merely

dissemination of knowledge, but to create awareness about the balance among human, economics, cultural customs and appreciation for natural resources. This urges instructors and students to contemplate on their local surroundings, as well as establish and assess alternative future plans. This called for a reformation of the education scheme and administrative code of conduct which will make way for "learning to know, to live together, to do and to be" (Delors, 1996) in the global community. A follow-up from WSSD was the UN Decade of Education for Sustainable Development (2005-2014) (UN for Education for Sustainable Development 2005-2014, 2005). Governments and educational agencies were urged to ensure teachers' professional and academic freedom to select best practices to meet the objectives of their respective education systems. In its 2015 report, despite the fact that more children have been enrolled in school, especially primary schools, it was reported that there is an imbalance in pupil-teacher proportion in many parts of the world (Global Monitoring Report Team, 2015). This translates that the need for teachers is at an alarming state. According to UNESCO Institute for Statistics, UIS, by 2030 the world requires approximately 26 million primary school teachers to teach every child (UNESCO Institute for Statistics, 2015).

Although there is an acute shortage of teachers around the globe, the quality of the selection of teachers should not be short changed. Teachers are the ones who mould the young minds of the children. Thus, what is taught in schools must be accurate and precise. Widdowson (2000) said that teachers should know their subject-matter very well as it provides the grounds for their authority and credibility in their profession. The development of a highly qualified and committed teaching force is one of the main concerns of the global educational reform (Cheng, Chow, & Tsui, 2001). This teaching force is necessary as mentioned in the United Nations' Education for Sustainable Development (UN for

Education for Sustainable Development 2005-2014, 2005). This is very much dependent on the efficiency of the pre-service and in-service teacher-education programmes at the institutions of teacher education (Maclean, 2001). Thus selection of candidates into teacher training programmes must be done carefully, so that all these demands can be fulfilled.

At the home front, education in Malaysia is becoming more competitive. Each year, the Malaysian Certificate of Education (SPM) shows an improvement in its results and more students are achieving straight As (Kulasagaran, 2013). The local universities have to be selective in their choice of candidates for the various programmes offered. This is also apparent for the teacher training programmes. Candidates who are interested in teaching have to sit for an entrance examination, Malaysian Educators Selection Inventory (MEdSI). MEdSI is a 300-multiple-choice-item test of personality, which will be able to test the interest of the candidates in the teaching profession, integrity and emotional aptitude. Upon passing this test, the candidates are then shortlisted for an interview, before being considered for the local public universities' teacher training programmes (Ministry of Higher Education, 2015).

The Ministry of Higher Education, which was set up on 27 March 2004, trains secondary school teachers via the public universities, while the Ministry of Education equips the primary school teachers in the Institute of Teacher Education (formerly known as Teacher Training College). However, all issues related to schools and teachers, regardless whether primary or secondary, (besides the training of the teachers) are handled by the Ministry of Education.

The Ministry of Education and the Ministry of Higher Education of Malaysia agreed at a coordination meeting on 29 August 2006 to make it compulsory for all candidates applying for any bachelor's degree in Teacher

Education to sit for this test and an interview to ensure the quality of teachers that are generated through both these ministries based on the Malaysian Teacher Standards (*Standard Guru Malaysia, SGM*) (Ministry of Education Malaysia, 2006a).

The SGM outlines the professional competence which is based on the National Mission, Educational Philosophy, Teacher Education Philosophy, Code of Ethics of the Teaching Profession and the Work Ethics of the Ministry of Education. The SGM has two components: Standards and Needs (Ministry of Education Malaysia, 2006a). The Standards, which are inter-related, are listed below:

1. High competency in the aspect of teaching professionalism
2. Knowledge and understanding of the teaching profession
3. Teaching and learning skills

The Needs component should be made available by the teacher education agencies and institutes to achieve the standards mentioned above. These are:

1. Entry qualification and selection procedures to the Teacher Education programmes
2. Training, interpretation and evaluation of the teacher education programme
3. Collaboration
4. Infrastructure and Info structure
5. Quality Assurance

So, with the SGM, it is hoped that better quality teachers will be produced by the teacher education programmes throughout Malaysia.

This is further supported by the following statement from the Educational Development Master Plan (*Pelan Induk Pembangunan Pendidikan*):

> "... as the most significant and costly resource in schools, teachers are central to school improvement efforts. Improving the efficiency and equity of schooling depends, in large measure, on ensuring that competent people want to work as teachers, that their teaching is of high quality, and that all students have access to high quality teaching."
>
> (Ministry of Education, 2006b; p106)

The Master Plan's main objective is to enhance teacher professionalism by tapping on teachers' knowledge, skills, experience, spiritual, social, intellectual and financial capitals. It is an effort to make teacher profession among the prestigious and respectful profession in the country (Ministry of Education Malaysia, 2006b). This effort is summarised in Figure 1.1.

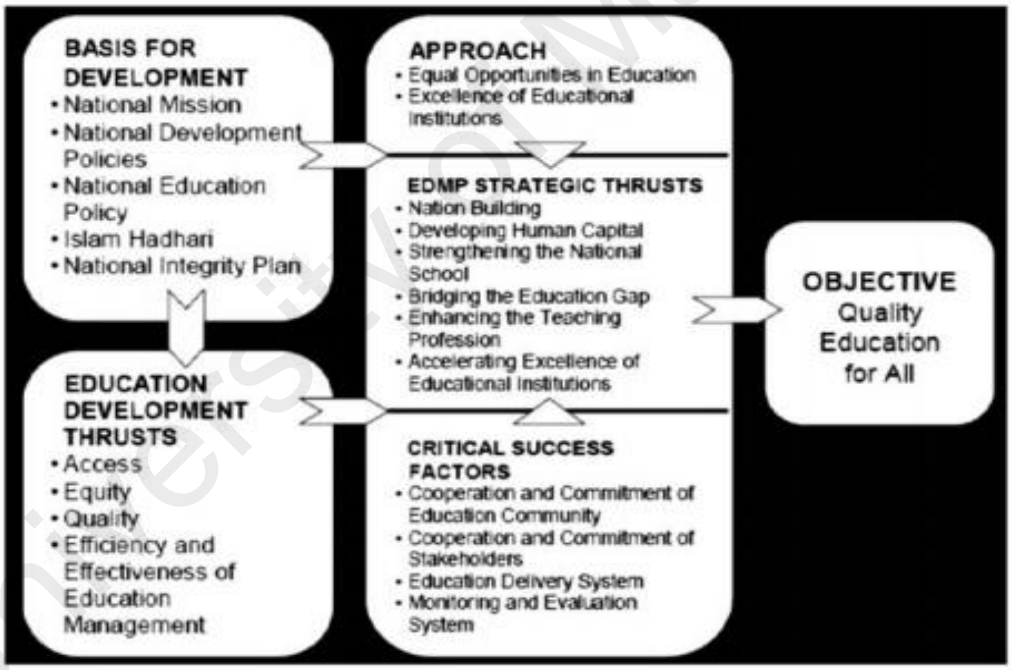

Figure 1.1. Malaysian Education Development Master Plan Conceptual Framework (Ministry of Education Malaysia, 2006b)

The recent development in raising the standards of teacher education was addressed in the National Education Blueprint 2013-2025. This latest blueprint aspires to transform the higher education system in two aspects: the education system and individual students. This balanced academic ecosystem is a vision to

provide capable research and academic staff to mould the young minds to become useful citizens in the country. Thus, the teaching profession is now meant for the highly competent individuals who are not only intelligent, but also passionate in teaching (Ministry of Education Malaysia, 2013).

The Teacher Education Division (TED), part of the Ministry of Education, watches over the teacher training programmes in Malaysia. In the quest of raising education in Malaysia to world class, it is hoped that by 2020, 100% of the secondary school teachers and 70% of primary school teachers are degree holders (Jamil, 2014). According to Jamil (2014), with highly qualified teachers, students will be exposed to quality education, making them better individuals. The biggest challenge faced by the TED is the provision for different medium schools in the Malaysian education system. There are four kinds of schools in Malaysia: the National Schools, with the use of Malay language as the official language; the Chinese Schools with Mandarin; the Tamil schools with Tamil language and the Islamic religious schools with both Malay and Arabic languages. Thus, the approach to teacher education at the universities has accommodated the developments and changes in the Malaysian education scenario. The provision of knowledge and tutelage for the teachers at the Institutes of Teacher Education or universities definitely will be helpful in any type of school. In addition, the Institutes of Teacher Education and universities impart English Language courses to all pre-service teachers as a requirement for teachers to be proficient in the language regardless of the school's medium of instruction (*State of Teacher Education in the Asia-Pacific Region*). This is also highlighted in the National Education Blueprint 2013-2025, in which English is recognized as an essential tool for success in the 21st century.

English is an international lingua for business and communication although it ranks second after Mandarin in terms of most widely spoken language (Wikipedia, 2015). The former Deputy Prime Minister Tan Sri Muhyiddin Yassin said that English is an important subject and had proposed to make it a compulsory pass for this subject at the SPM level. "English is a language of the world and business. It is an international means of communication," he told reporters ("Should a pass in English be made compulsory to pass SPM" *The Star*, 8 June 2009). This issue of whether passing SPM English should be made mandatory brought about mixed reactions from several people. The Malaysian English Language Teaching Association (MELTA) president Dr. S. Ganakumaran insists that the competency of the teachers must be upgraded before English is made a "passing subject" in SPM. This was also supported by the Parent Action Group for Education chairman, Datin Noor Azimah Abd Rahim, that compulsion of passing English can only materialise when the weaknesses in the teaching profession are addressed ("Boost quality of teaching", The Star, 9 June 2009). Due of the objection from the general public, this proposal was rejected. However, in 2013, with the initial launch of the Malaysian Education Blueprint 2013-2025, the minister once again announced that the Ministry of Education will implement this proposal, making English a compulsory pass subject in the 2016 SPM, together with *Bahasa Malaysia* (Malay Language) and *Sejarah* (History) (The Sun Daily, 14 January, 2015). The decision was to encourage the less proficient to be use English more frequently as English is an essential 21st century survival tool. In tandem with this, the ministry has implemented the policy of Empowering *Bahasa Malaysia* and Strengthening English Usage to raise the level of command of English among school students (Ministry of Education Malaysia, 2013).

The debate on the use of English Language in schools started prior to the new blueprint. As an outcome of the Ministers' Council Meeting in 2002, the teaching of Mathematics and Science in English at all school levels began in 2003 mainly to prepare a nationwide human capital that can embrace globalisation. This policy was opposed by many groups according to the research findings of Prof. Dr. Nor Hashimah Jalaluddin (*Pengajaran dan Pembelajaran Sains dan Matematik dalam Bahasa Inggeris*) where 75% of students agreed that it was problematic to study Mathematics and Science in English while 96% of the teachers said that students were not interested in these two subjects when taught in English. This survey findings concurred with the study conducted under Universiti Pendidikan Sultan Idris (Haron,Gapor, Masran, Ibrahim, & Nor, 2008) and as a result, this policy was abolished beginning 2008 and in total by 2012. However, due to the recognition of English as an international language and its versatility, the Dual Language Programme was brought forward in line with the second wave of the National Education Blueprint (Ministry of Education Malaysia, 2013). This programme gives provision to schools to choose to conduct Science, Technology, Engineering and Mathematics (STEM), in either both languages, Malay or English (Dual language programme to continue in national schools, says Education Ministry, 2016). For this programme, the Ministry of Education Malaysia has established an electronic English Language Teaching Centre (ELTC) as a support system for the teachers.

1.2 Rationale of the Study

Looking into the teachers' proficiency of English, in 2012, the Deputy Director General of the Ministry of Education, Datuk Dr. Khair Mohamad Yusof revealed that only one-third of students and one-third of the English teachers were

proficient in the English language in a survey conducted prior to the Preliminary Report Malaysian Education Blueprint (2013-2025) (Filmer, 2012). The main finding from this study showed that the correlations between the SPM English and the Cambridge English Language 1119 claimed that only one-third who scored high in SPM fulfilled the essentials in the 1119. Another finding was that only one-third of the teachers reached proficiency levels of C1 and C2 in the Cambridge Placement Test. This report has given a very clear picture of the status of English teachers' proficiency in Malaysia.

In addition, the MEdSI test is taken by candidates who apply for the TESL degree programs in public universities and it is merely a personality test. It does not test the proficiency level of the content area of the students. So this does not guarantee that a TESL undergraduate will enter the programme with sufficient knowledge of the English Language and be able to master the language within three to four years. As such, it is pertinent that the candidates applying for the course are filtered with a language proficiency test so that the entry behaviour for the programme is met.

It is important to note that constructing a language test is complex as it does not have any particular theme or content. It is mainly based on the language skills. Furthermore, it also has another aspect called the "native speaker" (Davies, 1990). In addition to this, there are four kinds of language tests, namely diagnostic, achievement, placement and proficiency tests (Hughes, 2003; Gronlund, 1982; Harrison, 1983). The diagnostic test attempts to find out what is the entry behaviour of a particular group or individuals. A teacher can determine strengths and weaknesses of the individuals. An achievement test assesses the mastery of the content that was taught prior to the test. The teacher would be able to determine what are the sections mastered and misunderstood. Meanwhile a

placement test is used mainly to segregate the students and stream them according to their proficiency levels. Proficiency tests gauge the general ability in certain aspects of the language without referring to any particular course. Hughes (2003) has mentioned that the focus of the proficiency test is on the current ability for a future requirement. As Bachman and Palmer (2010) and Kane (2013) have put it there must be appropriate use and interpretation of tests scores as well as evaluation of the consequences of those uses. Bachman and Palmer (2010) have coined "Assessment Use Argument" (AUA) as a conceptual framework (see Figure 1.2) to design a test with a justified intended use and evidence to link the AUA with interpretations of the testtakers' performance. As such a proficiency test would be appropriate for selection purpose.

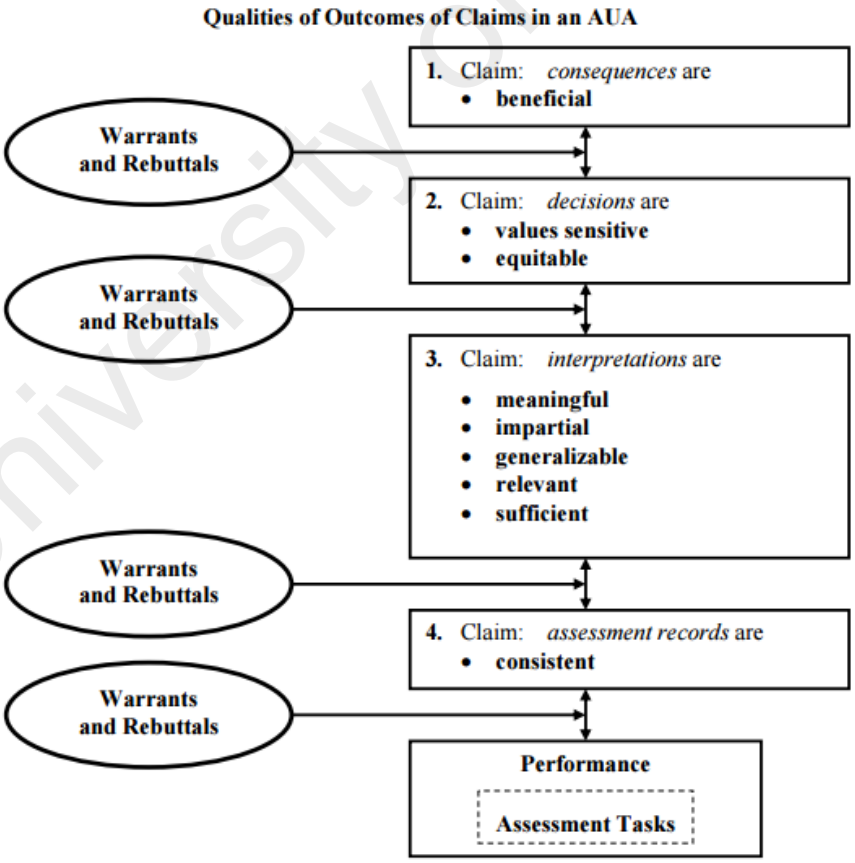

Figure 1.2. Structure of the Assessment Use Argument
(Bachman & Palmer, 2010, p. 104)

Traditionally reliability and validity are the main concerns of a language test and for a long time, reliability has taken the centre stage (Van der Walt & Steyn (Jr.), 2008). Besides reliability and validity, Weir (1991) adds another dimension, efficiency, which are feasibility and cost effectiveness in test construction, and ease of test administration. As such not many studies are done on validity in the past as compared to reliability (McNamara, 2005). One of the reasons for the lack of validation studies is the cost involved as it has to be administered over a period of time and the number of sample involved must be big enough to reduce errors (McNamara,2000). Validation studies are necessary especially for high stake tests as the results have many consequences on the stake holders.

Another issue in tests is equivalence. If a test has many batteries, they should be equal in terms of difficulty and discrimination levels, which may not happen in reality. According to McNamara (1996), one of the methods, although an unfair conduct, is to ensure the number of success is fixed. The test difficulty is dependent on the group ability, that is, a good student will score lower in a group of smart students than that good student placed amidst weaker students. This does not promote stability of the test. In fact, there is no clue if the test characteristics, particularly item difficulty will be the same for all students in the Classical Test Theory (CTT).

On the other hand, the Rasch analysis focuses on the latent trait of the student, which is denoted as measures (in logits) and not scores. This analysis is more robust as it considers the pattern of responses on all the items to estimate item difficulty. In fact if there is a missing response, an expected response is estimated from the pattern of existing responses (Bond & Fox, 2007). The results can show that students who have similar ability are placed on the same level.

Bachman (2005) adds on that the fact that latent traits are indicated, the analysis is inferential and not descriptive. Item difficulty is matched to the students' ability and item-ability maps (Linacre,2005) project probability of students' response with a certain difficulty level of the item (McNamara,1996) which are placed unidimensionally (Baghaei, 2008).

1.3 Statement of Problem

It is clear that the English Language is given much importance in Malaysia. It is a very important language globally for political, social and economic sustenance. There are many multinational companies operating in Malaysia which require their staff to be proficient in English. Thus, the English teachers/tutors/instructors are more in demand than any other language teachers/tutors/instructors (How large is the job market for English teachers abroad?). The training in this language teaching field must be comprehensive enough to meet the requirements of the job market. The TESL (Teaching English as a Second Language) teachers, in particular, have to be well trained so that they are competent enough to teach students who will be going out to work in companies like the multinational ones which place importance in English. So, the selection of the TESL candidates also has to be stringent.

According to the Ministry of Higher Education, tertiary studies are reasonable enough in cost for SPM leavers to consider other alternatives instead of Form Six. As such, admission into a matriculation or foundation course which leads to an undergraduate study seems to be the preferred path (Kaur, 2008). It has also become a norm nowadays that those who aspire to be an English teacher take up pre-TESL after SPM. Typically of a public university, through its Faculty of Education extends TESL Foundation Programme to SPM leavers

(Narayanan & Jamaluddin, 2012). The increasing number of interested applicants has made the selection process more rigorous. As such, the faculty has sketched a flow chart as seen Figure 1.3:

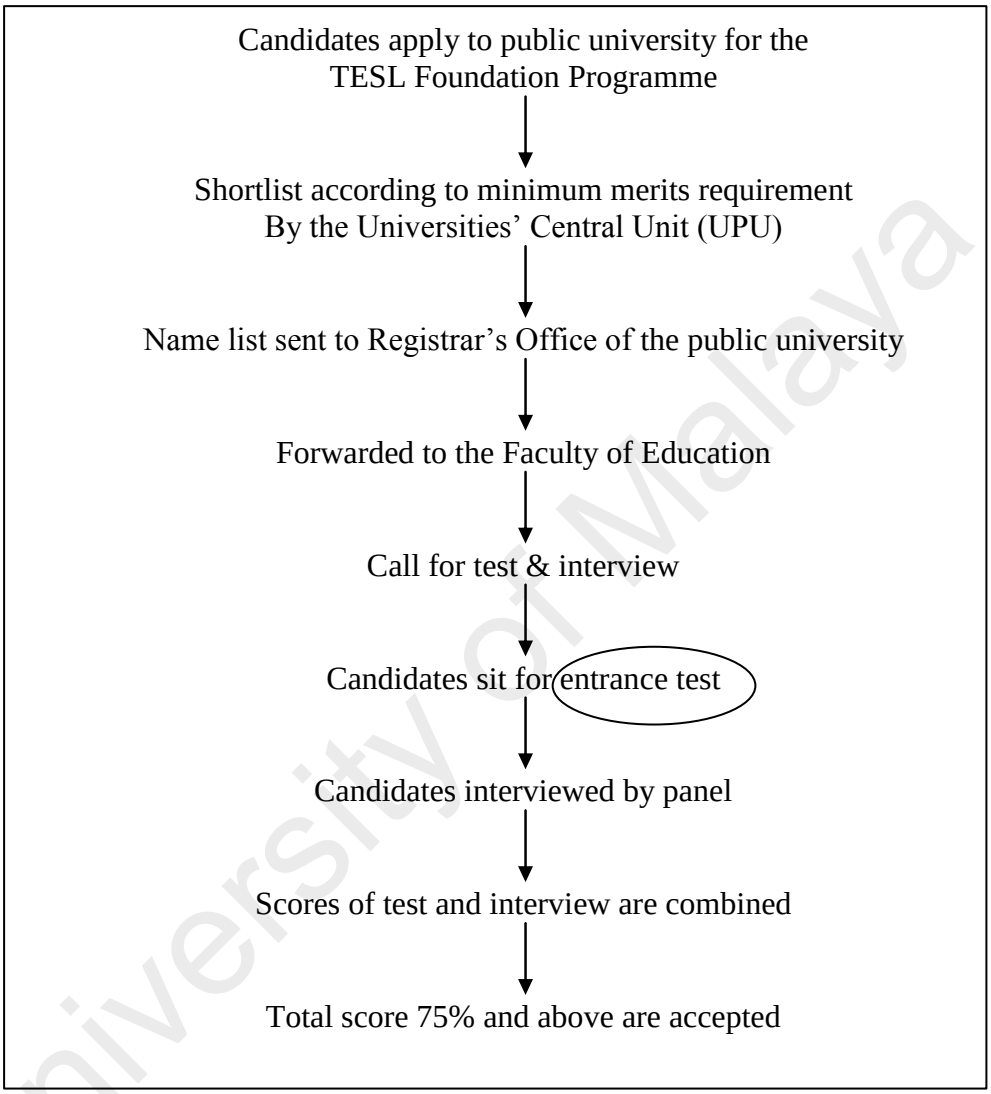

Figure 1.3. Flowchart of selection process for TESL Foundation Programme

Based on Figure 1.3, the entrance testwas authenticated. However, the existing entrance test was found to be faulty, thus this study moved into a different direction, that is to construct and validate a prognostic test for the TESL Foundation Programme. Accepted applicants should be at least at an intermediate level as the course does not address basic skills of the language.

The test, being high stake, has been taken lightly. A high stake test must be highly reliable and valid (American Educational Research Association, 2000). The process of reliability and validity must be stringent. In fact, the process of validating a test is the most important process in the development of a test (Walt & Steyn, 2008). In the past, there have been only a handful of studies that had embarked on the area of validation (Davies & Elder, 2005).

Consequently, a discussion was held with the Head of Programme (HOP) of the TESL Foundation Programme. It was resolved that there should be a revised selection test to be constructed and validated for 2010 intake, with a proper blueprint. This test needs to improve the accuracy of test scores to select appropriate candidates for the programme (i.e. intended purpose). The test scores will be norm- referenced as it is used to compare test takers.

Unfortunately, due to departmental decision with the appointment of a new HOP, the test that was constructed (ATET version 1) was not used for the 2010 intake. The new HOP was not in favour of using a different format and content for the test. He insisted that the length of the test should only be for one hour and there must be an essay question. Despite informing the finding of the existing test (to increase the number of questions), the number of items had to be retained. The format of the test for this study incorporated some verbal aptitude items modelled after the SAT as well as MUET format reading passages with some modified items. As this study's test could not be administered with the actual target group, the test items were experimented with the group who were already accepted into the programme. With this notion, this research chose to design and endorse a test, targeting on a post secondary school level, which can segregate the applicants and serve as an exemplary to create new batteries of the test. The logistics behind the test execution and the marking should assist the

TESL Foundation programme personnel to quickly key in the scores to make decisions for admitting applicants into the TESL Foundation programme.

As English is the main focus of the foundation programme, the test is an English Language Test. Besides this, the administrative problems involved during the process of the selection have to be minimised. The applicants are given an hour for the test and then adjourn to their respective rooms for the interview. So, the panel of interviewers will have to mark the test paper (30 multiple choice questions and 1 essay), ask questions and check the documents that support the application, including the merits obtained in the SPM examination and co-curricular involvement before deciding whether the student is selected or not for the programme.

This study developed a test which can be scored very quickly, based on the rubrics. The test scores were analysed for reliability and validity of the items. The validation of this entrance test has made way for future tests to be developed according to this blueprint. The items range from easy to difficult questions as well as considering all levels of Bloom's Taxonomy.

1.4 Purpose of the Study

This study has constructed a prognostic test (Farhady, 1983), whereby the selection test involves decision-making, whether to accept or not to accept a candidate into the programme. The content being Notional Functional classifies the test to be Functional-Communicative type. This test being high-stake would also require some standardization, which utilises scientific techniques in analysing the test (Spolsky, 1978).

Cloze tests are contextual in nature. This allows language to be tested in an integrative approach. As Farhady (1983) puts it, cloze test are able to test an

array of language skills. Thus inclusion of close test in the test would be a good idea to check of candidates' diversity of language skills.

This approach was also considered when dealing with reading passages. The questions that were asked had contextual clues. Among the items that were taken into account include inferencing, stated and implied main ideas and drawing conclusions.

The writing component is able to elicit candidates' ability on the general knowledge, grammatical accuracy and writing skills. This integrative approach (Spolsky, 1978) is much sought after in the current trend in language testing.

Besides the content matter, the analysis of the results have also taken a trending direction. The Rasch analysis utilising the mathematical modelling (logistic probability) is sample independent (Narayanan & Jamaluddin, 2012) and is able to compare the measures of person and items on the same ruler, an equal interval scale. The analysis identifies data that are misfits (Linacre, 2005). Constructing and validating an entrance test for the TESL Foundation would be feasible, functional and spot on. Thus, the intent of the research is to create and endorse an entrance test for the TESL Foundation Programme called TESL Foundation Entrance Test (ATET).

1.5 Objectives of the Study

With such a purpose to construct and validate a test, this study was conducted to fulfil the following objectives:

i. To develop an English Language test for the TESL Foundation programme that has interpretive and validity argument.
ii. To identify the number of raters for the essay component
iii. To determine the cut-off point for the entry to the TESL Foundation programme

1.6 Research Questions

The overarching question for this study is "Can the properties of the Rasch Model be exploited to develop the TESL Foundation Entrance Examination for a public university?" There are two parts of this study, first being the construction of the test (instrument) and the second, validation of the test. The quantitative component of this study will attempt to answer the following questions:

i. How to construct an English Language test for the TESL Foundation programme that has interpretive and validity argument?
ii. What is the suggested number of raters for the essay component?
iii. What is the cut-off point to be accepted into the TESL Foundation programme?

1.7 Significance of the Study

At the end of this study, the findings will be reported to the faculty. This presentation will be a comprehensive guide for anyone who heads the programme to administer the entrance test to select the candidates for this foundation programme.

In line with the accredited certification to ISO 9001:2008 for all academic processes and procedures, of one of the public universities' Faculty of Education, this careful validation of the selection test will enhance the quality of the selection process in the TESL Foundation programme. The construction of the test and test specifications can be a guide to construct parallel tests and the items can be saved in an item bank. In future these items can generate many batteries of the test which will have similar difficulty levels.

The validation process is rigorous and comprehensive using the Conceptual Framework in Figure 2.1. As such, once validation is endorsed, the test is automatically endorsed as reliable as well.

The instructors will be assured of the proficiency level of their students as the instructors would not have to deal with basic language skills but are able to focus on polishing the students' knowledge of the language. More time can be spent with thorough exercises and activities.

The test would be able to filter better candidates that the sponsor of this TESL Foundation programme, the Ministry of Higher Education, would not have to worry about attrition. There would be positive returns of investment.

The fact that this test focuses on the multiple-choice questions for all skills including writing, the marking will be more reliable. In any case that the interviewer is a non-TESL optionist, the results will always remain stable. In short reliability is guaranteed.

All these multiple-choice items can be digitized to run an online test from the construction of this test. In any case of shortage of time or too many candidates per interview session, tests can be conducted online and results would be instant. By doing so, interviewers can concentrate 100% on the interview.

1.8 Limitation of the Study

This study was carried out with the TESL Foundation students at a public university who were already chosen into the programme. The test score may not be generalised for other programme or faculty or university applicants. Another limitation is that the students who did not get into the programme were not tested with the ATET, leaving that portion of the scale untapped. This takes a toll on the profile of the normative approach. The number of sample is also a limitation as the research is limited to the number available. Thus this has an implication in the choice of model used. The three parameter model requires at least a thousand

samples (Henning, 1987), while the one parameter model requires only 300 applicants.

The innovations in terms of the format of test items and types of items are limited because the unfamiliarity of items might jeopardize the possibility of being selected (Kunnan, 2000). To address this issue, the ATET takes a slightly different approach in the stem, but not in the stimuli (reading passages and grammar items). Candidates might not do well on the test if they are unsure how to answer the questions. However, as this study also addressed fairness in testing, the test was conducted on the last day of the orientation week before the start of classes to ensure no learning had taken place, i. e. there is no washback effect.

In addition, the raters for the essay component for this study were of different experiences. Rater 1 has 3 years of teaching experience while Rater 2 and 3 have more than 15 years of teaching experience. The diversity in terms of teaching experience has affected the results for inter-rater reliability, and seems as one of the limitations of the study as it is inevitable to get hold of raters with similar background, especially teaching experience.

1.9 Operational Definition

1.9.1 TESL Foundation Entrance Test

According to thefreedictionary.com, an entrance examination determines a candidate's preparation for a course of studies. Farhady (1983) classifies a test that involves decision-making as prognostic. He claims that if the decision is made "on the acceptance or non-acceptance of students into a certain programme, it is referred as a selection test". Brown (2002) added on that entrance examination must be appropriate, particularly when testing English, and constructed with a great deal of

quality". Brown also mentioned that "if the entrance examination is norm-referenced, they should be either aptitude tests or proficiency tests."

TESL, abbreviated for Teaching English as a Second Language, focuses on the five main skills, i.e. Reading, Writing, Grammar, Listening and Speaking (Brown, 2002). However, in the process of testing for an entrance test, the logistics might not be feasible to test Listening and Speaking. These two skills are addressed during the interview, whereby candidates are evaluated in terms of the language fluency as well logical thinking.

Reviewing these definitions, the operational definition for a TESL Foundation entrance test in this study would be an English Language test, based on Reading, Grammar and Writing skills, that would enable selection of candidates into the TESL Foundation programme.

1.9.2 Test Construction

According to the Dictionary of Psychology, test construction is the way to meet the aspects of test standardization like validity, dependability and norms. According to Wikipedia, it is about how items in a psychological measure are generated and selected. Meanwhile Kline (2015) says that test construction is when the test items either have appropriate and favourable features to fit a Rasch or similar model with much accuracy or manipulate it well enough to fit in accordance to the test takers' abilities. . According to Kline, factor-analytic test construction is far better than criterion-keyed methods to create a unifactorial test. Nunnally (1978) in Kline (2015) promotes item-analytic test construction followed by factor analysis of the short set of selected items. In this study,

the process in which items are subjected to a Rasch analysis together with subject matter experts' advice. Thus after considering the psychometric properties, items are then finalised and put together as a test, according to the classification of the content.

1.9.3 Test Validation

Test validity, coined by Kelley (1927) in Baghaei (2008) is that the test measures what it is supposed to measure. It requires empirical evidence and it depends on results of several studies done longitudinally. According to Baghaei (2008), test validation, using the Rasch Model, ensures unidimensionality and psychometric properties. Kane (2006) says that there must be test utility, i.e. the test must be useful.

The classic way to define validity is how far does a test exactly observed what it is purported to be observed. It can be misguiding if this observation is not fulfilled, which in return defeats the purpose of testing in the first place. At the earlier years, validity was classified as predictive validity, content validity, construct validity, concurrent validity (Gipps, 1994).

- Predictive validity tells if what is being tested can precisely foresee a particular future achievement.
- Content validity includes all relevant and crucial characteristics that can ensure favourable achievement.
- Construct validity tells if the test is in tandem with the domains or skills that are being tested.
- Concurrent validity provides comparable outcomes between a particular test and another test which has similar domains/skills.

However, such a focus on the traditional way of looking at the different kinds of validity has the probability of misguiding test setters to only have proof for few of the types of validity and not for all the four types. This can be very misleading for test-setters (Gipps, 1994).

As such, this study defines test validation as a process of ensuring the test has all the psychometric properties with empirical evidence in accordance to the Rasch Model and is confirmed having predictive and construct validity.

1.9.4 Public University

Public universities are universities that are funded by the government. Application to these universities are centralised through an independent body called the Universities' Central Unit (UPU). Thus, gaining admissions into a public university is competitive. There are all together twenty public higher learning institutions in the country. All these higher learning institutions have autonomous governance and conform to the Universities and University Colleges Act 1971 and are compliant to the Malaysian Qualifications Agency (MQA) (Malaysian Universities Guide, Retrieved from http://www.universitymalaysia.net).

1.10 Summary

This research developed and validated an English Language proficiency test to be used for the selection of candidates into the TESL Foundation programme at a public university. The study is divided into two main parts: (i) construction of the test and (ii) the validation of the test. There were three research questions (RQ) which steered the research: How to construct a valid test that measures the proficiency in English based on the Rasch Model? What is the suggested number

of raters for the essay component? What is the cut-off point to be accepted into the programme? The significance of this study is to create new batteries of the test, using the same Table of Specification. The gap of the field of study is that not many validations have been undertaken and none have utilized the mapping of the Kane's Argument-based validity on the Rasch Validity Model. This study was carried with some limitations at hand. This study was carried out after contemplating on numerous studies and theories. The following chapter will discuss the related theories, past research, theoretical and conceptual framework that was utilized in this study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews theories and past studies related to language testing and decides on a theoretical framework. Further scrutiny on relevant studies towards the methodology has mooted the conceptual framework.

2.2 Related Theories and Models

There are several theories and models related to language testing.

2.2.1 Classical Test Theory

The Classical Test Theory was originated from Spearman whose concept is known today as Classical True Score Model (Crocker and Algina, 1986). According to Magnusson (1967) and Lord and Novick (1968), this model served as the starting point for developing various mathematical procedures for test data analysis. The True Score Model is based on the formula: $X = T + E$,where X is a particular test taker's observed score, T is the true score and E represents the error of measurement (Crocker and Algina, 1986). Guilford and Fruchter (1978) have defined true score as the score that test takers would achieve if the instrument used was perfect and conditions were ideal. As such the error of measurement would be entirely eliminated. However, practically, this is not possible. It is also unlikely to match a test taker's average score over all acceptable occasions, test forms, and administrators. A test taker's score usually would be different on other occasions, on other test

forms, or with different administrators. The Classical Test Theory can only detect one source of error at any one time (e.g. differences in scores for different occasions can be assessed with test-retest reliability).

2.2.2 Generalizability Theory

Generalizability Theory (G-Theory) is a method of estimating the relative magnitudes of various components of error variation and for indicating the most efficient strategy for achieving desired measurement precision (Shavelson & Webb, 1991). G-Theory concerns solely with the variance in what is observed and the other unwanted main effect, interaction or random error variances. G-Theory utilises Fisher ANOVA to raw scores to estimate variance components and to calculate reliability coefficients (Linacre,1993). He said that a pilot test should be administered first to estimate all the variance components. Then a Decision Study (D-Study) will be done to see if a desired level of reliability can be obtain based on those G-Study variance components. Linacre also lists a couple of advantages for Rasch-based G-Theory. Firstly, the precision (standard error) of a measure can be predicted solely from the number of replications, without any preliminary G-Study. Secondly, the statistical aspects of the D-Study are minimal. The ratio of 'true' standard deviation to the average standard error indicates how many distinct measurement levels the test design can discriminate in this measure distribution.

Cronbach (1972) had introduced G-Theory as a statistical framework which extends beyond CTT by distinguishing the many different sources of error that may affect a particular measurement. By

this, it is possible to improve the test or its design to reduce the errors and increase the precision of measurement.

Shavelson & Webb (1991) point out that researchers can estimate what proportion of total variance in the results is due to factors like setting, time, items and raters. They call these factors as "facets". The facet that the researcher wants to examine will serve as the object of measurement for the purpose of the analysis, while the other facets are treated as sources of measurement error. From the G-Study, the D-Study can be done to find what will be the result if the circumstances are changed for the object of measurement. Another difference between CTT and G-Theory according to them is that the G-Study takes into account how the consistency of outcomes may change if a measure is used to make absolute or relative decisions.

2.2.3 Item Response Theory (IRT)

The Classical Test Theory (CTT) has one drawback, that is the results of the analysis of a test is very much sample-dependent. This makes it difficult to provide a fixed item difficulty measure (Alderson et al, 1995). This is where the Item Response Theory, IRT, comes in use. The results from the analysis of a test is independent of the sample. This also means that two different tests can be equated, although they are taken by two different sample groups.

IRT does not contradict with the basic principles of CTT. In fact, IRT proposes an alternative statistical analysis centered on the items, besides presenting new technological resources for the psychological and educational evaluation. IRT is an umbrella of statistical models that

attempts to measure the abilities, attitudes, interests, knowledge or proficiencies of respondents as well as specific psychometric characteristics of test items. IRT attempts to model the relationship between unobserved variable, usually ability of examinees and the probability of the examinee responding correctly to any particular test item (Harris, 1989). Hambleton (2000) stated that item response theory places the ability of the respondent and the difficulty of the item on the same measurement scale in order to make comparisons between respondents' abilities and items. The ability of the respondent is labeled "b" while the test item characteristics are described by the difficulty (b), discrimination (a), and pseudo-chance (c) parameters. The Rasch Model considers difficulty as a variable, while discrimination and the pseudo-chance parameters are considered as constant. It provides unbiased, efficient, sufficient, and consistent estimate of separate person and item calibrations (Schumacker, 2005). The formula is given below:

$$P_i = \frac{e^{(bn-Di)}}{1 + e^{(bn-Di)}} \quad , i = 1,2,3,..,n.$$

Where:

P_i = probability of examinee with ability b_n answering item *i* correctly,

D_i = difficulty parameter of item *i*,

n = number of items, $e \approx 2.718$

The One Parameter Logistic Model (1PL IRT Model) is descriptive. The computational is a simple approximation to the Normal Ogive Model. Meanwhile, the Rasch Model is prescriptive. It is distribution-free person ability estimates and distribution-free item difficulty estimates on an additive latent variable (Linacre, 2005). The Rasch Dichotomous

Model is utilized when each person in the sample is parameterized for item estimation, while the 1PL IRT Model item estimation is parameterized by a mean and standard deviation for the person sample. Item characteristics curve (ICC) in Rasch is modeled to be parallel with a slope of 1 (the natural logistic ogive), while for the 1PL IRT, ICCs are modeled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive). Thus, the 1PL IRT Model has d= 1.7, while for the Rasch Model is 1.

Note the difference in the formula as well as the symbols used for the One Parameter Logistic Model,

$$P_i(\theta) = \frac{e^{1.7(\theta - b_i)}}{1 + e^{1.7(\theta - b_i)}}, \quad i = 1,2,3,..,n.$$

Where:

P_i = probability of examinee with ability θ answering item *i* correctly,

b_i = difficulty parameter of item *i*,

n = number of items, $e \approx 2.718$

The Two Parameter Logistic Model, considers difficulty and discrimination as variable and psedo-chance is fixed as constant. The formula is given below:

$$P_i(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_i)}}, \quad i = 1,2,3,..,n.$$

Where:

$P_i(\theta)$, b_i , *n*, and *e* are defined the same as in the 1-PL model.

a_i = discrimination parameter of item *i*.

The Three Parameter Logistic Model, considers all the three as variable and the formula is given below:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-ai(\theta - bi)}}, i = 1,2,3,..,n.$$

Where:

$P_i(\theta)$, $b\,i$, n, and e are defined the same as in the 1-PL model.

ai is the same as 2-PL model, c_i is the pseudo-chance parameter (Baker, 1985)

IRT, also known as the latent trait theory was introduced by Lord in the 1950s. It was only in the 1970s that IRT had generated interest in the field of measurement (Clapham & Carson (eds.), 1997). In 1977, a special edition of the Journal of Educational Measurement on 'Applications of latent trait model' was dedicated to IRT. Hambleton (1989) contributed to the understanding of IRT and its use in the field of education. Henning (1987) and Bachman (1990) showed the use of IRT in language testing. Most of these contributors limited their studies to the one-parameter logistic model, which is unidimensional for dichotomously scored data.

2.2.4 Rasch Model

"This psychometric model was discovered by Georg Rasch, which is an interaction between the ability of the examinees and the item difficulty. The mathematical theory for this Rasch Model is somewhat like the variation of the 1PL IRT model."

$$P(\chi_{ni} = 1) = \frac{e^{(\theta n - \beta i)}}{1 + e^{(\theta n - \beta i)}}, i = 1,2,3,..,n.$$

Where:

χ_{ni} = score of examinee n for item *i* with $\chi_{ni} = 1$ for a correct answer $\chi_{ni} = 0$ for wrong answer

θ_n = ability of examinee

β_i = difficulty of item *i*,

n = number of items, $e \approx 2.718$

"In this model the probability of a correct answer is modeled as a logistic function of the difference between the person and the item. Thus, the higher a person's ability, the higher the probability to get a correct answer on a item".

