PSYCHOMETRIC PROPERTIES OF A YEAR-END FORM FOUR CHEMISTRY PAPER 1 IN SELECTED SCHOOLS IN PETALING UTAMA DISTRICT

ROSE ENNE EMELLIA BINTI MOHAMED RAZALI

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

PSYCHOMETRIC PROPERTIES OF A YEAR-END FORM FOUR CHEMISTRY PAPER 1 IN SELECTED SCHOOLS IN PETALING UTAMA DISTRICT

ROSE ENNE EMELLIA BINTI MOHAMED RAZALI

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF EDUCATION (MEASUREMENT AND EVALUATION)

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020

UNIVERSITY MALAYA
ORIGINAL WORK LITERARY WORK DECLARATION

Name of Candidate: **ROSE ENNE EMELLIA BINTI MOHAMED RAZALI**
Registration Matric No: **POM170001**
Name of Degree: **MASTER OF MEASUREMENT AND EVALUATION**
Title of Dissertation ("this work"):
PSYCHOMETRIC PROPERTIES OF A YEAR-END FORM FOUR CHEMISTRY PAPER 1 IN SELECTED SCHOOLS IN PETALING UTAMA DISTRICT

Field of Study:

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 of 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 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

The low-performance of science stream students in Chemistry especially in high-stakes testing, has alarmed the stakeholders. This problem has led to the need of analyzing the instruments used to measure the achievement of students. The present study examined the psychometric properties of multiple-choice questions in the Year-end Form Four Chemistry Paper 1 using the Rasch Model in the Klang Valley. A quantitative research design has been used for this study. The sample comprised of 435 Form Four Pure Science students from four randomly selected secondary schools in the Petaling Utama district. The Rasch analysis was conducted in two stages. The first stage was to prove that there was no violation of the unidimensionality assumption and the data fitness to the model. The second stage of the analysis focussed on the validity and reliability test paper. The Principal Component Analysis (PCA) result reveals that Chemistry Paper 1 was unidimensional. The raw variance measures observed by the data was almost equivalent to the raw variance of the model indicated the data set fitted the Rasch Model. Moreover, the eigenvalue of the first contrast showed a strength of 2 items which proved no secondary dimension existed in the Chemistry items. In terms of infit and outfit statistics, the results showed that all items are in the acceptable range. These comprehensive analyses demonstrated not only the fitness of the data set for the Rasch model but also supported the concept of unidimensionality. The dimension, however, was held only at a reasonably acceptable level. The unexplained variance in the first contrast of the data set signifies 3% of the noise level compared to the variance. This indicated that there were no residual factors in measuring the ability of the students. The local independence assumption was met

and held across the Chemistry Paper 1 as residual correlations for each pair of test items were less than 0.7. This indicated no redundancy of the test items. In the second stage, estimates of reliability were 0.99 for items and 0.87 for persons with the separation index is 8.95 and 2.54 respectively. Thus, means that the ability of the students were efficiently distinguished and the items separation is broad. The person-item map exhibited that most of the test items measured students' abilities only in a specific range of measurement continuums. The Rasch statistical analysis by point measure correlation (PTMEA Corr.) established that all test items were positively correlated with the measured constructs and moved in one direction. Thus, there was no misfitting item. The result of the distractor analysis showed that all distractors of the test items were efficient. The DIF analysis examined the students' answers based on the gender. It identified six items that were found to be difficult for the male students compared to the female students. In conclusion, the Year-end Form Four Examination of Chemistry Paper 1 instrument has good psychometric properties and is capable of producing valid and reliable scores to measure the achievement of the students. Besides that, the Year-end Form Four Chemistry Paper 1 instrument has the same standard as the actual Chemistry Paper 1 of Sijil Pelajaran Malaysia (SPM).

CIRI-CIRI PSIKOMETRIK PEPERIKSAAN AKHIR TAHUN TINGKATAN EMPAT KIMIA KERTAS 1 DI BEBERAPA BUAH SEKOLAH DALAM DAERAH PETALING UTAMA