Each item in an instrument has its own characteristics. From the Rasch Model, Item Characteristic Curve can be drawn to show the location of an item at which probability that $\chi_{ni} = 1$ is equal to 0.5. It represents the proportions of persons who answered an item correctly.

Another point to note in is that the items are all the same type, where the term unidimensionality is used. It refers to the fact that the computation applies to only one attribute of an object (Bond & Fox, 2007). Besides unidimensionality, within this Rasch context, the scale works the same way at all times, regardless of which group is being tested. This model is sample independent.

2.2.5 Multi Faceted Rasch Model

The Multi Faceted Rasch Model (MFRM) denotes the probability that "an examinee (*n*) will receive a rating in a particular category (k) by a rater (j)

on a task, (*i*)". MFRM belongs to the Rasch Models and has picked up its application in the field of measurement. The basic MFRM model is a follows:

$$\ln\left[\frac{P_{nkji}}{P_{njik-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_{k-1}$$

Where

P_{nkji} = probability of examinee *n* receiving a rating *k* from the rater *j* on the criterion *i*,

P_{njik-1} = probability of examinee n receiving a rating *k-1* from the rater *j* on the criterion *i*

θ_n = proficiency of examinee, *n*

β_i = difficulty of criterion i

α_j = severity of rater *j*

τ_{k-1} = difficulty of receiving a rating of *k* relative to *k-1*

2.3 Theoretical Framework of the Study

IRT originates from the Probability Theory. It shows the probability of a specified response modelled as a function of person and item parameters. In educational tests, person parameters pertain to the probability of a person getting an item right, while the item parameters pertain to the difficulty of the items (Alderson et. al., 1995). The higher a person's ability compared to the difficulty of an item, the higher the probability to getting a correct response for that item in the Rasch model. The purpose of utilising the model is to get measurements from categorical response data.

Usually, models are used with the objective of describing a set of data.

Parameters are modified, accepted or rejected based on how well they fit the data. However, the objective of using the Rasch model is to get the data which fit the model (Andrich, 2004).

Yen (1992) adds that the idea of generalising and predicting from one test occasion to another has been practised for a long time in educational measurement. The proportion of students getting an item right (*p* value) is how item difficulty is described traditionally. IRT, however, describes item difficulty in a manner that is stable over groups of students and interacts with the ability level of the group. The process allows detailed predictions of how much *p*-values will change when different students are tested. To produce these descriptions and predictions efficiently, a statistical model is required. The core of each IRT model is that it "defines the probability of a student's correct response to an item as a function of the ability of the student and properties of the item. This function is called the item characteristic function and a graph of it is called the item characteristic curve (ICC)" (p. 658).

According Schumacker (2005), IRT has its advantages and disadvantage compared to the Classical Test Models. The advantages include: item statistics are independent of the sample from which they were estimated, examinee scores are independent of test difficulty, item analysis accommodates matching test items to examinee's knowledge level and are reported on the same scale, and test analysis doesn't require strict parallel tests for assessing reliability. On the other hand, the disadvantages include IRT models are more complex, the model-outputs are more technical and difficult to understand, and IRT models require a larger sample to get more accurate and stable parameter estimates, although Rasch Models require small to moderate samples.

On the other hand, the Rasch analysis focuses on the latent trait of the

student, which is denoted as measures (in logits) and not scores. This analysis is more robust as it considers the pattern of responses on all the items to estimate item difficulty. In fact if there is a missing response, an expected response is estimated from the pattern of existing responses (Bond & Fox, 2007). The results can show that students who have similar ability are placed on the same level. Bachman (2005) adds on that the fact that latent traits are indicated, the analysis is inferential and not descriptive. Item difficulty is matched to the students' ability and item-ability maps (Linacre,2005) project probability of students' response with a certain difficulty level of the item (McNamara,1996) which are placed unidimensionally (Baghaei, 2008).

Thus this study is based on the Item Response Theory as the underlying theory and the approach is based on the dichotmous Rasch unidimensional measurement model. The dichotomous Rasch Model is clarified as:

$$P(\chi_{ni} = 1) = \frac{e^{(\theta n - \beta i)}}{1 + e^{(\theta n - \beta i)}} \quad , i = 1,2,3,..,n.$$

Where:

χ_{ni} = score of examinee n for item *i* with $\chi_{ni} = 1$ for a correct answer $\chi_{ni} = 0$ for wrong answer

θ_n = ability of examinee

β_i = difficulty of item *i*,

n = number of items, $e \approx 2.718$

2.4 Testing

"Testing is a universal feature of social life" (McNamara, 2000, p.3).

From ancient times to this modern era, people are tested for their abilities or knowledge. Testing is used in many forms and areas. For example, drug testing,

DNA tests, blood tests, software testing, machineries' testing, achievement tests and personality tests.

A test, according to the Merriam-Webster Dictionary, is a procedure used for critical examination, observation, or evaluation. This will determine the presence, quality, or truth of something. In terms of a series of questions or exercises, a test measures the skill, knowledge, intelligence, capacities, or aptitudes of an individual or group.

> "The purpose of evaluation is to make a judgment about the quality or worth of something- an educational program, worker performance or proficiency, or student attainments. That is what we attempt to do when we evaluate students' achievements, employees' productivity, or prospective practitioners' competencies."
>
> (Ebel & Frisbie, 1991, p23)

Ebel and Frisbie (1991) add that a test is a type of measurement technique. They claim that all tests are measures, and all measures are included in the set of qualitative and quantitative techniques of evaluation.

Clapham (1997) said that testing is a subordinate of assessment. Tests are designed for large number of students, often for gatekeeping purposes and may be imposed by outside authorities.

According to Airasian & Terrasi (1994), the main purpose of testing to get information to assist in making inferences about a test taker's performance on a behaviour domain of interest. They claim that this purpose can be acquired if there is good judgment at each of the five steps in the testing process: construction, selection, administration, scoring, and interpretation.

Popham (2000) looks at educational measurement, educational testing and educational assessment as synonymous as a "process by which educators use students' responses to specially created or naturally occurring stimuli in order to make inferences about students' knowledge, skills or affective status" (p.3).

2.4.1 Language Testing

Lado (1961) said that language testing in the early theories measure candidates' knowledge of the grammatical system, vocabulary and aspects of pronunciation. These aspects were tested in isolation and at different times. This is called Discreet Point Testing.

Harris (1969) had listed the principal educational uses of language tests, i.e. (1) To determine readiness for instructional programmes, (2) To stream individuals in appropriate language classes, (3) To diagnose the individual's specific strengths and weaknesses, (4) To measure aptitude for learning, (5) To measure the extend of student achievement of instructional goals, and (6) To evaluate the effectiveness of instruction. These categories can be classified under three parts : General Proficiency (categories 1-3), Aptitude (category 4) and Achievement (categories 5 & 6) (Davies et. al., 1991).

Spolsky (1978) has divided language testing into three main trends, the pre-scientific, the psychometric-structuralist, and the integrative-sociolinguistic. The pre-scientific trend belongs to the traditional approach of language testing and the types of tests involve translation, composition or isolated sentences. Harris (1970) shows the domination of the grammar translation method, which focuses on listening skills, grammar, vocabulary and reading. This method requires candidates to translate from English to foreign language and from the foreign language to English. Moller (1982) also adds that the texts are usually long, and test forms include sentence completion, dictation, reading aloud, compositions on literary topics at advanced level, and grammatical items. The characteristics of the tests during this period, according to Moller

(1982) and Spolsky (1978), lacks statistical means for objectivity and reliability, very much dependent on teacher's discretion in terms of scoring, highly subjective assessment and does not identify clearly the type of proficiency that is tested.

The next trend in language testing is the psychometric-structuralist trend (Spolsky, 1978). In this trend, the measurement experts exerted their belief that measurements must be precise and scientific. Issues like objectivity, reliability and validity were of primary concern (Ingram, 1968). Spolsky (1978) adds that the objective test techniques were developed. The multiple choice items were widely used and the analytical scoring procedures were introduced. This trend is where psychologists and testers interacted with the linguists. In terms of language testing, contrastive analysis was used to identify problems in second language acquisition. Lado (1961) claims that these problems are unique and are not similar to the native language learning. All these resulted in discrete point tests in language testing.

The third trend is the psycholinguistic-sociolinguistic trend (Spolsky, 1978). The test in this trend considers the entire communicative effect of the message rather than discrete items. The dictation and cloze procedures were introduced in this trend. Carroll (1961) supports this trend as he says that the measurement of the knowledge in language is not based on structural and lexical items, but the overall communicative ability of the testees.

Language testing began to incorporate IRT-based approaches only from the mid-eighties onwards even though such approaches have been applied in other areas of educational measurement as early as late sixties.

(Szabo, 2008)

Henning, Hudson and Turner (1985) tried to test if the Rasch model can be used in the case of seemingly multidimensional data. They had analysed a test which had four subtests: listening, reading, grammar, vocabulary and writing (error detection). The data was analysed as one set well as in separate subtests. The difficulty estimates and the fit statistics from both these analysis were compared and they were almost identical. Thus, the unidimensionality was not violated.

Henning (1984) presented a comparative study of traditional and IRT-based English Language test analyses. The traditional statistical components were items' facility values, variance and point biserial correlation indices, while IRT statistics included Rasch-based item difficulty estimates and fit statistics. Henning found that the reliability had increased when item were deleted on the basis of IRT-related statistics.

According to Henning (1987), as the teaching staff and facilities are restricted, tests are often used to select candidates for a programme from the pool of applicants. Aptitude tests are used to select candidates who can sustain in a course/programme (Carroll, 1965). Meanwhile, the achievement tests are used to measure what has been learnt during a course/programme, usually complying to the objectives of the course/programme (Mehrens and Lehmann, 1975). Proficiency tests are used to measure, globally, the ability candidates have before entering a particular course/programme (Henning, 1987). The type of test used will depend on its purpose.

There are two families of tests in language testing: (a) norm-

referenced tests, which help make program level decisions, i.e. proficiency and placement decisions and (b) criterion-referenced tests, which help make classroom-level decisions, i.e. diagnostic and achievement decisions (Brown, 2005). Table 2.1 below shows the differences between these two families.

Table 2.1
Norm-referenced and Criterion-referenced Test Differences (Brown, 2005, p 3)

Characteristics	Norm-Referenced	Criterion-Reference
Type of Interpretation	Relative (a student's performance is compared to those of all other students in percentile terms)	Absolute (a student's performance is compared only to the amount, or percentage, of material learned)
Type of Measurement	To measure general language abilities or proficiencies	To measure specific objectives-based language points
Purpose of Testing	Spread students out along a continuum of general abilities or proficiencies	Assess the amount of material known or learned by each student
Distribution of Scores	Normal distribution of scores around the mean	Varies; often non-normal. Students who know the material should score 100%
Test Structure	A few relatively long subtests with a variety of item contents	A series of short, well-defined subtests with similar item contents
Knowledge of Questions	Students have little or no idea of what content to expect in test items	Students know exactly what content to expect in test items

2.5 Selection Decision

The literature is full of articles on admission policies and the selection decisions being made on the basis of standardized test scores and grade point averages (Imber, 2002; Micceri, 2001; and Perfetto, 2002). Lei, Bassiri and Schultz, (2001) found that a college GPA was an unreliable predictor of student achievement. Since we assume that norm referenced tests are valid measures, the tendency is to put more weight on those results concerning student achievement.

Opponents of standardized achievement testing would argue otherwise. For example, Bennett, Wesley and Dana-Wesley (1999) suggested that a college admission model should be developed to encompass GPA, rank in class and a district performance index or a similar predictor as an alternative to standardized test scores. A formula index based on these predictors would afford some protection in selectivity issues.

2.6 Test Construction

The multiple-choice item (MCQ) is one of the mis usually found to be the preferred format in the field of educational measurement. An MCQ item has three main sections which are the stem (that asks the question or sets forward the problem), a few distractors (which are errorneous in nature) and the appropriate answer.

According to Haladyna et.al. (2002), there are certain guidelines that has to be followed when constructing multiple choice items, i.e. (i) avoid unfamiliar terminology. The focus should be on the subject matter and not on comprehending the words used in the items. (ii) avoid vague qualitative modifiers (eg. many, much, important) to avoid confusion. (iii) avoid complex word arrangement which may make it difficult to comprehend the question. (iv) avoid double negatives. (v) avoid trick statements with misleading word or spelling anomaly. (vi) paraphrase from source material (vii) avoid strongly worded statements, which more often are distracters. (viii) for true and false items, avoid too many of 'false' or 'true' statements. The test should have a few 'false' items more than the 'true' statements. However they should not be arranged in any particular pattern. (ix) item stem should be as short and precise. This is to test the comprehension of the question, rather than reading skills. (x)

distracters should be equally plausible and attractive. (xi) all options should be grammatically synchronised with the stem's. (xii) the grammar, length, and precision should be parallel for all options (xiii) avoid stems that provide answers to other items (xiv) avoid options that have the same meaning (xv) avoid presenting items that follow the same sequence in the passage (xvi) avoid colloquial language; be specific (xvii) avoid including unnecessary information in the stem. Be precise. (xviii) avoid the use of non-relevant source of difficulty (xix) avoid items that require extremely specific answers (xx) Include as much as possible in the stem, so that the options do not repeat parts of the question. (xxi) Use the "none of the above' option when the answer is totally correct. Avoid "none of the above" option when stem asks for the 'best' answer. (xxii) the use of "all of the above" may help those who have partial knowledge. (xxiii) having compound options may increase item difficulty. (xxiv) the difficulty of the items increases when options are more homogeneous.

However, Millman & Greene (1993) say that test experts normally write items based on their wisdom and common sense. This has been a worrying issue since Cronbach's warning in the 70s that achievement tests lack scholarly attention. Haladyna (1989) and Haladyna et. al (2002) claim that the knowledge of multiple choice item writing is quite limited and has room for improvement. The foundation to the guidelines of item construction, although have been supported by experimental and quasi-experimental research, is still based on recommendations of experts alongside with the validity and reliability of the tests.

Brown (2005) has outlined that test construction should begin with a discussion with the teachers who teach the course to find out what kind of items should be tested. He said that it is more difficult to construct multiple-choice

items, fill-in items and cloze passages compared to essay questions. The process is rather time-consuming. In his test evaluation checklist, Brown (2005) has listed out 6 test characteristics to be considered while constructing a test. They are item description (receptive or productive), norms (standardization sample, subtests and type of standardised scores), descriptive information (central tendency, dispersion and item characteristics), reliability, validity, and actual practicality of the test.

Employing the Item Response Theory, IRT, an IRT model can be utilized to construct a group of items that have arithmetical characteristics. Measures of differential item function could be obtained to check for biasness of each item in the test (Yen, 1992). The psychometric properties (like number of correct score means, standard deviations, standard errors of measurement, reliabilities and item *p*-value) can be readily predicted when items taken out from that pool, for different groups of students and even when those students have not taken that test.

2.7 Table of Specification

Every test should be based on some type of objectives. It depends on the purpose of the test. The Bloom's Taxonomy (Bloom, 1959) lists the educational objectives for the cognitive domain. Valette (1977) claims that the intellectual operations employed are of different levels to answer questions in a test. The following levels have been identified: knowledge (bringing to mind the suitable material), comprehension (understanding the denotation of the material), application (applying the knowledge of the elements of language to production of the correct oral or written response), analysis (breaking down a task into parts to make precise relationships between ideas employing connotation and inference),

synthesis (organizing ideas to produce an oral or written response) and evaluation (making quantitative and qualitative judgments about material). These levels demand increasingly higher cognitive abilities as one moves from knowledge to evaluation. Knowledge, Comprehension and Application are considered as lower level cognitive skills while Analysis, Synthesis and Evaluation are more advanced level cognitive skills which demand more advanced control of the language.

Three steps are involved in creating a Table of Specifications: 1) choosing the measurement goals and domain to be covered, 2) breaking the domain into key or fairly independent parts- concepts, terms, procedures, applications, and 3) constructing the table (Chase, 1999).

Brown (2005) insists that test setters should write out the Table of Specification. It should consist of a general description of the item, a sample item, stimulus attributes, responses attributes, and supplemental lists.

Teachers who do not use conventional construction guidelines for paper/pencil test development will not be assessing student achievement well. Their tests will likely have poor content validity, "cause for concern because each assessment instrument depends on its validity more than on any other factor." (Ooster, 2003, p. 40).

The Table of Specifications are also referred to as the "test blueprint," "master chart," "matrix of content and behaviors," "prescription," "recipe," "road map," "test specifications," or "formal specifications" (Bloom, Hastings, & Madaus, 1971; Carey, 1988; Gredler, 1999; Kubiszyn & Borich, 2003; Linn & Grunland, 2000; Mehrens & Lehman, 1973. Ooster. 2003). The blueprint is meant to insure content validity. Content validity is the most important factor in constructing an achievement test. (Notar, 2004).

2.8 Validity

Valid, derived from the Latin "validus" means "strong". In psychometrics, test validity means "the degree to which evidence and theory support the interpretations of test scores" (AERA, 1999)

Lado (1961) said that if a test measures what it is suppose to measure than the test is valid. Cronbach (1971) agreed with him and asked further what an instrument really measures, which then is the information that provides construct validity. Zeller (1990) adds on that valid measurement is essential to successful scientific activity. According to Weir (2005), validity actually resides in test scores and it is multifaceted. This simply means that different types of evidence is needed to support any claims for the validity of scores on a particular test. Test validation is the process of generating proof to support the inferences concerning trait from test scores. Weir claims that testing should be concerned with evidence-based validity.

Psychometricians in the recent past have been quick to point out that although we speak informally of valid tests, "[i]t's not the test itself that can be valid or invalid but, rather, the inference that's based on a student's test performance" (Popham, 2003, p.43). This thinking has expanded into not just inferences made from test scores, but to the social consequences of test use (Shepard, 1997). This definition of validity has been challenged recently, however, in a paper by Borsboom, Mellenbergh and van Heerden (2004). The argument is

> "If something does not exist, then one cannot measure it. If it exists but it does not causally produce variations in the outcomes of the measurement procedure, then one is either measuring nothing at all or something different altogether. Thus, a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure."
>
> (Borsboom et al., 2004, p1061)

Validity is the strength of the conclusions, inferences or propositions. More formally, Cook and Campbell (1979) define validity as the "best available approximation to the truth or falsity of a given inference, proposition or conclusion." According to them there are four types of validity commonly examined in social research: conclusion validity (relationship between program and the observed outcome), internal validity (causal relationship), construct validity (how the causal relationship was operationalised), and external validity (generalizability of results to other settings).

Content validity, according to Brown (2005) is to establish an argument that the test is a representative sample of the content for the test claims to measure. The process many take many forms. Test setters should talk to teachers teaching respective components of the test to decide how the test should be designed to measure. Then the different types of items are outlined and organised, leading to item specification with corresponding testing objectives drawn out. Item specifications include a general description, a sample item, stimulus attributes, response attributes and supplemental lists. The match between the item specifications and the item itself can be verified as part of the argument for content validity.

Construct Validity is the process of examining evidence that helps to justify the use of the test scores or a given interpretation (Messick, 1989). The term *construct*, according to Ebel and Frisbie (1991), refers to a psychological construct, an aspect of human behaviour that cannot be measured or observed directly (eg. reading comprehension, motivation and achievement). They said that construct validation is a process where evidence is gathered to support that the given test actually measures the psychological construct that the test setters had intended it to measure.

Zeller (1990) said that construct validity is present when a particular measure corresponds to the other measures consistent with theoretically derived hypotheses concerning relationships among the concepts. There are 6 steps in establishing construct validity: (a) construction of a theory by defining the concepts and anticipating relationship between them, (b) selecting indicators that symbolises each concept in the theory, (c) ascertaining the dimensional nature of these indicators, (d) constructing scales for these indicators, (e) computing the correlations among these scales, and (f) comparing these empirical correlations with the initial anticipated relationships among the defined concepts. Zeller (1990) also adds on that "it is impossible to validate a measure of concept unless there exists a theoretical network that surrounds the concept"(p 258). Alderson, Clapham and Wall (1995) said that to assess "construct validity of a test is to correlate the different test components with each other" (p.183-184). The assumption for having different components in a test is that each component measures something different from the other, and thus giving an overall picture of the language ability that is measured. With this, it is also expected that the correlations will be quite low. If the correlations are quite high, then that one of the components has to be dropped off the test. Alderson et al (1995) concluded that construct validation procedure is to hypothesise the relationships among test components considering the requirements of the underpinning theory, and then to compare these hypotheses with the correlation coefficients.

Another approach to construct validation is some form of factor analysis by reducing the more complex matrix of correlation coefficients to a more manageable proportion by statistical means. There are two main types – exploratory factor analysis (EFA) and the confirmatory factor analysis (CFA). EFA is where the factors that emerge are looked at and scrutiny of the factors that

relate to the test are then labelled. CFA on the other hand, starts with a prediction of components which will relate to each other. The test of goodness of fit is carried out to see if these predictions are true (Alderson et.al., 1995)

There are two types of criterion-related validity: concurrent validity and predictive validity (Zeller, 1990). Concurrent validity is when two test scores which were administered at about the same time are correlated. This shows that the criterion variable exists in the present. Meanwhile, predictive validity is when the two measures are administered at two different times. The purpose of the test has to be predictive. As such, the criterion variable will not exist until a later point of time. The correlation coefficients between the two scores are studied. The higher the coefficients, then there is external validity for that particular test, be it concurrent or predictive validity (Brown, 2005).

Although classical models list content validity, criterion validity and construct validity as part of providing evidence of the validity of a test, the modern view presents validity as a single construct. Messick (1989, 1996a, 1996b) argues that the traditional conception of validity is fragmented and incomplete because it fails to take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use. His modern approach views validity as a unified concept which places a heavier emphasis on how a test is used. He proposed that evidence to support (or question) the validity of an interpretation can be categorized into one of five categories: test content, response processes, internal structure, relations to other variables and consequences of testing. These five aspects must be viewed as interdependent and complementary forms of validity evidence and not viewed as separate and substitutable validity types. This framework has a four-way classification described by two facets: (1) the source of justification of the

testing, which takes into account either evidence or consequence or both, and (2) the function or outcome of the testing, which considers either test interpretation or use or both. The framework is illustrated Table 2.2.

Table 2.2
Facets of Validity (Messick, 1988)

	Function of outcome of testing	
Source of justification	Test interpretation	Test use
Evidential Basis	Construct validity	Construct validity + Relevance utility
Consequential Basis	Construct validity + Value implications	Construct validity + Relevance utility + Social consequences

Messick's unified concept is further enhanced by Kane (1999). Based on Toulmin's model, Kane introduced the argument-based approach in validation. This approach covers two parts: interpretive argument and validity argument. The interpretive argument is where an interpretation and the use of the test scores are proposed, while the validity argument evaluates the interpretive argument. Briggs (2004) relates "the argument-based approach to validation to the 6 principles for scientific investigation in education by the National Research Council (2001):

1. Pose significant questions that can be investigated empirically
2. Link research to relevant theory
3. Use methods that permit direct investigation of the question
4. Provide an explicit and coherent chain of reasoning

5. Replicate and generalize across studies
6. Disclose research to encourage professional scrutiny and critique."

The benefit of an argument-based validation is its ability to pinpoint a particular point of the interpretive argument and to the interactions with the characteristics of the measurement procedure (Kane, 2004).

Chapelle et al (2010) investigated whether an argument-based approach to validity makes a difference. Their conclusion was that an argument-based approach to validity introduces some new and useful concepts and practices. They point out clearly the differences in the key aspects in the validation process between the 1999 AERA/APA/ACME Standards for Educational and Psychological Testing [hereafter, "Standards"] and Kane's argument-based validation (2006), summarised in Table 2.3.

Table 2.3
Key Aspects in the Process of Validation in the Standards (1999) and in Educational Measurement (Kane, 2006)

Four Aspects Characterizing Approaches to Validity	**Standards (1999)**	**Kane (2006)**
Framing the intended score interpretation	A construct	An interpretive argument
Outlining the essential research	Propositions consistent with the intended interpretation	Inferences and their assumptions
Structuring research results into a validity argument	Listing types of evidence	Series of inferences linking grounds with conclusions
Challenging the validity argument	Counterevidence for propositions	Refuting the argument

Chapelle et al (2010) has listed down 13 propositions to produce supporting evidence which serve as hypotheses about score interpretations for TOEFL, which also provide guidance about the types of validity evidence required: (1) language skills (listening, speaking, reading and writing) are defined independently and in combination are necessary for students to succeed in advanced academic settings, (2) content domain of the tasks on the TOEFL requires certain language skills that students need in North American university settings, (3) each of the skills is composed of a set a subskills, (4) Each skill's score show internal consistency, (5) each of the skills is distinct to be measured independently, but the skills are related by some core competencies, (6) test performance is not affected by test-taking processes irrelevant to the constructs of interest, (7) test scores are arrived at through judgments of appropriate aspects of learners' performance, (8) test performance is not affected by examinees' familiarity with computer use, (9) test performance is not affected inappropriately by background knowledge of the topics represented on the test, (10) the test assesses second language abilities independent of general cognitive abilities, (11) criterion measures can validly assess the linguistic aspects of academic success, (12) test scores are positively related to criterion measures of success and (13) use of the test will result in positive washback in ESL/EFL instruction, such as increased emphasis on speaking and writing and focus on academic language. Their examination of the Standards and related materials showed the need for more explicit guidance on how to formulate an intended interpretation and the propositions that can point to the types of evidence that would at the end contribute to the TOEFL validity argument.

Kane's approach to propositions is to connect them to the inferences in the interpretive argument through two types of statements: warrants and assumption.

Warrants are the rule or established procedure to show the evidence. Meanwhile assumptions are the appropriate rubric for providing the relevant evidence. Figure 2.4 shows the 6 inferences, each with a warrant and assumptions that structure the basis for the TOEFL interpretive argument. Each of the inferences is used to move from grounds to a claim; each claim becomes grounds for a subsequent claim. In short, Table 2.4 focuses on the warrants and assumptions which need to be generated by the researcher to guide the validity research (Chapelle et al, 2008).

Table 2.4
Summary of the Inferences, Warrants in the TOEFL Validity Argument with their Underlying Assumptions (Chapelle et. al., 2010)

Inference	**Warrant Licensing the Inference**	**Assumptions Underlying Inferences**
Domain description	Observations of performance on the TOEFL reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use in the English-medium institutions of higher education	1. Critical English language skills, knowledge, and processes needed for study in English-medium colleges and universities can be identified. 2. Assessment tasks that require important skills and are representative of the academic domain can be simulated.
Evaluation	Observations of performance on TOEFL tasks are evaluated to provide observed scores reflective of targeted language abilities.	1. Rubrics for scoring responses are appropriate for providing evidence of targeted language abilities. 2. Task administration conditions are appropriate for providing evidence of targeted language abilities. 3. The statistical characteristics of items, measures, and test forms are appropriate for norm-referenced decisions.
Generalization	Observed scores are estimates of expected scores over the	1. A sufficient number of tasks are included on the test to provide

	relevant parallel versions of tasks and test forms and across raters.	stable estimates of test takers' performances. 2. Configuration of tasks on measures is appropriate for intended interpretation. 3. Appropriate scaling and equating procedures for test scores are used. 4. Task and test specifications are well defined so that parallel tasks and test forms are created.
Explanation	Expected scores are attributed to a construct of academic language proficiency.	1. The linguistic knowledge, processes, and strategies required to successfully complete tasks vary across tasks in keeping with theoretical expectations. 2. Task difficulty is systematically influenced by task characteristics. 3. Performance on new test measures relates to performance on other test-based measures of language proficiency as expected theoretically. 4. The internal structure of the test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components. 5. Test performance varies according to amount and quality of experience in learning English.
Extrapolation	The construct of academic language proficiency as assessed by TOEFL accounts for the quality of linguistic performance in English-medium institutions of higher education.	Performance on the test is related to other criteria of language proficiency in the academic context.
Utilization	Estimates of the quality of performance in the English-medium institutions of higher education obtained from the TOEFL are useful for making decisions about admissions and appropriate curricula for test takers.	1. The meaning of test scores is clearly interpretable by admissions officers, test takers, and teachers. 2. The test will have a positive influence on how English is taught.

2.9 Reliability

Ebel & Frisbie (1991) define reliability as the consistency of the measurement. Under the same condition with the same subjects, the instrument should measure the same way. It is important to note that reliability is not measured but estimated.

Thorndike (1982) said that reliability has 3 aspects to be concerned about: the basic rationale, the procedures for data collection, and the statistical procedures for data analysis. He also adds that the Classical Reliability Model sees the test scores as having two additive parts, the "true" score and a random "error". Both these additives are unrelated. The true score is defined as the value that the average of repeated measurements with the "same" measure approaches as the number of measurements is increased without limit.

Crocker and Algina (1986) defines reliability in a practical manner, i.e. the degree of the individual's deviation, which remain relatively consistent over the times of the repetition of the test, be it the same form or the equivalent. Thus, the measurement errors which are kept minimum should be the main concern in the construction of any test.

Brown (2005) defines test reliability as the extent to which results are considered consistent or stable. The degree of consistency can be estimated by calculating a reliability coefficient. This coefficient ranges from +1.00 to 0.00. When converted to percentage, the reliability estimate becomes percentage in consistency and the balance is the random variance. For example, if the reliability coefficient is 0.95, then the scores are 95% consistent, with 5% measurement error or random variance. He adds that there are three strategies used to estimate reliability: test-retest, equivalent forms and internal-consistency strategies.

Test-retest method is suitable for estimating stability of a test over a period of time. The method is to administer the same test twice to the same group. The time between the two test should not be too short (students can remember the items) or too long (where learning from the language programme will take place). Then a Pearson product-moment coefficient between the two sets of scores will be calculated (Brown, 2005).

The equivalent forms reliability is the same as the test-retest method, except that for the second administration of the test, a parallel form of the first test is used. The items should be written similarly and equivalent forms should have equal means and standard deviation, and the two forms correlate equally with some third measure (Brown, 2005).

Internal consistency reliability strategies estimate the consistency of a test using the information provided within the test itself. This can be done through the split-half method, where items are divided into two equal parts, usually even-numbered items separated from odd-numbered. They are scored separately and correlation coefficient is calculated. Usually a longer test is found to have a more reliable estimate. The Spearman-Brown prophecy formula is used to calculate the full test reliability estimate (Brown, 2005).

2.10 Goodness of Fit

Fit statistics presents the degree to which a given IRT model adequately fits the empirical data (Smith, 2002). Tests for goodness of fit must be performed to ensure that the appropriate model is applied, and all IRT software packages provide the goodness of fit statistics. He also said that there is no single universal fit statistic that is the best to detect all measurement disturbances. According to Smith, there are 3 types of fit statistics: total fit, within fit, and between fit.

These differ in their purpose, and in the manner in which they summarize the squared standardized residuals. The term misfit is used to identify when a model fails to adequately fit the data. The total fit statistic describes misfit due to the interactions of any item/person combination. This is used when identifying random types of measurement disturbances between a target and focal group. The between fit statistic compares logical groups like gender, ethnic or age to detect item bias and is used for systematic measurement disturbances. The within fit statistic is similar to the between fit statistic except that within fit statistic is summed over the group of interest, not the entire respondent sample (as for between fit statistics). No single fit statistic function can describe the various types of misfit or is best for all conditions within the three categories.

However, the goodness of fit in the Rasch analysis insists that the item and person data fit the Rasch Model (Bond and Fox, 2007)

The fit statistic should be selected based on the specific type of misfit that is of interest. Each of these types of fit statistics can be calculated as either weighted or unweighted. The weighted calculation attempts to reduce the variation introduced by wide ranges of person abilities or item difficulties.

Goodness of fit indices is largely dependent on the sample size. For some indices, such as the likelihood ratio chi-square, large sample sizes can distort the statistic, artificially inflating its value and leading to erroneous assumptions about the data (Byrne, 2001). Small sample sizes are also problematic because of the lack of statistical power (Hambleton, Swaminathan, & Rogers, 1991). If the sample size is between 100 and 1000, the chi-square can be an appropriate goodness-of-fit indicator. An additional advantage of the chi-square is that of a known distribution. According to Linacre (2015), there are suggested ranges of Infit and Outfit Mean-square (MNSQ) as seen in Table 2.5 for different types of

tests.

Table 2.5
Reasonable Item Mean Square Ranges for Infit and Outfit

Type of Test	Range
MCQ (High stakes)	0.8 - 1.2
MCQ (Run of the mill)	0.7 - 1.3
Rating scale (survey)	0.6 - 1.4
Clinical observation	0.5 - 1.7
Judged (agreement encouraged)	0.4 - 1.2

2.11 Differential Item Function

When two groups with approximately similar ability show a different probability of correct response, the items then are said to have Differential Item Function (DIF) (Thissen et al., 1993). DIF has been tested for many high stake tests/examinations like the Scholastic Aptitude Test, Graduate Record Exam and Graduate Management Administration Test, particularly establishing if there is gender DIF in the test. Studies have highlighted that males tend to do better on technical related items, especially reading comprehension items than their female counterparts (Lawrence et.al., 1988; O'Neill et.al., 1993).

Traditionally, checking if a test favours a particular gender was known as gender bias test. DIF has a broader meaning. It covers diverse, consequential implicatons for both test development and test use (Osterlind, 1998). DIF studies adds on to the test validity. Kane (2006) mentioned that DIF is able to validate the interpretation of test scores and claims made on the scores. As such, conducting a simple DIF investigation will definitely help to strengthen the test validation.

Besides this, Kunnan (2000) said that test setters should ensure that tests are well constructed, taking into consideration the candidates' difference in race,

gender, or ethnicity. He calls this test fairness. In short a DIF study will automatically ensure a test is fair to its candidates.

2.12 Conceptual Framework of the Study

This study is divided into two main parts: (i) construction of the test and (ii) the validation of the test.

The theoretical framework for the test construction is mainly based on the Rasch Model, which is derived from the Item Response Theory (IRT) or the Latent Trait Theory. IRT refers to three probabilistic measurement models: the 1-parameter, 2-parameter and 3-parameter model. The three factors considered in this theory are the difficulty level, the discrimination level and the pseudo factor. All three models can be specified from a single probabilistic function for the occurrence of a right answer by a person to an item. There are two main considerations in IRT: (i) the scale is unidimensional (i.e. it measures only one trait or attribute), and (ii) at any given level of the trait, the probability of endorsing one item is not related to the probability of endorsing any other item (also known as local independence). If a scale meets these two considerations, then the Item Characteristic Curve (ICC) is drawn for each item. In addition, the results from the analysis of a test are independent of the sample (Clapham & Carson (eds.), 1997).

This study utilizes the Rasch Model which is almost similar to the 1-parameter model. This model takes the difficulty factor as the single item parameter and fixes the other two factors, discrimination and pseudo/others, as constants. However, the difference between the Rasch Model and the 1-parameter Logistics Model lies in the slope that is parallel to the ICCs. The 1-parameter Model has ICCs parallel with a slope of 1.7 (approximating the slope

of the cumulative normal ogive) while the Rasch Model is with a slope of 1 (the natural logistic ogive) (Linacre, 2005).

The Rasch model for dichotomous data is represented as:

$$P_i = \frac{e^{(bn-Di)}}{1 + e^{(bn-Di)}}, \; i = 1,2,3,..,n.$$

Where:

P_i = probability of examinee with ability b_n answering item *i* correctly,

D_i = difficulty parameter of item *i*,

n = number of items, $e \approx 2.718$

(Linacre, 2005)

The content of the test is based on the Notional-functional Theory. This is to synchronize with the KBSM English Language Secondary School Syllabus, which is theoretically based on the communicative approach. Its main aim is to provide students with communicational ability and competence in using language forms and structure accurately (Ratnawati, 1996). The notional-functional theory focuses on three different types of meaning students need to convey: functional (the social purpose of the utterance), modal (the degree of likelihood) and conceptual (categories of communicative function). Therefore, the content of a communicative approach test should have authentic texts, targeting language competence and focus on function over form (Hawkey, 2004). The test will take the integrative sociolinguistic approach.

The validation part of this study is based on Kane's Argument-based validity framework and the Rasch Validity Model. According to Kane (1992), the argument-based approach has a practical prominence. He said that validation research involves a "systematic effort to improve (1) the accuracy of conclusions based on test scores, (2) the appropriateness of the uses made of these scores, and (3) the quality of the data-collection procedures designed to support the proposed

conclusions and uses." Kane (2006) states that to establish test validity, the focus must be on the gathering different types of supporting evidence. Froelich (2009) has summarised Kane's ideas into a process. Firstly, a substantive analysis of the test should be carried out by content area specialists. This will ensure the match between the subject specification and the goals of the educational programme (content validity). Secondly, other standardised tests (external consideration) which have similar constructs should be correlated with the test (criterion validity). Finally, a psychometric analysis of the test and its items should be done (construct validity). Kane (1992) has pointed out that the use of the test scores must be appropriate, particularly for high stake and standardised tests. Aryadoust (2009) said that Kane's framework has two phases: interpretive argument, which is in the form of statements, followed by validity argument, which is to investigate the usefulness of the interpretive argument.

The Rasch Validity Model (Wright & Stone, 1988) considers fit and order validity. Fit validity deals with the consistency of response patterns from (1) analysis of residuals, which is the difference between the Rasch Model and the responses (response validity), (2) analysis of item fit, which assists the test revision (item function), and (3) analysis of person fit, which diagnoses the testees whose performance do not fit the expectations (person performance validity). Meanwhile, order validity has two categories: meaning validity (from the calibration of test variables) and utility validity (from the calibration of persons to show criterion validity).

The model below (Figure 1.4) shows the conceptual framework to use Rasch-based measurement to build on Kane's validity argument. According to Aryadoust (2009) the validation process begins with observation, then moves on to generalization, explanation and finally extrapolation (the four major

inferences). Warrants are any data to back up these inferences, while backings are the theoretical assumptions behind the warrants.

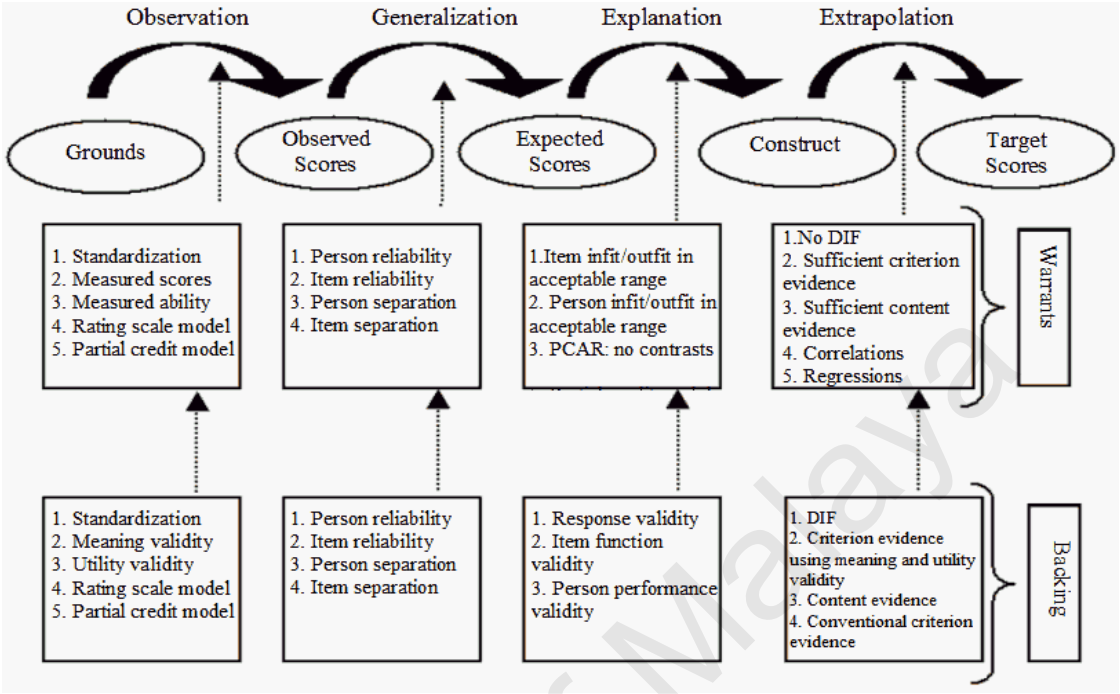

Figure 2.1. Supporting validity arguments using Rasch analysis

This study will employ the above model for the validation process, which encompasses both the Rasch Validity Model and Kane's Argument-based Validity.

2.13 Summary

In a nutshell, the literature review shows that the dichotomous Rasch Model can be used to construct a test and validated through the argument-based approach (Kane, 2004). Thus this study utilized these two, The Rasch Model and Argument-based Validation for the construction and validation of the ATEE. Several tests were conducted to ensure there are various types of validity and fairness in the test.

CHAPTER 3

METHODOLOGY

3.1 Introduction

The aim of this study is to develop and validate an instrument, TESL Foundation Entrance Test (hereafter ATET) for selection of candidates for the TESL Foundation Programme. This study will scrutinize the psychometric properties of the instrument used. As the original selection test was found to be faulty in its construct validity (see 3.2), a new set of test was reconstructed.

Initially, the data collection technique was to validate an existing test for the 2010 cohort. It was recycled from the previous Pre-TESL entrance test. There were three sections, Reading (3 passages with 24 MCQ items), Grammar (6 MCQ items) and Writing (1 prompt requiring a 250-word argumentative essay). The scoring of the test, both the MCQ and essay, was done by the interviewers who were made up from TESL and non-TESL lecturers. There was no calibration in the marking of the essays. The test score was added to the scores of the interview and these were totalled with the merit scores (based on the SPM results) that were pre-calculated by the Student Intake Division of the public university. Since the standardization and fairness in the selection process was questionable, the test scores for the 2010 intake was analysed using Winsteps and the results for the MCQ is seen in Table 3.1. A sample of 134 was taken to investigate the psychometric properties of the test.

Table 3.1
Overall Winsteps Results for the Original Pre TESL Entrance Test

Persons	134 INPUT		134 MEASURED		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	25.1	30.0	75.64	6.97	1.00	.1	.94	.2
S.D.	2.3	.2	8.53	1.95	.18	.5	.67	.7
REAL RMSE	7.23	ADJ.SD	4.52	SEPARATION	.62	Person	RELIABILITY	.28
Items	30 INPUT		30 MEASURED		INFIT		OUTFIT	
MEAN	117.3	133.8	50.00	3.88	1.00	.1	.94	.0
S.D.	14.5	.5	12.54	2.11	.05	.3	.26	.7
REAL RMSE	4.42	ADJ.SD	11.74	SEPARATION	2.66	Item	RELIABILITY	.88

According to Linacre in Winsteps@Rasch Measurement Computer Program User's Guide (2016), person separation classifies people and low person separation, less than 2 with person reliability less than 0.8 indicates that the test is not able to discriminate between high and low performers. Linacre suggests more items should be added. Meanwhile he also mentions that item separation which is 2.66 shows there are only 2 levels of item difficulties and item reliability more than 0.9 denotes that the sample size is large enough to confirm the item hierarchy of the instrument, which is the construct validity. From Table 3.1, it can be concluded that there were not enough items as the person reliability is only 0.28 and the test cannot discriminate the sample as the person separation is 0.62, while the sample number is almost adequate as item reliability shows 0.88 with 2 categories of difficulty levels for the items in this test. Based on this analysis, it can be concluded that the test has no construct validity.

3.2 Research Design

This research is basically a mixed method validation study, conducted predominantly quantitatively and triangulated qualitatively. Creswell (2007) says that conducting mixed methods research would give a better insight to the problem. It is not merely a combination of quantitative and qualitative data,

but the integration, linking and merging the data to provide a better understanding of the research findings. It gives a multi level perspective of the problem.

Mixed method was chosed for this study as the textual information adds meaning to the statistics, and the numeric data beefs up the precision of the information. It is also understood that the mixed methods works in a complementary manner, i.e. where there is weakness in a particular method, the other method is seen as a strength to compensate what is lacking.

This study was divided into two parts, the development of the test and validation. If the data is tested for reliability, it does not necessarily have to spell validity. However, when a certain validity is obtained, it is understood that the outcome is reliable. The use of the conceptual framework allows these two parts, reliability and validity, to be addressed simultaneously. One of the methods was the use of statistical packages to analyse the test scores quantitatively and corresponding interpretations were obtained. Meanwhile the other method used was in the form of interview the subject matter experts (SME) to help make decisions based on the statistical interpretation. For example, if the coefficient was low, but the SME suggests to retain the item, then the item is retained. This human factor intervention is important as the subject for this study are students.

3.3 Population and Sample of the Study

The population for this study is the entire group of shortlisted candidates who turn up for the interview and entrance test for the TESL Foundation Programme. These candidates are 18 year-old Bumiputera who have just obtained their SPM results and are shortlisted based on the requirements set by

the University Central Unit better known with the acronym, UPU (*Unit Pusat Universiti*), which is, a pass in the SPM or its equivalent (approved by the Malaysian government) with at least 5 credits, including in Malay Language (*Bahasa Melayu*) and Mathematics or Additional Mathematics and at least a Grade 2A or A- in the English Language (https://asasi.public university.edu.my/ v4/index.php/ programmes/asasi-tesl).

The purposive sampling technique was applied in this study. The sample was taken according to the cohorts for each year, who were placed at 3 different campuses: Shah Alam, Melaka and Kuantan. The sample for the first test was a total of 120 students for the 2010 batch (Sample A), with 69% females and 31% males. The improvised version 2 was administered to the next cohort in 2011 with a sample of 285 students (Sample B), with 73% females and 27% males. The final improvised version 3 was given to the 2013 cohort with a sample of 285 students (Sample C), with 69% females and 31% males. Figure 3.1 is the demographic profile of the respondents.

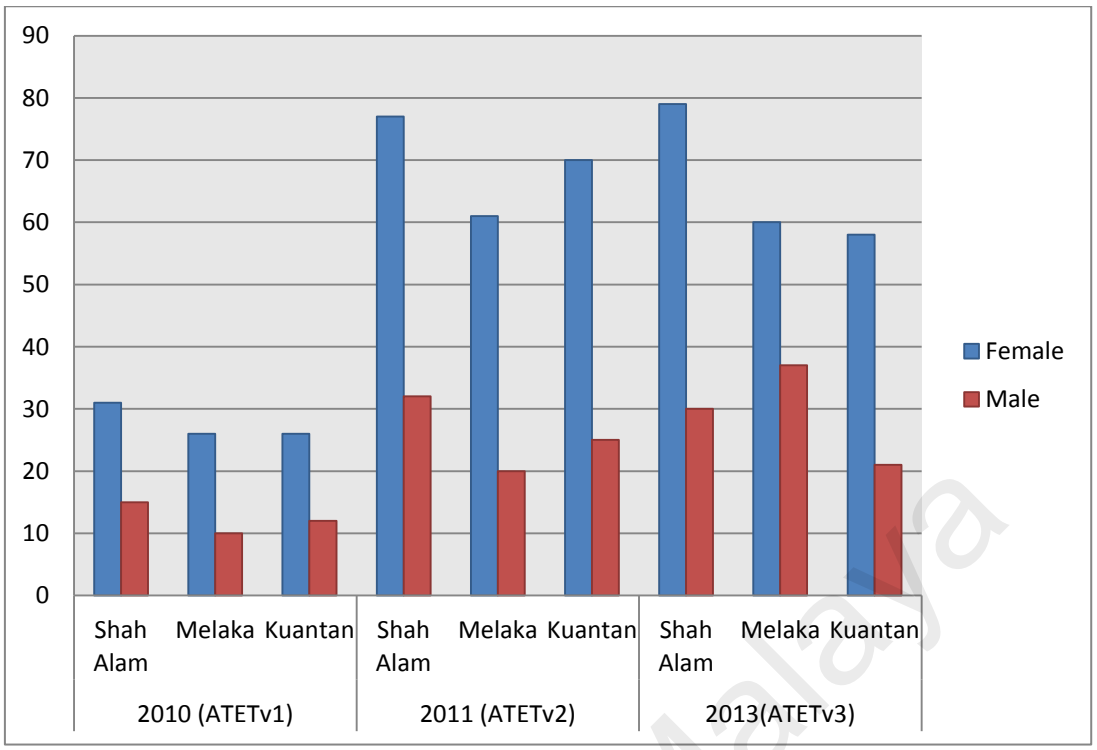

Figure 3.1. Demographic pofile of the samples

These students upon getting through the interview and entrance test were randomly placed in either one of the three campuses, not according to the proximity of their homes to the given campus. Therefore, the placement of the students did not have any barrier on this study.

To address the fairness of the test used in this study, the test was given to the respective samples at the end of the orientation programme organised by the university, i.e. before classes began to avoid the interference of newly learnt knowledge which might alter the expected results.

3.4 Sampling Method

This study employed the purposive sampling method. This was based on the nature of the study. This type of sampling is known as judgmental and discriminatory and sampling ratio is not of any interest. One of the seven

types of purposive sampling method is the Total Population Sampling. The entire population that have similar features and due to the limited number, the entire population was studied.

3.5 Instrument of the Study

The main instrument for this study is a English Language test, ATET, which is hoped to be able to discriminate the sample in order to select candidates for the TESL Foundation course. A table of specifications for the test was developed. This blueprint was used closely with the SPM English list of vocabulary to determine the items to be constructed. To conduct this study, several tests were refered to, particularly high stake standardised tests. Among them was the Malaysian University Entrance Test (MUET) and the Scholastic Aptitude Test (SAT) verbal aptitude test. The format of the questions were adapted from the SAT and MUET. Document analysis was done to get a clearer picture of the types of items that can be tested. Then when items were generated, a checklist together with the testpaper were given to the SME who verified the items. Then interviews were conducted to find out more about their comments. This was done for the MCQ part of the test (60 items all together). For the Essay section, the rubrics for the essay was given to the three raters who scored the essays.

All these items were developed using a checklist adapted from Brown (2005) (see Appendix A) according to the format of the items. Candidates have one hour to complete the test. As this test had three versions, each version is explained separately.

3.5.1 Checklist

The checklist used for this study was mainly to develop the test. This checklist was also given to the SMEs to ensure the items were generated accordingly. This checklist was adapted from Brown (2005) in Appendix A. This checklist focused on the process of constructing a test, satring with eight items on the receptive response items. The next part were four items to check the MCQ format in terms of stem, distractors and answer. The third part was for the Cloze Test which encompassed the filling in the blanks type. The final part of the checklist was on the Essay section.

3.5.2 Interview

The questions that were asked followed a general protocol, but in the midst of the interview, the SMEs had different views on different sections. Thus the questions were changed according to their answers. The basic questions were about their working experience, their educational background, and then about the test. Each section was probed in detail. (See Appendix J for details of the notes).

3.5.3 ATETv1

The ATET version 1 (ATETv1) was used as the pilot test (Refer to Appendix B for Table of Specification and Appendix C for ATETv1). Questions included 30 multiple choice questions which were 10 reading comprehension items, 10 items in the cloze test (1 passage) and 10 grammar items.

The Reading Comprehension section had 2 passages. The first

passage is about the experience of a teacher, which is narrative in genre. The passage was adapted from The Sunday Star article. This passage has 4 questions. Meanwhile the second passage is about one of the predicaments faced by the *Orang Asli* (indigenous people) in Malaysia. This argumentative type article was adapted from a press statement from the internet and has 6 multiple choice questions. The items for this Reading Comprehension section were designed to test skills like main idea, supporting details, inference and vocabulary, which was in line with the MUET format. The cognitive domain covered in this section were Comprehension, Application, Analysis and Evaluation in accordance to the Bloom's Taxonomy (see Appendix B for details).

The second section was a cloze passage with 10 blanks. The passage was about a personal experience of a student. The blanks tested mainly vocabulary and grammar. The third section tested grammar which was made up of 5 error identification items and 5 sentence completion items. The items for error identification were at sentence level and had four parts which were underlined. One of the underlined parts was incorrect and needed to be identified. Meanwhile, the sentence completion items focussed on the missing grammar item to be filled in the blank provided. This was constructed according to the SAT format. This test also had an essay question, where students did not have any choice but to answer one expository essay with a simple prompt.

3.5.4 ATETv2

The ATET version 2 (ATETv2) had 50 items all together. This test consists of 4 sections of multiple choice questions, with four options for each answer: A, B, C, and D. The first section is the Reading Comprehension with 20 questions in total. This section is tested through 3 passages, ranging from easy to higher intermediate level of difficulty in ascending order. The first two passages were similar to ATETv1 Reading Comprehension which slight changes in the questions. The third passage is about Flower Remedies and herbal value, adapted from a MUET revision guide. The items for this Reading Comprehension section were also designed like ATETv1 to test skills like main idea, supporting details, inference and vocabulary. The cognitive domain covered in this section is Comprehension, Application, Analysis and Evaluation in accordance to the Bloom's Taxonomy. All the items were formatted to follow the MUET format for the Reading Comprehension.

The second section is the Grammar section with 30 multiple choice items. These items were tested in three parts. The first part is a cloze passage with 10 blanks. The passage is about a personal experience of a student, similar passage as in ATETv1 cloze passage, with focus on grammar. This part is in accordance to the MUET format. The second part is the Sentence Completion, with 5 items. The items require candidates to use mechanics of writing and grammatical clues to complete the sentences. The answers are in phrase level. This part is in accordance to the SAT Sentence Completion format. The third part is Error Identification with 10

items. This part requires candidates to identify which underlined part of the sentence contains an error. These items test grammar rules. This was done according to the SPM English and the SAT format.

The fourth section is Writing with 5 items. This section tests the ability of candidates to edit a short passage in terms of its coherence and cohesion with 5 items. This short passage was adapted from a part of a student's essay on child abuse. The format follows the Paragraph Improvement of the SAT.

The fifth section is the Essay section with one item. The question was an expository type with a short prompt. The prompt was something students can relate to easily (students should not have problems coming up with relevant points for the essay) as the main aim of this section is to gauge if students can write a coherent and cohesive essay with little grammatical mistakes (Refer to Appendix D for Table of Specification and Appendix E for ATETv2).

3.5.5 ATETv3

The ATET version 3(ATETv3) and the final one has an improved Reading Comprehension section (Refer to Appendix F for Table of Specification and Appendix G for ATETv2). Section A has 2 new reading passages. The first one is adapted from https://www.gadoe.org/Curriculum-Instruction-and-Assessment /Assessment/Documents/Grade_ 7_Study_Guide_GA13_Final.pdf which is about the difference in characteristics of peppers and pepper, while the second passage is adapted with permission from: http://teacher.depaul.edu/Documents/TheTrainRideFiction7thgrade.pdf

. The permission via email is in Appendix I. The items for comprehension question, included literary elements, apart from the earlier version of comprehension type items which were main ideas, supporting details, author's perpose, inference, reference and vocabulary items.

Section B, had some changes. The cloze passage was changed in the grammar section , to give a more current exposure of theme, with 10 items. The blanks required different parts of speech like preposition, conjunction, infinitive, modal and nouns. The sentence completion items were added to a total of 10 items. These items tested the knowledge of sentence structure and contextual clues to determine the answers which dealt with subject and predicate, conditionals, pronouns, subject verb agreement, and relative clause. Menwhile the 10 error identification items were retained and these were related to subject-verb agreement, superlatives and comparatives, pronouns and nouns, preposition, gerund and infinitive as well as determiners.

Section C, the Writing section had 10 items. The passage on child abuse was retained with items testing coherence and cohesion through comma splice, paraphrase, compound sentence, run-ons and transitions. 5 new items were added to assess the ability to revise sentences (at sentence level) through capitalization, direct/indirect speech, comma splice,run-ons and compound sentence. The format was in accordance to the Sentence and Paragraph Improvement sections of the SAT.

The answers to all the the three versions of the test is in Appendix H. The differences in each version is summarised in Table 3.2.

Table 3.2
The Differences in Content for Three Versions of the Test

Test Versions	ATET v1(Pilot)	ATET v2	ATET v3
No of Items	30	50	60
Reading Comprehension (no of passages & items)	2 passages (10 items)	3 passages (20 items)	2 passages (20 items)
Grammar	Cloze Test (10 items) Error Identification (5 items) Sentence Completion (5 items)	Cloze Test (10 items) Error Identification (10 items) Sentence Completion (5 items)	Cloze Test (new passage - 10 items) Error Identification (10 items) Sentence Completion (10 items)
Writing	1 essay question	Paragraph Improvement (5 items) 1 essay question	Sentence Improvement (5 items) Paragraph Improvement (5 items) 1 essay question

3.6 Reliability and Validity of the Instruments

Data analysis was done for all versions whereby using the conceptual framework, the Winsteps computer software was utilized. Generally, the Cronbach Alpha was obtained for each version, ATETv1 was 0.62, ATETv2 was 0.65 and finally ATETv3 was 0.8. As noticed, the reliability became stronger as the number of items were increased and refined in difficulty levels. The analysis also showed high inter-rater reliability for the essay component as well as strong construct and predictive validity for the MCQs. The details of the results are in Chapter 4 for reliability and validity of the instuments.

3.7 Procedure of the Study

This study was conducted over four years. It was a tedious process as it involved construction of items according to the Rasch Model, which was labelled as version 1@pilot study. The enhancement was done and version 2 was administered with a bigger target group with revised items. Although the target was the entire group who attend the interview session, but due to certain constraints, the data was not allowed to be collected from this group. So it was administered to the group after they had gained admissions into the programme , at the end of the orientation week before classes began. This test items were verified by the subject matteer experts who gave some good insight over the items and areas of the test. Finally version 3 was carried out after a revamp of items and areas, based on the data analysis results of ATETv2 and subject matter experts' advice.

3.7.1 Item Generation

The construction of the test started with some discussions among the faculty members who were responsible for the TESL Foundation programme as well as the panel for the selection process. A table of specification was constructed and after approval from the designated faculty members, the test items were developed (see Appendix B). The approach used to develop the items is based on the Item Response Theory (IRT). The test, which initially had about 30 items was administered on 120 candidates in the three campuses for the pilot study, to ensure the items were clear, no ambiguity and to determine the difficulty level of the items. The data was analysed using the Winsteps software. The Infit MNSQ was examined and items were

either dropped, or modified to construct the ATET. The items were also reworded to mock the SPM format, which is more familiar to the candidates. Finally, some items were dropped and some revised based on the data analysis. There was one question for the Essay section, which was scrutinized for clarity in the question as well as inter-rater reliability. Thus the ATET version 2 (ATETv2) was constructed. This test is triangulated by expert advice. This was done by asking a panel of experts in the faculty to check the final content of the test with a checklist (see Appendix A) as well as the information was further verified through an interview.

The ATETv2 was administered on the sample in the three campuses to answer the research questions that have been posed. The Rasch Model was employed to describe and explain the items in the ATETv2, using the Item Characteristic Curve, Infit MNSQ, Item-Person map and difficulty level.

An improvement to the ATETv2 was ATET version 3 (ATETv3). Modification was done to the items after considering the four fit statistics as well as subject matter expert's advice.

3.7.2 Test Validation

To ensure the test is reliable, to test its stability, the Summary Statistics in Winsteps is focussed. The Cronbach Alpha (KR-20) Person Raw Score reports the reliability factor. The test validation process will employ the model as mentioned in Chapter 2, Figure 2.1. This model has four steps: observation, generalization, explanation and extrapolation.

The process begins with the observation inference. This is where the raw scores are converted into measured scores and ability, which is the standardization of the scoring process. This ensures the unanimity of the scoring procedure. Conversion of raw scores to interval or measured scores in the Rasch analysis is necessary as the distance between measured scores is real and item difficulty can be directly compared with person ability or trait levels. The data is analysed using Winsteps software, where θ will denote the person ability and trait level in logit.

The next step is to generalize the observed scores into expected scores for person and item reliability, and person and item separation indexes are scrutinized. The theories behind these will the basis for the generalisation. This is done using Winsteps software where the person and item separation indexes is reported.

The third step is the explanation inference. This is the theoretical construct under measurement. Through the data analysis, item and person infit and outfit estimates are studied. This is explained with the theoretical concepts of fit validity. The study of the item and person fit provides information about construct-irrelevant factors, where the residuals is analysed. All of these are provided by the analysis in the Winsteps software.

The final step is the extrapolation inference. This part will utilize Kane's (1992) criterion-referenced evidence. A Multiple Regression, using SPSS was done among ATETv3 and MUET This is done in two ways. As there should be external consideration, the scores from the test is correlated with the GPA for the first semester for three subjects,

Reading, Writing and Grammar. and the MUET results (which tests similar areas as the ATETv3). This also to conclude predictive validity.

3.7.3 Essay Component

The raters were given the rubrics to mark the essay scripts. This was the rubrics used in the TESL Foundation programme. There were three raters and each one gave their scores separately. The facets, raters and person (candidates), were the two facets used in the MFRM analysis using a computer software, FACETS. In addition, the EduG software was utilized to decide on the optimum number of raters and inter-rater reliability.

3.8 Summary

This study has two parts: the first being the development of the test and the second, the validation of the test. The development of the test went through three rounds of improvisation. Data was collected from the existing cohorts of 2010, 2011 and 2013 batches, and analyzed mainly using Winsteps, FACETS, EduG and SPSS.

CHAPTER 4

RESULTS

4.1 Introduction

The chapter will present the findings according to the procedure of the item construction and validation as suggested by Aryadoust (2009). It will present the data analysed for each version of the ATET and how it lands itself with the final version for the multiple choice questions (MCQ). At each stage the psychometric properties will be shown as empirical evidence and triangulated by Subject Matter Experts' advice. Following this is the presentation of the results for the Essay section. This is to ensure the optimal number of raters needed for the writing assessment. The results from both the MCQ and Essay sections will then be related to the research questions that were posed in Chapter 1. Apart from this, the test results, i.e. the final stage (ATETv3) will be correlated with the SPM, MUET and the Final Semester examination (GPA). This is to establish that there is extrapolation evidence. Finally, there will be a summary of the results according to the Research Questions (RQ). The answers to the RQs will be provided at the end of this chapter. The data presented will be accompanied with a brief description.

4.2 ATET version 1 (ATETv1)

This test scores were analysed to see if the items and persons were reliable and then checking for the validity.

4.2.1 Summary Statistics

The number of persons involved in this first part of the study was 120 and there were 30 items all together. The Data Points for ATETv1 are 3600 (product of persons and items). Table 4.1 shows the person reliability is 0.58, which is lesser than 0.8. This indicates that the number of items are not enough to gauge the ability of the persons and/or suggests improvement of items and/or distractors. This is further confirmed by the person stratification, $H_p = (4G+1)/3$, with G being the Person Separation. ATETv1 has $H_p = [4(1.17)+1]/3 = 1.89$. This implies that there are less than 2 categories of persons. More items may be needed to be able to discriminate the low and high achievers.

Table 4.1
Summary Statistics of ATETv1

SUMMARY OF 120 MEASURED Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	21.1	30.0	1.68	.54	1.00	-.1	1.11	.1
S.D.	3.1	.0	.92	.05	.46	1.6	1.55	.9
MAX.	28.0	30.0	4.11	.79	2.56	3.9	9.90	3.8
MIN.	13.0	30.0	-.46	.50	.36	-2.9	.22	-.9
REAL RMSE	.60	ADJ.SD	.70	SEPARATION	1.17	Person RELIABILITY		.58
MODEL RMSE	.54	ADJ.SD	.74	SEPARATION	1.36	Person RELIABILITY		.65
S.E. OF Person MEAN = .08								

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .62

SUMMARY OF 30 MEASURED Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	84.3	120.0	.00	.39	.98	.0	1.20	.4
S.D.	33.3	.0	2.15	.27	.18	1.8	.71	2.1
MAX.	119.0	120.0	3.11	1.01	1.52	4.9	3.25	6.8
MIN.	27.0	120.0	-3.48	.20	.65	-3.5	.18	-3.4
REAL RMSE	.48	ADJ.SD	2.09	SEPARATION	4.38	Item RELIABILITY		.95
MODEL RMSE	.47	ADJ.SD	2.10	SEPARATION	4.42	Item RELIABILITY		.95
S.E. OF Item MEAN = .40								

UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -.95
3600 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 2625.79 with 3451 d.f. p=1.0000

However, the item reliability is 0.95 which shows according to Winsteps (http://winsteps.com/winman/reliability.htm) that the number of persons is acceptable (more than 0.8) and large enough to confirm the item difficulty (=construct validity) of the instrument.. The Item Separation=4.38 (more than 3) indicates the sample selection is good enough to determine the item difficulty hierarchy.

There were no missing data. Missing data: if some persons have missing observations, these can considerably reduce precision, and so lower reliability estimates. Suggestion: omit person-records with missing data when estimating reliabilities (Linacre, 2014). Overall, the reliability factor of ATETv1 Cronbach Alpha is 0.62, which is lesser than the acceptable factor, 0.8. Person (sample, test) reliability depends chiefly on:

a) Sample ability variance. Wider ability range = higher person reliability. Thus person reliability is low.

b) Length of test (and rating scale length). Longer test = higher person reliability. From Table 8, person reliability is low.

c) Number of categories per item. More categories = higher person reliability. This indicates person reliability is low.

d) Sample-item targeting. Better targeting = higher person reliability

This further establishes person reliability is low.

Item reliability depends chiefly on:

a) Item difficulty variance. Wide difficulty range = high item reliability

Table 4.1 claims the item reliability is high.

b) Person sample size. Large sample = high item reliability

The study confirms the sample is large enough. Therefore, the conclusion drawn from the pilot test is that the number of items has to be increased but the sample selection can be retained. This requires another scrutiny of the Table of Specifications.

4.2.2 Variable Map

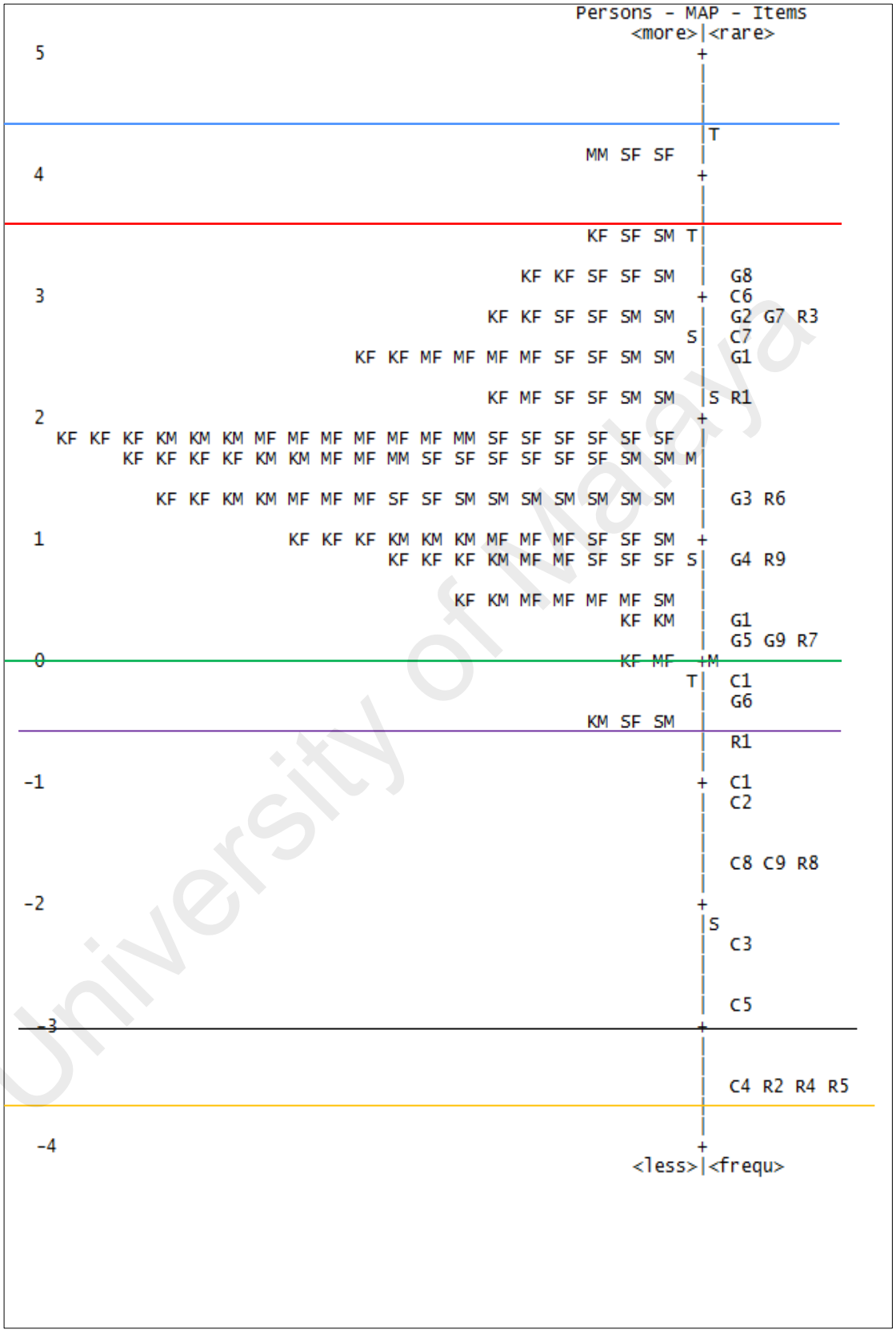

Figure 4.1. Variable map of ATETv1

The persons seems to have a normal distribution. The items do not have such distribution but they are varied in difficulty levels, easiest at the bottom of the scale (C4, R2, R4 and R5) while the most difficult at the top of the scale, G8. Looking at the persons, the maximum measure is 4.11 logit while the minimum measure is -0.46 logit. The measurement scale for person is 4.11+0.46= 4.57. While the maximum measure for items is 3.11 logit and minimum is -3.48. The measurement scale for items is 3.11+3.48 = 6.59. The difference between the item scale and person scale is 6.59-4.57 = 2.02. Notice, there is a gap between R1 and R6 and another gap above W8. This shows that the items are poorly defined. Thus items have to be added to fill up the gaps in the scale. The green line that is drawn at 0 logit is the mean. There are quite a number of easy items (below the green line) while there are only a few persons below this line. The bulk of the persons are above average. Thus there must be some new items that are between 0 and +4 logits.

4.2.3 Item Fit

Table 4.2
Item Fit Statistics of ATETv1

Person: REAL SEP.: 1.17 REL.: .58 ... Item: REAL SEP.: 4.38 REL.: .95

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
2	119	120	-3.48	1.01	1.03	.4	3.25	1.7	A -.08	.07	99.2	99.2	R2
19	114	120	-1.61	.43	.94	-.1	2.86	2.4	B .20	.17	95.0	95.0	C9
13	117	120	-2.35	.59	1.04	.3	2.49	1.6	C .01	.13	97.5	97.5	C3
30	40	120	2.47	.21	1.52	4.9	2.14	6.8	D -.24	.38	59.2	71.8	G10
12	111	120	-1.15	.36	.92	-.2	2.10	2.0	E .24	.21	92.5	92.5	C2
26	103	120	-.39	.27	1.08	.5	1.43	1.4	F .13	.27	85.0	85.9	G6
8	114	120	-1.61	.43	1.04	.2	1.33	.7	G .11	.17	95.0	95.0	R8
25	95	120	.13	.24	1.03	.3	1.32	1.4	H .23	.31	81.7	80.0	G5
7	93	120	.24	.23	1.09	.8	1.31	1.4	I .19	.31	78.3	78.5	R7
23	68	120	1.35	.20	1.16	2.3	1.22	2.1	J .20	.37	59.2	65.9	G3
21	90	120	.39	.22	1.00	.0	1.21	1.1	K .28	.32	77.5	76.4	G1
10	46	120	2.22	.20	1.19	2.2	1.19	1.7	L .20	.38	59.2	69.3	R10
3	31	120	2.90	.23	1.16	1.4	1.19	1.1	M .20	.37	75.0	76.9	R3
6	69	120	1.31	.20	1.08	1.2	1.16	1.5	N .27	.37	65.0	66.1	R6
24	80	120	.85	.21	1.00	.1	1.16	1.1	O .31	.35	75.8	70.3	G4
1	106	120	-.63	.30	1.10	.5	1.12	.4	O .14	.25	88.3	88.3	R1
18	114	120	-1.61	.43	1.00	.1	1.10	.4	n .16	.17	95.0	95.0	C8
9	82	120	.77	.21	1.08	.9	1.02	.2	m .28	.35	68.3	71.3	R9
4	119	120	-3.48	1.01	1.01	.3	.62	.0	l .09	.07	99.2	99.2	R4
14	119	120	-3.48	1.01	1.01	.3	.62	.0	k .09	.07	99.2	99.2	C4
29	95	120	.13	.24	.95	-.4	.97	-.1	j .34	.31	81.7	80.0	G9
15	118	120	-2.77	.72	.96	.2	.61	-.2	i .18	.10	98.3	98.3	C5
5	119	120	-3.48	1.01	.94	.3	.18	-.8	h .21	.07	99.2	99.2	R5
20	109	120	-.92	.33	.92	-.2	.79	-.4	g .31	.22	90.8	90.8	C10
11	100	120	-.18	.26	.88	-.7	.88	-.4	f .39	.28	85.8	83.7	C1
17	36	120	2.66	.22	.69	-3.4	.59	-3.3	e .69	.38	82.5	74.0	C7
16	30	120	2.95	.23	.67	-3.1	.54	-3.1	d .70	.37	84.2	77.5	C6
28	27	120	3.11	.24	.66	-2.9	.51	-3.0	c .70	.36	86.7	79.3	G8
27	33	120	2.80	.22	.66	-3.5	.54	-3.4	b .72	.37	85.0	75.6	G7
22	31	120	2.90	.23	.65	-3.5	.52	-3.4	a .72	.37	83.3	76.9	G2
MEAN	84.3	120.0	.00	.39	.98	.0	1.20	.4			84.1	83.6	
S.D.	33.3	.0	2.15	.27	.18	1.8	.71	2.1			12.3	11.0	

The results for Item Fit is seen in Table 4.2. Judging from the decision made according to Linacre (2015), the acceptable range for Infit Mean Square(MNSQ) is between 0.8 to 1.2. However, from this table an acceptable range can be calculated from the Total Infit MNSQ ± S.D. In Table 4.2, these figures are seen at the bottom of the table, in a dotted box

For ATETv1, the Infit MNSQ range is 0.98 ± 0.18, which is between 0.8 – 1.16. The items out of this range is circled in blue in the Infit MNSQ column, which are overfitting. Next is the z-score, ZSTD which should be between -2 and 2. The items out of this range is circled in green. Then, the Point Measure Correlation, PT-Measure Corr, is examined. The negative measures indicates that the responses are

opposite. The measures less than 0.2 shows that item might be misleading. These are circled in orange.

Upon examining this table, items that have all the above-mentioned deficiencies are taken out (in the red boxes) of the next version. The ones with the purple boxes and the others with lesser problems were scrutinised further. They require modification. The items that are overfitting, have no practical implication (Bond & Fox, 2007), thus can be taken off the next version.

Table 4.3
Item Fit Statistics: Measure Order of ATETv1

Person: REAL SEP.: 1.17 REL.: .58 ... Item: REAL SEP.: 4.38 REL.: .95

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
28	27	120	3.11	.24	.66	-2.9	.51	-3.0	.70	.36	86.7	79.3	G8
16	30	120	2.95	.23	.67	-3.1	.54	-3.1	.70	.37	84.2	77.5	C6
3	31	120	2.90	.23	1.16	1.4	1.19	1.1	.20	.37	75.0	76.9	R3
22	31	120	2.90	.23	.65	-3.5	.52	-3.4	.72	.37	83.3	76.9	G2
27	33	120	2.80	.22	.66	-3.5	.54	-3.4	.72	.37	85.0	75.6	G7
17	36	120	2.66	.22	.69	-3.4	.59	-3.3	.69	.38	82.5	74.0	C7
30	40	120	2.47	.21	1.52	4.9	2.14	6.8	-.24	.38	59.2	71.8	G10
10	46	120	2.22	.20	1.19	2.2	1.19	1.7	.20	.38	59.2	69.3	R10
23	68	120	1.35	.20	1.16	2.3	1.22	2.1	.20	.37	59.2	65.9	G3
6	69	120	1.31	.20	1.08	1.2	1.16	1.5	.27	.37	65.0	66.1	R6
24	80	120	.85	.21	1.00	.1	1.16	1.1	.31	.35	75.8	70.3	G4
9	82	120	.77	.21	1.08	.9	1.02	.2	.28	.35	68.3	71.3	R9
21	90	120	.39	.22	1.00	.0	1.21	1.1	.28	.32	77.5	76.4	G1
7	93	120	.24	.23	1.09	.8	1.31	1.4	.19	.31	78.3	78.5	R7
25	95	120	.13	.24	1.03	.3	1.32	1.4	.23	.31	81.7	80.0	G5
29	95	120	.13	.24	.95	-.4	.97	-.1	.34	.31	81.7	80.0	G9
11	100	120	-.18	.26	.88	-.7	.88	-.4	.39	.28	85.8	83.7	C1
26	103	120	-.39	.27	1.08	.5	1.43	1.4	.13	.27	85.0	85.9	G6
1	106	120	-.63	.30	1.10	.5	1.12	.4	.14	.25	88.3	88.3	R1
20	109	120	-.92	.33	.92	-.2	.79	-.4	.31	.22	90.8	90.8	C10
12	111	120	-1.15	.36	.92	-.2	2.10	2.0	.24	.21	92.5	92.5	C2
8	114	120	-1.61	.43	1.04	.2	1.33	.7	.11	.17	95.0	95.0	R8
18	114	120	-1.61	.43	1.00	.1	1.10	.4	.16	.17	95.0	95.0	C8
19	114	120	-1.61	.43	.94	-.1	2.86	2.4	.20	.17	95.0	95.0	C9
13	117	120	-2.35	.59	1.04	.3	2.49	1.6	.01	.13	97.5	97.5	C3
15	118	120	-2.77	.72	.96	.2	.61	-.2	.18	.10	98.3	98.3	C5
2	119	120	-3.48	1.01	1.03	.4	3.25	1.7	-.08	.07	99.2	99.2	R2
4	119	120	-3.48	1.01	1.01	.3	.62	.0	.09	.07	99.2	99.2	R4
5	119	120	-3.48	1.01	.94	.3	.18	-.8	.21	.07	99.2	99.2	R5
14	119	120	-3.48	1.01	1.01	.3	.62	.0	.09	.07	99.2	99.2	C4
MEAN	84.3	120.0	.00	.39	.98	.0	1.20	.4			84.1	83.6	
S.D.	33.3	.0	2.15	.27	.18	1.8	.71	2.1			12.3	11.0	

From Table 4.3, the items that have similar measures suggest they are of the same difficulty level and might be testing the same construct. These are items in the coloured boxes. Items 3 and 22 (red box) seem to have the same measure but they are from different components. Item 3 tests on Reading while Item 22 tests on Grammar. But after looking at

the earlier table, Item 22 seem to be problematic. Item 3 can be retained, but item 22 must be checked qualitatively. The placement of the adverb in item 22 seems to be confusing as suggested by the distractors.

Items 25 and 29 share the same measure (blue box). Both are testing Grammar, but they are testing in two different sections. Item 25, Sentence Completion, seems alright and can be retained, but Item 29, Error Identification, seems a bit problematic and needs to be checked. Items 8, 18 and 19 have the same measure (green box). Item 8 is on Reading, seems not much of a problem, as such it can be retained. Items 18 and 19 are from the Cloze and one of these items should be dropped.

Items 2, 4, 5, 14 (purple box) also have the same measure. Items 2, 4, 5 test Reading. Items 2 and 4 are from the same passage while Item 5 is from a different passage. Item 2 also has a negative point measure correlation, so this will be dropped. Item 4 seems to have some problems and need further investigation. While Item 5 can be retained. Item 14 from the grammar component has been highlighted in Table 10. This item can be dropped or modified.

4.2.4 Principal Component Analysis

Table 4.4
Principal Component Analysis of ATETv1

CONTRAST 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --			Modeled
Total raw variance in observations	=	54.8	100.0%		100.0%
Raw variance explained by measures	=	24.8	45.3%		44.5%
Raw variance explained by persons	=	5.4	9.9%		9.8%
Raw Variance explained by items	=	19.4	35.3%		34.7%
Raw unexplained variance (total)	=	30.0	54.7%	100.0%	55.5%
Unexplned variance in 1st contrast	=	5.4	9.8%	17.8%	

The Principal Component Analysis is to ensure unidimensionality. This analysis needs a minimum of 40% 'raw variance explained by measures' as a yardstick to unidimensionality of the instrument. In ATETv1, the raw variance explained by measures (orange box) is 45.3%, which fulfills the minimum requirement. In fact this is more than the modelled value, 44.5%.

Next is the unexplained variance in the 1st contrast should not be more than 15% (Linacre, 2015a). Table 4.04 (blue box) shows 9.8%. This indicates the noise level, which is acceptable.

Table 4.5
Largest Standardized Residual Correlations of ATETv1

RESIDUL CORRELN	ENTRY NUMBER	It	ENTRY NUMBER	It
.92	22	G2	27	G7
.90	16	C6	27	G7
.89	16	C6	22	G2
.88	16	C6	28	G8
.87	12	C2	19	C9
.80	27	G7	28	G8
.77	22	G2	28	G8
.76	17	C7	22	G2
.69	17	C7	27	G7
.67	16	C6	17	C7

Following this is the Table 4.5 , which points out the items that are noise makers, with a residual correlations that is more than 0.7. This further confirms the items that may be testing the same thing or may be confusing to the respondents.

Going back to Table 4.4, Items 22 and 27 are both from the Grammar section and Item 22 seems a bit problematic. Item 27 is all right. Items 16 and 27 are from two different sections, Cloze and Grammar respectively. This is similar to the next two pairs, Items 16 and 22, and Items 16 and 28. The Cloze item is testing grammar. As for Items 12 and 19, both from Cloze, and are testing two different grammatical items. Items 27 and 28, however, are from the same section of Grammar. One of these items should be dropped. Regarding Items 22 and 28, Item 22 from Table 4.4 has been suggested to be dropped off the instrument, while retaining Item 28. This is similar to the last pair, Items 17 and 22, where Item 17 can be retained.