ABSTRAK

Pencapaian yang rendah dalam kalangan pelajar aliran sains terutamanya dalam mata pelajaran Kimia amat membimbangkan pihak-pihak yang berkepentingan. Ini menyebabkan wujudnya keperluan dalam menganalisis instrumen pentaksiran yang digunakan untuk mengukur pencapaian murid-murid berkenaan. Kajian ini bertujuan untuk mengkaji ciri-ciri psikometrik soalan peperiksaan berbentuk aneka pilihan bagi Peperiksaan Akhir Tahun Tingkatan Empat Kima Kertas 1 yang digunakan di sekolah-sekolah menengah di Lembah Klang melalui Model Rasch. Reka bentuk kajian yang digunakan merupakan kajian kuantitatif. Bilangan sampel kajian ialah sebanyak 435 orang murid Sains Tulen Tingkatan Empat daripada empat buah sekolah menengah yang dipilih secara rawak di daerah Petaling Utama. Analisis Rasch dilakukan dalam dua peringkat iaitu, peringkat pertama adalah untuk memastikan bahawa andaian unidimensionaliti dipatuhi. Manakala, peringkat kedua analisis pula merujuk kepada kesahan dan kebolehpercayaan instrumen. Analisis Komponen Utama (PCA) menunjukkan bahawa instrumen Kimia adalah unidimensional. Ukuran varians mentah data yang diperoleh ialah 27.7% iaitu hampir setara dengan ukuran varians mentah model, 27.8%. Ini menunjukkan data yang diperoleh adalah sepadan dengan Model Rasch. Selain itu, nilai eigen daripada kontras pertama menunjukkan kekuatan 2 item. Ini membuktikan bahawa tiada dimensi sekunder dalam item Kimia. Berdasarkan kesepadanan statistik, nilai infit dan outfit menunjukkan bahawa semua item berada dalam julat di antara 0.7 hingga

1.30. Analisis komprehensif ini bukan sahaja menunjukkan kesepadanan set data dengan model Rasch tetapi juga menyokong idea unidimensionaliti pada masa yang sama. Walau bagaimanapun, dimensi yang ditunjukkan hanya berada pada tahap memuaskan. Varians yang tidak dapat dijelaskan dalam kontras pertama set data mendapati bahawa terdapat sebanyak 3% tahap kebisingan berbanding dengan varians. Hal ini menunjukkan bahawa tiada faktor sampingan wujud dalam mengukur keupayaan murid. Dapatan kajian turut menunjukkan instrumen Kimia Kertas 1 memenuhi andaian kebebasan setempat memandangkan korelasi residual untuk setiap pasangan item adalah kurang dari 0.7, yang bermaksud tiada pertindanan item. Pada peringkat kedua analisis, anggaran kebolehpercayaan ialah 0.99 untuk item dan 0.87 untuk individu dengan indeks pemisahan masing-masing ialah 8.95 dan 2.54. Dapatan kajian ini menunjukkan bahawa keupayaan murid dikelaskan dengan sangat baik manakala pemisahan item di sepanjang kontinum pengukuran pula adalah luas. Peta individual-item menunjukkan bahawa kebanyakan item ujian hanya mengukur keupayaan murid pada julat kontinum pengukuran yang tertentu sahaja. Analisis statistik Rasch menggunakan PTMEA Corr. menunjukkan bahawa semua item ujian mempunyai korelasi positif dengan konstruk yang diukur dan bergerak dalam satu arah yang sama. Oleh itu, tiada item ujian yang tidak sepadan. Hasil analisis distraktor menunjukkan bahawa semua distraktor item ujian berfungsi dengan baik. Analisis tambahan iaitu analisis DIF menunjukkan bahawa terdapat enam item yang sukar dijawab oleh murid lelaki berbanding murid perempuan. Secara keseluruhannya, dapat disimpulkan bahawa instrumen peperiksaan akhir tahun Tingkatan Empat Kimia Kertas 1 mempunyai ciri-ciri psikometrik yang baik dan mampu menghasilkan skor yang sah dan boleh dipercayai dalam mengukur pencapaian murid. Selain itu, instrumen Kimia Kertas 1 ini juga

mempunyai standard yang setara dengan instrumen Sijil Pelajaran Malaysia (SPM) Kimia Kertas 1.

ACKNOWLEDGMENTS

In the name of Allah, the Most Gracious, and the Most Merciful

Alhamdulillah, all praises to Allah, the Almighty, who bestowed me with health, abilities, and guidance in completing this dissertation successfully. My sincere appreciation goes to my supervisor, Dr. Harris Shah Abd. Hamid, whose invaluable guidance, constant support, and constructive comments throughout the dissertation have contributed to the success of this study. His timely and efficient contribution assists me in shaping this study into its final form. Not forgotten, my appreciation to my former supervisor, Dr. Shahrir Jamaluddin for his valuable advice and supervision at the initial stages of this study.

I am eternally grateful to Dr. Bambang Sumintono who in his way provided insight in turning theory into application. Most of all, I owe a heartfelt debt of gratitude to Dr. Adnan Fateh for his guidance and patience not only in teaching research methods but his never-ending encouragement, caring, and untiring effort. Special thanks, tribute, and appreciation to the Head of Department, Professor Dr. Loh Sau Cheong for her assistance in any way that I may have asked. I also would like to acknowledge the staff of the Faculty of Education, the University of Malaya for their full support and co-operation towards my postgraduate affairs.

Last but not least, I wish to express my sincere thanks to all my friends especially Ardiana, and Wan Izani for their kindness and moral support in making this study a success. My deepest gratitude goes to my beloved parents for their endless love, prayers, and encouragement. To my daughters, Anis Athirah and Anis Aina Adleena, for they have inspired me in their ways to finish my dissertation. To

those who indirectly contributed to the successful completion of this study, your kindness means a lot to me. Thank you very much.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

The Malaysia Education Blueprint 2013 – 2025 was launched in 2013 by the Ministry of Education (MOE) which provides a holistic development framework based on access, quality, equity, unity, and efficiency to enhance the quality of the Malaysian education system. The education system has to be reorganized if Malaysia intends to compete with industralized countries for instance, the United States and Japan. The education system should be able to produce knowledgeable young people who are able to think critically and creatively, have strong leadership skills, and able to communicate effectively globally (Kementerian Pendidikan Malaysia, 2013).

One of the best indicators for the future development of the country is the teaching and learning process which takes place in the classroom. The success of a country mainly depends on the individuals' knowledge, skills, and competencies. The changing nature of work and society implies that in the present world, the premium is not only about gaining knowledge among students, but also to analyze, synthesize and apply what they have learned to tackle new problems, discover new solutions, collaborate effectively and interact convincingly (Bereiter, 2013; Pellegrino, 2014).

A shift in teaching, learning and assessment is essential to attain this objective. This transformation would entail a revamp of the curriculum and the assessment systems so that profound learning competencies able to be promoted. In line with this transformation, the Ministry of Education (MOE) needs to reorganize the curriculum and the assessment system so that new skills can be emphasized and enable our education system to meet the demands of education in the 21st century.

1.2 Background of the Study

The education system is the linchpin of the development of a society and country for a better direction. Nurul Awanis Abdul Wahid, Hazlina Abdul Hamid, Stephanie Low, and Zariyawati Mohd Ashhari (2011) argued that the education system in Malaysia is highly centralized. This system serves as a key role in increasing the quality of human resources so that it is able to cope with global challenges by providing knowledge and skills to the present and future generations (Rubiah Sidin, Juriah Long, Khalid Abdullah, & Puteh Mohamed, 2001). This is because good education increases the chances of employment. In this demanding global competitive environment, Malaysia has made significant reforms to the existing education system to enhance student achievement so that these high aspirations are achieved.

1.2.1 Achievement Test in Malaysia

Students' achievement has always been a key concern of all stakeholders including the Malaysian government, educators, and parents alike especially in high-stakes accountability due to the success or failure might bring serious (Veloo, Rahimah Nor, & Rozalina Khalid, 2015; Zulkifli Mohd Nopiah et al., 2012) National examination such as the Sijil Pelajaran Malaysia (SPM) which is carried out in the final year of secondary school is capable of determining the direction and future of the students as well as influencing the students for a lifetime (S. Brown, Race, & Smith, 2005). The information obtained from the examination is mainly for certification purposes and selection into the next phases of education or into employment (Howie, Long, Sherman, & Venter, 2009).

The Malaysia Examination Syndicate (MES) is the assessment body that is responsible in managing assessment, examination and tests such as the Ujian Penilaian

Sekolah Rendah (UPSR), the Pentaksiran Tingkatan Tiga (PT3) and the Sijil Pelajaran Malaysia (SPM). It is also the responsibility of the MES to install the instruments and to evaluate the validity and reliability of the instruments in accordance with the standard that has been set. The assessed items in the instrument have to be relevant and meet the specification of the syllabus in order for the quality of the instrument to be assured (Lembaga Peperiksaan, 2009). The quality of the instrument in measuring the performance of students plays a vital role especially in national examinations. This is to ensure fairness to the students and to retain the credibility of the assessment institution.

Achievement tests are one of the best traditional assessment tools that are still widely used by educators to measure the student performance with the aim of knowing whether learning has taken place in the classroom (Pang & Lajium, 2008). In addition, the achievement test serves as a motivation and guidance to student learning (Ebel & Frisbie, 1991). Literature notes that achievement tests were developed for collecting the basic knowledge acquired only in schools and through the daily life of students (Heckman, Humphries, & Kautz, 2014; Heckman & Kautz, 2012). In essence, achievement tests are tests which are particularly developed to measure the level of learning and proficiency shown by students in specific content areas.

1.2.2 Grades and Scores of Achievement Tests

Borghans, Golsteyn, Heckman, and Humphries (2016) in their study found that grades and scores from achievement tests were widely used as a measure of cognition. Scoring and reporting in public examinations usually follow the norm-referenced procedure which emphasizes is given by comparing the performance of a student with another student in the examination (Kellaghan & Greaney, 2001). The grades provided

solely state that they are higher than or lower than other grades, rather than providing information on the level of knowledge and specific skills of students (Howie et al., 2009). Although the students are only given grades but the assessment data obtained can be used as evidence in relation to the estimation of how students acquire knowledge and skills. As for trial achievement test, literature shows that grades expected by students to obtain have a significant relationship with their actual performance (Maksy & Zheng, 2008).