4.2.5 Differential Item Functioning (DIF)

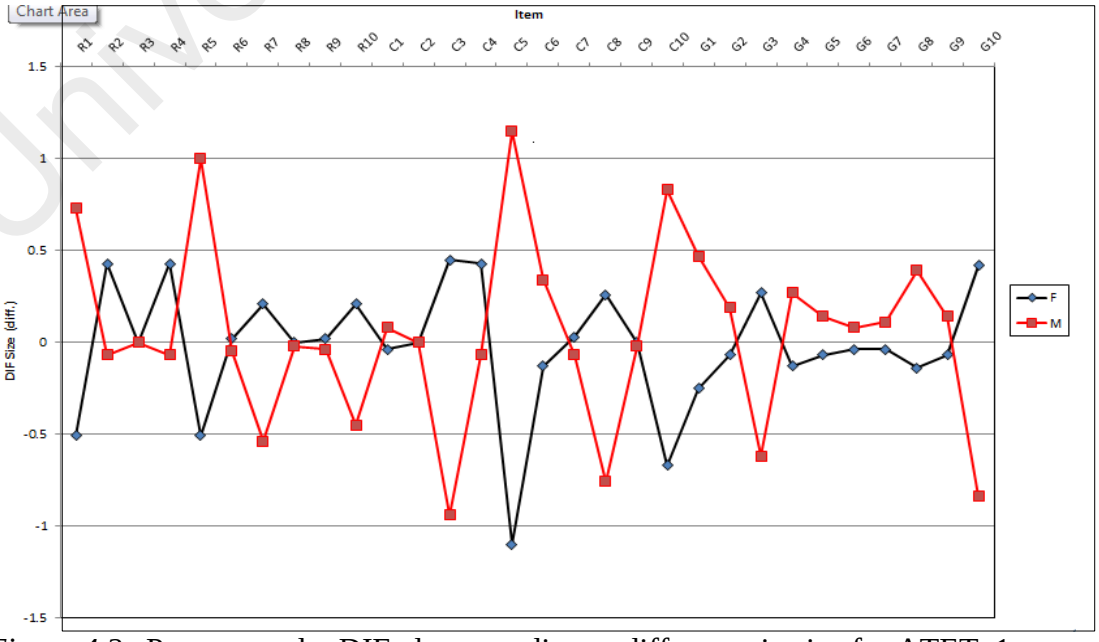

Figure 4.2. Person-gender DIF plot according to difference in size for ATETv1

This DIF analysis inspects the items in a test individually for signs of interactions with sample's gender, i.e. if the items are gender biased. For Figure 4.2, the limit is ±0.5 to see if the difference is significant. There are about 10 items that show this difference, R1, R5, R7, R8, C2, C3, C5, C6, C8 and C9. Take for example, item C5. The difference between female and male responses for this item about 1 and -1, which has exceeded the limit ±0.5. This shows there is a difference in the way the females respond to C5 compared to males. As such, it is recommended to revise these items or drop them.

4.2.6 Person Misfit Order

Table 4.6
Person Misfit Order of ATETv1

Person: REAL SEP.: 1.17 REL.: .58 ... Item: REAL SEP.: 4.38 REL.: .95

Person STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
17	24	30	2.48	.56	1.86	2.6	9.90	3.6	A .14	.52	60.0	84.0	SF017
42	28	30	4.11	.79	1.45	.9	9.90	3.8	B-.17	.30	93.3	93.3	SF042
46	26	30	3.17	.62	1.39	1.3	9.09	2.8	C .16	.43	86.7	86.7	KF046
56	25	30	2.81	.58	1.86	2.6	4.47	1.9	D .09	.47	76.7	84.4	SM056
113	24	30	2.48	.56	2.11	3.2	3.93	1.7	E .07	.52	66.7	84.0	MF113
79	18	30	.81	.51	1.21	.8	3.66	2.6	F .52	.66	76.7	81.6	KF079
104	18	30	.81	.51	1.11	.5	3.14	2.3	G .57	.66	83.3	81.6	KF104
78	21	30	1.60	.53	2.56	3.9	2.80	1.6	H .12	.61	50.0	83.6	KF078
63	20	30	1.33	.52	1.25	.9	2.68	1.6	I .51	.63	76.7	82.8	SM063
70	13	30	-.46	.51	1.62	2.1	2.20	1.8	J .46	.68	70.0	81.3	SM070
40	22	30	1.88	.53	2.18	3.2	1.64	.9	K .25	.58	60.0	84.1	KM040
34	23	30	2.18	.55	2.01	2.9	1.94	1.0	L .23	.55	63.3	84.0	SF034
112	22	30	1.88	.53	2.01	2.8	1.56	.8	M .30	.58	60.0	84.1	KM112
103	22	30	1.88	.53	1.91	2.6	1.47	.7	N .32	.58	66.7	84.1	MF103
116	18	30	.81	.51	1.48	1.6	1.90	1.3	O .48	.66	76.7	81.6	MF116
88	24	30	2.48	.56	1.77	2.4	1.60	.8	P .27	.52	66.7	84.0	MF088
47	24	30	2.48	.56	1.74	2.3	1.26	.6	Q .30	.52	66.7	84.0	KF047
77	23	30	2.18	.55	1.72	2.2	1.67	.9	R .33	.55	70.0	84.0	SM077
89	22	30	1.88	.53	1.66	2.0	1.44	.7	S .39	.58	66.7	84.1	MF089
90	25	30	2.81	.58	1.59	1.9	1.62	.8	T .27	.47	76.7	84.4	KF090
30	24	30	2.48	.56	1.58	1.9	1.20	.6	U .34	.52	66.7	84.0	KF030
68	13	30	-.46	.51	1.23	.9	1.52	1.0	V .58	.68	76.7	81.3	KM068
28	25	30	2.81	.58	1.52	1.7	1.50	.8	W .30	.47	76.7	84.4	SF028
41	26	30	3.17	.62	1.51	1.6	1.46	.8	X .24	.43	86.7	86.7	KF041
49	20	30	1.33	.52	1.50	1.6	1.17	.5	Y .49	.63	70.0	82.8	SM049
31	27	30	3.58	.68	1.39	1.1	1.48	.8	Z .21	.37	90.0	90.0	KF031
BETTER FITTING OMITTED													
55	20	30	1.33	.52	.59	-1.6	.48	-.4	z .75	.63	90.0	82.8	SM055
8	21	30	1.60	.53	.59	-1.6	.35	-.5	y .74	.61	90.0	83.6	MM008
94	22	30	1.88	.53	.57	-1.7	.34	-.4	x .72	.58	93.3	84.1	KF094
96	19	30	1.07	.51	.57	-1.7	.36	-.9	w .78	.64	90.0	82.3	SF096
82	19	30	1.07	.51	.57	-1.7	.35	-.9	v .78	.64	90.0	82.3	KF082
107	19	30	1.07	.51	.57	-1.7	.35	-.9	u .78	.64	90.0	82.3	MF107
13	22	30	1.88	.53	.54	-1.8	.32	-.4	t .73	.58	93.3	84.1	MF013
50	20	30	1.33	.52	.54	-1.9	.35	-.7	s .77	.63	90.0	82.8	KM050
45	21	30	1.60	.53	.53	-1.9	.37	-.5	r .75	.61	96.7	83.6	SF045
53	24	30	2.48	.56	.51	-2.1	.25	-.4	q .67	.52	100.0	84.0	SM053
15	23	30	2.18	.55	.49	-2.2	.27	-.3	p .71	.55	96.7	84.0	SM015
1	22	30	1.88	.53	.49	-2.1	.28	-.5	o .74	.58	93.3	84.1	SF001
32	22	30	1.88	.53	.49	-2.1	.28	-.5	n .74	.58	93.3	84.1	SF032
60	22	30	1.88	.53	.49	-2.1	.28	-.5	m .74	.58	93.3	84.1	KM060
81	22	30	1.88	.53	.49	-2.1	.28	-.5	l .74	.58	93.3	84.1	MF081
23	21	30	1.60	.53	.46	-2.2	.29	-.7	k .77	.61	96.7	83.6	SF023
61	20	30	1.33	.52	.46	-2.3	.29	-.9	j .79	.63	96.7	82.8	KM061
100	20	30	1.33	.52	.46	-2.3	.28	-.9	i .80	.63	90.0	82.8	SF100
83	23	30	2.18	.55	.43	-2.5	.24	-.4	h .73	.55	96.7	84.0	SF083
80	21	30	1.60	.53	.39	-2.7	.24	-.8	g .79	.61	96.7	83.6	KF080
91	21	30	1.60	.53	.39	-2.7	.24	-.8	f .79	.61	96.7	83.6	MF091
2	22	30	1.88	.53	.36	-2.9	.22	-.6	e .77	.58	100.0	84.1	MF002
11	22	30	1.88	.53	.36	-2.9	.22	-.6	d .77	.58	100.0	84.1	SF011
21	22	30	1.88	.53	.36	-2.9	.22	-.6	c .77	.58	100.0	84.1	SF021
101	22	30	1.88	.53	.36	-2.9	.22	-.6	b .77	.58	100.0	84.1	SF101
106	22	30	1.88	.53	.36	-2.9	.22	-.6	a .77	.58	100.0	84.1	KF106
MEAN	21.1	30.0	1.68	.54	1.00	-.1	1.11	.1			84.1	83.6	
S.D.	3.1	.0	.92	.05	.46	1.6	1.55	.9			10.3	2.3	

Similar to the Item Fit analysis, the Person Fit analysis starts off by checking the Infit MNSQ. The acceptable range is Total Infit MNSQ ± S.D., 1.00 ± 0.46, which is between 0.54 and 1.46. The persons who are out of the acceptable range are in the blue box in Table 4.6. Next is the PT-MEASURE CORR, which should be positive. Only Person SF042 has negative measure. This shows that this person is not behaving in the

expected manner. When these persons who did not fit the Infit MNSQ acceptable range and the PT-MEASURE CORR, these persons were deleted from the file, with a PDFILE command. The statistics is seen in

Table 4.7
Summary Statistics after Deleting the Misfit Persons

SUMMARY OF 83 MEASURED Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	20.5	30.0	1.29	.56	1.02	.1	1.00	.1
S.D.	3.3	.0	1.05	.05	.35	1.1	1.13	.8
MAX.	28.0	30.0	4.09	.80	1.85	2.3	9.90	3.5
MIN.	13.0	30.0	-.85	.52	.50	-1.8	.25	-1.1
REAL RMSE	.61	ADJ.SD	.85	SEPARATION	1.38	Person RELIABILITY		.66
MODEL RMSE	.57	ADJ.SD	.88	SEPARATION	1.56	Person RELIABILITY		.71
S.E. OF Person MEAN = .12								

DELETED: 37 Persons
VALID RESPONSES: 99.9%
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .67 (approximate due to missing data)

SUMMARY OF 28 MEASURED Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	54.9	83.0	.00	.40	.97	.0	1.05	.1
S.D.	24.8	.0	2.13	.21	.28	1.7	.90	2.0
MAX.	82.0	83.0	3.32	1.01	1.91	5.2	4.84	6.6
MIN.	13.0	83.0	-3.55	.24	.53	-2.8	.31	-2.9
REAL RMSE	.46	ADJ.SD	2.07	SEPARATION	4.48	Item RELIABILITY		.95
MODEL RMSE	.46	ADJ.SD	2.08	SEPARATION	4.56	Item RELIABILITY		.95
S.E. OF Item MEAN = .41								

MINIMUM EXTREME SCORE: 2 Items
UMEAN=.000 USCALE=1.000

Table 4.7 shows improvement in its person reliablity 0.66 compared to 0.58 in Table 4.1, but it is still not high enough to be acceptable for a high stake test (more than 0.8). As such, revision of items might give a higher person reliability instead of deleting the misfit persons.

4.2.7 Scalogram

The scalogram displays the Person response distribution according to items. Figure 4.3 shows the misfit Persons.

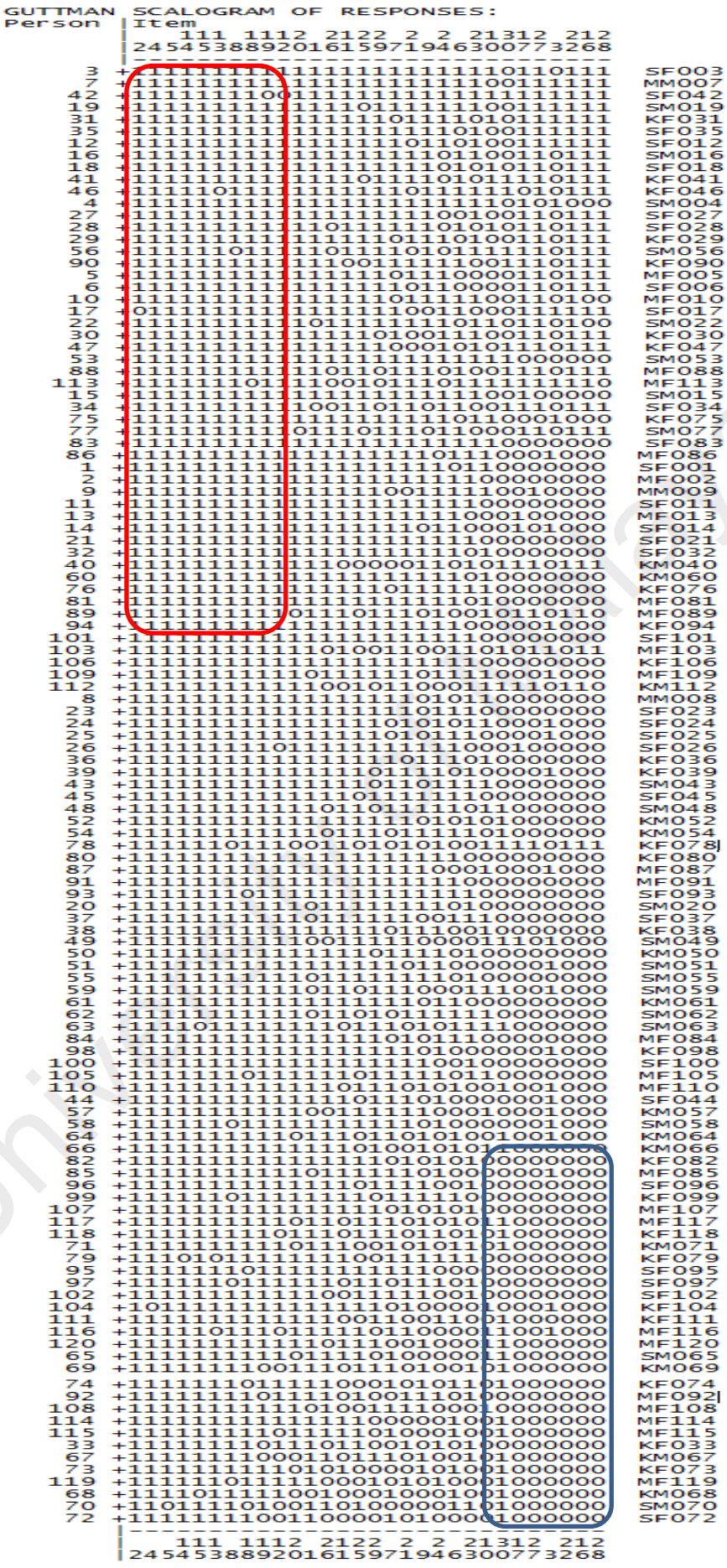

Figure 4.3. Guttman Scalogram of responses for ATETv1

The interpretation of Figure 4.3 is that the persons are ordered from high measure to low measure while the items are also arranged from low to high measure. Thus the Persons has 50% chances of getting most of the easier items correct (1) and 50% chances of getting the more difficult items wrong (0). The top left corner is where the more able persons respond to the easier items (from Figure 4.3, mostly "1"), while the top right corner, there should be more "0", but the responses show "1" (the red box) instead. This should be opposite in the bottom of the scalogram. The bottom right hand corner (blue box) should have more or almost all "0". But this is evident in Figure 4.3. So this clearly shows that there is a discrepency in the results compared to the expected pattern. Items need to be revised.

4.2.8 Item Characteristic Curve

The Item Characteristic Curve (ICC) plots the model-expected item characteristic curve. This is the Rasch model prediction for each measure relative to item difficulty. The steeper the ICC, the more discriminating it is between high and low achievers (Linacre, 2015).

The overall results for all items is seen in Figure 4.4. The horizontal black line show the 50% chances of getting an item correct on ATETv1. Drawing a line parallel to the vertical axis from the interception of the horizontal black line to the coloured sigmoid, the corresponding item location measure is shown on the horizontal axis. It can be seen that the items spread from the easiest being on the far left (R2), with the lowest measure while the most difficult on the far right (G8) with the highest measure.

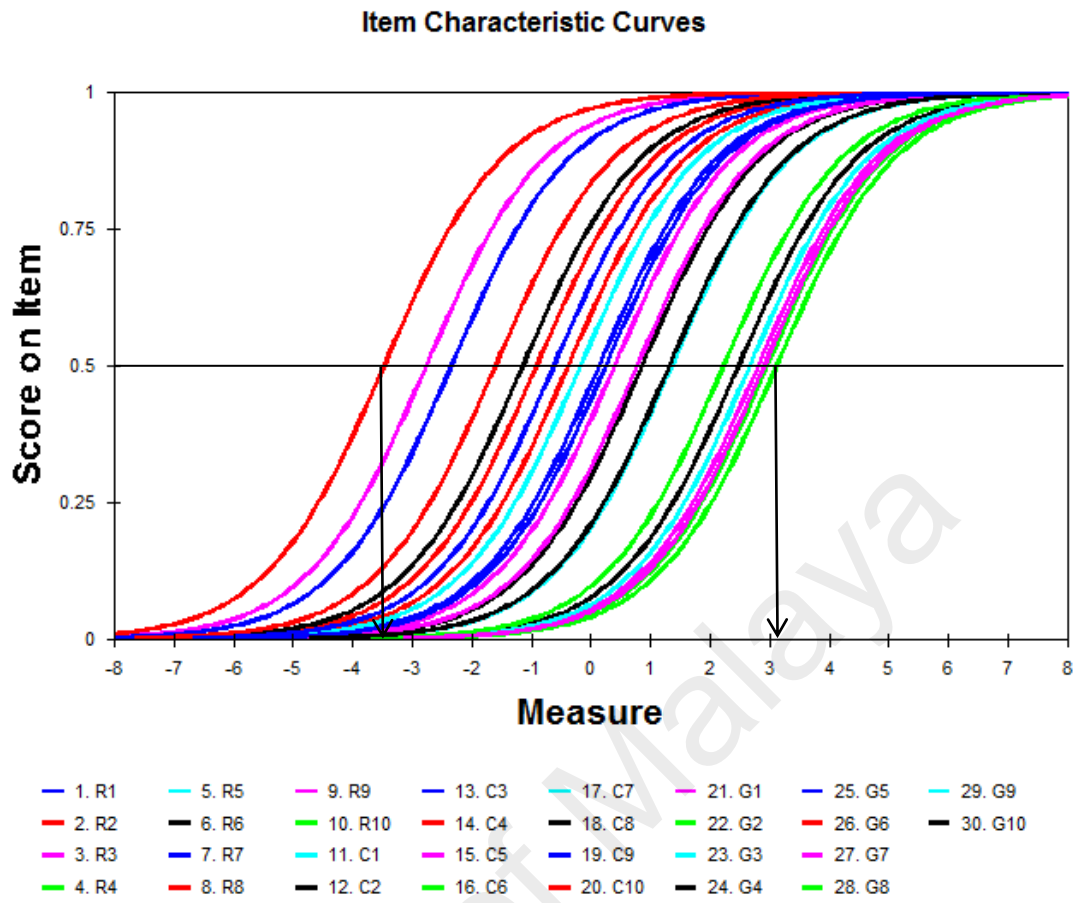

Figure 4.4. Item Characteristic Curve for all items in ATETv1

4.2.9 Bubble Chart

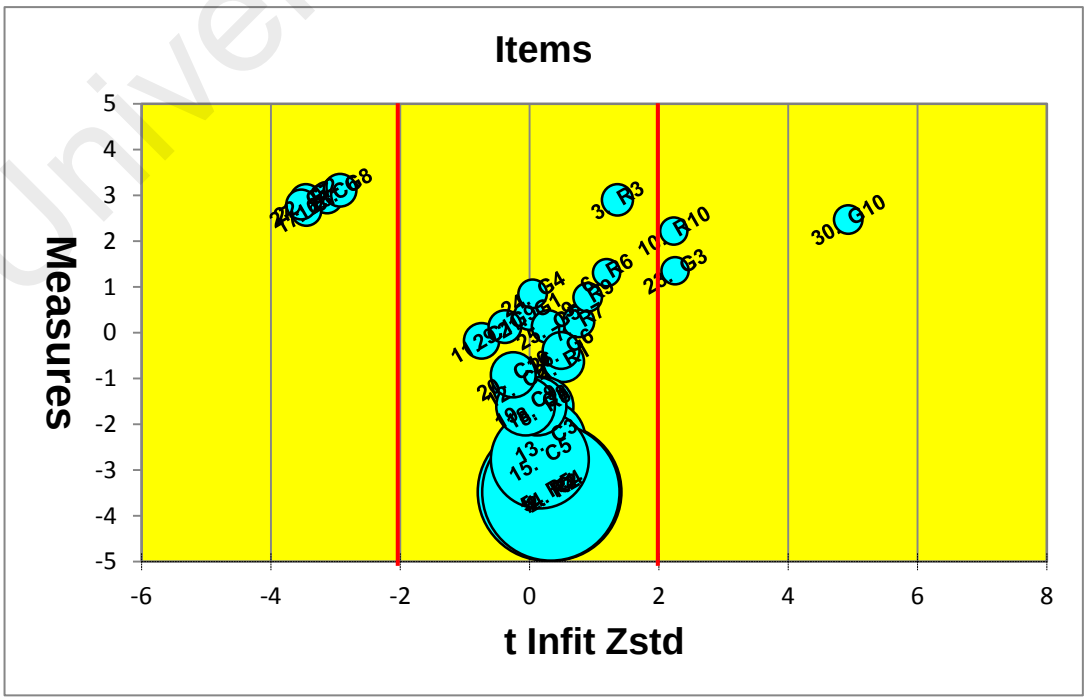

Figure 4.5 Bubble chart of items of ATETv1

Figure 4.5 shows the distribution of items according to their measures. The acceptable range for the z-score along the horizontal axis is between -2 and 2, indicated by two red lines. Generally items are within the range.except for items C6, C7, G2, G7, G8 and G10 which are totally out of the range. It is also noticed some items have huge bubbles as the size of the bubbles indicate the Standard Error of items, the bigger the bubbles, the higher the stadard error. This can be reduced by looking back at the items and making some modifications.

Taking all the results into consideration, ATETv1 was modified with more items added in Reading and Grammar as well as a whole new section on Wrting. The writing section tests writing skills using MCQs. This second version is called ATETv2. The test was given to another batch and the results were analysed with Winsteps. The following are the findings for ATET version 2 (ATETv2).

4.3 ATET version 2 (ATETv2)

This test was constructed based on the analysis of ATETv1 and the format was modelled after SAT and MUET. The following is the findings.

4.3.1 Summary Statistics

Table 4.8
Summary Statistics of ATETv2

SUMMARY OF 285 MEASURED Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	30.1	50.0	.54	.34	1.00	.0	1.00	.0
S.D.	5.2	.0	.68	.07	.16	1.2	.28	1.1
MAX.	49.0	50.0	4.70	1.05	1.44	3.3	2.02	3.1
MIN.	16.0	50.0	-.98	.32	.70	-2.7	.58	-2.1

REAL RMSE	.36	ADJ.SD	.57	SEPARATION	1.59	Person RELIABILITY	.72	
MODEL RMSE	.35	ADJ.SD	.58	SEPARATION	1.67	Person RELIABILITY	.74	
S.E. OF Person MEAN = .04								

Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .65

SUMMARY OF 50 MEASURED Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	171.6	285.0	.00	.15	1.00	.0	1.00	.0
S.D.	62.1	.0	1.21	.04	.07	1.9	.12	1.5
MAX.	272.0	285.0	3.60	.29	1.17	4.9	1.29	3.1
MIN.	16.0	285.0	-2.66	.12	.85	-5.4	.80	-3.2

REAL RMSE	.15	ADJ.SD	1.20	SEPARATION	7.84	Item RELIABILITY	.98	
MODEL RMSE	.15	ADJ.SD	1.20	SEPARATION	7.92	Item RELIABILITY	.98	
S.E. OF Item MEAN = .17								

UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -.99
14250 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 15245.52 with 13916 d.f. p=.0000

The number of persons involved in the second part of the study was 285 and there were 50 items all together. The Data Points for ATETv2 are 14250 (product of persons and items).

The person reliability is 0.72, which is lesser than 0.8, but higher than ATETv1, 0.58. This indicates that the number of items are still not enough to gauge the ability of the persons and/or suggests improvement of items and/or distractors. The person stratification is given as H_p = (4G+1)/3, with G being the Person Separation. ATETv1 has H_p = 1.89, but ATETv2 has H_p = [4(1.59)+1]/3 = 2.45. This implies that there are 2 categories of persons. This test is able to discriminate the low and high achievers.

However, the item reliability is 0.98, which is slightly higher than

in ATETv1, 0.95 and large enough to confirm the item difficulty (=construct validity) of the instrument.. The Item Separation=7.84 (more than 3) indicates the sample selection is good enough to determine the item difficulty hierarchy. There were no missing data. Overall, the reliability factor of ATETv2 Cronbach Alpha is 0.68, which is slightly better than in ATETv1, 0.62, but still not high enough (at least 0.8).

Person reliability depends chiefly on:

1) Sample ability variance. Wider ability range = higher person reliability.

Person reliability is 0.72 is rather low.

2) Length of test (and rating scale length). Longer test = higher person reliability, which is also low

3) Number of categories per item. More categories = higher person reliability, also considered low

4) Sample-item targeting. Better targeting = higher person reliability, also low.

Item reliability depends chiefly on

1) Item difficulty variance. Wide difficulty range = high item reliability.

Item reliability is 0.98, rather high.

2) Person sample size. Large sample = high item reliability, thus sample size is sufficient

Therefore, the conclusion drawn from the pilot test is that the number of items has to be increased but the sample selection and size can be retained. The Table of Specifications has to be revisited.

4.3.2 Variable Map

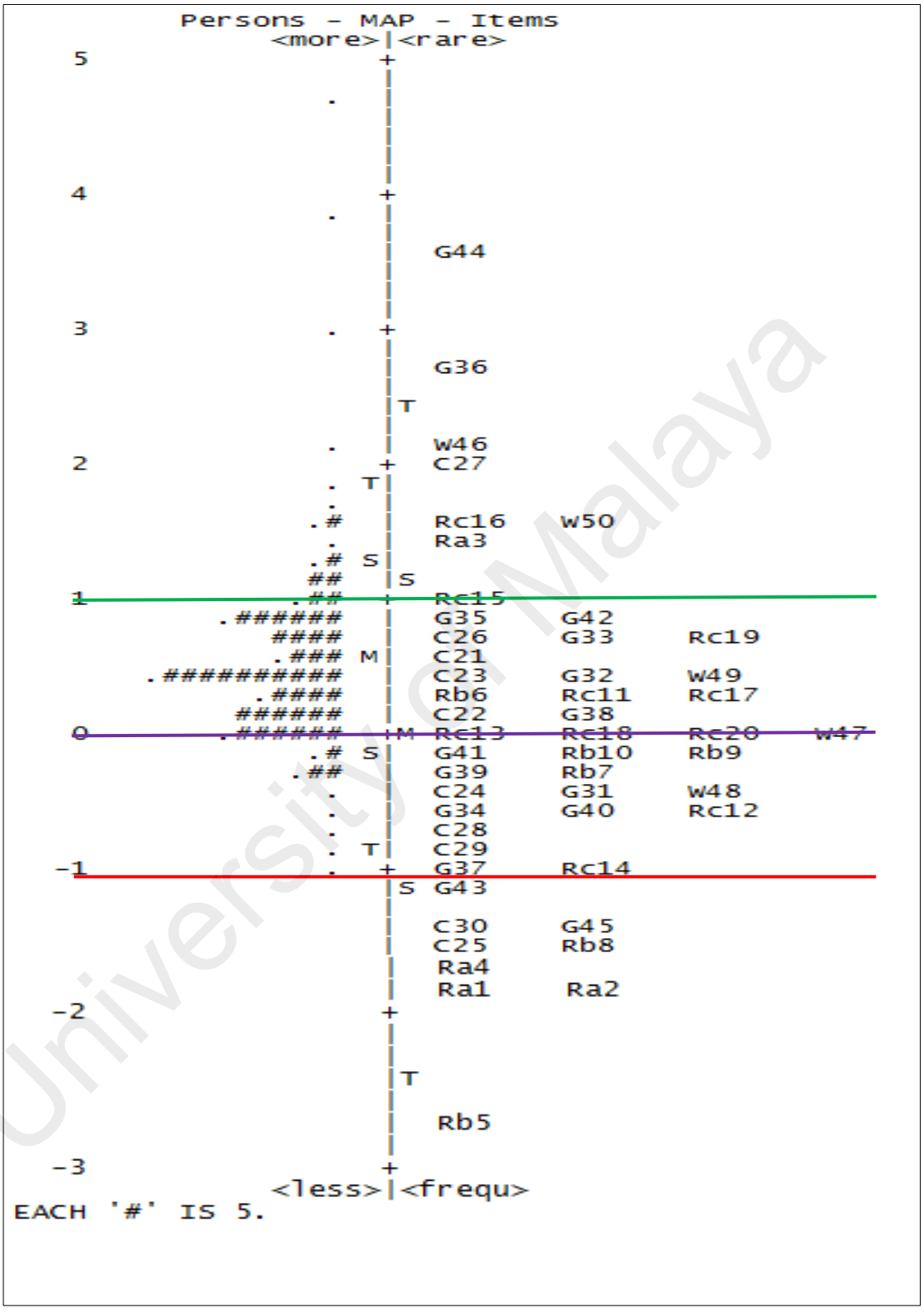

Figure 4.6. Variable map of ATETv2

Both the persons and items seem to have normal distributions. The items are spread from difficult to easy, but with more easy items than difficult ones. Items with logit of less than -1 appear as very easy items as there are no persons matching to this measure (below the red line). Meanwhile the difficult items are limited, logit 2 and above. Here it can be seen that there are persons within this range as well as some who have the measure of ability more than logit 4. There are no items with such high measure. It can be seen that difficult items ranging more than logit 1.5 are only 5 items (above green line): Ra3 and Rc16 (Reading Section, Passage 1 and 3 respectively, Questions 3 and 16), W46 and W50 (Writing Section, Questions 46 and 50), C27 (Cloze Passage, Question 27) and G36 and G44 (Grammar Section, Questions 36 and 44).

Looking at the persons, the maximum measure is 4.7 logit while the minimum measure is -0.98 logit. The measurement scale for person is 4.7+0.98= 5.68. While the maximum measure for items is 3.6 logit and minimum is -2.66. The measurement scale for items is 3.6+2.66 = 6.26. The difference between the item scale and person scale is 6.26-5.68 = 0.58.

The purple line that is drawn at 0 logit is the mean. There are quite a number of easy items (below the green line) while there are only a few persons below this line. The bulk of the persons are above average, just like ATETv1.

4.3.3 Item Fit

Table 4.9
Item Fit Statistics of ATETv2

Person: REAL SEP.: 1.59 REL.: .72 ... Item: REAL SEP.: 7.84 REL.: .98

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
14	231	285	-1.03	.15	1.04	.5	1.29	1.9	A .07	.18	81.1	81.1	Rc14
18	179	285	-.05	.13	1.17	4.1	1.25	3.0	B -.03	.23	62.8	64.5	Rc18
3	89	285	1.37	.13	1.08	1.4	1.23	3.1	C .14	.28	67.4	70.8	Ra3
13	173	285	.05	.13	1.14	3.8	1.21	2.8	D .02	.24	57.9	63.0	Rc13
6	159	285	.26	.12	1.08	2.6	1.20	3.0	E .12	.25	58.2	60.7	Rb6
9	181	285	-.08	.13	1.03	.7	1.19	2.3	F .17	.23	62.8	65.0	Rb9
17	155	285	.32	.12	1.15	4.9	1.16	2.6	G .03	.25	50.2	60.4	Rc17
19	130	285	.70	.11	1.16	4.9	1.15	2.7	H .04	.26	41.8	61.5	Rc19
11	158	285	.28	.12	1.06	2.1	1.11	1.8	I .14	.25	58.9	60.6	Rc11
12	209	285	-.57	.14	1.06	1.0	1.09	.9	J .11	.21	73.0	73.6	Rc12
42	123	285	.81	.12	1.08	2.4	1.07	1.3	K .15	.27	52.3	62.6	G42
25	250	285	-1.56	.18	1.03	.3	1.07	.5	L .09	.15	87.7	87.7	C25
15	112	285	.98	.13	1.05	1.3	1.07	1.2	M .19	.27	63.9	64.7	Rc15
44	16	285	3.60	.27	.90	-.4	1.06	.3	N .34	.27	95.4	94.8	G44
43	233	285	-1.08	.16	1.02	.3	1.06	.5	O .13	.18	81.8	81.8	G43
29	225	285	-.89	.15	1.05	.7	1.00	.1	P .13	.19	78.6	79.0	C29
37	228	285	-.96	.15	1.03	.4	1.04	.3	Q .14	.19	79.6	80.0	G37
22	163	285	.20	.12	1.04	1.2	1.03	.5	R .19	.25	58.2	61.2	C22
45	246	285	-1.43	.18	1.01	.1	1.04	.3	S .13	.16	86.3	86.3	G45
48	200	285	-.40	.13	1.02	.3	1.01	.1	T .19	.22	70.9	70.7	W48
32	150	285	.40	.12	1.01	.3	.99	-.1	U .25	.25	60.0	60.2	G32
40	212	285	-.62	.14	1.01	.1	1.00	.0	V .19	.20	74.0	74.6	G40
1	257	285	-1.82	.20	1.00	.1	.97	-.1	W .14	.14	90.2	90.2	Ra1
10	181	285	-.08	.13	1.00	.0	.98	-.2	X .24	.23	63.5	65.0	Rb10
30	246	285	-1.43	.18	.99	.0	.98	.0	Y .17	.16	86.3	86.3	C30
28	216	285	-.70	.14	.99	-.1	.96	-.3	y .22	.20	76.1	75.9	C28
41	188	285	-.20	.13	.99	-.3	.95	-.6	x .26	.23	66.7	67.0	G41
2	258	285	-1.86	.20	.98	-.1	.92	-.3	w .17	.13	90.5	90.5	Ra2
33	129	285	.72	.12	.98	-.6	.96	-.8	v .30	.26	59.6	61.7	G33
49	152	285	.37	.12	.98	-.8	.97	-.5	u .29	.25	61.1	60.2	W49
50	79	285	1.55	.14	.97	-.4	.96	-.5	t .32	.28	74.4	73.7	W50
5	272	285	-2.66	.29	.97	.0	.84	-.5	s .16	.10	95.4	95.4	Rb5
7	193	285	-.28	.13	.97	-.6	.95	-.5	r .27	.22	68.8	68.5	Rb7
36	32	285	2.76	.20	.97	-.1	.88	-.6	q .34	.28	89.5	89.5	G36
39	197	285	-.35	.13	.96	-.7	.96	-.4	p .27	.22	71.2	69.7	G39
4	254	285	-1.70	.19	.96	-.3	.80	-1.0	o .24	.14	89.1	89.1	Ra4
8	250	285	-1.56	.18	.95	-.3	.83	-.9	n .24	.15	87.7	87.7	Rb8
38	166	285	.15	.12	.95	-1.5	.93	-1.1	m .32	.24	64.2	61.6	G38
31	204	285	-.47	.14	.95	-.9	.90	-1.0	l .30	.21	73.0	71.9	G31
47	173	285	.05	.13	.95	-1.5	.93	-1.0	k .32	.24	64.9	63.0	W47
16	77	285	1.59	.14	.95	-.8	.94	-.7	j .35	.28	73.3	74.4	Rc16
27	55	285	2.06	.16	.93	-.7	.90	-.8	i .38	.28	82.5	81.7	C27
35	118	285	.89	.13	.93	-1.9	.92	-1.4	h .36	.27	73.0	63.5	G35
46	53	285	2.11	.16	.93	-.7	.93	-.6	g .38	.28	82.8	82.4	W46
24	205	285	-.49	.14	.91	-1.5	.84	-1.5	f .35	.21	74.0	72.3	C24
34	210	285	-.59	.14	.91	-1.4	.85	-1.4	e .35	.21	74.7	73.9	G34
21	138	285	.58	.12	.91	-3.3	.89	-2.1	d .40	.26	70.2	60.7	C21
26	128	285	.73	.12	.89	-3.6	.88	-2.3	c .42	.26	76.8	61.8	C26
20	180	285	-.07	.13	.86	-3.6	.83	-2.4	b .44	.23	68.8	64.7	Rc20
23	149	285	.41	.12	.85	-5.4	.83	-3.2	a .47	.25	70.9	60.2	C23
MEAN	171.6	285.0	.00	.15	1.00	.0	1.00	.0			72.4	72.6	
S.D.	62.1	.0	1.21	.04	.07	1.9	.12	1.5			12.1	10.8	

The results for Item Fit is seen in Table 4.9. Judging from the decision made from the Literature Review, the acceptable range for Infit Mean Square(MNSQ) is between 0.8 to 1.2. However, from this table an acceptable range can be calculated from the Total Infit MNSQ ± S.D. In Table 4.9, these figures are seen at the bottom of the table, in a red dotted box. For ATETv2, the Infit MNSQ range is 1.00 ± 0.07, which is between 0.93 – 1.07. The items out of this range is circled in blue in the Infit MNSQ column.

Next is the z-score, ZSTD which should be between 2 and -2. The items out of this range is circled in green. Then, the Point Measure Correlation, PT-Measure Corr, is examined. The negative measure in item Rc18 indicates that the responses are opposite. This is circled in orange.

Upon examining this table, items that have all the above-mentioned deficiencies are taken out (in the red boxes) of the next version. The ones with the purple boxes and the others with lesser problems were scrutinised further. They require improvisation.

Table 4.10
Item Fit Statistics: Measure Order of ATETv2

Person: REAL SEP.: 1.59 REL.: .72 ... Item: REAL SEP.: 7.84 REL.: .98

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
44	16	285	3.60	.27	.90	-.4	1.06	.3	.34	.27	95.4	94.8	G44
36	32	285	2.76	.20	.97	-.1	.88	-.6	.34	.28	89.5	89.5	G36
46	53	285	2.11	.16	.93	-.7	.93	-.6	.38	.28	82.8	82.4	W46
27	55	285	2.06	.16	.93	-.7	.90	-.8	.38	.28	82.5	81.7	C27
16	77	285	1.59	.14	.95	-.8	.94	-.7	.35	.28	73.3	74.4	Rc16
50	79	285	1.55	.14	.97	-.4	.96	-.5	.32	.28	74.4	73.7	W50
3	89	285	1.37	.13	1.08	1.4	1.23	3.1	.14	.28	67.4	70.8	Ra3
15	112	285	.98	.13	1.05	1.3	1.07	1.2	.19	.27	63.9	64.7	Rc15
35	118	285	.89	.13	.93	-1.9	.92	-1.4	.36	.27	73.0	63.5	G35
42	123	285	.81	.12	1.08	2.4	1.07	1.3	.15	.27	52.3	62.6	G42
26	128	285	.73	.12	.89	-3.6	.88	-2.3	.42	.26	76.8	61.8	C26
33	129	285	.72	.12	.98	-.6	.96	-.8	.30	.26	59.6	61.7	G33
19	130	285	.70	.12	1.16	4.9	1.15	2.7	.04	.26	41.8	61.5	Rc19
21	138	285	.58	.12	.91	-3.3	.89	-2.1	.40	.26	70.2	60.7	C21
23	149	285	.41	.12	.85	-5.4	.83	-3.2	.47	.25	70.9	60.2	C23
32	150	285	.40	.12	1.01	.3	.99	-.1	.25	.25	60.0	60.2	G32
49	152	285	.37	.12	.98	-.8	.97	-.5	.29	.25	61.1	60.2	W49
17	155	285	.32	.12	1.15	4.9	1.16	2.6	.03	.25	50.2	60.4	Rc17
11	158	285	.28	.12	1.06	2.1	1.11	1.8	.14	.25	58.9	60.6	Rc11
6	159	285	.26	.12	1.08	2.6	1.20	3.0	.12	.25	58.2	60.7	Rb6
22	163	285	.20	.12	1.04	1.2	1.03	.5	.19	.25	58.2	61.2	C22
38	166	285	.15	.12	.95	-1.5	.93	-1.1	.32	.24	64.2	61.6	G38
13	173	285	.05	.13	1.14	3.8	1.21	2.8	.02	.24	57.9	63.0	Rc13
47	173	285	.05	.13	.95	-1.5	.93	-1.0	.32	.24	64.9	63.0	W47
18	179	285	-.05	.13	1.17	4.1	1.25	3.0	-.03	.23	62.8	64.5	Rc18
20	180	285	-.07	.13	.86	-3.6	.83	-2.4	.44	.23	68.8	64.7	Rc20
9	181	285	-.08	.13	1.03	.7	1.19	2.3	.17	.23	62.8	65.0	Rb9
10	181	285	-.08	.13	1.00	.0	.98	-.2	.24	.23	63.5	65.0	Rb10
41	188	285	-.20	.13	.99	-.3	.95	-.6	.26	.23	66.7	67.0	G41
7	193	285	-.28	.13	.97	-.6	.95	-.5	.27	.22	68.8	68.5	Rb7
39	197	285	-.35	.13	.96	-.7	.96	-.4	.27	.22	71.2	69.7	G39
48	200	285	-.40	.13	1.02	.3	1.01	.1	.19	.22	70.9	70.7	W48
31	204	285	-.47	.14	.95	-.9	.90	-1.0	.30	.21	73.0	71.9	G31
24	205	285	-.49	.14	.91	-1.5	.84	-1.5	.35	.21	74.0	72.3	C24
12	209	285	-.57	.14	1.06	1.0	1.09	.9	.11	.21	73.0	73.6	Rc12
34	210	285	-.59	.14	.91	-1.4	.85	-1.4	.35	.21	74.7	73.9	G34
40	212	285	-.62	.14	1.01	.1	1.00	.0	.19	.20	74.0	74.6	G40
28	216	285	-.70	.14	.99	-.1	.96	-.3	.22	.20	76.1	75.9	C28
29	225	285	-.89	.15	1.05	.7	1.00	.1	.13	.19	78.6	79.0	C29
37	228	285	-.96	.15	1.03	.4	1.04	.3	.14	.19	79.6	80.0	G37
14	231	285	-1.03	.15	1.04	.5	1.29	1.9	.07	.18	81.1	81.1	Rc14
43	233	285	-1.08	.16	1.02	.3	1.06	.5	.13	.18	81.8	81.8	G43
30	246	285	-1.43	.18	.99	.0	.98	.0	.17	.16	86.3	86.3	C30
45	246	285	-1.43	.18	1.01	.1	1.04	.3	.13	.16	86.3	86.3	G45
8	250	285	-1.56	.18	.95	-.3	.83	-.9	.24	.15	87.7	87.7	Rb8
25	250	285	-1.56	.18	1.03	.3	1.07	.5	.09	.15	87.7	87.7	C25
4	254	285	-1.70	.19	.96	-.3	.80	-1.0	.24	.14	89.1	89.1	Ra4
1	257	285	-1.82	.20	1.00	.1	.97	-.1	.14	.14	90.2	90.2	Ra1
2	258	285	-1.86	.20	.98	-.1	.92	-.3	.17	.13	90.5	90.5	Ra2
5	272	285	-2.66	.29	.97	.0	.84	-.5	.16	.10	95.4	95.4	Rb5
MEAN	171.6	285.0	.00	.15	1.00	.0	1.00	.0			72.4	72.6	
S.D.	62.1	.0	1.21	.04	.07	1.9	.12	1.5			12.1	10.8	

From Table 4.10, the items that have similar measures suggest they are of the same difficulty level and might be testing the same construct. These are items in the coloured boxes. Items Rc13 and W47 (green box) seem to have the same measure but they are from different components, Reading and Writing respectively. Items Rb9 and Rb10 (orange box) have the same measure. These two items are from the same Reading section and from the same passage. Suggestion is to remove one of these items. Items C30 and G45 (blue box) also have the same measure. Item C30 is from the Cloze Passage while G45 is from Grammar. Another pair is Rb8 and C25 share the same measure but are from different sections, Reading and Cloze Passage respectively.

4.3.4 Principal Component Analysis

Table 4.11
Principal Component Analysis of ATETv2

CONTRAST 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --			Modeled
Total raw variance in observations	=	66.5	100.0%		100.0%
Raw variance explained by measures	=	16.5	24.8%		24.8%
Raw variance explained by persons	=	3.8	5.7%		5.7%
Raw Variance explained by items	=	12.6	19.0%		19.1%
Raw unexplained variance (total)	=	50.0	75.2%	100.0%	75.2%
Unexplned variance in 1st contrast	=	6.4	9.7%	12.9%	

This Principal Component Analysis on ATETv2 is definitely unidimensional. The raw variance explained by measures (orange box) is 24.8%, similar to the modeled value. Next is the 'unexplained variance in the 1st contrast should not be more than 15% (Linacre, 2015a). Table 4.11 (blue box) shows 9.7%. This indicates the noise level, which is acceptable.

Table 4.12
Largest Standardized Residual Correlations of ATETv2

RESIDUL CORRELN	ENTRY NUMBER	Item	ENTRY NUMBER	Item
.57	47	W47	49	W49
.56	17	Rc17	19	Rc19
.47	21	C21	26	C26
.46	13	Rc13	17	Rc17
-.52	17	Rc17	26	C26
-.51	19	Rc19	26	C26
-.49	17	Rc17	35	G35
-.48	17	Rc17	21	C21
-.47	19	Rc19	21	C21
-.47	19	Rc19	35	G35

Following this is the Table 4.12 , which points out the items that are noise makers, with a residual correlations that is less than 0.58. This further confirms the items that may be testing the same thing or may be confusing to the respondents. The positive residual correlations show that the items are from the same section. However, the negative correlations show that the items paired are from different sections. These items should be checked if they are really testing the same construct.

4.3.5 Differential Item Functioning (DIF)

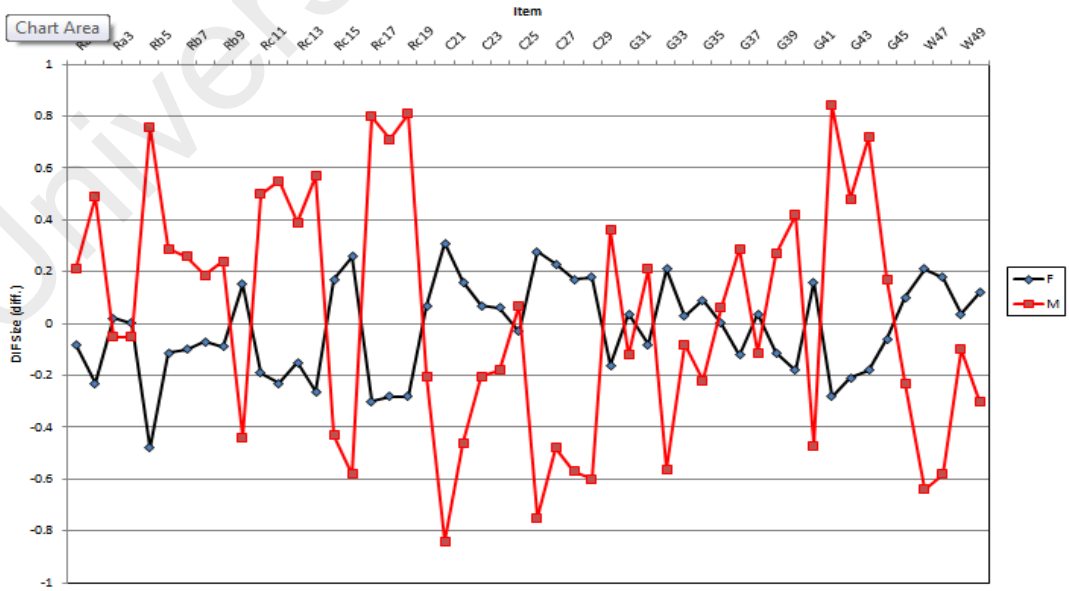

Figure 4.7. Person-gender DIF plot according to difference in size for ATETv2

This DIF analysis inspects gender biasness. From Figure 4.7, the limit is ±0.5 to see if the difference is significant. There are about 27 items that show this difference, Ra2, Rb5, Rb10, Rc11, Rc12, Rc13, Rc14, Rc15, Rc16, Rc17, Rc18, Rc19, C21, C22, C26, C27, C28, C29, C30, G33, G40, G41, G42, G43, G44, W47, and W48. As such, it is recommended to revise these items or to find a reason why different gender behaves differently to these items.

4.3.6 Person Misfit Order

Table 4.13
Person Misfit Order of ATETv2

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
149	46	50	3.04	.57	1.26	.7	2.02	1.2	A .13	.35	90.0	92.5	KF149
280	20	50	-.55	.32	.93	-.5	1.99	3.1	B .43	.43	76.0	70.5	MF280
13	34	50	.92	.34	1.26	1.6	1.92	2.8	C .22	.46	64.0	75.0	MM013
168	48	50	3.91	.77	1.33	.7	1.89	1.0	D .02	.28	96.0	96.0	KF168
38	32	50	.70	.33	1.27	1.8	1.85	2.9	E .20	.46	66.0	72.7	SF038
103	35	50	1.04	.35	1.26	1.5	1.80	2.3	F .20	.45	76.0	76.4	KM103
30	18	50	-.76	.33	1.42	2.8	1.79	2.3	G .09	.42	56.0	72.1	MF030
67	29	50	.37	.32	1.33	2.4	1.74	2.9	H .18	.46	54.0	70.2	SF067
54	27	50	.17	.32	1.29	2.2	1.60	2.4	I .21	.46	56.0	69.2	KM054
49	28	50	.27	.32	1.40	2.9	1.59	2.4	J .14	.46	58.0	69.6	SM049
8	24	50	-.14	.32	1.29	2.3	1.57	2.3	K .20	.45	64.0	69.0	MM008
33	23	50	-.24	.32	1.44	3.3	1.56	2.2	L .11	.45	44.0	69.2	SF033
50	33	50	.81	.34	1.19	1.2	1.54	1.9	M .28	.46	74.0	73.8	SF050
172	32	50	.70	.33	1.12	.9	1.50	1.9	N .32	.46	70.0	72.7	SF172
47	17	50	-.87	.33	1.32	2.1	1.48	1.5	O .17	.42	58.0	72.9	SM047
17	28	50	.27	.32	1.33	2.4	1.48	2.0	P .20	.46	66.0	69.6	SF017
256	27	50	.17	.32	1.09	.7	1.47	2.0	Q .34	.46	72.0	69.2	MF256
52	31	50	.59	.33	1.35	2.4	1.46	1.8	R .19	.46	60.0	71.8	MF052
106	34	50	.92	.34	1.09	.6	1.46	1.6	S .34	.46	76.0	75.0	KM106
24	29	50	.37	.32	1.20	1.5	1.44	1.9	T .29	.46	58.0	70.2	SF024
195	28	50	.27	.32	1.05	.5	1.44	1.9	U .39	.46	66.0	69.6	KF195
2	32	50	.70	.33	.98	-.1	1.43	1.6	V .43	.46	78.0	72.7	MM002
272	26	50	.07	.32	.90	-.8	1.42	1.8	W .48	.45	80.0	69.0	SF272
37	22	50	-.34	.32	1.21	1.6	1.41	1.6	X .26	.44	68.0	69.5	SF037
44	26	50	.07	.32	1.29	2.3	1.39	1.7	Y .23	.45	56.0	69.0	SF044
32	25	50	-.04	.32	1.39	2.9	1.37	1.6	Z .18	.45	50.0	68.9	SF032
227	34	50	.92	.34	1.35	2.0	1.37	1.3	.21	.46	60.0	75.0	SF227
18	29	50	.37	.32	1.31	2.2	1.36	1.6	.23	.46	54.0	70.2	SF018
51	25	50	-.04	.32	1.28	2.2	1.36	1.6	.23	.45	58.0	68.9	SF051
105	29	50	.37	.32	1.29	2.1	1.31	1.4	.25	.46	58.0	70.2	KM105
108	24	50	-.14	.32	1.27	2.1	1.29	1.3	.25	.45	60.0	69.0	MM108
BETTER FITTING OMITTED													
236	28	50	.27	.32	.80	-1.7	.68	-1.6	z .61	.46	74.0	69.6	SF236
180	30	50	.48	.33	.80	-1.6	.71	-1.3	y .61	.46	80.0	71.0	KF180
266	23	50	-.24	.32	.80	-1.8	.70	-1.4	x .60	.45	80.0	69.2	SM266
198	29	50	.37	.32	.80	-1.7	.69	-1.5	w .62	.46	74.0	70.2	KF198
196	28	50	.27	.32	.79	-1.7	.69	-1.5	v .62	.46	78.0	69.6	SF196
265	25	50	-.04	.32	.79	-1.9	.68	-1.6	u .62	.45	78.0	68.9	KM265
185	33	50	.81	.34	.78	-1.5	.70	-1.2	t .61	.46	82.0	73.8	KF185
170	31	50	.59	.33	.78	-1.7	.68	-1.5	s .62	.46	76.0	71.8	MF170
261	26	50	.07	.32	.78	-1.9	.68	-1.6	r .62	.45	76.0	69.0	SF261
273	26	50	.07	.32	.78	-1.9	.67	-1.6	q .62	.45	76.0	69.0	MF273
283	22	50	-.34	.32	.78	-1.9	.67	-1.5	p .62	.44	72.0	69.5	KF283
239	26	50	.07	.32	.78	-2.0	.67	-1.7	o .63	.45	76.0	69.0	SM239
264	25	50	-.04	.32	.77	-2.0	.67	-1.7	n .63	.45	78.0	68.9	MF264
275	27	50	.17	.32	.77	-2.0	.66	-1.7	m .63	.46	76.0	69.2	KM275
251	27	50	.17	.32	.76	-2.1	.66	-1.7	l .64	.46	80.0	69.2	KF251
210	36	50	1.17	.36	.76	-1.4	.62	-1.3	k .63	.45	82.0	77.7	SF210
231	30	50	.48	.33	.76	-1.9	.65	-1.7	j .64	.46	80.0	71.0	MF231
206	31	50	.59	.33	.76	-1.9	.68	-1.5	i .64	.46	84.0	71.8	KF206
222	27	50	.17	.32	.75	-2.2	.68	-1.6	h .64	.46	84.0	69.2	KF222
253	28	50	.27	.32	.75	-2.2	.65	-1.8	g .65	.46	82.0	69.6	KF253
277	24	50	-.14	.32	.75	-2.3	.64	-1.8	f .64	.45	80.0	69.0	SF277
233	26	50	.07	.32	.74	-2.3	.64	-1.9	e .65	.45	76.0	69.0	KF233
267	26	50	.07	.32	.73	-2.4	.67	-1.7	d .65	.45	88.0	69.0	KF267
216	29	50	.37	.32	.72	-2.4	.62	-1.9	c .67	.46	82.0	70.2	KF216
221	32	50	.70	.33	.72	-2.1	.60	-1.8	b .67	.46	82.0	72.7	MF221
242	27	50	.17	.32	.70	-2.7	.60	-2.1	a .68	.46	84.0	69.2	SF242
MEAN	30.1	50.0	.54	.34	1.00	.0	1.00	.0			72.4	72.6	
S.D.	5.2	.0	.68	.07	.16	1.2	.28	1.1			7.8	4.7	

Similar to the Item Fit analysis, the Person Fit analysis starts off by checking the Infit MNSQ. The acceptable range is Total Infit MNSQ ± S.D., 1.00 ± 0.16, which is between 0.84 and 1.16. The persons who are out of the acceptable range are in the blue box in Table 4.13. Next is the PT-MEASURE CORR, which should be positive. According to Table 4.13, all Point Measure Correlations are positive. The persons who did not fit the Infit MNSQ acceptable range (51 persons) were deleted from the file, with a PDFILE command. The statistics is seen in Table 4.14.

Table 4.14
Summary Statistics after Deleting the Misfit Persons Person Misfit Order of ATETv2

INPUT: 285 Persons 50 Items MEASURED: 234 Persons 50 Items 2 CATS 3.66.0

SUMMARY OF 234 MEASURED Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	30.5	50.0	.60	.34	1.00	.0	.98	.0
S.D.	5.0	.0	.65	.07	.12	.8	.22	.8
MAX.	49.0	50.0	4.74	1.06	1.34	2.1	2.08	3.1
MIN.	16.0	50.0	-.98	.32	.80	-1.3	.59	-1.2

REAL RMSE	.36	ADJ.SD	.54	SEPARATION	1.52	Person RELIABILITY		.70
MODEL RMSE	.35	ADJ.SD	.55	SEPARATION	1.58	Person RELIABILITY		.71
S.E. OF Person MEAN = .04								

DELETED: 51 Persons
Person RAW SCORE-TO-MEASURE CORRELATION = .96
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .63

SUMMARY OF 50 MEASURED Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	142.9	234.0	.00	.17	1.00	.1	.98	.0
S.D.	51.4	.0	1.27	.05	.07	1.8	.12	1.3
MAX.	229.0	234.0	3.73	.45	1.21	5.6	1.31	3.8
MIN.	12.0	234.0	-3.39	.14	.86	-4.1	.77	-2.5

REAL RMSE	.18	ADJ.SD	1.25	SEPARATION	7.07	Item RELIABILITY		.98
MODEL RMSE	.18	ADJ.SD	1.25	SEPARATION	7.14	Item RELIABILITY		.98
S.E. OF Item MEAN = .18								

UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -.98
11700 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 12395.80 with 11417 d.f. p=.0000

Table 4.14 shows the person reliablity 0.70 is lesser than before deleting the persons. Therefore it is decided that the further analysis will not consider deleting the persons as 285 persons give a better

The scalogram displays the Person response distribution according to items. Figure 4.8 shows the misfit Persons.

The interpretation of Figure 4.8 is that the persons are ordered from high measure to low measure while the items are also arranged from low to high measure. Thus the Persons has 50% chances of getting most of the easier items correct (1) and 50% chances of getting the more difficult items wrong (0). The top left corner is where the more able persons respond to the easier items (from Figure 4.8, mostly "1"), while the top right corner, there should be more "0", but the responses show "1" (the red box) insteaad. This should be opposite in the bottom of the scalogram. The bottom right hand corner (blue box) should have more or almost all "0". But this is evident in Figure 4.8 . So this clearly shows that there is a discrepency in the results compared to the expected pattern. Items need to be revised.

4.3.8 Item Characteristic Curve

The Item Characteristic Curve (ICC) plots the model-expected item characteristic curve. This is the Rasch model prediction for each measure relative to item difficulty. The steeper the ICC, the more discriminating it is between high and low achievers (Linacre, 2015).

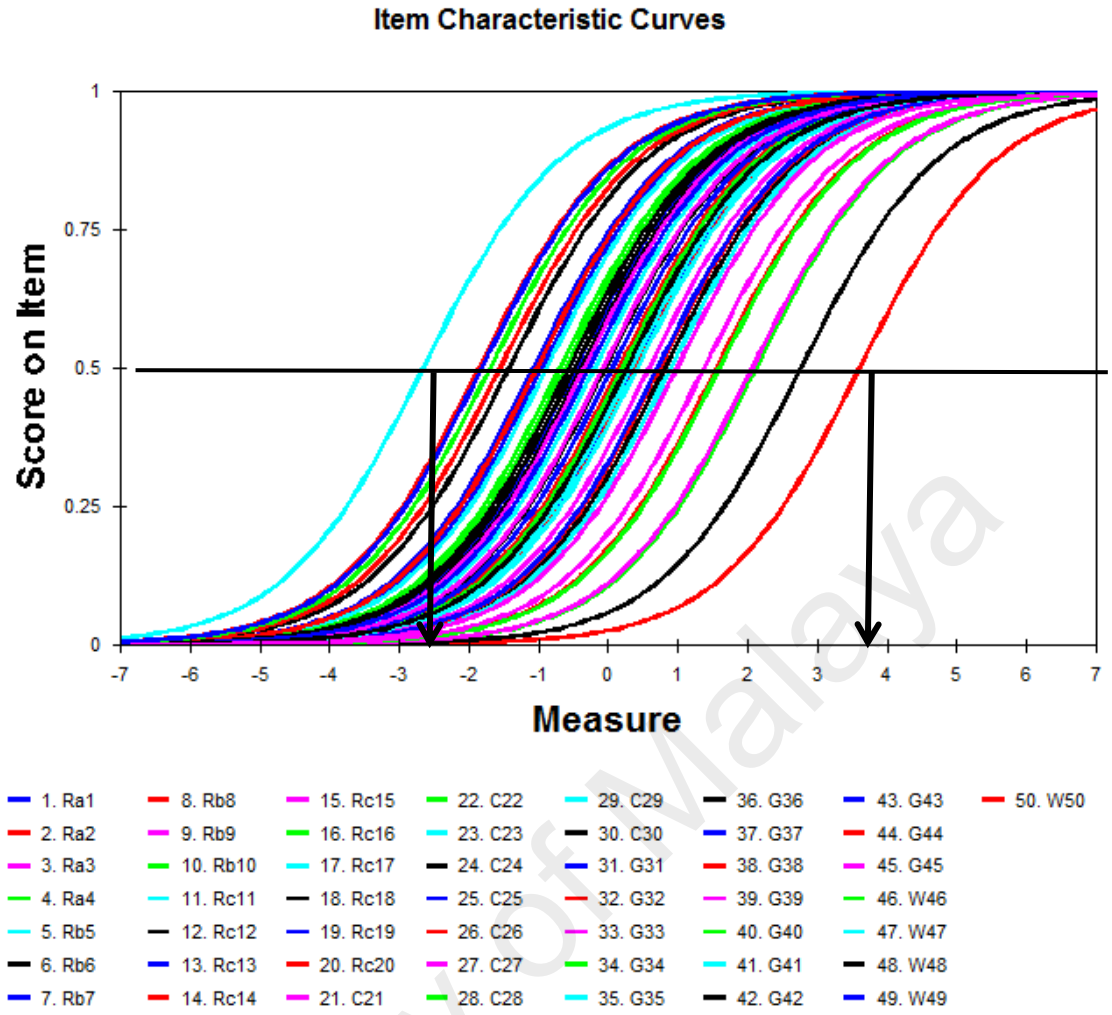

Figure 4.9. Item Characteristics Curve for all 50 items in ATETv2

The overall results for all items is seen in Figure 4.9. The horizontal black line show the 50% chances of getting an item correct on ATETv2. Drawing a line parallel to the vertical axis from the interception of the horizontal black line to the coloured sigmoid, the corresponding item location measure is shown on the horizontal axis. It can be seen that the items spread from the easiest being on the far left (R5), with the lowest measure while the most difficult on the far right (G44) with the highest measure. The overall result is scrutinised according to the three sections, Reading, Grammar and Writing.

Figure 4.9 shows the results of ICC for the Reading section. Figure 4.9 shows that Rb5 is the easiest (far left) and Rc16 the most difficult (far right). The difficulty level of the items in this section seems well distributed. However, items Ra1 and Ra2 are of the same curve, pointing that they share the same characteristics. The same with Rb9 and Rb10; Rb6, Rc11 and Rc17; Rc13, Rc18 and Rc20. Thus the graph for this section appears as only 14 ICCs, with 4 overlapping set of items.

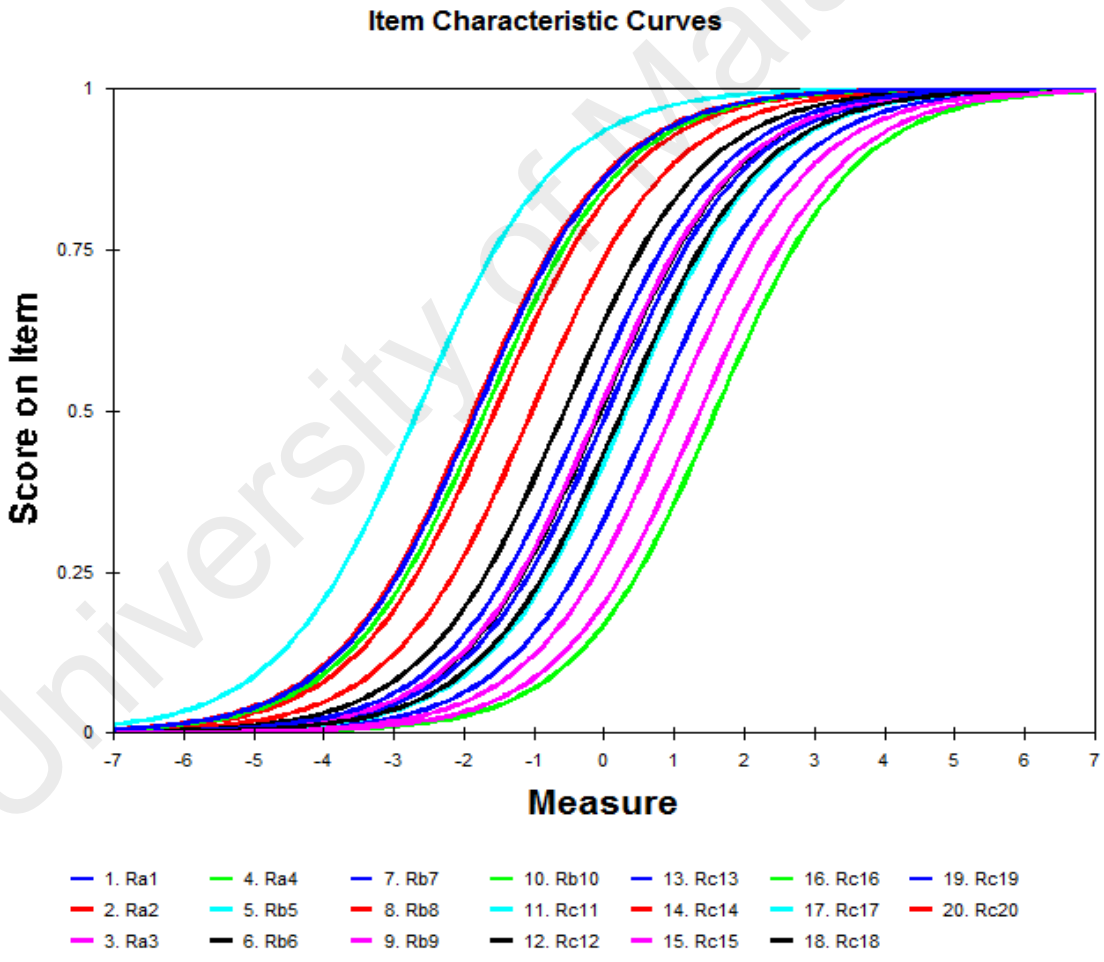

Figure 4.10. Item Characteristics Curve for Reading section

Meanwhile for the cloze items, the characteristic curve is seen in Figure 4.11.

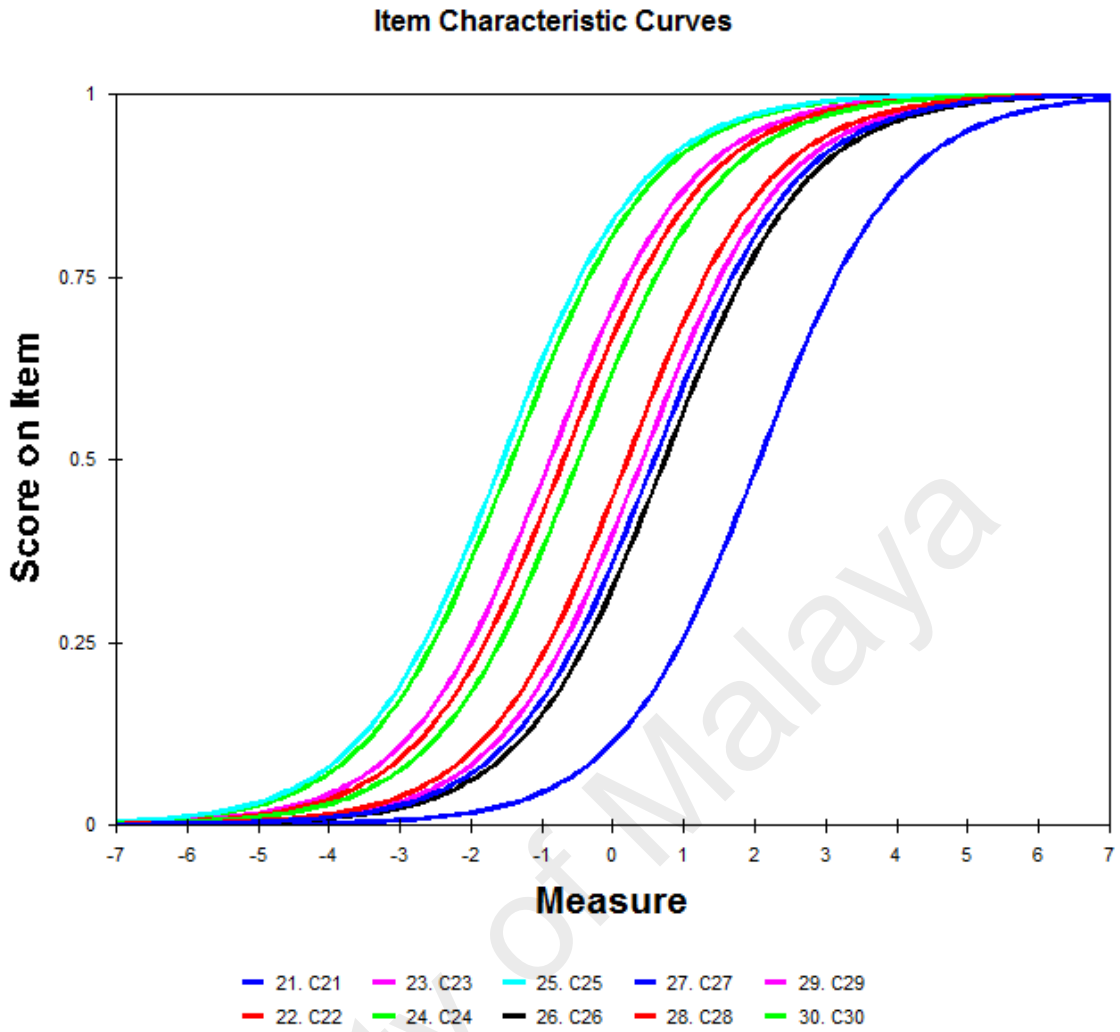

Figure 4.11. Item Characteristics Curve for the Cloze section

For this Cloze Section, there are no overlapping sigmoids in Figure 4.11. They are also well spread in difficulty levels, C25 being the easiest and C27 being the most difficult item in the Cloze Section.

The third sction is the Grammar component which is shown collectively in Figure 4.12.

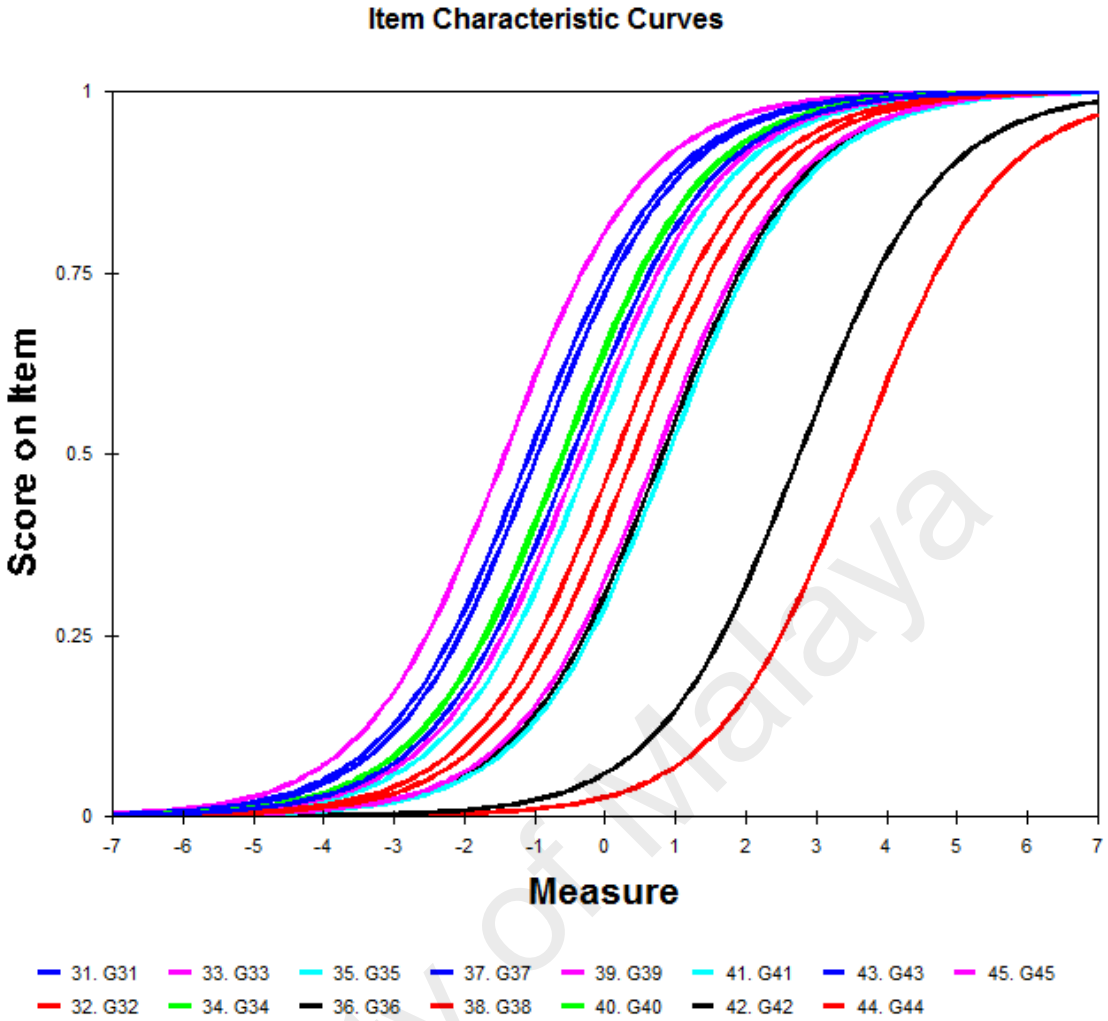

Figure 4.12. Item Characteristics Curve for Grammar section

In Figure 4.12, items that are overlapping are G34 with G40 and G35 with G42. The rest of the items are distributed fairly well.

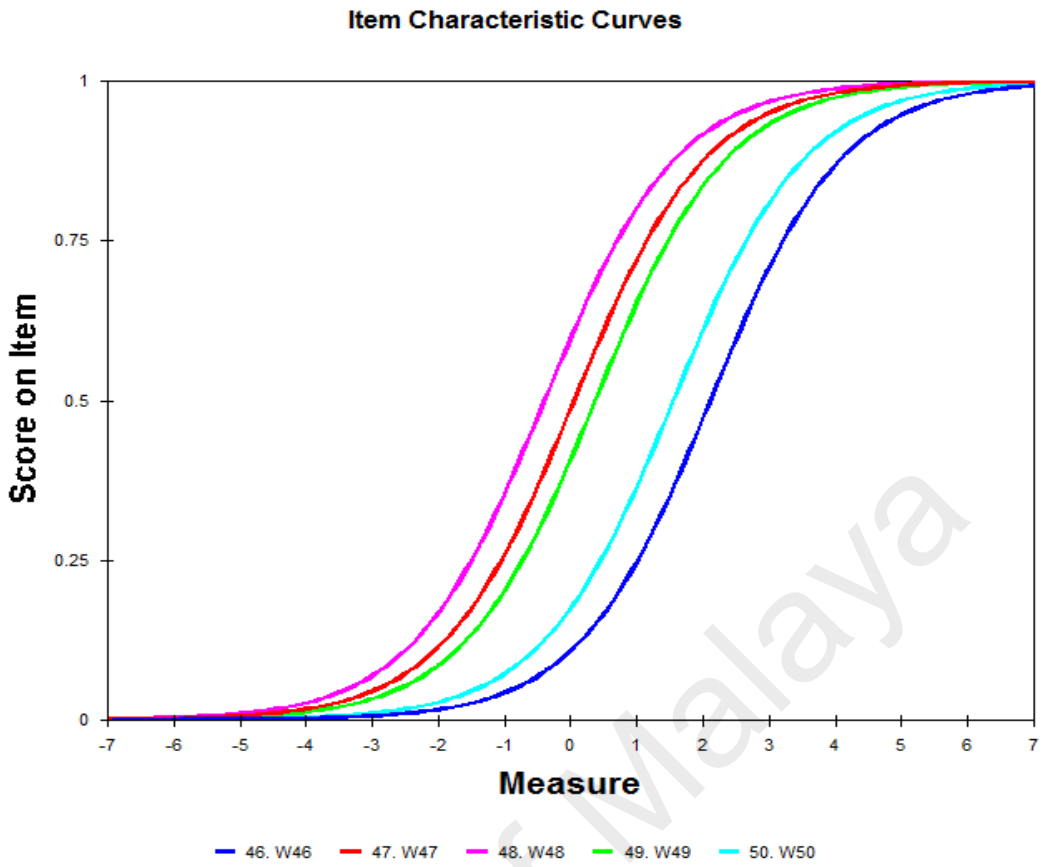

Figure 4.13. Item Characteristics Curve for Writing section

The final section of the ATETv2 is Writing. There were only 5 items which tested paragraph improvement depicted in Figure 4.13. It shows that all items are well spread. There are no overlaps for this section.

4.3.9 Bubble Chart

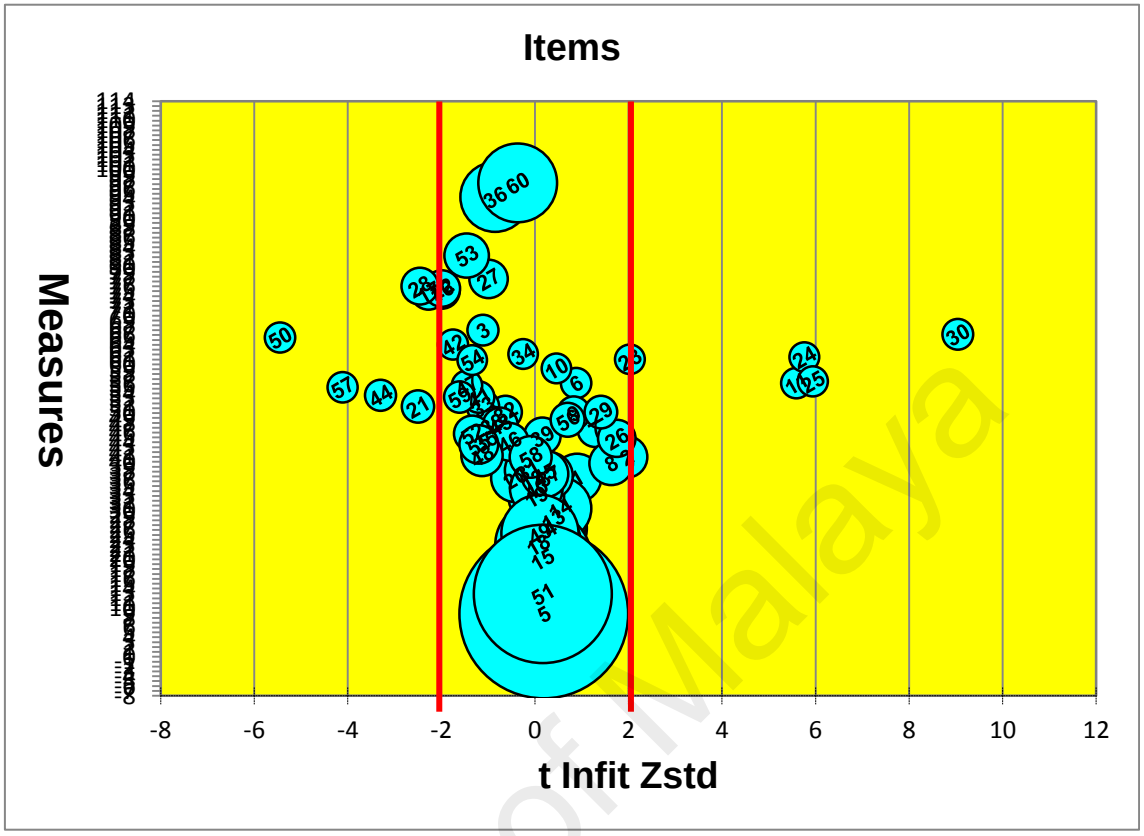

Figure 4.14 Bubble chart of items of ATETv2

Figure 4.14 shows the distribution of items in its spread of measures. There are still items which are out of the acceptable z-score range, i.e. between 2 and -2 (the boundaries are the red vertical lines). 39 items are within the range. The items out of the range are Ra3, Rb6, Rb9, Rc13, Rc17, Rc18, Rc19, Rc20, C21, C23,and C26. However, when scrutinised, all items in the Grammar and Writing sections were within the acceptable range.

4.3.10 Subject Matter Expert Feedback

ATETv2 was given to three senior lecturers from the TESL Department of the Faculty of Education, Public university. This was a blind review and the three of them had no idea what this test was

about. All three of them are subject matter experts and are part of the faculty's Vetting Committee. However, only two of them gave constructive feedback. To get further explaination on the feedback, a short interview was conducted with the two lecturers separately. Their names are not revealed and will be addressed as SME A and SME B. They profiles and are in Appendix J. The outcome of the interview and comments will follow the sequence of the test paper, ATETv2.

Both subject matter experts felt that the number of passages should be reduced and to target at least higher intermediate to advanced level. Both of them also advised to include items related to literary elements, contextual meaning of difficult words, inferencing and drawing conclusions. These skills, according to them, will be taught in the programme and it will be good to see how many of the candidates have such an exposure. Number of items were asked to be retained after they were shown the Winsteps analysis of ATETv2.

For the Grammar section, the Cloze passage was found to be easy and suggested to choose a different passage with some contextual clues and items that would encourage them to think. Both the SMEs conveyed their concern for the Sentence Completion section. SME A asked to reduce the items for this section,but SME B asked to increase the number of items. The Winsteps analysis was shown again to emphasise the need to increase the items, but the question was where, that is for which section. Only then SME A was convinced that there is a need to increase the items and the Sentence Completion would be the best section, but with a varied difficulty level. There were no comments of the Error Analysis section as they felt it was acceptable

and had a range of difficulty levels.

Both the SMEs felt that writing an essay was sufficient. However, for the reliability of the marking during interviews, the MCQs were welcomed. There was a stern caution about the guessing factor that might taint the results. As such both SMEs felt that the essay section must be retained for triangulation purpose. Refer to Appendix J for the Interview Notes.

4.4 ATET version 3 (ATETv3)

4.4.1 Summary Statistics

Table 4.15
Summary Statistics of ATETv3

SUMMARY OF 285 MEASURED Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	36.8	60.0	.58	.34	.98	.0	.99	.0
S.D.	7.0	.0	.98	.11	.19	1.1	.61	1.0
MAX.	59.0	60.0	6.70	1.42	1.71	3.2	8.53	6.7
MIN.	21.0	60.0	-1.03	.31	.20	-2.2	.01	-1.5
REAL RMSE	.37	ADJ.SD	.91	SEPARATION	2.46	Person RELIABILITY		.86
MODEL RMSE	.36	ADJ.SD	.91	SEPARATION	2.53	Person RELIABILITY		.87
S.E. OF Person MEAN = .06								

Person RAW SCORE-TO-MEASURE CORRELATION = .94
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .80

SUMMARY OF 60 MEASURED Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	174.7	285.0	.00	.18	1.01	.0	.99	-.2
S.D.	72.9	.0	1.82	.13	.09	1.2	.18	.8
MAX.	277.0	285.0	7.58	1.09	1.39	2.5	1.75	1.8
MIN.	1.0	285.0	-3.24	.13	.85	-3.3	.63	-1.9
REAL RMSE	.24	ADJ.SD	1.81	SEPARATION	7.51	Item RELIABILITY		.98
MODEL RMSE	.22	ADJ.SD	1.81	SEPARATION	8.16	Item RELIABILITY		.99
S.E. OF Item MEAN = .24								

UMEAN=.000 USCALE=1.000
Item RAW SCORE-TO-MEASURE CORRELATION = -.95
17100 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 16057.62 with 16756 d.f. p=.9999

Table 4.15 shows the final version of the test after taking into consideration all the feedback from the data analysis done on ATETv2 and Subject Matter Expert's advice. There were in total 285 candidates who sat for this test. The test had a total of 60 items, all in multiple

choice type, with 4 options each. Compared to ATETv2 which had 50 items, 10 extra items were included in the ATETv3.