1.3 Rationale of the Study

In secondary school, the Chemistry subject is often seen as a boring subject due to the abstract concepts and unfamiliar language setting (Childs, Hayes, & O'dwyer, 2015). Furthermore, Chemistry is also seen as irrelevant to real-life phenomena that cause many students to lose their interest and are put-off the subject (Childs et al., 2015). However, in a technological society, it is critical to have an understanding of the fundamental chemical and scientific ideas that are vital for life. This understanding is also important for student in addressing the problems and issues of daily life. Therefore, teachers play a significant role in assessing student performance to ensure their understanding reaches the targeted outcomes as set by the Chemistry curriculum.

A comprehensive assessment is required to assess student performance, therefore all types of evaluation must meet the essential requirements of validity, reliability, and usability (Miller, Linn, & Gronlund, 2013). An ideal measurement can be used for various purposes (valid) and accurate (reliable) by using the test scores. It has a stable frame of reference for comparing various students, and offers a linear measure that can give significance to scores, and detect misfits. Hence, a valid instrument is vital in measurement as it is able to provide reliable data for a meaningful

analysis. The valid instrument is also able to generate useful information that can be used specifically in decision making. Many researchers have proved that the Rasch model is a suitable approach for examining and validating the educational instrument. However, many teachers still favor in enrolment using the traditional method of CTT to measure student performance.

1.3.1 Chemistry in Education

Recently, there is a matter of significant concern on students' enrolment in science and technology at different levels of education system as the economy is increasingly powered by complex knowledge and advanced cognitive skills. The Organization for Economic Cooperation and Development Global Science Forum (2006) found that the most impacted fields by lack of student interest are subjects with lots of theoretical content for instance mathematics, physics and chemistry. According to Ruhaiza Rusmin (2015), the decline in student enrolment in the Science stream has taken place not only in Malaysia but it is happening throughout the world. In general, Chemistry and other Science subjects are very important to students as they act as a catalyst for national development in producing experts in various fields such as engineering, and technology (Dani Asmadi Ibrahim, Azraai Othman, & OthmanTalib, 2015). Chemistry is also known as the main science because it connects physics with other natural sciences such as biology and geology (Veloo et al., 2015).

In Malaysia, Chemistry is one of the elective science subjects taught in upper secondary schools for science stream students and its curriculum is designed to provide scientific knowledge and prepare students for tertiary level and develop student's abilities in solving the related problems (Siti Salbiah Omar, Jamalludin Harun, Johari Surif, Noor Dayana Abd Halim, & Suraiya Muhamad, 2016). Essentially, the main

objective of the chemistry curriculum is to prepare students with the knowledge and skills in chemistry and technology for solving problems and making decisions on the basis of scientific attitudes and noble values in their daily lives. Additionally, the curriculum also seeks to develop a responsible, dynamic and progressive society with a culture of science and technology that values nature and works to preserve and conserve the environment (Curriculum Development Centre, 2005).

Diagnostic tests such as TIMSS have shown that only 6% of the Malaysian students have mastered the science content that contributes 20% of the Chemistry domain. This finding clearly demonstrates that Chemistry is at a critical level among students (Martin, Mullis, Foy, & Stanco, 2012). Literature has proved that students who have poor mathematic skills and lack of basic knowledge as well as weak in mastering the concepts involved (Aziz Nordin, 2007; Chan, Zaleha Ismail, & Sumintono, 2014) have faced difficulties in learning Chemistry.

Most of the students found that Chemistry is a challenging subject as various issues such as topics that are typically related to or focused on the structure of matter which is closely related to the abstract concepts (N. Grove & Bretz, 2012; Sirhan, 2007). This abstract nature of chemistry requires a high-level skills set because of the reliance on mathematical equations in explaining these phenomenon (Fensham, 1988; Harris, 2003; Taber, 2002; Zoller, 1990). In addition, the confusion caused by teachers who move quickly between macro, sub-micro and symbolic representations also regards Chemistry as a difficult subject (Bucat & Mocerino, 2009; Johnstone, 1991; Mocerino, Chandrasegaran, & Treagust, 2009; Van Driel, Jong, & Verloop, 2002). As there is an abundant amount of abstract concepts in Chemistry, students have to provide a lot of significant time and effort to learn the subject (C. Wu & Foos, 2010). Furthermore, the finding of a research has shown that most of the issues in learning

and comprehending Chemistry tend to be triggered by a view of Chemistry instruction. which is mainly seen as academic and unrelated to the students' daily life (Treagust, Duit, & Nieswandt, 2000). Chemistry would draw students' interest, curiosity and understanding if teachers demonstrate that science is a human enterprise (Cardellini, 2012).