The person distribution in Table 4.15 shows the mean person measure is 0.58 logits. The observed Person S.D. is 0.98 logits, so the observed variance is $(0.98)^2 = 0.96$. The square root of the mean error variance is RMSE (root mean square error), The true RMSE is between "Real SE" and the "Model SE". Thus, the model error variance is $(0.36)^2 = 0.13$. In Table 4.15, "Adj SD" (Adjusted for Error standard deviation) is 0.91.

In terms of reliability for serious decision-making, according to Linacre (2015), reliability should be at least 0.8. This is exactly the result of the Cronbach Alpha, which shows the reliability of the total test, ATETv3 is 0.80, compared to Cronbach Alpha for ATETv2 was 0.65. This increment is seen upon adding on 10 more items to the ATETv3. This is because, according to Linacre (2015), the person measurement precision is by increasing the number of items on the test.

The person stratification is given as $H_p = (4G+1)/3$, with G being the Person Separation. ATETv2 had $H_p = 2.45$, but ATETv3 has $H_p = [4(2.46)+1]/3 = 3.61$. This implies that there are 3 categories of persons. This test is able to discriminate the low, intermediate and high achievers.

In terms of the measures, for Persons it ranges between -1.03 and 6.70 logits whiles the Items measures between -3.24 and 7.58 logits. This shows that there is wide ability of candidates and wide range of items.

4.4.2 Variable Map

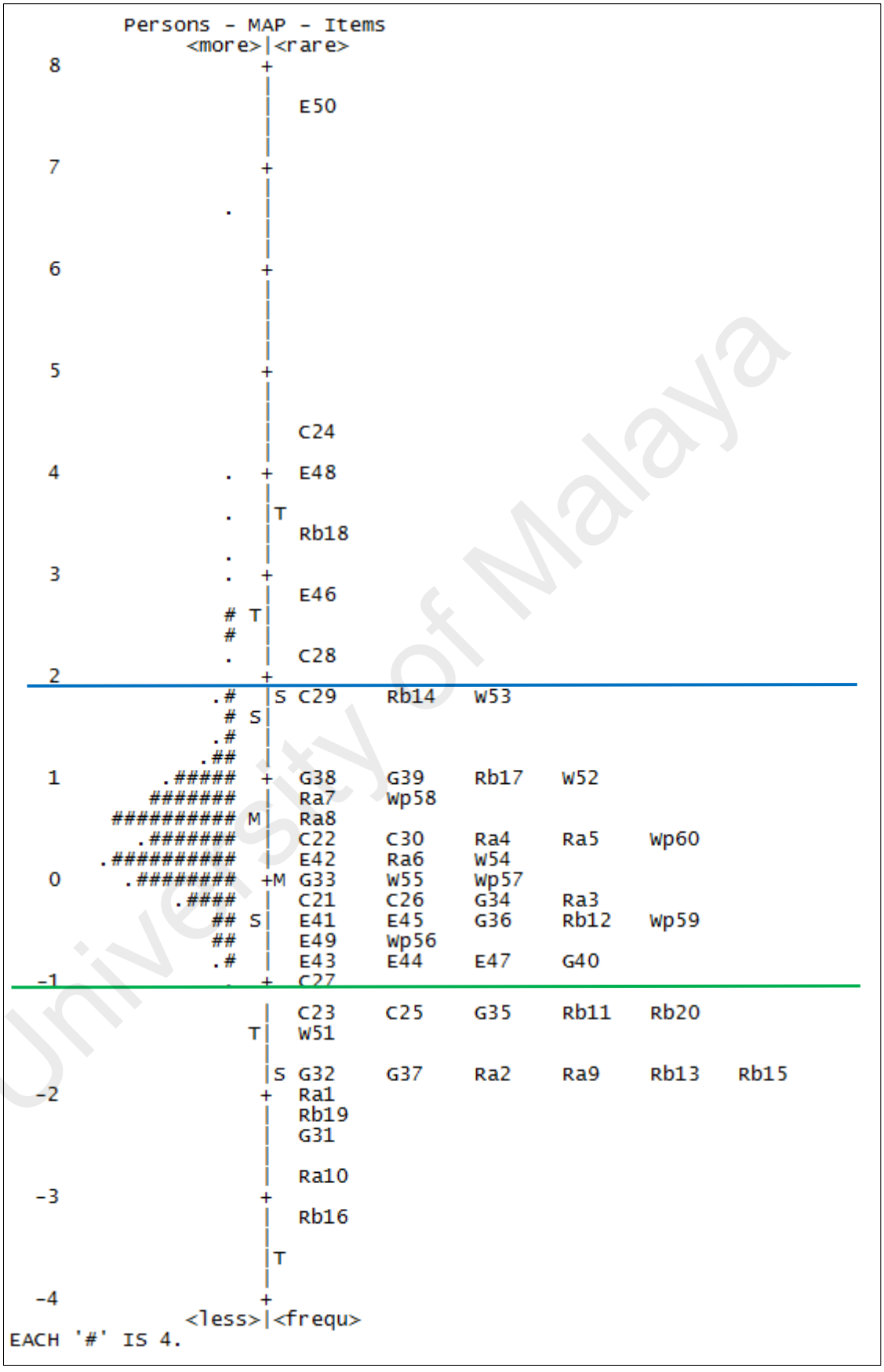

Figure 4.15. The variable map of ATETv3

Figure 4.15 shows the Person measurement scale is 6.70+1.03 = 7.73 in length. Meanwhile, the Item measurement scale is 7.58+3.24 = 10.82 logits in length. This means that the items are a wider range and is able to capture any candidate that takes this test. The easiest item is item no 16, which a reading comprehension question from the second passage and the most difficult item is item no. 50, which tests Error Identification.

The map in Figure 4.15 shows there is a gap between items 24 and 50. The highest person ability seems to be between items 24 and 50. However the purpose of this tes was to select candidates into the program. Thus there are 3 clear cut divisions of candidates from this map, low, intermediate and high achievers. Low is defined as any persons below the green line (at -1 logits), intermediate is between the green and blue line while high is above the blue line (at 1.75 logits). For the selection purpose, all intermediate and high achievers are accepted into the programme, i.e. all candidates whose ability is above -1 logit. It is also noted that there are no persons below -1 logits, although there are 17 items below the green line.

4.4.3 Item Fit

Table 4.16
Item Fit Statistics of ATETv3

Item STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
48	14	285	4.06	.31	1.21	.9	1.75	1.7	A .21	.40	95.8	95.7	E48
50	1	285	7.58	1.09	1.39	.7	1.66	1.2	B .05	.25	99.6	99.6	E50
18	22	285	3.46	.25	1.03	.2	1.33	1.2	C .35	.41	94.0	93.1	Rb18
46	36	285	2.79	.20	1.23	1.6	1.16	.9	D .25	.41	85.6	88.9	E46
14	69	285	1.83	.15	1.19	2.1	1.22	1.8	E .22	.39	75.1	79.5	Rb14
24	11	285	4.38	.35	1.07	.3	1.21	.6	F .31	.40	96.8	96.7	C24
56	212	285	-.67	.14	1.10	1.5	1.13	1.0	G .13	.24	72.3	74.6	Wp56
39	113	285	.99	.13	1.12	2.5	1.07	.7	H .25	.35	59.6	67.7	G39
4	154	285	.33	.13	1.08	2.4	1.05	.6	I .23	.31	54.7	62.7	Ra4
47	217	285	-.77	.14	1.08	1.2	1.03	.3	J .16	.23	74.0	76.3	E47
11	233	285	-1.13	.16	1.04	.4	1.08	.5	K .16	.20	81.8	81.8	Rb11
34	187	285	-.21	.13	1.07	1.5	1.08	.8	L .20	.27	64.9	67.4	G34
27	227	285	-.99	.15	1.00	.0	1.07	.5	M .20	.21	79.3	79.7	C27
40	222	285	-.87	.15	1.07	1.0	1.06	.4	N .15	.22	77.9	77.9	G40
54	157	285	.28	.13	1.03	1.0	1.07	.8	O .27	.30	62.1	62.7	W54
42	160	285	.23	.13	1.07	1.9	1.03	.4	P .24	.30	58.2	62.9	E42
52	110	285	1.04	.13	1.07	1.3	1.01	.2	Q .30	.35	61.4	68.4	W52
59	198	285	-.40	.13	1.06	1.1	1.02	.3	R .20	.26	67.0	70.4	Wp59
22	145	285	.47	.13	1.05	1.6	1.02	.3	S .27	.32	61.8	62.9	C22
9	256	285	-1.84	.20	1.01	.1	1.05	.3	T .13	.15	89.8	89.8	Ra9
45	199	285	-.42	.13	1.05	1.0	.99	.0	U .21	.25	67.4	70.7	E45
20	232	285	-1.11	.16	1.04	.5	1.02	.2	V .16	.20	81.4	81.4	Rb20
33	180	285	-.09	.13	1.04	1.0	.99	-.1	W .25	.28	65.6	65.7	G33
26	185	285	-.18	.13	1.03	.8	.98	-.2	X .25	.27	65.3	66.9	C26
19	263	285	-2.15	.23	1.03	.2	.98	.0	Y .10	.14	92.3	92.3	Rb19
38	116	285	.94	.13	1.02	.5	1.03	.3	Z .32	.35	66.3	67.0	G38
BETTER FITTING OMITTED													
13	254	285	-1.76	.19	1.00	.1	.91	-.3	z .16	.16	89.1	89.1	Rb13
43	218	285	-.79	.14	1.00	.0	.92	-.5	y .24	.23	75.8	76.6	E43
10	272	285	-2.73	.29	.99	.1	.79	-.5	x .13	.10	95.4	95.4	Ra10
16	277	285	-3.24	.36	.99	.1	.69	-.7	w .12	.08	97.2	97.2	Rb16
1	258	285	-1.92	.21	.98	.0	.89	-.3	v .17	.15	90.5	90.5	Ra1
8	133	285	.66	.13	.98	-.5	.97	-.3	u .35	.33	68.1	64.1	Ra8
55	173	285	.02	.13	.98	-.6	.93	-.8	t .31	.29	64.9	64.3	W55
44	215	285	-.73	.14	.98	-.3	.87	-1.0	s .27	.23	74.7	75.6	E44
41	203	285	-.49	.14	.97	-.5	.88	-1.0	r .29	.25	71.9	71.9	E41
7	119	285	.89	.13	.97	-.7	.93	-.7	q .38	.34	69.5	66.4	Ra7
2	257	285	-1.88	.20	.97	-.1	.96	-.1	p .19	.15	90.2	90.2	Ra2
29	73	285	1.75	.15	.97	-.4	.91	-.7	o .43	.38	77.9	78.3	C29
12	203	285	-.49	.14	.96	-.7	.94	-.5	n .28	.25	74.0	71.9	Rb12
3	183	285	-.14	.13	.95	-1.1	.88	-1.2	m .33	.28	68.1	66.4	Ra3
51	240	285	-1.31	.17	.95	-.5	.85	-.8	l .25	.19	84.2	84.2	W51
49	212	285	-.67	.14	.94	-.9	.86	-1.1	k .30	.24	76.5	74.6	E49
60	146	285	.46	.13	.94	-1.9	.88	-1.3	j .39	.32	67.0	62.8	Wp60
28	58	285	2.10	.16	.93	-.7	.85	-1.1	i .47	.39	83.9	82.7	C28
30	152	285	.36	.13	.92	-2.5	.92	-.9	h .39	.31	68.1	62.7	C30
57	175	285	-.01	.13	.92	-2.2	.84	-1.8	g .38	.28	63.9	64.6	Wp57
5	149	285	.41	.13	.91	-2.8	.92	-.9	f .40	.31	73.0	62.7	Ra5
17	110	285	1.04	.13	.90	-2.0	.87	-1.2	e .44	.35	74.0	68.4	Rb17
36	193	285	-.31	.13	.90	-2.0	.82	-1.7	d .37	.26	68.4	69.0	G36
32	253	285	-1.73	.19	.90	-.7	.63	-1.8	c .30	.16	88.8	88.8	G32
25	238	285	-1.26	.16	.89	-1.1	.76	-1.4	b .32	.19	83.5	83.5	C25
21	189	285	-.24	.13	.85	-3.3	.81	-1.9	a .41	.27	74.7	67.9	C21
MEAN	174.7	285.0	.00	.18	1.01	.0	.99	-.2			76.7	77.0	
S.D.	72.9	.0	1.82	.13	.09	1.2	.18	.8			11.7	11.3	

From Table 4.16 (the red dotted box at the bottom of the table), the acceptable range calculated from the Total Infit MNSQ ± S.D. is 1.01 ± 0.09, which ranges from 0.92 to 1.10. The items that are out of this ranges are seen in the blue circles in the Infit MNSQ column.

However, the Linacre (2015) suggests the acceptable range for the Infit Mean Square (MNSQ) is between 0.8 and 1.2. Thus the 2 items that out of the range (0.8-1.2) are items 48 and 50, which are also the most difficult questions. Looking at the ZSTD in the Outfit column in Table 4.16, there are no items that have a z-score more then 2 or less than -2. The Point Measure Correlation, PT-Measure Corr is checked and there are no negative values found in this column. Thus there are no issues related to opposite responses.

The Item Measure table gives the information about each item. It is also convenient to detect any Item Outliers or Misfits.

Table 4.17
Item Fit Statistics: Measure Order of ATETv3

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
1	258	285	-1.92	.21	.98	.0	.89	-.3	.17	.15	90.5	90.5	Ra1
2	257	285	-1.88	.20	.97	-.1	.96	-.1	.19	.15	90.2	90.2	Ra2
3	183	285	-.14	.13	.95	-1.1	.88	-1.2	.33	.28	68.1	66.4	Ra3
4	154	285	.33	.13	1.08	2.4	1.05	.6	.23	.31	54.7	62.7	Ra4
5	149	285	.41	.13	.91	-2.8	.92	-.9	.40	.31	73.0	62.7	Ra5
6	163	285	.19	.13	1.00	.1	.96	-.4	.30	.30	59.6	63.0	Ra6
7	119	285	.89	.13	.97	-.7	.93	-.7	.38	.34	69.5	66.4	Ra7
8	133	285	.66	.13	.98	-.5	.97	-.3	.35	.33	68.1	64.1	Ra8
9	256	285	-1.84	.20	1.01	.1	1.05	.3	.13	.15	89.8	89.8	Ra9
10	272	285	-2.73	.29	.99	.1	.79	-.5	.13	.10	95.4	95.4	Ra10
11	233	285	-1.13	.16	1.04	.4	1.08	.5	.16	.20	81.8	81.8	Rb11
12	203	285	-.49	.14	.96	-.7	.94	-.5	.28	.25	74.0	71.9	Rb12
13	254	285	-1.76	.19	1.00	.1	.91	-.3	.16	.16	89.1	89.1	Rb13
14	69	285	1.83	.15	1.19	2.1	1.22	1.8	.22	.39	75.1	79.5	Rb14
15	255	285	-1.80	.20	.95	-.3	1.01	.1	.19	.16	89.5	89.5	Rb15
16	277	285	-3.24	.36	.99	.1	.69	-.7	.12	.08	97.2	97.2	Rb16
17	110	285	1.04	.13	.90	-2.0	.87	-1.2	.44	.35	74.0	68.4	Rb17
18	22	285	3.46	.25	1.03	.2	1.33	1.2	.35	.41	94.0	93.1	Rb18
19	263	285	-2.15	.23	1.03	.2	.98	.0	.10	.14	92.3	92.3	Rb19
20	232	285	-1.11	.16	1.04	.5	1.02	.2	.16	.20	81.4	81.4	Rb20
21	189	285	-.24	.13	.85	-3.3	.81	-1.9	.41	.27	74.7	67.9	C21
22	145	285	.47	.13	1.05	1.6	1.02	.3	.27	.32	61.8	62.9	C22
23	239	285	-1.29	.17	1.01	.1	.94	-.3	.19	.19	83.9	83.9	C23
24	11	285	4.38	.35	1.07	.3	1.21	.6	.31	.40	96.8	96.7	C24
25	238	285	-1.26	.16	.89	-1.1	.76	-1.4	.32	.19	83.5	83.5	C25
26	185	285	-.18	.13	1.03	.8	.98	-.2	.25	.27	65.3	66.9	C26
27	227	285	-.99	.15	1.00	.0	1.07	.5	.20	.21	79.3	79.7	C27
28	58	285	2.10	.16	.93	-.7	.85	-1.1	.47	.39	83.9	82.7	C28
29	73	285	1.75	.15	.97	-.4	.91	-.7	.43	.38	77.9	78.3	C29
30	152	285	.36	.13	.92	-2.5	.92	-.9	.39	.31	68.1	62.7	C30
31	266	285	-2.32	.24	1.01	.1	.87	-.4	.13	.13	93.3	93.3	G31
32	253	285	-1.73	.19	.90	-.7	.63	-1.8	.30	.16	88.8	88.8	G32
33	180	285	-.09	.13	1.04	1.0	.99	-.1	.25	.28	65.6	65.7	G33
34	187	285	-.21	.13	1.07	1.5	1.08	.8	.20	.27	64.9	67.4	G34
35	233	285	-1.13	.16	1.01	.1	.95	-.2	.20	.20	81.8	81.8	G35
36	193	285	-.31	.13	.90	-2.0	.82	-1.7	.37	.26	68.4	69.0	G36
37	254	285	-1.76	.19	.99	.0	1.02	.2	.16	.16	89.1	89.1	G37
38	116	285	.94	.13	1.02	.5	1.03	.3	.32	.35	66.3	67.0	G38
39	113	285	.99	.13	1.12	2.5	1.07	.7	.25	.35	59.6	67.7	G39
40	222	285	-.87	.15	1.07	1.0	1.06	.4	.15	.22	77.9	77.9	G40
41	203	285	-.49	.14	.97	-.5	.88	-1.0	.29	.25	71.9	71.9	E41
42	160	285	.23	.13	1.07	1.9	1.03	.4	.24	.30	58.2	62.9	E42
43	218	285	-.79	.14	1.00	.0	.92	-.5	.24	.23	75.8	76.6	E43
44	215	285	-.73	.14	.98	-.3	.87	-1.0	.27	.23	74.7	75.6	E44
45	199	285	-.42	.13	1.05	1.0	.99	.0	.21	.25	67.4	70.7	E45
46	36	285	2.79	.20	1.23	1.6	1.16	.9	.25	.41	85.6	88.9	E46
47	217	285	-.77	.14	1.08	1.2	1.03	.3	.16	.23	74.0	76.3	E47
48	14	285	4.06	.31	1.21	.9	1.75	1.7	.21	.40	95.8	95.7	E48
49	212	285	-.67	.14	.94	-.9	.86	-1.1	.30	.24	76.5	74.6	E49
50	1	285	7.58	1.09	1.39	.7	1.66	1.2	.05	.25	99.6	99.6	E50
51	240	285	-1.31	.17	.95	-.5	.85	-.8	.25	.19	84.2	84.2	W51
52	110	285	1.04	.13	1.07	1.3	1.01	.2	.30	.35	61.4	68.4	W52
53	70	285	1.81	.15	.97	-.3	1.01	.1	.40	.39	81.1	79.2	W53
54	157	285	.28	.13	1.03	1.0	1.07	.8	.27	.30	62.1	62.7	W54
55	173	285	.02	.13	.98	-.6	.93	-.8	.31	.29	64.9	64.3	W55
56	212	285	-.67	.14	1.10	1.5	1.13	1.0	.13	.24	72.3	74.6	Wp56
57	175	285	-.01	.13	.92	-2.2	.84	-1.8	.38	.28	63.9	64.6	Wp57
58	128	285	.74	.13	1.00	.1	.95	-.5	.34	.33	61.4	64.8	Wp58
59	198	285	-.40	.13	1.06	1.1	1.02	.3	.20	.26	67.0	70.4	Wp59
60	146	285	.46	.13	.94	-1.9	.88	-1.3	.39	.32	67.0	62.8	Wp60
MEAN	174.7	285.0	.00	.18	1.01	.0	.99	-.2			76.7	77.0	
S.D.	72.9	.0	1.82	.13	.09	1.2	.18	.8			11.7	11.3	

Items that have the same measures are in the same coloured boxes in Table 4.17. Items 17 and 52 have the same measure (green box), but 17 tests Reading while 52 tests Writing skills. Items 12 and 41 (red box) are from Reading and Grammar respectively. Items 49 and 56 (light blue) are from Grammar and Writing sections respectively. Meanwhile the last 2 pairs, Items 11 and 35 as well as 13 and 37 are from similar sections. 11 and 13 are from Reading while 35 and 37 are from Cloze section. However, after looking at the Infit MNSQ and the Outfit ZSTD, and each pair is from two different sections. So, the same measure for each pair does not really matter. So all items will be retained.

4.4.4 Principal Component Analysis

The Principal Component Analysis would provide evidence that the instrument is unidimensional.

Table 4.18
Principal Component Analysis of ATETv3

CONTRAST 1 FROM PRINCIPAL COMPONENT ANALYSIS OF
Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Empirical --			Modeled
Total raw variance in observations	=	91.6	100.0%		100.0%
Raw variance explained by measures	=	31.6	34.5%		34.6%
Raw variance explained by persons	=	7.6	8.3%		8.3%
Raw Variance explained by items	=	24.0	26.2%		26.3%
Raw unexplained variance (total)	=	60.0	65.5%	100.0%	65.4%
Unexplned variance in 1st contrast	=	4.4	4.8%	7.3%	

In Table 4.18, the raw variance explained by measures (orange box) is 34.5%, similar to the modeled value 34.6% (the orange boxes). This value should be between 40-60%, which is usually affected by the noise level. However, the unexplained variance in the 1st contrast is 4.8% (blue box), which is less then 15%, an acceptable value for noise level.

Table 4.19
Largest Standardized Residual Correlations of ATETv3

RESIDUL CORRELN	ENTRY NUMBER	Item	ENTRY NUMBER	Item
.56	14	Rb14	38	G38
.43	5	Ra5	21	C21
.43	5	Ra5	36	G36
.42	1	Ra1	15	Rb15
.40	14	Rb14	48	E48
.40	46	E46	48	E48
.39	32	G32	36	G36
.38	4	Ra4	39	G39
-.36	13	Rb13	24	C24
-.36	39	G39	41	E41

The noise is caused by the items listed in Table 4.19. However, it is considered dismissable as the residual correlations is lesser than 0.7. As such these items are considered as different type of items. Table 4.19 shows that only 2 pairs are testing the same type, E46 & E48, and G32 & G36. This explains why the pairs despite having the large standardised residual correlations, the correlation values are lower than 0.7.

4.4.5 Differential Item Functioning (DIF)

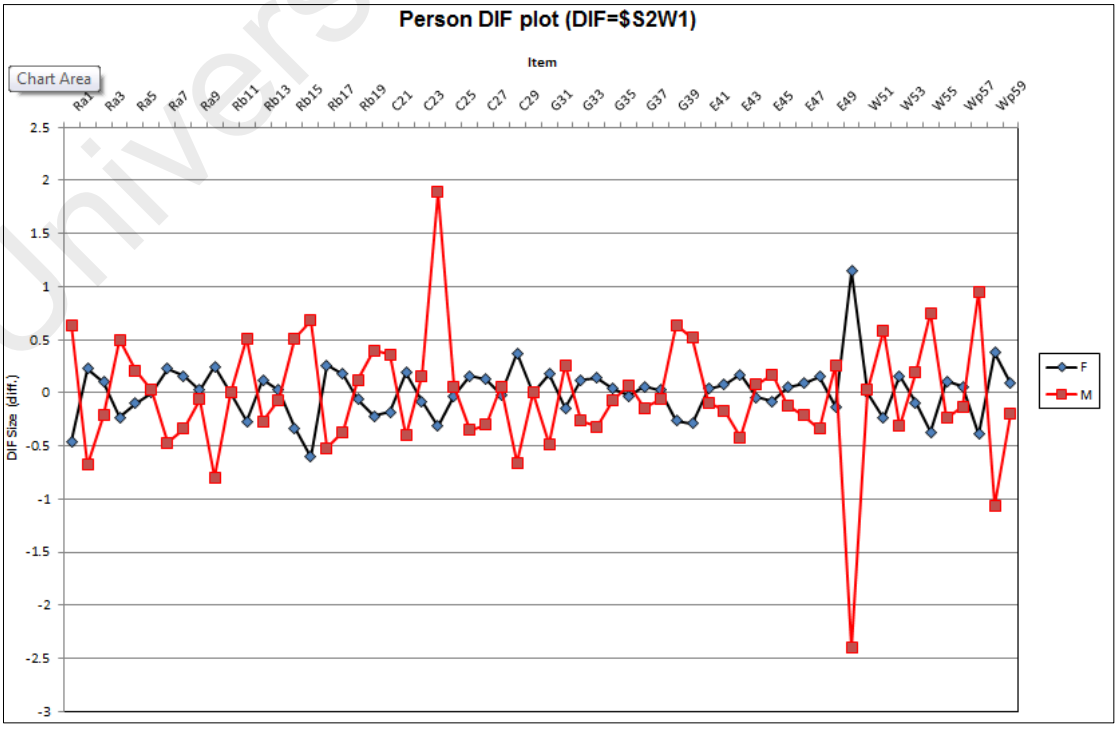

Figure 4.16. Person-gender DIF plot according to difference in size for ATETv3

This analysis is to find out if this test is gender biased. Figure 4.16 shows items Ra1, Ra2, Ra4, Ra7, Ra10, Rb12. Rb15, Rb16, Rb17, Rb18, Rb20, C21, C22, C24, C26, C29, G31, G39, G40, E43, E50, W52, W55, Wp59 have the limit approximately ±0.5, which shows that these items are not gender biased. Judging from Figure 4.16, there are 2 items that have a big difference in size, items 24 and 50. However as mentioned earlier the candidates may not be familiar with the type of grammar items. As such these items will be retained. Furthermore, compared to versions 1 and 2, this test is not gender biased as it is seen that Figure 4.16 has the graph or males and females nearing 0. Thus,there is no difference in their ways of responding to the items, which also supports test fairness.

4.4.6 Person Misfit Order

Table 4.20
Person Misfit Order of ATETv3

Person STATISTICS: MISFIT ORDER

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Person
112	45	60	1.42	.37	.96	-.1	8.53	6.7	A .50	.56	83.3	83.4	KM112
110	27	60	-.46	.31	1.43	3.1	3.14	4.2	B .22	.48	61.7	72.6	MM110
25	32	60	.01	.31	1.47	3.2	2.87	4.2	C .23	.50	58.3	73.2	KF025
228	40	60	.82	.33	1.51	2.8	2.62	3.2	D .26	.54	66.7	78.0	KF228
216	32	60	.01	.31	1.20	1.5	2.53	3.6	E .36	.50	68.3	73.2	KF216
18	34	60	.20	.31	1.46	3.0	2.53	3.6	F .26	.51	58.3	73.9	SF018
153	26	60	-.55	.31	.90	-.8	2.45	3.1	G .47	.47	80.0	72.8	KF153
202	39	60	.71	.33	1.39	2.3	2.31	2.9	H .32	.53	66.7	77.2	KF202
38	46	60	1.56	.38	1.71	2.7	2.18	1.9	I .25	.56	75.0	84.4	SF038
29	37	60	.50	.32	1.36	2.3	2.02	2.5	J .33	.53	71.7	75.6	SF029
156	27	60	-.46	.31	1.26	2.0	1.94	2.3	K .33	.48	65.0	72.6	KF156
17	37	60	.50	.32	1.46	2.8	1.90	2.3	L .31	.53	58.3	75.6	MM017
155	27	60	-.46	.31	1.15	1.2	1.78	2.0	M .38	.48	71.7	72.6	MF155
78	29	60	-.27	.31	1.24	1.8	1.76	2.0	N .36	.49	63.3	72.6	MM078
258	29	60	-.27	.31	1.24	1.8	1.76	2.0	O .36	.49	63.3	72.6	SF258
11	43	60	1.17	.35	1.25	1.3	1.73	1.6	P .41	.55	73.3	81.1	KM011
43	37	60	.50	.32	.94	-.4	1.67	1.8	Q .52	.53	81.7	75.6	SF043
275	42	60	1.05	.34	1.18	1.0	1.62	1.5	R .43	.55	75.0	79.9	MF275
217	37	60	.50	.32	.99	.0	1.60	1.7	S .50	.53	78.3	75.6	SF217
123	44	60	1.29	.36	1.33	1.6	1.58	1.3	T .40	.55	78.3	82.3	MM123
151	26	60	-.55	.31	1.35	2.5	1.58	1.5	U .29	.47	63.3	72.8	SF151
254	45	60	1.42	.37	1.19	.9	1.52	1.1	V .46	.56	80.0	83.4	SF254
272	34	60	.20	.31	1.20	1.5	1.51	1.5	W .40	.51	71.7	73.9	KM272
14	42	60	1.05	.34	1.33	1.7	1.50	1.2	X .39	.55	75.0	79.9	KM014
108	28	60	-.36	.31	1.41	2.9	1.49	1.4	Y .29	.48	56.7	72.5	MM108
149	28	60	-.36	.31	1.41	2.9	1.49	1.4	Z .29	.48	56.7	72.5	KF149
BETTER FITTING OMITTED													
117	37	60	.50	.32	.78	-1.5	.61	-1.3	z .63	.53	81.7	75.6	SM117
243	51	60	2.42	.46	.78	-.6	.47	-.7	y .66	.57	93.3	90.1	SM243
142	34	60	.20	.31	.77	-1.8	.73	-.9	x .61	.51	81.7	73.9	SF142
174	39	60	.71	.33	.77	-1.5	.59	-1.3	w .64	.53	80.0	77.2	SF174
82	31	60	-.08	.31	.77	-2.0	.61	-1.3	v .61	.50	78.3	72.9	KF082
262	32	60	.01	.31	.76	-1.9	.61	-1.4	u .62	.50	78.3	73.2	SF262
269	34	60	.20	.31	.76	-1.9	.62	-1.3	t .63	.51	81.7	73.9	SM269
246	55	60	3.54	.61	.75	-.5	.53	-.1	s .63	.57	96.7	94.4	KF246
197	34	60	.20	.31	.75	-2.0	.65	-1.2	r .62	.51	85.0	73.9	MF197
187	34	60	.20	.31	.74	-2.1	.58	-1.5	q .64	.51	81.7	73.9	MF187
121	24	60	-.74	.31	.73	-2.2	.59	-1.2	p .59	.46	81.7	73.2	SM121
160	46	60	1.56	.38	.73	-1.3	.54	-.9	o .67	.56	91.7	84.4	SF160
241	51	60	2.42	.46	.69	-1.0	.37	-.9	n .70	.57	93.3	90.1	SF241
231	55	60	3.54	.61	.64	-.8	.17	-.8	m .69	.57	96.7	94.4	MF231
233	55	60	3.54	.61	.64	-.8	.17	-.8	l .69	.57	96.7	94.4	SF233
237	53	60	2.90	.52	.62	-1.1	.36	-.7	k .70	.57	95.0	92.3	SF237
236	53	60	2.90	.52	.62	-1.1	.35	-.7	j .70	.57	95.0	92.3	SM236
247	52	60	2.65	.49	.61	-1.2	.27	-1.1	i .72	.57	95.0	91.3	KF247
248	52	60	2.65	.49	.61	-1.2	.27	-1.1	h .72	.57	95.0	91.3	SF248
244	51	60	2.42	.46	.59	-1.5	.26	-1.2	g .73	.57	93.3	90.1	MM244
232	52	60	2.65	.49	.57	-1.4	.25	-1.1	f .73	.57	95.0	91.3	KF232
235	56	60	3.95	.68	.55	-1.0	.11	-.8	e .69	.56	98.3	95.3	MF235
242	53	60	2.90	.52	.53	-1.4	.19	-1.2	d .74	.57	95.0	92.3	MF242
249	54	60	3.19	.56	.50	-1.4	.17	-1.1	c .73	.57	96.7	93.4	SF249
238	59	60	6.70	1.42	.20	-.8	.01	-1.4	b .54	.45	100.0	99.0	SF238
245	59	60	6.70	1.42	.20	-.8	.01	-1.4	a .54	.45	100.0	99.0	MF245
MEAN	36.8	60.0	.58	.34	.98	.0	.99	.0			76.7	77.0	
S.D.	7.0	.0	.98	.11	.19	1.1	.61	1.0			7.9	5.4	

Checking the Infit MNSQ in Table 4.20, the acceptable range is Total Infit MNSQ ± S.D., 0.98 ± 0.19, which is between 0.79 and 1.17. The persons who are out of the acceptable range are in the blue box in Table 4.20. Then the Outfit ZSTD is looked into. The value should be between -2 and 2, The ones beyond this range is seen in the red boxes. Next is the PT-MEASURE CORR, which should be positive. All Persons have positive measures. This shows that all persons are behaving in the expected manner. When these persons who did not fit the Infit MNSQ acceptable range and the Outfit ZSTD, these persons were deleted from the file, with a PDFILE command. The statistics is seen in Table 4.21.

Table 4.21
Summary Statistics after Deleting the Misfit Persons in ATETv3

SUMMARY OF 269 MEASURED (NON-EXTREME) Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	36.8	60.0	.68	.34	.99	.0	.96	.0
S.D.	6.7	.0	.87	.07	.17	1.0	.38	.8
MAX.	56.0	60.0	4.30	.73	1.81	3.1	3.29	4.3
MIN.	21.0	60.0	-.93	.31	.50	-2.1	.10	-1.3

REAL RMSE	.36	ADJ.SD	.80	SEPARATION	2.23	Person RELIABILITY	.83	
MODEL RMSE	.35	ADJ.SD	.80	SEPARATION	2.31	Person RELIABILITY	.84	
S.E. OF Person MEAN = .05								

MAXIMUM EXTREME SCORE: 2 Persons
DELETED: 14 Persons
VALID RESPONSES: 99.9%

SUMMARY OF 271 MEASURED (EXTREME AND NON-EXTREME) Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	37.0	59.0	.73	.35				
S.D.	7.0	.0	1.03	.15				
MAX.	59.0	59.0	7.20	1.88				
MIN.	21.0	59.0	-.93	.31				

REAL RMSE	.39	ADJ.SD	.96	SEPARATION	2.45	Person RELIABILITY	.86	
MODEL RMSE	.38	ADJ.SD	.96	SEPARATION	2.52	Person RELIABILITY	.86	
S.E. OF Person MEAN = .06								

Person RAW SCORE-TO-MEASURE CORRELATION = .94 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE RELIABILITY = .80 (approximate due to missing data)

Table 4.22
Summary Statistics after Deleting the Misfit Item in ATETv3

SUMMARY OF 59 MEASURED (NON-EXTREME) Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	169.9	271.0	.00	.17	1.01	.0	.96	-.2
S.D.	67.6	.0	1.65	.07	.07	1.2	.12	.8
MAX.	264.0	271.0	4.99	.43	1.26	2.9	1.16	1.4
MIN.	8.0	271.0	-3.20	.13	.86	-3.3	.62	-2.1

REAL RMSE	.19	ADJ.SD	1.64	SEPARATION	8.60	Item	RELIABILITY	.99
MODEL RMSE	.19	ADJ.SD	1.64	SEPARATION	8.78	Item	RELIABILITY	.99
S.E. OF Item MEAN = .22								

MAXIMUM EXTREME SCORE: 1 Items
UMEAN=.000 USCALE=1.000

SUMMARY OF 60 MEASURED (EXTREME AND NON-EXTREME) Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	165.1	269.0	.14	.20				
S.D.	70.4	.0	1.94	.22				
MAX.	262.0	269.0	8.11	1.83				
MIN.	.0	269.0	-3.20	.13				

REAL RMSE	.30	ADJ.SD	1.91	SEPARATION	6.33	Item	RELIABILITY	.98
MODEL RMSE	.30	ADJ.SD	1.91	SEPARATION	6.38	Item	RELIABILITY	.98
S.E. OF Item MEAN = .25								

Item RAW SCORE-TO-MEASURE CORRELATION = -.95 (approximate due to missing data)
15871 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 14955.06 with 15544 d.f. p=.9996

Table 4.22 (after deleting 1 items) shows similar results before the deletion as in Table 4.15. Therefore it is not necessary to delete any items from this analysis. As such further analysis will use all 285 persons and 60 items, although there are extreme persons and items.

The interpretation of Figure 4.17 is that the persons are ordered from high measure to low measure while the items are also arranged from low to high measure. Thus the Persons has 50% chances of getting most of the easier items correct (1) and 50% chances of getting the more difficult items wrong (0). The top left corner is where the more able persons respond to the easier items (from Figure 4.17, mostly "1", the blue box), while the bottom right hand corner (red box) should have more or almost all "0". But this is evident in Figure 4.17. So this clearly shows that there is no discrepency in the results compared to the expected pattern. Items can be retained.

4.4.8 Item Characteristic Curve

The Item Characteristic Curve (ICC) plots the model-expected item characteristic curve. This is the Rasch model prediction for each measure relative to item difficulty. The steeper the ICC, the more discriminating it is between high and low achievers (Linacre, 2015). Figure 4.18 shows the overall results of all the 60 items in ATETv3.

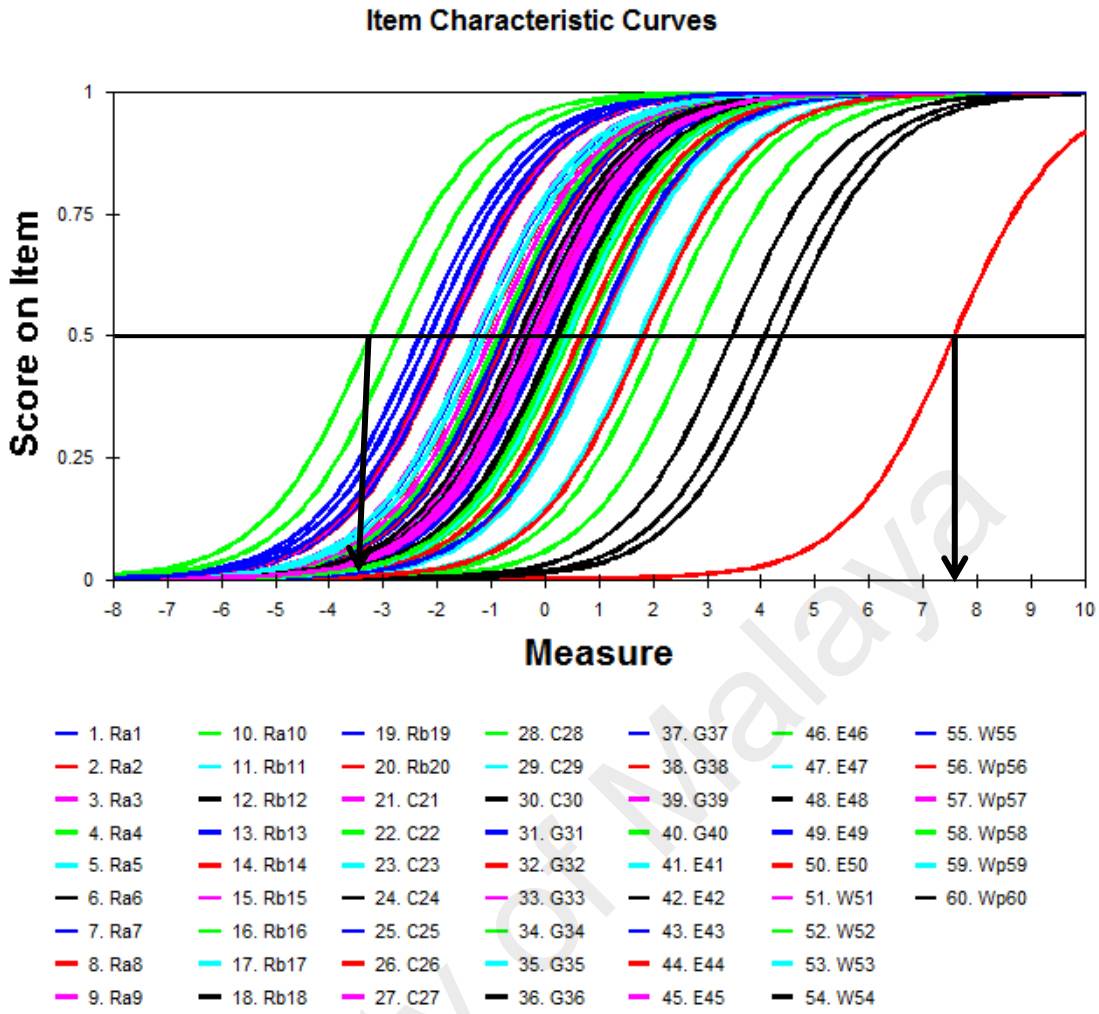

Figure 4.18 Item Characteristic Curve for all 60 items in ATETv3

The horizontal black line show the 50% chances of getting an item correct on ATETv3. Drawing a line parallel to the vertical axis from the interception of the horizontal black line to the coloured sigmoid, the corresponding item location measure is shown on the horizontal axis. It can be seen that the items spread from the easiest being on the far left (Rb16), with the lowest measure while the most difficult on the far right (E50) with the highest measure.

The overall result is scrutinised according to the three sections, Reading, Grammar and Writing. Figure 4.19 shows the results of ICC for the Reading section

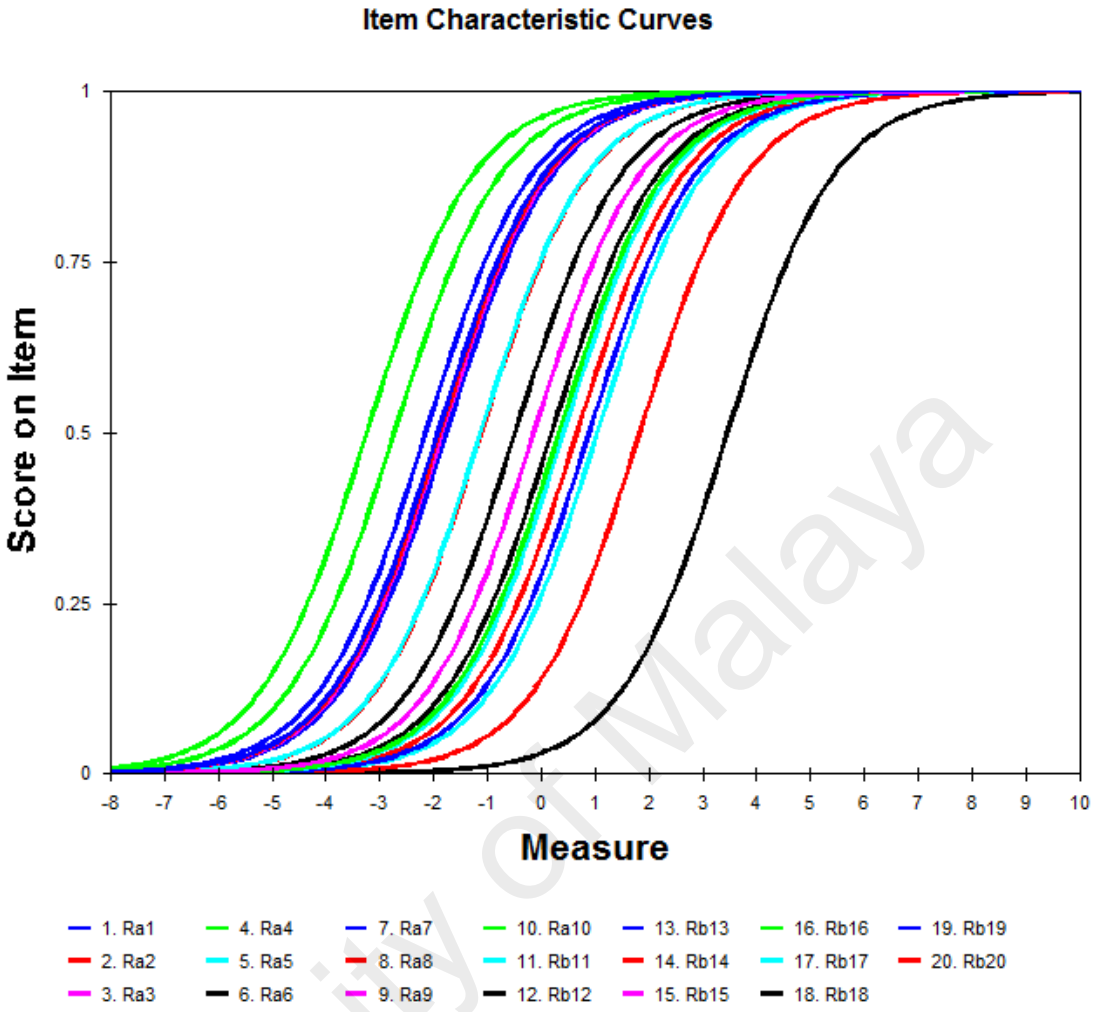

Figure 4.19 Item Characteristic Curve for Reading section of ATETv3

Figure 4.19 shows that Rb16 is the easiest (far left) and Rb18 the most difficult (far right). The difficulty level of the items in this section seems well distributed. However, items Ra1 (-1.92 logits), Ra2 (-1.88 logits) , Ra9 (-1.84 logits), Rb13 (-1.76 logits) and Rb15 (-1.80 logits) seem to appear very near each other. Another group of items which have almost similar measures are items Ra 4 (0.33 logits), Ra5 (0.41 logits) and Ra6 (0.14 logits). The ICCs for the above mentioned items appear as a thicker curve.

Meanwhile for the cloze items, the characteristic curve is seen in Figure 4.20.

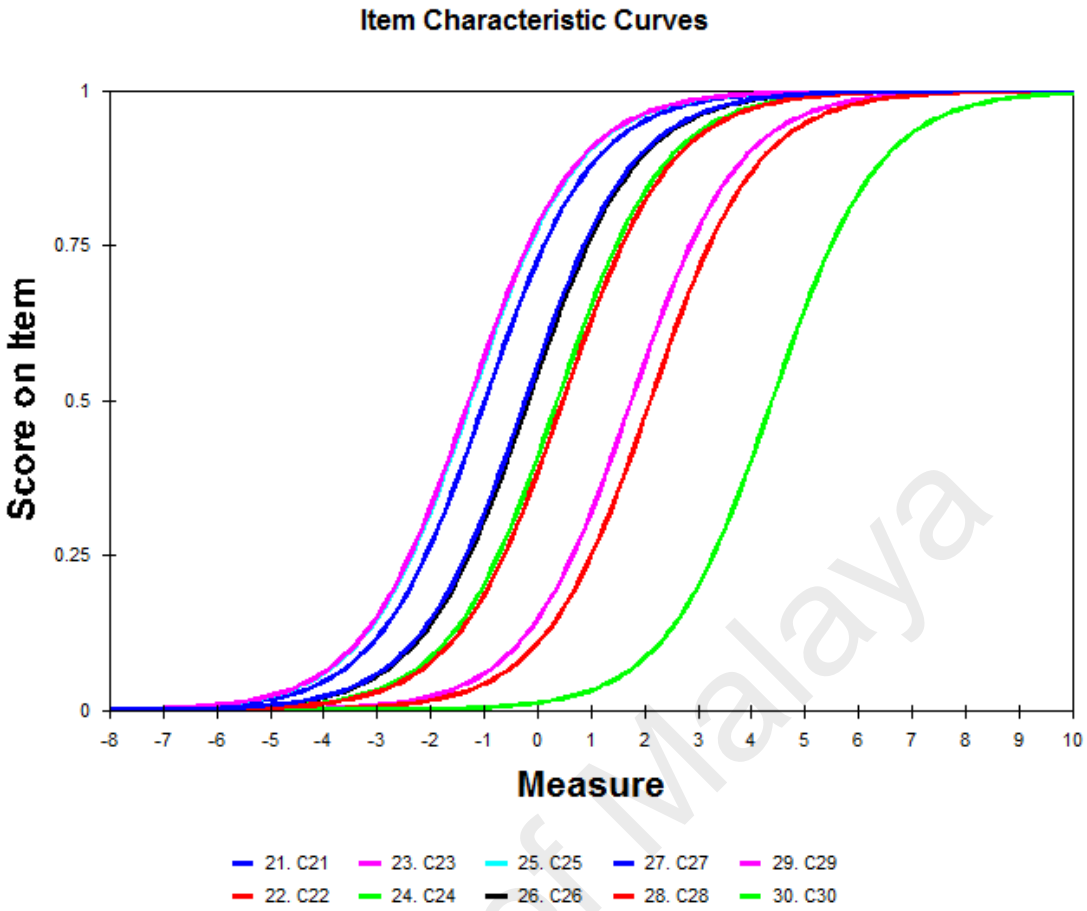

Figure 4.20 Item Characteristic Curve for the Cloze section of ATETv3

In this Cloze Section, there are 2 items that have almost the same measures, i.e. items C23 (-1.29 logits) and C25 (-1.26 logits). Therefore, the sigmoid appears as one. Figure 4.20 also show that the easiest item in this section is C23 and the most difficult one is C24.

Next is the Grammar section, which culminates from items G31 right up to E50. However, for analysis purpose, these items are subdivided into two parts: G31 to G40 are grammar items in sentence completion form while E41 to E50 are identification of errors in grammar. Figure 4.21 shows the overall results for the Grammar component.

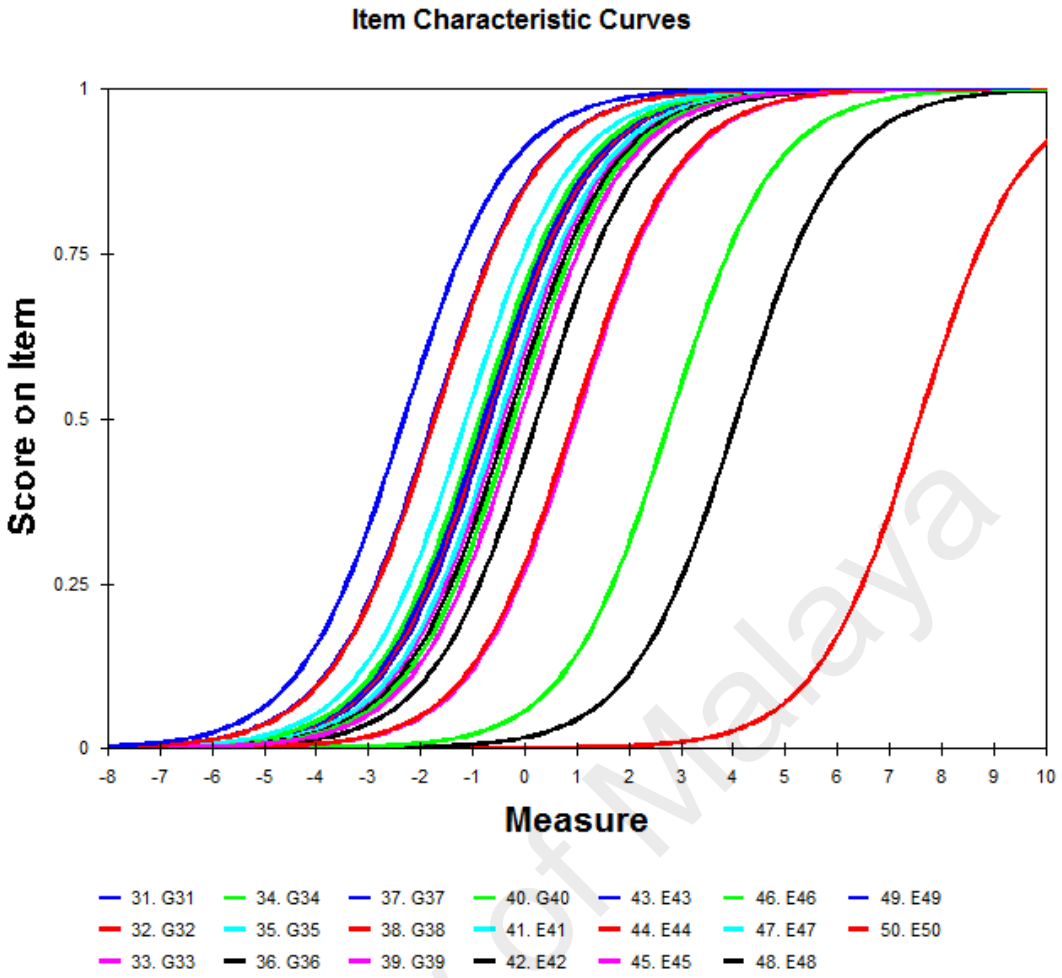

Figure 4.21. Item Characteristic Curve for Grammar section of ATETv3

Although there are no overlapping sigmoids for this section, there are a number of items grouped with minimal difference in measures. Items G38 (0.94 logits) and G39 (0.99 logits) are near to each other. Another group of almost similar sigmoids are for items G34 (-0.21 logits), G36 (-0.31 logits), E41 (-0.49 logits) and E45 (-0.42 logits). The items G40 (-0.87 logits), E43 (-0.79 logits), E44 (-0.73 logits) and E47 (-0.77 logits) are clustered near to each other just like the other clustered sigmoids of items G32 (-1.73 logits) and G37 (-1.76 logits).

In terms of difficulty levels, for the Sentence Completion section, the easiest item is G31 and the most difficult is G39. For the Error Identification section, the easiest item is E43 and the most difficult is

E50.

It is also noticed that there is an obvious gap between items E49 and E50. However, for selection of candidates into the programme, what is required is only the cut off point. There need not be a range of items in the difficult level of items.

Final section of the ATETv3 is the Writing section with subdivision of Paragraph Improvement, marked by W and Paraphrasing, marked by Wp. Figure 4.22 shows the overall results of the Writing component.

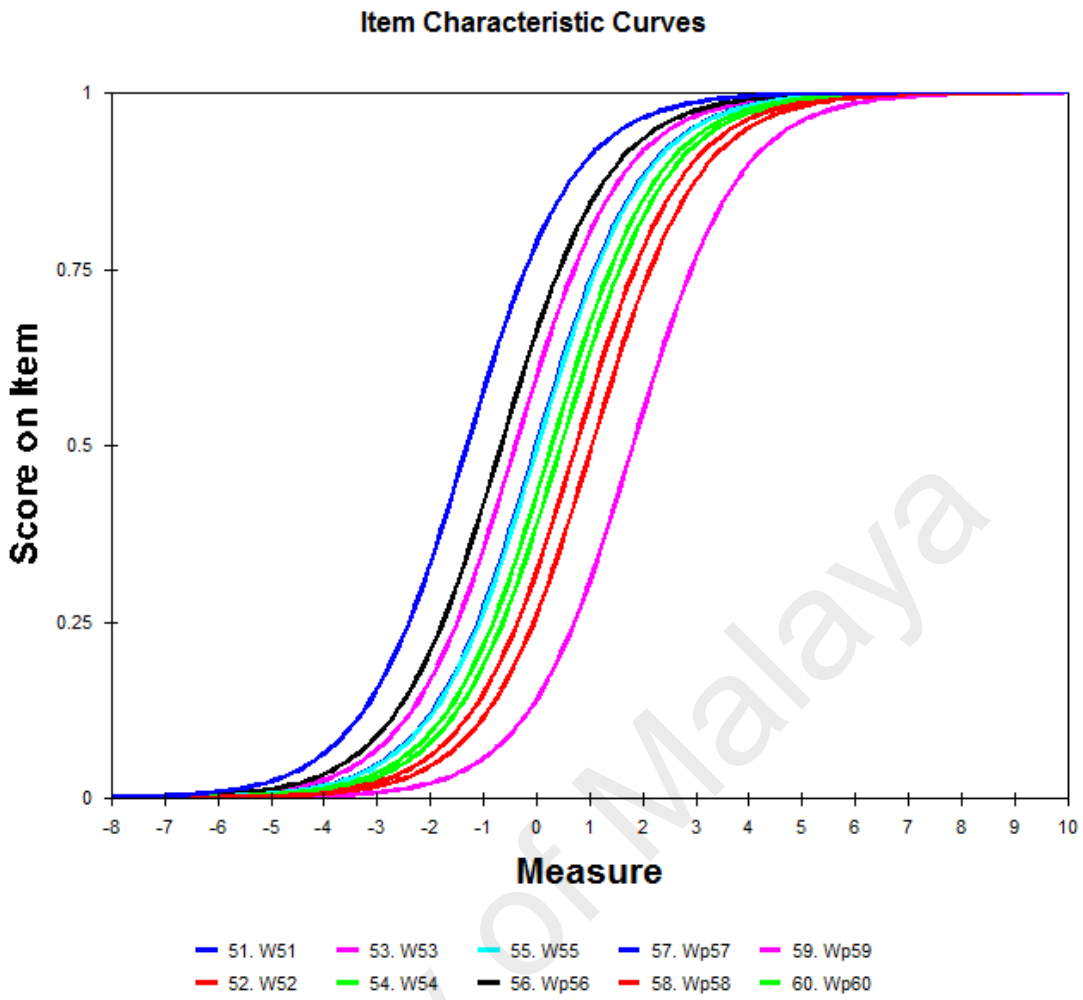

Figure 4.22 Item Characteristic Curve for Writing section of ATETv3

Figure 4.22 shows that all items are well spread. There are two items almost similar in measures, W55 (0.02 logits), and Wp57 (-0.01 logits).

4.4.9 Bubble Chart

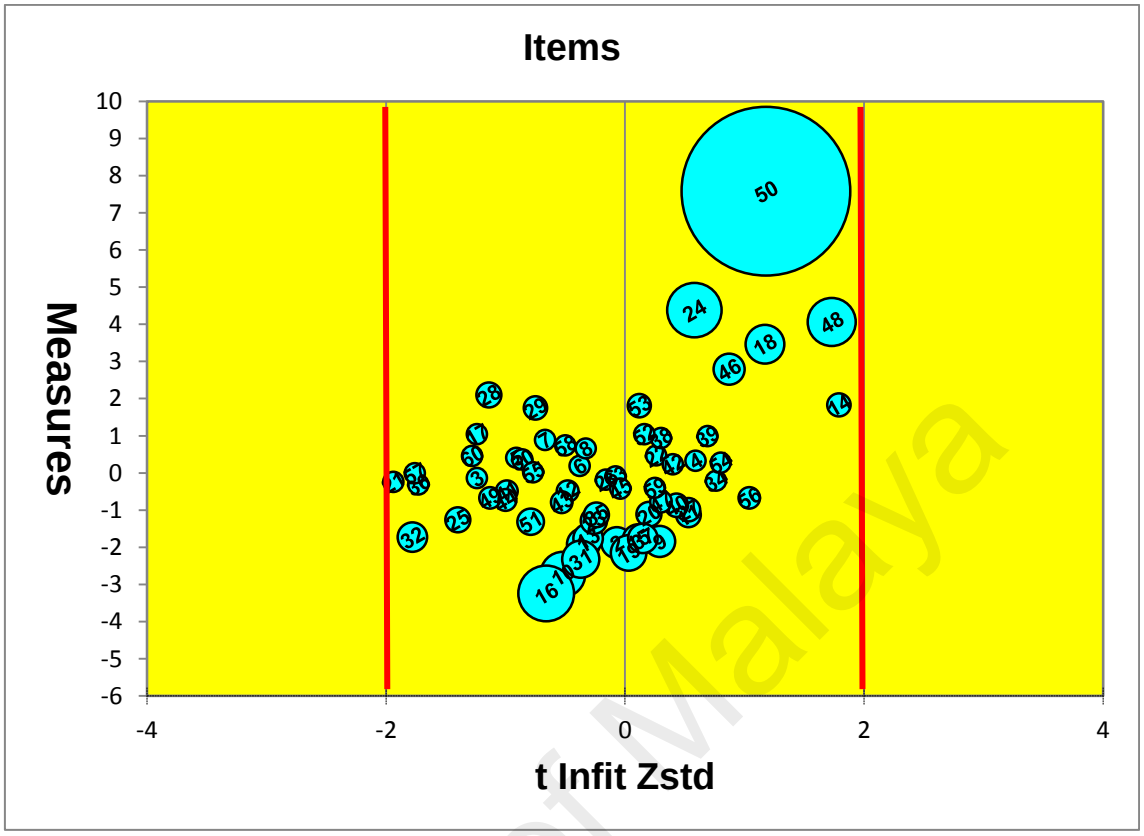

Figure 4.23. Bubble chart of items of ATETv3

Figure 4.23 shows that all items are within the acceptable limit, i.e. -2 < t < 2, represented by the red lines. The 60 items in ATETv3 are well spread vertically as well. It is also noticed that the bubbles are rather small compared to ATETv2, which indicated that the standard error has reduced with the modification of items. However, the bigger bubbles are the more difficult items, where a lot more of guessing must have taken place, thus increasing the size of the error.

4.5 Essay Component

There was one essay question, which requires candidates to give their opinion on an issue. They were not expected to take sides, but were required to provide evidence to justify their opinion in about 350 words. Marks were awarded according to content, language and organisation.

Three raters were selected. Rater 1 is a lecturer with 3 years teaching experience, while Rater 2 and 3 are lecturers who have more than 15 years of teaching experience (Raters' profiles are in Appendix L). The following, Table 4.23, is a general descriptive statistics of the scores for the three versions of ATET.

Table 4.23
Descriptive Statistics of the Essay Scores According to the Raters and Versions of the Test

ATETv1

	N	Minimum	Maximum	Mean	Std. Deviation
Rater 1	120	7	18	12.50	1.997
Rater 2	120	10	17	13.27	1.719
Rater 3	120	10	17	13.17	1.731

ATETv2

	N	Minimum	Maximum	Mean	Std. Deviation
Rater 1	285	6	19	11.76	2.117
Rater 2	285	6	18	11.51	2.413
Rater 3	285	8	17	11.92	2.079

ATETv3

	N	Minimum	Maximum	Mean	Std. Deviation
Rater 1	285	6	19	11.82	2.115
Rater 2	285	6	18	11.55	2.421
Rater 3	285	8	17	11.95	2.088

Table 4.23 shows that the Mean and the Standard Deviation for all the three tests are almost the same for Raters 2 and 3. These results were subjected to a correlational analysis among the three Raters which is depicted in Table 4.24.

Table 4.24

Pearson Product Moment Correlations of the Essay Scores According to the Raters and Versions of the Test

ATETv1

		Rater 1	Rater 2	Rater 3
Rater 1	Pearson Correlation		.777**	.816**
Rater 2	Pearson Correlation			.954**

**. Correlation is significant at the 0.01 level (2-tailed).

ATETv2

		Rater 1	Rater 2	Rater 3
Rater 1	Pearson Correlation		.364**	.439**
Rater 2	Pearson Correlation			.918**

**. Correlation is significant at the 0.01 level (2-tailed).

ATETv3

		Rater 1	Rater 2	Rater 3
Rater 1	Pearson Correlation		378**	.455*
Rater 2	Pearson Correlation			.919**

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4.24 displays the strong and significant correlations between Rater 2 and Rater 3 in all the versions (approximately 0.9 for all the three versions of the test). However, Rater 1 did not correlate strongly with Raters 2 and 3.

A software, EduG 6.1, was utilised to find out what was the optimum number of raters for the Essay section. This software is based on the Generalizability Theory. Table 34 shows the result of the analysis using EduG 6.1. With reference to the Analysis of Variance in Table 4.25, the bulk of the error comes from the persons (more than 58.1%) and not the raters (0.7%). This is confirmed by the next part, the measurement design, where the G-coefficient is 0.81.

Table 4.25
Optimum Number of Raters for Essay Item

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
Person	P	285	INF	
Rater	R	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	3017.19	284	10.62	2.86	2.86	2.86	58.1	0.30
R	23.83	2	11.92	0.03	0.03	0.03	0.7	0.03
PR	1155.50	568	2.03	2.03	2.03	2.03	41.2	0.12
Total	4196.53	854					100%	

G Study Table
(Measurement design P/R)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	2.86		.....		.....	
	.....	R	.....		0.01	1.7
	.....	PR	0.68	100.0	0.68	98.3
Sum of variances	2.86		0.68	100%	0.69	100%
Standard deviation	1.69		Relative SE: 0.82		Absolute SE: 0.83	
Coef_G relative	0.81					
Coef_G absolute	0.81					

Grand mean for levels used: 11.77
Variance error of the mean for levels used: 0.02
Standard error of the grand mean: 0.15

This study was also aimed to disclose the optimum number of raters to ensure reliability. The Decision Study or the D-study was done and Table 4.26 displays the results. Looking at Table 4.28, the G-coefficient touches 0.8 with recommended 4 raters.

Table 4.26
Decision Study for the Optimum Number of Raters

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	19	INF	19	INF	19	INF	19	INF	19	INF	19	INF
R	3	INF	1	INF	2	INF	3	INF	4	INF	5	INF
Observ.		57		19		38		57		76		95
Coef_G rel.		0.78		0.54		0.70		0.78		0.82		0.85
rounded		0.78		0.54		0.70		0.78		0.82		0.85
Coef_G abs.		0.77		0.53		0.69		0.77		0.82		0.85
rounded		0.77		0.53		0.69		0.77		0.82		0.85
Rel. Err. Var.		0.64		1.92		0.96		0.64		0.48		0.38
Rel. Std. Err. of M.		0.80		1.38		0.98		0.80		0.69		0.62
Abs. Err. Var.		0.66		1.98		0.99		0.66		0.50		0.40
Abs. Std. Err. of M.		0.81		1.41		1.00		0.81		0.70		0.63

However, it is noticed that Rater 2 and 3 seem to correlate significantly high, another D-study was done with only 2 raters. Table 4.27 shows this result.

Table 4.27
Decision Study for the Two Experienced Raters

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	285	INF	285	INF	285	INF	285	INF	285	INF
R	2	INF	1	INF	3	INF	4	INF	5	INF
Observ.		570		285		855		1140		1425
Coef_G rel.		0.95		0.91		0.97		0.98		0.98
Rounded		0.95		0.91		0.97		0.98		0.98
Coef_G abs.		0.94		0.89		0.96		0.97		0.98
Rounded		0.94		0.89		0.96		0.97		0.98
Rel. Err. Var.		0.23		0.47		0.16		0.12		0.09
Rel. Std. Err. of M.		0.48		0.68		0.39		0.34		0.31
Abs. Err. Var.		0.27		0.55		0.18		0.14		0.11
Abs. Std. Err. of M.		0.52		0.74		0.43		0.37		0.33

From Table 4.27, even with 1 rater, the G-coefficient is 0.91. This translates that one rater is more than enough to ensure reliability provided the rater is an experienced one. Thus, the selection of raters seemed more

favourable if they are fully trained raters.

Three separate analysis were done to see the correlation between the multiple choice items that test writing skills in ATETv3 and the essay scores of Rater 1, Rater 2 and Rater 3, Table 4.28 summarises the relationship.

Table 4.28
Correlation between Writing Skill Items on ATETv3 and Raters

	Pearson Correlation
	Writing ATETv3
Rater 1	0.227*
Rater 2	0.724*
Rater 3	0.656*
Average Essay Score	0.629*

N=285
*Significant at the 0.01 level (2 tailed)

Table 4.28 shows there is higher correlation between the multiple choice items and the essays marked by Rater 2 and Rater 3. With the average of the three raters' scores, the correlation is quite significant. This proves further that the more experience the rater has, the more reliable the scoring becomes.

To triangulate the inter-rater reliability, FACETS was employed. The ATETv3 essay question was used to run the analysis with 3 raters. Table 4.29 shows the summary of the results.

Table 4.29
Summary of Examinee Measure

Model, Populn: S.D. 6.82	Separation 6.41	Strata 8.89	Reliability .98
Model, Sample: S.D. 6.83	Separation 6.43	Strata 8.90	Reliability .98

This table shows that the person reliability is more than 0.8. Thus, there is high person reliability (0.98) for this essay question. Then the rater measures are scrutinized. Table 4.30 shows this results.

Table 4.30

Raters Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Raters
3208	285	11.26	11.36	2.87	.10	.82	-2.2	.86	-1.4	-10.0	.98	.97	2 2
3319	285	11.65	11.79	1.67	.11	.74	-3.1	.63	-4.2	-10.0	.97	.97	3 3
3898	285	13.68	13.84	-4.55	.11	.98	-.1	.88	-1.0	-10.0	.97	.97	1 1
3475.0	285.0	12.19	12.33	.00	.11	.85	-1.9	.79	-2.3		.97		Mean (Count: 3)
302.5	.0	1.06	1.08	3.25	.00	.10	1.3	.12	1.4		.00		S.D. (Population)
370.5	.0	1.30	1.32	3.98	.00	.12	1.5	.14	1.8		.00		S.D. (Sample)

Model, Populn: RMSE .11 Adj (True) S.D. 3.25 Separation 30.86 Strata 41.48 Reliability 1.00
Model, Sample: RMSE .11 Adj (True) S.D. 3.98 Separation 37.80 Strata 50.73 Reliability 1.00
Model, Fixed (all same) chi-square: 2802.8 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

The table above shows that Rater 2 is the strictest while Rater 1 is most lenient. The inter-rater relaiability is 1. However, when looking closely into the measure, Raters 2 and 3 have positive measures 1.67 and 2.87 logits respectively while Rater 3 has -4.55 logits. This also triangulates the fact that Rater 1 has less experience compared to Raters 2 and 3.

4.6 Criterion Validity

The criterion validity was established with the Malaysian University English Test (MUET) and the First Semester's Grade Point Average (GPA). This is because the content of the ATETv3 is similar to the MUET content namely Reading and Writing as well as the various papers offered in the first semester, Reading, Writing and Grammar. For this analysis, the results of the final version of the test, ATETv3, was used. Table 4.31 shows the correlational analysis of these two scores.

Table 4.31
Correlations between ATETv3 and MUET

	Pearson Correlation	
	MUET	GPA
ATETv3	0.826**	0.779**
N	285	285

** Significant at the 0.01level (2-tailed)

The result of this correlation shows the Pearson Correlation is more than 0.8 for MUET and close to 0.8 for the GPA. This indicates that ATETv3 has a good criterion validity.

Together with this, a multiple linear regression was carried out. This was to check if the results of the ATETv3 can actually predict the MUET bands and the GPA results for three papers, Reading, Writing and Grammar. This was done using SPSS. Results are in Table 4.32.

Table 4.32
Description of ATETv3, MUET and GPA

Variable	**Mean**	**Std. Deviation**
ATETv3 (independent variable)	35.85	6.956
MUET (dependent variable)	3.76	0.659
GPA (dependent variable)	2.90	0.445

Table 4.32 shows the description of both independent and dependent variables with their respective means and standard deviation (SD). The SD is quite big for the ATETv3 as it covers a range of items with three different skills while the SD for MUET is small as the MUET results had only 5 bands for this sample (Band 2 to Band 6) and SD for GPA is the smallest ranging from 1 to 4.

Then, a multiple linear regression was run to predict the MUET results and GPA from the ATETv3. This variable statistically significantly predict

MUET results, F(2, 282) = 405.984, p < 0.000 with an R^2 of 0.742 as seen in Table 4.33.

Table 4.33
Multiple Linear Regression

Model: 1	R^2 = 0.742	
	β	Sig.
(Constant)	-2.771	.051
MUET	5.822	.000
GPA	5.756	.000

df Regression = 2
df Residual = 282

Table 4.33 shows the R^2 =0.742, which translates to 74.2% of the candidates can predict their MUET bands and GPA from the ATETv3 scores.

The Reading and Writing components of MUET, GPA and ATETv3 were correlated. The result in Table 4.34 shows significant coefficients for all corresponding components.

Table 4.34
Correlation of Corresponding Components of MUET and ATETv3

	Pearson Correlation			
	RMUET	WMUET	ReadingGPA	Writing GPA
RATETv3	0.817**		0.726**	
WATETv3		0.802**		0.735**
N=285				

** Significant at the 0.01 level (2-tailed)

The results is definitely favourable and in conclusion, it is proven that the ATETv3 has a positive predictive ability.

4.7 Construct Validity

The construct validity is determined by the dimensionality analysis. This is determined by the PCAR for ATETv3

Table 4.35
Table of Standardized Residuals Variance (in Eigenvalue units)

		Empirical		Modeled
Total raw variance in observations	=	91.6	100.0%	100.0%
Raw variance explained by measures	=	31.6	34.5%	34.6%
Raw variance explained by persons	=	7.6	8.3%	8.3%
Raw Variance explained by items	=	24.0	26.2%	26.3%
Raw unexplained variance (total)	=	60.0	65.5% 100.0%	65.4%
Unexplned variance in 1st contrast	=	**4.4**	**4.8**% 7.3%	
Unexplned variance in 2nd contrast	=	3.6	3.9% 6.0%	
Unexplned variance in 3rd contrast	=	2.5	2.8% 4.2%	
Unexplned variance in 4th contrast	=	2.5	2.7% 4.2%	
Unexplned variance in 5th contrast	=	2.5	2.7% 4.1%	

The eigenvalue for the unexplained variance in the 1st contrast is 4.4 (4.8%). According to Linacre (2015b), anything more than 2 has a potential for a dimension. However, from Table 4.36, the loading of the 3 items seems heaviest, but they are less than 0.7. According to Linacre (2015b), if the loading is more than 0.7, there is a possibility that there might another dimension that exists.

Table 4.36
Standardized Residual Loadings for Items

CON-TRAST	LOADING	INFIT MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	Item	LOADING	INFIT MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	Item
1	.67	.99	1.12	1.07	A 39	G39	-.54	.47	1.05	1.02	a 22	C22
1	.62	1.04	1.07	1.01	B 52	W52	-.51	.66	.98	.97	b 8	Ra8
1	.61	.74	1.00	.95	C 58	Wp58	-.40	1.04	.90	.87	c 17	Rb17
1	.51	.33	1.08	1.05	D 4	Ra4	-.33	.89	.97	.93	d 7	Ra7
1	.40	-.49	.96	.94	E 12	Rb12	-.28	2.10	.93	.85	e 28	C28
1	.37	-.67	.94	.86	F 49	E49	-.27	.41	.91	.92	f 5	Ra5
1	.36	-1.26	.89	.76	G 25	C25	-.27	.23	1.07	1.03	g 42	E42
1	.35	-1.31	.95	.85	H 51	W51	-.27	-1.76	1.00	.91	h 13	Rb13
1	.34	.02	.98	.93	I 55	W55	-.26	-.67	1.10	1.13	i 56	Wp56
1	.31	-.42	1.05	.99	J 45	E45	-.26	-.77	1.08	1.03	j 47	E47
1	.29	-.87	1.07	1.06	K 40	G40	-.25	-.18	1.03	.98	k 26	C26
1	.24	-1.84	1.01	1.05	L 9	Ra9	-.24	-.49	.97	.88	l 41	E41
1	.23	-1.11	1.04	1.02	M 20	Rb20	-.24	3.46	1.03	1.33	m 18	Rb18
1	.21	-3.24	.99	.69	N 16	Rb16	-.23	-1.88	.97	.96	n 2	Ra2
1	.21	-1.80	.95	1.01	O 15	Rb15	-.21	.28	1.03	1.07	o 54	W54
1	.17	4.38	1.07	1.21	P 24	C24	-.19	.94	1.02	1.03	p 38	G38
1	.14	-1.92	.98	.89	Q 1	Ra1	-.18	-.21	1.07	1.08	q 34	G34
1	.08	-.24	.85	.81	R 21	C21	-.17	1.81	.97	1.01	r 53	W53
1	.08	-1.13	1.04	1.08	S 11	Rb11	-.16	.46	.94	.88	s 60	Wp60
1	.07	-.99	1.00	1.07	T 27	C27	-.15	4.06	1.21	1.75	t 48	E48
1	.06	-.09	1.04	.99	U 33	G33	-.14	-2.73	.99	.79	u 10	Ra10
1	.06	-1.29	1.01	.94	V 23	C23	-.14	1.75	.97	.91	v 29	C29
1	.05	-1.76	.99	1.02	W 37	G37	-.14	-.01	.92	.84	w 57	Wp57
1	.04	-1.13	1.01	.95	X 35	G35	-.13	2.79	1.23	1.16	x 46	E46
1	.03	.36	.92	.92	Y 30	C30	-.13	-.31	.90	.82	y 36	G36
1	.01	-1.73	.90	.63	Z 32	G32	-.12	-2.32	1.01	.87	z 31	G31
1	.00	-2.15	1.03	.98	19	Rb19	-.10	-.79	1.00	.92	43	E43
1	.00	.19	1.00	.96	6	Ra6	-.08	-.40	1.06	1.02	59	Wp59
							-.07	-.73	.98	.87	44	E44
							-.06	1.83	1.19	1.22	14	Rb14
							-.02	-.14	.95	.88	3	Ra3
							-.02	7.58	1.39	1.66	50	E50

The dimensionality analysis, i.e. ATETv3 is unidimensional substantiates the construct validity.

4.8 Summary

This chapter has presented the results according to the three different versions of ATET and the essay component. The relevant findings are elicited to relate to the corresponding Research Questions. The summary of the findings, discussion and conclusions will be discussed in Chapter 5.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Introduction

This chapter narrates the discussion by going through the entire process from the preliminary parts of the study to the findings from Chapter 4, relating to the issues observed during the study as well as the Literature in Chapter 2. Following this, conclusions will be drawn from the discussion.

5.2 Summary of Findings

RQ1: How to construct a valid test that measures the proficiency in English based on the Rasch Model?

To answer this question, there were several evidence that were presented. The Item Fit, Person Fit and the Principal Component Analysis (PCA) tables for ATETv3 are the relevant findings. Table 5.1 summarises these findings.

Table 5.1
Summary of Fit and Order Validity

	ATETv3
Item Fit	0.8 < Infit MNSQ < 1.2 -2 < t < 2
Person Fit	0.79 < Total Infit MNSQ ± S.D. < 1.17 -2 < t < 2
PCA	Variance explained by measures, Raw value=modelled value, 34.6% Unexplained variance in the 1st contrast= 4.8% (<15%)

Besides the acceptable psychometric properties, the Variable Map (Figure 4.15) shows a normal distribution of items with a good spread of difficulty levels. The ICCs in Figure 4.18 also concurs to the range of

distribution. Finally the Bubble Chart in Figure 4.23 confirms ATETv3 fits the Rasch model well especially having all items in the range of $-2 < t < 2$. The bubbles in the chart are smaller in size, indicating lesser errors, compared to the Bubble Chart of ATETv2 and visually displays the vertical and horizontal spread of items in ATETv3.

Table 5.2
Psychometric Properties of ATETv3

	ATETv3
Person Reliability	0.86
Item Reliability	0.98
Person Separation	2.46
Item Separation	7.51
Cronbach Alpha	0.80

Then, there was an investigation if there is a correlation between the test MUET and GPA? According to Table 4.31, there is a very strong and significant relationship between ATETv3 and MUET, with a 0.826 coefficient as well as between ATETv3 and GPA, with a 0.779 coefficient. In fact all the corresponding section also showed strong relationship between the test, MUET scores and the GPA.

With all of these analyses, the test was found to have interpretive and validity argument. There is no gender biasness in this test as pointed out by the GDIF analysis (Figure 4.16). ATETv3 seems to have all the necessary psychometric properties, thus making it have interpretive and validity argument.

So, on the whole, to what extent can the test be constructed using IRT? Judging from all of the above findings, this ATETv3 can definitely be constructed using IRT, The Rasch Model in particular as it has all the necessary psychometric properties for the data to fit the Rasch Model.

RQ2. What is the suggested number of raters for the essay component?

This analysis was done with SPSS, FACETS and EduG 6.1. Tables 4.24 and 4.25 were referred to for the suggested number of raters which was 3-4 raters. However, after reanalysing using only the two experienced raters and correlating them, Table 4.26 suggests even with one rater, the coefficient is very high, about 0.9. This is also confirmed with FACETS where the inter-rater reliability was high despite having a less experienced rater. As such, there can be just one rater, but he/she must be an experienced person. If there is a new rater, there must be at least 2-3 other experienced raters to guide.

RQ3. What is the cut-off point to be accepted into the programme?

At the end of the study, the cut-off point is seen in the Variable Map in Figure 4.15, which is logit -1. This is confirmed by the Bubble map where the smaller the bubble, the lesser the error, seen at logit -1.

5.3 Discussion

This study was first planned to examine the blueprint (table of specification) of the existing selection test used at the TESL Foundation programme and to validate it. Unfortunately, there was no blueprint. The Rasch analysis was run with a sample size of 134 and the Summary Statistics (Table 3.1) shows low psychometric properties on the test, in which candidates cannot be discriminated well and will not assist in selecting the more competent candidates. These findings however could

not be used as a guideline as there were a few demands by the Head of TESL Foundation Programme (HOP) at that time that had to be met: the number of items had to be retained, format of items had to be parallel to SPM and MUET and the time limit was to be maintained at 1 hour.

5.3.1 ATETv1

With all these ideas in mind, a new Table of Specifications (Appendix B) and the ATETv1 was constructed. Despite the construction of ATETv1 was prepared on time to test the candidates, the test was not used for the real selection as there were four teams of lecturers in the Faculty of Education who were assigned to prepare one set of the test.

Three of the tests were taken wholesale from various MUET revision books, while the fourth set was adapted from some websites, although the sources were not mentioned (see Appendix K). The reading comprehension passages of Set 1, Set 2 and Set 3 were informative type while Set 4 dealt with a short story. The items in Set 4 were different because there were items related to literary elements. Thus in terms of parallelism, Set 4 was totally different. The same happened for the Cloze Test passage. Set 4 was different from Sets 1, 2 and 3. The issue here is that all these sets were used for the four interview sessions and Set 4 was definitely unfair. This goes back to Kunnan (2000) who talks about test fairness. At face validity, the items of all the sets have to be similar in nature. As these results were not available for analysis, the difficulty level of all

the sets could not be determined to see if they were parallel tests.

With all these obstacles, the ATETv1, which was set earlier, was then allowed to be administered at the beginning of the semester. The HOP used the test results, the total number of right answers, to place the students into various classes. The positive remark that came out of this placement is that the ATETv1 was able to put students of similar proficiency level in the same class. This was found out after chatting with a few of the lecturers teaching the TESL Foundation programme, who were pleased with the homogenous groups. The other plus point was that lecturers knew exactly at which level to pitch a particular lesson as the students had similar ability. So teaching became less tedious.

Upon looking at the analysis, it was obvious that the number of items were lacking as reported by the Person Reliability (Linacre, 2015). Thus ATETv2 was constructed with 50 items.