There are three levels of chemical knowledge which are the macroscopic, the sub-microscopic and the symbolic. Macro representations are the observable properties of substances, sub-micro representations are models of atoms, electron density, clouds of molecules and symbolic representations are the symbols to represent atoms and chemical equations. This triplet relationship of macro, sub-micro and symbol representations is a key model for chemistry education and also known as chemical knowledge 'triplet' (Talanquer, 2011). This idea has become highly influential and widely useful in the chemistry education (Taber, 2013).

1.3.2 Assessing Students' Ability

Student assessment is a common practice in schools which links student performance to specific learning objectives with the aim of evaluating whether student have learned what has been taught. The assessment also provides useful information about progress of the students. Therefore, the students' ability to acquire knowledge and comprehend the subject that is being taught need to be accounted in the instruction.

A previous research notes that understanding of students is the most important aspect of learning (Sun & Chen, 2009). Thus, a good achievement test should be developed so that it can measure the performance of students according to their level of ability (Zulkifli Mohd Nopiah et al., 2012). Performance based on assessment can be referred as a set of strategies for applying knowledge, skills and work habits through

meaningful performance of tasks that engage students. It provides information on how a student understands and applies the knowledge (Brualdi, 1998).

1.4 Problem Statements

In Malaysia, science stream students are introduced to Chemistry in their upper secondary schools. The Chemistry curriculum is intended to develop and prepare students with Chemistry knowledge and skills to pursue their studies in Chemistry and relevant disciplines at higher education level, in addition to apply the knowledge and expertise gained for the development of the country (Siti Salbiah Omar et al., 2016).

Nevertheless, an analysis of the Sijil Pelajaran Malaysia (SPM) 2018 results that were released recently demonstrated that students had low grades in Chemistry subject. The percentage of students who achieved excellent results in 2018 recorded an increase of 0.1 while the percentage of students who passed with the minimum grade recorded the highest incline of 3.4. This analysis indicates that only a few students have achieved excellent results in Chemistry subjects. Meanwhile, the majority of the students who sat for the Chemistry examination only passed with minimum levels. Ultimately, the SPM result analysis of Chemistry concluded that the achievement of students below expectation. The persistently low student achievement in Chemistry especially in high-stakes tests such as SPM continues to attract the attention of major stakeholders in education because the performance of the students determines their admission to pre-university studies and later to universities.

Table 1.1
Results Analysis of the Sijil Pelajaran Malaysia (SPM) 2018

Year	Excellent (A^+, A, A^-)	Good (B^+,B,C^+,C)	Pass (D, E)	Fail (G)	A^+ - E
		Percentage (%)			
2017	19.6	46.0	30.6	3.8	96.2
2018	19.7	43.3	34.0	3.0	97.0
Difference	0.1	-2.7	3.4	-0.8	0.8

Source: Lembaga Peperiksaan (2019)

Assessment is a crucial aspect of education due to its ability to measure students' achievement (Hamilton & Klein, 1999). There are two types of assessments conducted in schools, namely formative assessments, and summative assessments. Formative assessment is primarily observed student learning and offers continuous feedback. Summative assessment, on the other hand, is intended to assess student learning by comparing it with some standard or baseline at the end of an instructional unit. The year-end examination is one of the summative assessments that is similar to the SPM trial examination and SPM examination.

The year-end Chemistry Paper 1 test items are constructed by teachers who are experts in Chemistry subject using the Table of Specification (TOS) to guarantee the test measures the content and thinking skills that it aims to measure (Fives & Barnes, 2018). Although the teachers are content experts, yet some of the them may lack knowledge in test development. Therefore, some of the items constructed may be flawed, biased, or not reliable to measure the performance of students. Students' performance can only be measured correctly (Banta, 2007; Figlio & Lucas, 2004; Fuchs, Fuchs, & Kazdan, 1999) with a reliable and valid measurement tool (Baghaei, 2008). Apart from the precise measurement tool, a reliable instrument is crucial in

determining what it intends to measure is measured. Therefore, the validity and reliability of the year-end exam Chemistry Paper 1 needs to be validated.

Validity and reliability act as key indicators in determining the quality of the measuring instrument as well as to maintain the accuracy of the instrument (Kimberlin & Winterstein, 2008; Siti Rahayah Ariffin, Bishanani Omar, Anita Isa, & Sharida Sharif, 2010). The quality of the year-end examination of the Form Four Chemistry Paper 1 depends on the quality of each item. Hence, it is very important to understand how well each item works to ensure the overall test measures the construct.

High-quality test items required an extensive amount of time and effort to be produced particularly multiple-choice questions as this form of test involves options for answer and distractors. The year-end examination of Form Four Chemistry Paper 1 comprises of 50 multiple-choice items with four options. Empirical studies have revealed that three options may be feasible and suitable for most ability and achievement tests (T. M. Haladyna & Downing, 2016; T. M. Haladyna, Downing, & Rodriguez, 2002; Rodriguez, 2005). Therefore, there might be a tendency for the Chemistry items and the Chemistry test to be biased or flawed or have the implausible distractors.

The Rasch model is able to detect the presence of these faults based on the answers of students. Rasch model emphasizes on the calibration of students' ability and the difficulty of the item, estimation of the model fitness, unidimensionality evaluation, and distractor analysis. These are the indicators used to measure the quality of the test items and their pertinence with the trait measured (Baghaei, 2008). The quality parameters or problematic and good items are an important stage in developing valid and reliable test items for measuring the true ability of students. By using the Rasch model, different versions of the instruments can be developed so that they can

be targeted to all students without taking into account the type completed and could be expressed on the same yardstick (Boone & Noltemeyer, 2017).

In test development, it is vital to guarantee the test is unidimensional because it implies that the items solely measure a single capability (Wright & Masters, 1982). However, educators often disregard in evaluating this criterion as it involves complicated procedures. Therefore, the unidimensionality of the year-end exam Chemistry Paper 1 is tested to ascertain to what extent the test meets the measurement criteria of the Rasch model. If the Chemistry items in the present study match the model, it implies that unidimensionality is supported and explains how well is the content validity (Sick, 2011; Wright, 1996). Additionally, other validity indices that are produced are person and item reliabilities, item difficulty, and item fit (Wang, 2008). The validation process was developed to ensure the quality of the year-end examination of Chemistry paper as a suitable and valid instrument in assessing the student achievement.

To date, most of the educators in Malaysia are still employing the Classical Test Theory (CTT) in analyzing test results (Adibah Abdul Latif, Ibnatul Jalilah, Nor Amin, Wilfredo Libunao, & Yusri., 2016). Many educators including the Malaysia Examination Syndicate (MES) use CTT due to its simplicity and easy to apply (Hambleton & Jones, 1993) not to mention its ability to provide overall and descriptive summaries (Royal, 2010). CTT is used to predict the test result by considering several parameters like students' abilities and item difficulty. Principally, the assumption of the CTT is based on the linear relationship in which the observed score (X) is equivalent to the true score (T) added with the measurement error (E), hence the equation is: $X = T + E$ (Alagumalai, Curtis, & Hungi, 2006; Sumintono & Widhiarso, 2015). The CTT posits that each person has a true score, but only the observed score,

which are the raw scores are real, however the true score and measurement error are latent. Furthermore, true score only exists in theory (De Ayala, 2013; Hambleton & Jones, 1993; Lord, 1980). Therefore, the test scores obtained in the achievement test consist of true scores and measurement error.

In schools, the most traditional and well-established method used by the teachers is summing up the raw score of the students' correct answer, and the accumulated raw scores are taken into account as a measurement of students' ability. An increase in test scores is often used as a piece of explicit evidence to infer that students have learned more and vice versa. However, cumulative raw scores only consider discrete observations, and not measuring the performance of students (Wright & Mok, 2004).

Measurement experts warn against the traditional method of using raw score in the analysis of assessment data because the raw score forms a linear scale instead of an interval scale which cannot be generalized and treating linear scale as interval scale will lead to the misinterpretation of test quality and students' achievement since the raw score for the students and items does not fit each other (Embretson, 1996; Wright & Masters, 1982). Hence, if the scale is formed from the raw score, it is incorrect to compare the abilities of students directly (Embretson, 1996). On the contrary, it is sample dependent.

According to Bond and Fox (2015), raw scores can estimate students' abilities in a hierarchy, but it is incapable of determining how these abilities differ from other students. Raw scores cannot accurately differentiate between those who are more capable and less capable. Bond and Fox (2015) added that the abilities of students are unexplained or inaccurately explained when students obtained 0% or 100% in their achievement test because the raw score and the percentage of correct responses are not

always linear. For instance, a student who earned 100% in Chemistry is considered to have mastered the knowledge that has been taught. While the students who earned 0% are considered weak or do not have the knowledge at all. The difference in students' achievement cannot be stated clearly if an uneven scale is used.

Educators typically tend to tally up the scores from different tests parts or across different items in a test and make the same assumption from that accumulated scores (Henning, 2016). Consequently, there might be more than a single factor that has a significant effect on the covariance for all items in their respective constructs. However, for unidimensionality measurement, each item needs to be measured to ensure that the items work together to form a single basic pattern in the measurement matrix (McNamara, 1996). Through this measurements, instruments are able to measure high quality latent trait such as students' abilities (Sick, 2010). In essence, a set of items assessed only one latent trait for unidimensionality (Finch & French, 2018).