Besides the number of items, the findings in Table 4.02 shows there are some faulty items. One of the reasons being they are not within the acceptable Infit MNSQ range (0.8 – 1.16), the ZSTD is not between -2 and 2 and the PTMeasure Corr is negative. Tables 4.2 to 4.6 refer to the items that were faulty. As such, items 2 and 30 were dropped. Items 4, 16, 17, 18, 19, 22, 27 and 28 were checked and improved.

In terms of the Misfit Person, only one person had a negative PTMeasure Corr., which shows that he/she was not

behaving in the manner that was expected. On the whole, the choice of sample had no problem because they had been selected for the course already.

The ICC curves were used to determine the probability of giving the correct answer across different levels of proficiency. The ATETv1 showed the ICCs spread in their proficiency levels. Some items that were overlapped were assumed to share the same characteristics. All these characteristics added to the statistical analysis, were used to make good judgement about the items for the improvised version.

5.3.2 ATETv2

Items were generated carefully, taking into consideration the factors from ATETv1. As the TESL Foundation programme does demand a certain level of proficiency as entry behaviour, a decision was made to make changes to the reading passages and to try out multiple choice questions to test writing skills. This idea came about after reading Stiggins (1982) that

there are direct and indirect methods of testing writing skills. The direct method is the rating of essays while indirect method is testing through multiple choice items. His findings show a consistent and solid relationship between the two approaches at various levels of education. Nevertheless, Hamp-Lyons (2007) cautioned that the inter-rater reliability was a main issue for entrance and placement tests. According to Dewar (2008), writing skills is a stable aptitude within the writer.

Thus the source of disparity points towards the rater.

The findings for this study concurs with these literature. The analysis done on inter-raters found that if the raters are experienced (taught Writing skills for more than 15 years), then the Pearson coefficient is higher. So this shows that inter-rater reliability is high.

With this notion, the items that were included in ATETv2 were 5 items based on a passage (adapted from student's essay) to improve the coherence and cohesion of the passage. This was a challenging feat as this is not a familiar format to the students, who are SPM leavers.

The reading section was also improvised. A new passage (informative) was added and items from the two passages in ATETv1 were revised. This made up for the extra questions on ATETv2

The Person Separation for ATETv2 was more than 2, which shows that the test is able to discriminate lower proficiency students from the higher proficiency ones. Wright and Stone (1988) said that the separation and strata index must be more than 2.

On the whole Cronbach Alpha was 0.68 for ATETv2, slightly better than ATETv1, 0.62. But this is not a strong coefficient for a high stake test like the ATET.

Further scrutiny of the data analysis, person reliability, 0.72 seems to point that there should be more items to be included in the test. Again, looking at the Infit MNSQ, the Outfit

ZSTD and the PTMeasure Corr in Table 4.10, there are a few items which are faulty and in Table 4.15, there are a few persons who are out of the acceptable Infit and Outfit range. However, the strange thing about deleting the persons who were not within acceptable range of Infit MNSQ and Outfit ZSTD, the measures became lower for person reliability. Thus the number of persons were retained. As such the items had to be revised.

The ICC curves for ATETv2 were examined. The items were spread well. Despite this, when each section was probed, there were a number of items that overlapped, i.e. they share the same measure. However, with further probing, the Reading section was found to be testing different constructs. According to the findings there were 4 groups of overlapping curves. Ra1 was testing inference, while Ra2 was drawing conclusion. Rb9 was testing main ideas while Rb10 was testing inference/literary element. Rb6 was drawing conclusion, Rc11 was paraphrasing an idea while Rc17 was a higher order thinking skill (HOTS)/drawing conclusion. Rc13 tests on main idea, Rc18 is drawing conclusion/vocabulary and Rc20 is drawing conclusion. It can be seen that although the graphs of these items overlap, the constructs arre fairly different from each other within the same groups. The Subject Matter Expert advice points out that the passages need not have a range of proficiency level (ATETv2 has 3 passages, from very easy to advanced level). Both the experts suggested to focus on the more challenging passages as the filteration of candidates would be ensure better quality ones

gaining admission to the TESL Foundation programme.

Next is the Cloze Section. After items were modified from ATETv1, the ATETv2 shows no overlapping of items. However, when looking on the steepness of the slopes (mostly gentle slopes), it is realised that the items were fairly easy (Bond & Fox, 2007) as the 50% chances of getting the Cloze items correct seems to have measures between -1.56 to 2.06.

The Grammar section sees a steeper set of ICC slopes compared to the Cloze ICCs. This means they can discrimate candidates better. Items G34 and G40 are definitely confusing items. G34 has an issue with subject and object placement, while G40 is about the use of pronoun when more than 1 person of the same gender is mentioned in the same sentence. The constructs are not the same, but share the same measures. Similarly for G35 and G42. G35 tests on conditionals while G42 tests on nouns/vocabulary. These items would have caused the candidates to guess as they are not tested at the SPM level.

The Cloze and Grammar sections were reported separately although most of the Cloze blanks require grammatical knowledge, but on the whole, it draws the students' reading and grammar ability. The contextual clues (from reading skills) will definitely help in determining the appropriate grammar option (Weir, 2005).

The multiple choice questions (MCQ) for the Writing section had spread well according to the ICCs. It is interesting to note that this was the first attempt of testing writing

skills through MCQs. The total score for this section (5 items) were correlated with the raters' average score for the essays using SPSS.

The average essay score when correlated with the MCQs of ATETv2 for the writing section is moderate (Field, 2013). This shows a moderate but significant relationship between the Direct and Indirect method of testing writing skills (Stiggins, 1982).

The Bubble Chart helped to looked into the items in a more careful manner. All Grammar and Writing items were within the boundaries beteen -2 and 2. The problem were with the reading and cloze items that were not within the acceptable ZSTD boundaries. This confirms the discussion above about the improvement for the reading and cloze sections, both by the empirical data as well as Subject Matter Experts' advice.

5.3.3 ATETv3

Taking the discussion of 5.3.2, ATETv3 was constructed. This test version saw an addition of 10 items, making it a total of 60 items on the ATETv3. Table 4.17 shows a tremendous improvement in the Person Reliability, 0.86. Table 4.18 shows that only items 48 and 50 are out of the Infit MNSQ range as suggested by Linacre (2015). The Outfit ZSTD has no items outside the -2 to 2 boundary. Thus there no outliers. All PTMeasure Corr have positive values. Thus, all these have proven the number of items are optimum to discriminate the

ability of the candidates.

In terms of Persons Fit Order, there were a few misfits according to Table 4.23. However after deleting these persons, Table 4.24 shows similar results as prior to the person deletion. Thus this test proves that it is sample independent (Andrich, 2004).

Looking into the details of the ATETv3, the ICCs are referred. There is a good spread of items. As this is a selection test, there is not much concern about the discrimination of advanced students from intermendiate or high intermediate students.

The ICCs for the Reading section seems much steeper than of ATETv2. The reading passages for ATETv3 were changed. Two fairly high-intermediate passages were used. It was taken froma website with permission (Appendix I). Items were added from the ones provided in the website to suit the nature of this proficiency test.

The Cloze section was also changed with a more current and interesting issue. Thus the focus of grammar would be camouflaged by the content. This was good as it is less intimidating. It can be seen in the ICCs that cross from easy to difficult items.

The Grammar section had 5 extra items to test Sentence Completion, making it 20 items all together, 10 for Sentence Completion and 10 for Error Identification. Just like the Cloze section, this Grammar section had a variety of difficulty level.

However as said earlier, the most difficult items were merely added to challege the candidates but not to be of any criteria for selection.

The Writing Section as seen as moderately correlated with the Essay Section of ATETv2, was enhanced with 5 items to test paraphrasing, combining sentences as well as direct/indirect speech. This was done parallel to the SAT Sentence Improvement section. Thus the MCQ Writing section had all together 10 items, 5 from Paragraph Improvement and 5 from Sentence Improvement. They were fairly difficult judging from the steepness of the slope.

Besides this, the correlation between items in the multiple choice section for writing skills and the essay scores for ATETv3 (Table 4.30) are significantly high between ATETv3 and Raters 2 (r=0.724) and 3 (r=0.656). Stiggins (1982) claims if the Pearson r is more that 0.6 for writing ability, it is considered strong relationship. This is evident that writing skills can be tested in the Indirect way, which is more objective, which is reliable.

5.3.4 Construction of a Valid Test

On the whole, it is seen as the data fits the model and if the acceptable range of the suggested literature review for Infit MNSQ is between 0.8 and 1.2 (Linacre, 2015), thus only items 48 and 50 seem problematic. However, after looking at these two items' stems, they have confused the students as these items

are not taught within the Malaysian school syllabus for English, namely the Grammar component. Thus the guessing factor may have disfigured the data, appearing to be problematic.

Item 48 relates to problems with subject-verb agreement. The word "place" should be "places". As it appears in option C as part of the answer, this is the faulty part. The confusion is when the word "students" is focused, it is assumed as the subject. But here, the subject is Delima College, which is singular..

Meanwhile Item 50 relates to issues with comparative. The word "little" should be "less" As such option A is where the error is. Nevertheless these items (48 and 50) were retained as this would fall into the higher order thinking skill, where students should be able to apply what they have learnt (subject verb agreement and comparative) in a new context.

The Principal Component Analysis also confirmed that the test had order validity as the raw variance explained by measures is almost the same as the modelled value (34.6%). With the acceptable noise level of 4.8% in the unexplained variance in the 1st contrast, the test has definitely unidimensionality.

There is a certain amount of fairness in this test as the items although tested at a higher level, but it involves skills that have been taught at the secondary school level. As long as a candidate has gone through the school system, he/she will not find the questions awkward. Even the items in the Reading Comprehension passage that deals with literary elements have

been exposed in the secondary school. The only difference is such items are testing higher order thinking skills.

The writing component on the other hand tests writing skills using multiple choice answers. This is to demonstrate if the student is able to rephrase or paraphrase using different sentence stuctures and vocabulary. It is necessary as students would be expected to do self and peer review of essays in the TESL Foundation programme. This would not be claimed as unfairness, although the format may be something unfamiliar, as these writing skills (paraphrasing, ellipsis etc) are part and parcel of the writing component in the secondary school. Again these application type of questions are relevant. Therefore, there is definitely fit and order validity for ATETv3.

According to the Rasch Analysis (Linacre, 2015), the Person and Item Reliability should be more than 0.8. ATETv3 has Person Reliability of 0.86 and Item Reliability of 0.98. This means the number of items and persons is sufficient to yield the results that are needed for this analysis. The person separation indicates how many groups of test takers can this test discriminate. ATETv3 shows 2.46, which means that we have about 2-3 groups of test takers. This is confirmed by the Variable Map (Figure 4.15). The lower line illustrates that there are no test takers below Logit -1. This is explained by the fact that all the students who have been accepted into the program had gone through an interview. Thus stringent scoring in the test and strict gauging of personality during interviews has seen

better students in the program. However, there is a clear group in the upper tier of the variable map. This indicates that there is a handful of advanced students within the group. Although this study embarked on a journey to only find the cut-off point to the program, this test utility (Kane, 1992) was added-on. Instructors can actually use the test scores for placement of classes as the results are able to discriminate weak from better students.

Besides the Person and Item Reliability, the Variable map and Person Separation, the ICCs as well as the Bubble Chart prove the acceptable psychometric properties of the ATETv3. The spread of the items, besides the fact that there are more items than ATETv2, is greater if the ICCs are compared between ATETv2 and ATETv3. Thus the test is able to discriminate good and weak candidates. Apart from this, the Bubble Chart displays measures and fit values graphically. The smaller bubbles in the Bubble chart (Figure 4.32) show there are lesser standard error for those items. However, the bigger bubbles are the ones which are more difficult items and the error could be caused by the guessing factor. Thus there are some candidates who do not 'behave' or 'perform' according to the expectation, thus adding on to the error in the measures.

One of the reasons for this error is the unfamiliarity of the format of the items. Although the reading passages are fairly simple, the questions may not have answers that can be directly lifted from the passages. They require a bit of analysis and

evaluation before deriving the answer which is also a higher order thinking skill.

Item 18 for example, tests on the application of literary elements. Students have learnt about literary elements in school, but were not assessed at SPM level. So, this appeared tricky to the students as the options given were technical jargons. Students might have got the elements mixed up.

The same justification applies to the second most difficult item for the reading section, which is item 14. This also tests on literary elements. As items 18 and 14 ranked the most difficult, there is no doubt students who were unsure of the answers would have guessed. With all these analysis, ATETv3 obviously fits the Rasch Model.

Originally the test was supposed to be correlated with the SPM results. However, due to the weak, but significant correlations, another test was correlated to the ATETv3 scores. This was the Malaysian University English Test (MUET). MUET is an English Language test which tests on the four skills, reading, writing, listening and speaking. Most students take this test towards the end of Semester 1. In other words, they have gone through the first part of the TESL Foundation programme. As such, there is a lot of learning and exposure to language that has taken place in that semester. These skills are also tested in the ATET. However, for ATET, the only exposure they had is the secondary school English. But in terms of the ranking and ability, there is resemblance. There is

statistically significant correlation between ATETv3 and MUET, proving that ATETv3 has predictive validity. This is one of the psychometric properties of a good test.

The ATETv3 was correlated with a test that has similar content. This was done with the Grade Point Average (GPA) for the first semester's final examinations. The subjects for the GPA include Reading, Writing, and Grammar. These are also the three components of the ATETv3. These two scores of ATETv3 and the GPA seem to correlate significantly high, again proving the ATETv3 has construct validity.

According to Kane (1992), "the interpretive arguments associated with most test-score interpretations involve multiple inferences and assumptions." ATETv3 has established the psychometric properties of. It also has proven the predictive and construct validity. In addition, this test is not gender biased as proven statistically by the GDIF analysis. Therefore the ATETv3 has multiple inferences and assumptions, leading to having interpretive and validity argument.

5.3.5 Raters of the Essay Component

The essay component was treated differently. The scores were analysed separately and not together with the MCQ scores. There were three different raters for each set of essays. There were two experienced raters and one newbie. Correlations show the two experienced raters correlated significantly high, while with the inexperienced rater, the coefficients were low. Thus a

G-study and D-study were done in two parts, one with all the 3 raters and the other with only the 2 experienced raters. The outcome showed that as long as the rater is experienced (or trained), even one rater is sufficient. But with inexperienced and experienced raters together, at least 3 to 4 raters are required. As this test is conducted during the interview and due to cost constraints, there are only 2 interviewers in each panel. Thus the suggestion is to ensure in every panel, there must be at least one experienced TESL lecturer. This lecturer in particular has to grade the essays for the respective panel. This way, the reliability of the essay scoring can be guaranteed and fairness of the test can be safeguarded.

5.3.6 Cut-off Point for Admissions into the TESL Foundation Programme

After establishing the psychometric properties of the test, the Variable Map for ATETv3 was scrutinized again to determine the cut-off point. It was decided that it should be at logit -1. So, all candidates who score logit -1 and above will have a chance in getting into the programme. The entrance test is not the only factor that determines the offer into the TESL Foundation programme. There is also the interview and the SPM results that are added on to the ranking of candidates to be considered for the programme.

5.4 Implications of the Study

5.4.1 Theoretical Implications

Judging from the findings and the discussion prior to this research question, it can be concluded for this overarching question that Item Response Theory, particularly the Rasch Analysis can be used not only to construct, but also to validate a test. However, it takes time to construct a test from scratch as it involves several levels of check-and-balance and fine-tuning before it is ready to go. The two-in-one point of using the Rasch Model is that as it is being constructed, the validation process happens simultaneously. Thus by the time the items are consolidated, it is rest assured that the items are all validated as well.

The other plus point of using the Rasch Model or even the general IRT is that it is sample independent and validated at the item level and not as a whole test. Thus these items can be put into an item bank and several parallel items can be written with the same difficulty level. Thus the items in the item bank will increase in number as well as variations of the same difficulty level will be there (with parallel items).

5.4.2 Practical Implications

This study has given way to some practical implications. Among them are the fact the test can be developed and validated simultaneosly using the Conceptual Framework can definitely reduce the time needed to ensure validity. Once the test is

validated, it is automatically reliable.

Another implication is that as the ATETv3 includes the testing of writing skills in MCQ format and the items have been proven valid, this test can be administered without the essay component. This would speed up the selection process and less time consuming. Thus, during the interviews, the interviewers can fully concentrate on the interview and not to worry about the scoring of the essays.

5.5 Recommendations from the Study

This study has proven that Writing skills can be tested using the MCQ format. So, the essay component can be omitted. However, if there is persistence to include the essay section, then the raters must be trained for calibration. This can be done by training the raters prior to the selection date. As the results have shown that experienced can score the scripts well and do not need a second marker to verify, such a calibration exercise would benefit, particularly the raters with less teaching experience.

Besides, the MCQ items, have to be tested at least with 60 items all together. The more the items, the higher the reliability. Due to the time constraint, 60 items would suffice with three sections, Reading, Grammar and Writing. If there is no essay component, then the time is recommended to be reduced to one hour.

5.6 Recommendations for Future Research

This test which set off to merely select candidates into the programme is now used by the programme as a Diagnostic Test, administered on the first day of class before teaching begins to see the entry behaviour of the

students. This has worked well the since 2013 as the test has different sections, focussing on different skills, to point out the strengths and weaknesses of the student . Initially batches from 2013 to 2015, this test was also used to place students into groups according to the total scores on the test. However, feedback from lecturers teaching this programme showed that some students, particularly the good ones, were arrogant and conceited. This made it difficult for lecturers teaching such groups as there was not much learning taking place.

The only reason why this test is not considered for the actual entrance test is the length of the test. Currently the entrance test has only 30 items, with 4 variations, which are adapted from the IELTS practice questions. Also seemingly challenging, the fact is these tests versions are not standardised, thus difficulty level of the test has not been established. This has brought in another element, fairness of the test. As such it is highly recommended that the ATETv3 has to have at least 4 batteries. Only then the ATET can be considered for actual entrance test. The reason behind the 4 versions is to reduce copying and leaking of questions, particularly the essay component.

Another recommendation for the ATETv3 is to include the listening component into the test. It is not necessary to include the speaking component as this will be judged at the interview. Besides, the items in the test must culminate the foundation to the skills that are taught in the first semester at the TESL Foundation programme. This includes all the reading, writing, listening and speaking skills and not forgetting the grammar items. Then the correlations done with the GPA will have exact components to be compared with for the construct validity. This would

also ensure readiness as the *TESL Foundation* programme prepares students to move from intermediate to advanced level in the proficiency of the language.

In terms of administration, this test for this study was done at different campuses, Melaka, Kuantan and Shah Alam. However, in December 2015, the entire Foundation programmes at various faculties have been centralised in the new campus in Dengkil. This campus is called the Centre of Foundation studies, with 4 programmes namely Asasi Science, Asasi Engineering, Asasi Law and TESL Foundation. As the ATETv3 has been used as a Diagnostic Test in all the 3 campuses, it was also recommended in Dengkil. Unfortunately in Dengkil, there is no one hall that can accommodate all the 536 students to give a standardised test. Thus the test was digitized and administered as an online test using the i-learn platform, Public university's online learning platform for the May 2016 intake. This has actually taken ATETv3 to greater heights. Thus another recommendation would be to come up with an item bank and have several batteries of the test generated by the system. This would also avoid the leaking of questions by candidates.

With these recommendations, it is hoped that the ATET would be enhanced and serve appropriate purpose to the stakeholders.

5.7 Conclusion

There are some conclusions that can be drawn from the discussion in 5.3. The final version of the entrance test, ATETv3 definitely has content validity as Froelich (2009) said there is a match between the subject

specifications and the goals of the educational programme, which is done by Subject Matter Experts. The external standardised tests that have similar construct (GPA and MUET) were correlated and they correlate well with ATETv3 particularly MUET. This supports criterion validity. The construct validity was established through the thorough analysis using Winsteps, FACETSm EduG and SPSS.

The most important utility of this test's scores' is to select students into the TESL Foundation programme, which is determined by the cut-off point, at -1 logit. The test scores is interpreted as the proficiency in the English Language and the readiness of the candidate to be able to go through the TESL Foundation course, which will eventually lead to a TESL degree level course or any language related degree level course. This is in accordance to Kane's interpretive argument (1992). The conclusion is candidates who score more than the cut-off score are prepared for the programme and should be given admission into the programme. The candidates who score lower than the cut-off point should be rejected. The basis for accepting this interpretation is the validity of this test. The test is proven to have gone through fine-tuning in a longitudinal way. Over the years between 2009 and 2012, the items were checked according to the different sections, i.e. reading, grammar and writing skills were looked at to culminate the proficiency in the language. As such ATETv3's validity argument is sound.

With the construction and validation of items for an entrance test using the Rasch Model, it is confirmed that the items in the test have fit and order validity, apart from the data fitting the model. In this study, the ATETv3 seems to fit the model and has the interpretive and validity

argument as it is able to predict future test results, making it more robust. An additional advantage of this ATETv3 is that it can discriminate weak students and advanced students from the intermediate. So, ATETv3 has another utility, i.e. for placement or streaming. Although this study did not delve into this area, it became part of the outcome of the utility of the test. This would make it easier for the instructors to teach according to the students' abilities. The students will not feel intimidated by the more advanced students if placed randomly and put in the same class. This would definitely help the teaching and learning process.

In short, the test going through a tedious and long check-and-balance process of IRT, employing the Rasch Analysis with Kane's Argument-based Approach has confirmed that ATETv3 is a useful tool to select candidates into the *TESL Foundation* programme, particularly at Universiti Teknologi MARA.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Airasian, P.W., & Terrasi, S. (1994). Test administration. *International Encyclopedia of Education* , 11, 6311-6315.

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice*, 12, 28-33.

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

Andrich, D. (2004). Controversy and the Rasch model: A Characteristic of Incompatible Paradigms? *Medical Care*, 42, 1-16.

Aryadoust S.V. (2009) *Mapping Rasch-Based Measurement onto the Argument-Based Validity Framework.* Rasch Measurement Transactions, 23(1), 1192-3.

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.

Bachman, L.F. (2004) *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bachman, L .F. and Palmer, A. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language Assessment in Practice.* Oxford: Oxford University Press.

Baghaei, P. (2008). Rasch Model as a Construct Validation Tool. Retrieved November 22, 2010, from *Rasch Measurement Transactions*. http://www.rasch.org/rmt/rmt221a .html

Baghaei, P. (2008, Summer). The Rasch Model as a Construct Validation Tool. *Transactions of the Rasch Measurement SIG, 22*(1)

.
Baker, F. (1985). *The Basics of Item Response Theory*. http://echo.edres.org:8080/irt/ Retrieved on 5 February 2009.

Bennett, D. T., Wesley, H., & Dana-Wesley, M. (1999). Planning for imminent change in college admissions: Research on alternative admissions criteria. *Journal of College Student Retention*, 1(1), 83-92.

Bloom, B.S. (ed.), Engelhart, M.D., Furst, E.J. Hill, W.H. and Krathwohl, D.R. (1956). *Taxonomy of Educational Objectives I: Cognitive Domain*. New York: McKay

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill, Inc

Bond, T., & Fox, C. (2007). *Applying the Rasch Model: Fumdamental Measurement in the Human Sciences* (2nd ed.). New York: Routledge.

"Boost quality of teaching", The Star, 9 June 2009

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of Validity. *Psychological Review*, 111(4), 1061-1071.

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement: InterdisciplinaryResearch and Perspectives*, *2*(3), 171–174.

Brown, J. D. (2002). English language entrance examination: A progress report. *Curriculum Innovation, Testing and Evaluation: Proceedings of the 1st Annual JALT Pan-SIG Conference.* Kyoto: Kyoto Institute of Technology

Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*. New York: McGraw Hill.

Byrne, B.M. (2001). *Structural Equation Modeling with Amos*. Mahwah, New Jersey: Lawrence Erlbaum.

Carey, L. M. (1988). *Measuring and Evaluating School Learning*. Boston: Allyn, and Bacon, Inc.

Chapelle, C. A. (2008). The TOEFL validity argument. *Building a validity argument for the Test of English as a Foreign Language*.. C. A. Chapelle, M. K. Enright, & J. Jamieson (Eds.). London: Routledge.

Chapelle, C. A., Enright, M. K. and Jamieson, J. (2010). Does an Argument-Based Approach to Validity Make a Difference? *Educational Measurement: Issues and Practice*, 29, 3–13.

Chase, C. I. (1964). Relative length of option and response set in multiple choice items. *Educational and Psychological Measurement*, 24(4), 861-866.

Chase, C.I. (1999). *Contemporary assessment for educators*. New York: Longman.

Cheng, Y.C., Chow, K.W. & Tsui, K.T. (Eds.) (2001). *New teacher education for the future: International perspectives*. Hong Kong/The Netherlands: Hong Kong Institute of Education/Kluwer Academic.

Clapham, C. & Corson, D. (eds.) (1997). *Encyclopedia of Language and Education, 7, Language Testing and Assessment*, ix-xi. Netherlands: Kluwer Academic Publishers.

Clegg, V. & Cashin, W. (1986). Improving Multiple Choice Tests. *Idea Paper , 16.*

Cook, T. & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues in field settings.* Boston: Houghton Mifflin.

Crocker, L. &Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Harcourt Brace Jovanovich.

Cronbach, L. J. (1971). Test validity. *Educational Measurement*, 2nd ed. Thorndike, R. L. (ed.), American Council on Education, Washington, DC,

Cronbach, L.J., Glesser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behaviourial measurements: Theory of generalzsability for scores and profiles*. New York: John Wiley.

Dahan, H. (2002). *Language Testing: The Construction and Vali*dation. Kuala Lumpur: University of Malaya Press.

Davies, A. (1990). *Principles of Language Testing.* Oxford: Blackwell.

Davies et al (1999). *Dictionary of Language Testing: Studies in Language Testing 7*. UK: Cambridge University Press.

Davies, A., & Elder, C. (2005). Validity and validation in Language Testing. *Handbook of Research in Second Language Teaching and Learning.* E. Hinkel (Ed.). Mahwah, New Jersey: Lawrence Erlbaum.

Delors, J. (1996). *Learning: The Treasure Within.* Paris: UNESCO.

Devine, M. & Yaghlian, N. *"Construction of Objective Tests,"*

Dual language programme to continue in national schools, says Education Ministry. (2016, March 26). *The Malay Mail*.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Encarta Dictionary. Microsoft Encarta 2006.

http://uk.encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx. Retrieved 16 Nov 2009.

Farhady, H. (1983). New Directions for ESL proficiency testing. *Issues in language testing research* . J. J. W. Oller (Ed.).. Rowley: Newbury House.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics, 4th Edition.* UK: Sage Publications.

Filmer, A. (2012, September 30). *Using the Universal Language.* Retrieved from The Star: http://www.thestar.com.my/news/education/2012/09/30/using-the -universal -language

Froelich, A. G. (2009). Methods from Item Response Theory: Going Beyond Traditional Validity and Reliability in Standardizing Assessments. *Quality Research in Literacy and Science Education*, IV, 287-301.

Gipps, C. (1994). *Beyong Testing: Towards a Theory of Educational Assessment.* London: RoutledgeFalmer.

Global Monitoring Report Team. (2015). *EFA Global Monitoring Report. Education For All 2000-2015: Achienvements and Challenges.* France:

UNESCO.

Gredler, G. R. (1999), Book reviews. *Psychology in the Schools*, 36, 171–173

Gronlund, N. E. (1982). *Constructing Achievement Tests.* New Jersey : Prentice-Hall

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.

Haladyna, T. M. & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, 2(1), 51-78.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Mahwah: Lawrence Erlbaum.

Haladyna, T.M. & Downing, S.M. & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. *Educational measurement* (3rd ed.). R.L. Linn (Ed.). New York: American Council on Education/Macmillan.

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 and Practice*, 12(3), 38-47.

Hambleton, R.K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38(9), 60-65.

Hambleton, R. K.., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Haron, I., Gapor, L., Masran, M. N., Ibrahim, A. H., & Nor, M. M. (n.d.). *Kesan dasar Pengajaran Matematik dan Sains dalam Bahasa Inggeris.* Retrieved from http://www.scribd.com/doc/11492280/PPSMI-Kajian-UPSI-April-2008-Penuh

Harris, D. P. (1969). *Testing English as a Second Language*. New York: McGraw-Hill.

Harris, D. P. (1970). *The Linguistics Of Language Testing*. Davies, A. (ed.), 36-45.

Harris, D. (1989). *Comparison of 1-, 2-, and 3-Parameter IRT Models*. ITEMS: Module 7, National Council on Measurement in Education.

Harrison, A. (1983). *A Language Testing Handbook.* London: Macmillan Press.

Hawkey, R. A. (2004). *The CELS: Developing a Modular Approach to Testing English Language Skills* (Vol. 16). Cambridge: Cambridge University Press & Cambridge English Language Assessment

Henning, G. (1984). [Review of Arthur Hughes & Don Porter (Eds.), Current Developments in Language Testing]. *Language Testing*, 1, 237-41.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, Mass.: Newberry House.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality. *Language Testing*, 2, 141-54.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation,*

Research. Cambridge, Massachusetts: Newberry House.

How large is the job market for English teachers abroad? (n.d.). Retrieved November 12, 2016, from International TEFL Academy: https://www.internationalteflacademy.com/faq/bid/102201/how-large-is-the-job-market-for-english-teachers-abroad

http://educationmalaysia.gov.my/education.php?article=studyop

https://asasi.uitm.edu.my/v4/index.php/programmes/asasi-tesl

Hughes, A. (2003). *Testing for Language Teachers* (2nd Ed. ed.). Cambridge: Cambridge University Press

Imber, M. (2002). The problem with grading. *American Schools Board Journal*, 189(6), 40-41,47.

Ingram, D. E. (1977). Basic concepts in testing. *Testing and experimental methods. The Edinburgh Course in Applied Linguistics*. Allen, J. P. B. & Davies, A. (Eds.). 4,11-37). London: Oxford University Press.

Inter-Agency Commission, W. (1990). *The Final Report of the World Conference of Education For All: Meeting Basic Learning Needs.* New York: UNICEF House.

Jabatan Pendidikan Tinggi, Kementerian Pendidikan Tinggi. (2015, December 7). Retrieved September 30, 2016, from Pekeliling Dasar& Prosedur Kemasukan MATRIK_ASASI 1617 07.12.2015.pdf: http://upu.mohe.gov.my/web/Pekeliling%20Dasar&Prosedur%20Kemasukan%20MATRIK_ASASI%201617%2007.12.2015.pdf

Jacobs, L. & Chase, C. (1992). *Developing and Using Tests Effectively*

Jamil, H. (2014). Teacher is Matter for Education Quality: A Transformation of Policy for Enhancing the Teaching Profession in Malaysia. *Journal of International Cooperation in Education, 16*(2), 181-196

Jie, L. & Fu, F. (2003). *Differential Performance by Gender in Foreign Language Testing*. Poster presentation 2003 Annual Meeting of NCME. Chicago.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535

Kane, M. (2004). Certification testing as an illustration of argument based validation.

Measurement, *2*(3), 135–170.

Kane, M. T. (2006). Validation. *Educational Measurement*(4 ed.). In R. Brennan (Ed.). Westport, CT: Greenwood Publish

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73

Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17.

Kaur, S. (2008, May 27). *More than 30,000 to enter varsity.* Retrieved October 3, 2012, from The Star: http://www.thestar.com.my/story/?file=/2008/5/27/nation/20080527192240&sec=nation

Kline, P. (2015). *Handbook of Test Construction(Psychology Revivals): Introduction to Psychometric Design.* London: Routledge.

Kulasagaran, P. (2013, March 22). *The Star Online.* Retrieved September 30, 2016, from Nation: http://www.thestar.com.my/news/nation/2013/03/22/more-spm-students-score-straight-as/

Kunnan, A. J. (2000) Fairness and Justice for all. *Studies in Language Testing 9: Fairness and Validation in Language Assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge: Cambridge University.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. New York: McGraw-Hill.

Language Testing and Evaluation, Vol. 10. Frankfurt: Peter Lang.

Lawrence, I. M., Curley, W.E. & McHale, F.J. (1988). Differential item functioning for males and females on SAT verbal reading subscore items. *Report No 88-4*. New York: College Entrance Examination Board.

Lei, P. W., Bassiri, D., & Schultz, E. M. (2001). *Alternatives to the grade point average as a measure of academic achievement in college*. (Report No. TM033669) American College Testing Program. (ERIC Reproduction Service No. ED462407)

Linacre, J.M. (1993). Rasch-based Generalisability Theory. *Rasch Measurement Transactions*, 7(1), 283-284.

Linacre, J. M. (2005). *Rasch Dichotomous Model vs One-parameter Logistic Model.* Retrieved January 20, 2012, from Rasch Measurement Transactions: http://www.rasch.org/rmt/rmt193h.htm

Linacre, J.M. (2015a). Winsteps .

Linacre, J. M. (2015b). *Winsteps Help for Rasch Analysis*. Retrieved from: https://www.winsteps.com/winman

Linacre, J. M. (2016). Winsteps@Rasch Measurement Computer Program User's Guide. Beaverton, Oregon

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching (8th ed.).* Upper Saddle River, NJ: Prentice-Hall.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company

Maclean, R. (2001) Educational Change in Asia: An Overview. *Journal of Educational Change*, 2(3), 189-192. Special issue on educational change in Asia.

Mandela, N. (n.d.). *BrainyQuote.com*. Retrieved September 30, 2016, from https://www.brainyquote.com/quotes/quotes/n/nelsonmand157855.html

McNamara, T. (2000). *Language Testing*. Widdowson (Ed.) Oxford Introduction to

Language Study. Oxford: Oxford University Press

McNamara, T. (2005). Language Testing. Davies, A.& Elder, C. (Eds.), *The Handbook of Applied Linguistics*,763-782. UK: Wiley-Blackwell.

Mehrens, W. A., & Lehman, I. J. (1973). *Measurement and Evaluation in Education and Psychology*. (4th ed.) Chicago: Holt, Rinehart and Winston, Inc.

Mehrens, W.A., and Lehman, I.J. (1991) *Measurement and Evaluation in Education and Psychology*, (4th Ed.). Orlando, Florida: Holt, Rinehart and Winston Inc.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. *Test Validity* . Wainer, H. & Braun, H.I. (Eds.). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. *Educational Measurement* (3rd edition). R.L. Linn (ed.). NY: American Council on Education and Macmillan Publishing Company.

Messick, S. (1996). Validity and washback in language testing. *Language Testing,* 13 (3), 241–256.

Micceri, T. (2001). *Facts and Fantasies regarding admission standards*. Report No. HE034083. Paper presented at the Annual Meeting of the Association for Institutional Research (Long Beach, CA, June 3-6, 2001. (ERIC Reproduction Service No. ED453757)

Ministry of Education Malaysia. (2006a). *Standard Guru Malaysia.*

Ministry of Education Malaysia. (2006b). *Education Development Master Plan 2006-2010 (PPIP 2006-2010).* Putrajaya.

Ministry of Education Malaysia. (2007). *Pelan Induk Pembangunan Pendidikan.*

Ministry of Education Malaysia. (2013). *Malaysia Education Blueprint 2013-2025.*

Putrajaya.

Ministry of Higher Education. (2015, December 7). Retrieved September 30, 2016, from Bahagian Pengurusan Kemasukan Pelajar: http://upu.mohe.gov.my/web/Pekeliling%20Dasar&Prosedur%20Kemasukan%20MATRIK_ASASI%201617%2007.12.2015.pdf

Mohd. Asraf, R. (1996). The English Language Syllabus for the Year 2000 and Beyond – Lessons from the Views of Teachers. *The English Teacher*, XXV, 1-5.

Moller, A. D. (1982). *A Study in the validation of Proficiency Tests of English as a Foreign Language*. Unpublished Ph.D thesis. University of Edinburgh. Department of Linguistics.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement*, *16*, 159-176.

Narayanan, G., & Jamaluddin, S. (2012). *Seminar MEDC 2012.* Retrieved January 10, 2013, from Malaysian Education Deans' Council: www.medc.com.my/medc/ seminar_medc/fromCD/pdf/23.pdf

Notar, C. E., Zuelke, D.C., Wilson, J.D. & Yunker, B.D. (2004). The Table of Specifications: Insuring Accountability in Teacher Made Tests. *Journal of Instructional Psychology.*

O'Neill, K.A., McPeek, W.M. & Wild, C.L. (1993). Differential Item Functioning on the Graduate Management Admission Tests. *ETS Research Report 93*. Princeton: ETS.

Oosterhof, A. (2002). *Classroom Applications of Educational Measurement (3rd ed)*. Columbus, OH: Merrill Prentice Hall.

Osterlind, S. J. (1998). *Constructing test items*. Boston: Klumer Academic.

Pengajaran dan Pembelajaran Sains dan Matematik dalam Bahasa Inggeris. (n.d.). Retrieved February 17, 2017, from Wikipedia: https://ms.wikipedia.org/wiki/Pengajaran_dan_Pembelajaran_Sains_dan_Matematik_dalam_Bahasa_Inggeris#Kajian_Prof_Dr_Nor_Hashimah_Jalaludin

Perfetto, G. (2002). Predicting academic success in the admissions process: Placing an empirical approach in a larger process. *College Board Review*, 196, 30-35.

Popham, W. J. (2000). *Modern Educational Measurement: Practical Guidelines for Educational Leaders* (3rd ed.). Boston: Allyn and Bacon.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Ratnawati, M. A. (1996). The English Language Syllabus for the Year 2000 and Beyond - Lessons from the Views of Teachers. *The English Teacher.*

Rodriguez, M. C. (1997). *The art & science of item writing: A meta-analysis of multiple-choice item format effects*. Paper presented at the annual meeting of the American Educational Research Association.

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language testing, 9*(1), 12-29.

Schumacker, R. E. (2005). *Item Response Theory*. http://www.appliedmeasurementassociates.com/White%20Papers/ITEM%20RESPONSE%20THEORY.pdf Retrieved on 5 August 2010.

Standard Guru Malaysia, Kementerian Pelajaran Malaysia, 2006

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*. Newbury: Sage.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues & Practice,* 16(2), 5-8.

"Should a pass in English be made compulsory to pass SPM" The Star, 8 June 2009.

Sick, J. (2008) Rasch Measurement in Language Education. http://jalt.org/test/sic_1.htm. Retrieved on 16 January 2011.

Spolsky, B. (1973). What does it mean to know a language; or how do you get someone to preform his competency? *Focus on the Learner:Pragmatic Perspective for the Language Teacher.* In J. W. Richards (Ed.). Massa: Newbury House Publishers Inc.

Spolsky, B. (1975). Language testing - the problem of validation. In L. Palmer & B. Spolsky (Eds.). *Papers on Language Testing 1967.*

Spolsky, B. (1978). *Advances in Language Testing: Series 2, Approaches to Language Testing.* Arlington: Centre for Applied Linguistics.

Stiggins, R. (1982). A Comparison of Direct and Indirect Writing Assessment Methods. *Research in the Teaching of English, 16*(2), 101-114

Szabo, G. (2008). *Applying Item Response Theory in Language Test Item Bank Building*.

test. (2010). In *Merriam-Webster Online Dictionary*. Retrieved March 23, 2010, from http://www.merriam-webster.com/dictionary/test

Thissen, D., Steinberg, L. and Wainer, H. (1993). Detection of Differential Item Functioning using the Parameters of Item Response Models. *Differential Item Functioning.* In Holland,P.W. & Wainer, H.(Eds.). NJ: Lawrence Erlbaum.

Thorndike, R.L. (1982) Reliability. *International Encyclopedia of Educational Evaluation*. In Walberg, H.J. & Haertel, G.D. (eds). New York: Pergamon.

UNESCO. (2000). *Final Report of World Education Forum 2000.* Paris: UNESCO.

UNESCO. (2015). *World Education Forum 2015 Final Report.* Paris: UNESCO.

UNESCO Institute for Statistics. (2015). Retrieved September 30, 2016, from UNESCO eAtlas of Teachers: http://www.uis.unesco.org/Education/Documents/fs33-2015-teachers.pdf

(1992). *United Nations Sustainable Development Agenda 21.* New York: United Nations.

(2005). *UN for Education for Sustainable Development 2005-2014.* Paris: UNESCO.

Valette, R. M. (1977) *Modern Language Testing*. USA: Harcourt Brace Jovanovich, Inc.

Van der Walt, J., & Steyn (Jr.), H. S. (2008). The Validation of language tests. *Stellenbosch Papers in Linguistics, 38*, 191-204.

Walt, J. L., & Steyn, H. S. (2008). The Validation of Language Tests. *Stellenbosch paper in Linguistics, 38*, 191-204.

Walvoord, B. & Anderson, V. (1998) *Effective Grading* Cornell University Office of Instructional Support, http://www.clt.cornell.edu/campus/teach /faculty/Materials /TestConstructionManual.pd

Weir, C. (1991). *Communicative Language Testing.* New York: Prentice Hall.

Weir, C. J. (2005). *Language Testing and Validation: An evidence-based approach*. Hampshire, UK: Palgrave-Macmillan

Widdowson, H.G. (2000). On the limitations of linguistics applied. *Applied Linguistics* 21(1), 3-5.

.

Wikipedia. (2015). Retrieved July 23, 2015, from List of languages by total number of speakers: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

World Education Forum – UNESCO Dakar Senegal, April 2000.

Wright, B. D., & Stone, M. H. (1988). *Validity in Rasch measurement*. University of Chicago: Research Memorandum No. 55.

Wright, B. D., & Stone, M. H. (1999). *Measurement Essentials 2nd Ed.* Wilmington, DE: Wide Range, Inc.

Yen, W. M. (1992). Item Response Theory. *Encyclopedia of Educational Research* (6th Ed.). In M. Alkin (ed.). NY: Macmillan.

Zeller (1990). Validity. *International Encyclopedia of Educational Evaluation*. Walberg, H.J. & Haertel, G.D. (eds). New York: Pergamon.