In year-end examination of Form Four Chemistry Paper 1, the correct response of students on an item probably depends on the answers of other questions. Therefore, the abilities of students are measured inaccurately while educators are inclined to make wrong interpretations on their students' abilities due to the summated raw scores obtained in the achievement test. These summated raw scores exhibited local dependency which failed to detect any significant source of covariance between items and as a result lead to biased inferences (Edwards, Houts, & Cai, 2018). Local independence, which is a fundamental assumption, should therefore be checked to determine the correlation between any set of items in the same construct for a fixed level of the trait. (Cappelleri, Jason Lundy, & Hays, 2014).

The year-end examination of Form Four Chemistry scores can be analyzed and interpreted in detail to gain a full picture of the students' performance by using a modern measurement method which is IRT. Although the IRT model has numerous variations, this study only focuses on the Rasch Model. The likelihood of a student answering a test which depends on his ability and the item difficulties can be predicted by using the Rasch model (Bond & Fox, 2015; DeMars, 2010).

Validity and reliability are the two aspects employed to evaluate the quality of the assessment tools (R. Cohen, Swerdlik, & Sturman, 2013). The tests that are used should produce valid, reliable, and accurate evidence of the purposes they serve and for whom they are intended. Hence, it is essential to measure the validity and reliability in test development (Pada, Kartowagiran, & Subali, 2016).

The Rasch model is an alternative scaling approach to the CTT in the educational instrument development (Prieto, Alonso, & Lamarca, 2003; Sumintono, 2018). In comparison, CTT mainly depends on the correlation principle (Ganglmair-Wooliscroft & Lawson, 2003). However, the basis of the Rasch model is the rigid mathematical model of a theoretical relationship (Bond & Fox, 2015). In terms of parametric, Rasch emphasizes on the difficulty of the item and the ability of the student, but the CTT concerns are limited to item difficulty and item discrimination indices. Therefore, the Rasch model is able to measure precisely the ability of students.

Furthermore, the Rasch model also offers a comprehensive insight on the quality of the test items. The Rasch analysis can be very worthwhile in completing this evidence by providing information on the quality of the test at particular measurement points of the scale (Zanon, Hutz, Yoo, & Hambleton, 2016). The Rasch model is unintended to replace the traditional methods that were prominent and crucial.

Nevertheless, the Rasch model is a valuable tool which could be employed to improve the quality of psychological evaluation.

1.5 Limitation of the Study

The present study only covers year-end examination of Form Four Chemistry Paper 1 that is multiple-choice question items in nature. Therefore, the results of this study cannot be generalized for entire year-end examination of Form Four Chemistry Paper.

1.6 Delimitation of the Study

This study is delimited to Form Four pure science stream students who have learned Chemistry as one of the elective subjects in the selected schools of Petaling Utama district only due to the time and financial constraints.

1.7 Purpose of the Study

This study is conducted to determine the psychometrics properties of the year-end Form Four Chemistry Paper 1 of the selected schools in Petaling Utama district.

1.8 Objectives of the Study

The objectives of this study are to:

1. determine to what extent does the data set of year-end Form Four Chemistry Paper 1 fit the Rasch Model.
2. determine the student reliability and item reliability of the year-end Form Four Chemistry Paper 1.
3. determine the item validity of the year-end Form Four Chemistry Paper 1 exam.
4. identify the appropriateness between item difficulty of year-end Form Four

Chemistry Paper 1 and Form Four Science students' ability.

1.9 Research Questions

This study seeks the answers of the following questions:

1. *1.10 Significance of the Study* To what extent does the data set of year-end of Form Four Chemistry Paper 1 fit the Rasch Model?
2. What are the student reliability and item reliability of the year-end of Form Four Chemistry Paper 1?
3. What are the item validity of the the final examination of Form Four Chemistry Paper 1 exam?
4. What are the appropriateness between item difficulty of year-end of Form Four Chemistry Paper 1 and Form Four Science students' ability?

This study is significant in the education field because item analysis is the process of gathering, summarizing and employing students' response to evaluate the quality of the test items and is one of the most crucial aspects in test construction (Quaigrain, Arhin, & King Fai Hui, 2017). It is important to:

a) Chemistry teachers

The Year-End Examination of Form Four Chemistry Paper 1 is a summative assessment conducted at school level to measure student's current status conclusively at a single time point. By using year-end exam as a platform, the data obtained becomes a meaningful source to determine whether the content and objectives of Chemistry have been mastered. Literatures states that summative assessment which is also known as assessment of learning can be used to determine the future of students (Guskey, 2003) and can affect the students for life (S. Brown et al., 2005).

Improvements and modifications can be made to the lessons based on assessment information.

b) School Administrators

The answers of students are a method used by teachers to address students' difficulties in responding the achievement test and to help teachers improve their skills in constructing quality items (Denny, Luxton-Reilly, & Simon, 2008; Krause & Kelly, 2011). Therefore, it is important for teachers to have the knowledge in the testing techniques so that the progress of the students can be evaluated reliably and validly. In addition, teachers need to have knowledge in determining the validity and reliability of tests as they are often involved in the construction of test items (Magno, 2009). A teacher should have knowledge on what good test items are and if the test items could depict the achievement of students in regards to the learning objectives.

Based on the findings of the present study, school administrators will be able to identify the level of knowledge of Chemistry teachers in the assessment especially in test development. Similarly, the use of statistical analysis of the tests can enhance teaching strategies as well as the construction of tests (Hingorjo & Jaleel, 2012). The school administrator may take follow-up actions in assisting teachers by conducting in-house training or knowledge-sharing sessions in collaboration with the Malaysia Examination Syndicate (MES) to strengthen test development knowledge and skills. According to Xu and Liu (2009), teacher knowledge in assessments and evaluation must be updated as evaluations are complex, dynamic and ongoing activities. Teachers who are knowledgeable in the assessment can produce good and quality instruments in measuring the achievement of students and help the school achieve the targets set by DEP.

c) District Education Department (DEP)

The findings of the study may assist the DEP in determining the suitability of items for bank item storage because the analysis of this study provides information regarding the features of the items constructed including their quality and functionality (S. Downing & Haladyna, 2006; Sadia Mahzabin, Roszilah Hamid, & Shahrizan Baharom, 2015; Waugh & Gronlund, 2013). As DEP Petaling Utama is the one that provides the achievement test, the findings of this study can be used to guarantee items in of appropriate standards to be included in a performance test or whether the items require modification (Azrilah Abd Aziz et al., 2008; Quaigrain et al., 2017). Item modified through complete item analysis can produce the most effective and stringent instruments (Boone & Noltemeyer, 2017).

The poor and defective item replacement is a more economical process in order to obtain good items for future testing rather than discard the items and construct new items to replace them that can take much longer time than revise the existing ones (Crocker & Algina, 2006; Lange, Lehmann, & Mehrens, 2005). Literature notes that the use of poor items affects the reliability of test scores where low reliability indicates that the scores obtained by the students may be unreliable and the information obtained with it is less or worthless (Ainol Mardziah Zubairi & Noor Lide Abu Kassim, 2006; Kaplan & Saccuzzo, 2017; Urbina, 2004).

1.11 Operational Definition

The following terms will guide the reader in better understanding the terms used in this study.

1.11.1 Psychometric Properties

Psychometric properties are characteristics and other measures of human characteristics of tests that identify and describe features of an instrument, such as its reliability or suitability for use in certain circumstance. Ginty (2013) and Portney and Watkins (2013) defined psychometric properties as the construction and validation of measuring instruments that need to be extensively evaluated whether they are reliable and valid forms of measurement.

1.11.2 Year-end Form Four Chemistry Paper 1

The year-end Form Four Chemistry Paper 1 is a summative assessment of multiple-choice questions which is conducted at the final year of school to measure the amount of Chemistry knowledge acquired by the Form Four students. The Chemistry Paper 1 is constructed by test developers using test of specification in collaboration with the Majlis Pengetua Semenanjung Malaysia (Selangor) and District Education Department. This test paper is set for form four science stream students.

1.11.3 Selected schools

Selected schools refer to the public secondary schools that are fully funded by the government which have been chose by the simple random sampling technique.

1.11.4 Fit Statistics

Two forms of fit statistics, infit and outfit are typically used. The mean square statistics (MNSQ) and standardized statistics (ZSTD) could be used to report both statistics. Outfit is equal to a sum of squared standardized residuals divided by its degree of freedom which could vary from 0 to infinity (Osborne, 2008). Values exceed

1.0 signify higher variability than expected. Values less than 1.0 indicate small variability than expected in the data (R. M. Smith, 2004; Wright, Linacre, Gustafson, & Martin-Lof, 1994; Wright & Masters, 1982). Outfit statistics are unweighted and therefore more sensitive to anomalous responses by either very high-ability or very low-ability persons, especially on tests with a wide variety of item difficulties and person abilities (Mead, 2008). Infit is weighed to lessen the impact of less informative, non-target responses and overweighs the completely unexpected responses by persons close to the center of the item distribution.

Similar to the outfit mean square statistics, the infit mean square statistic is expected to be 1.0 and could range from 0 to infinity (R. M. Smith, 2004; Wright et al., 1994; Wright & Masters, 1982). To assist with the asymmetrical distribution, mean square fit statistics could be converted into standardized fit statistics using a Wilson-Hilferty cube root transformation (A. B. Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). Standardized fit statistics greater than 0 suggest unexpected data variability (Linacre, 2002). The standardized fit statistics range for acceptance is -2.0 to 2.0 (Linacre, 2002), while -3.0 to 3.0 might be warranted for larger samples.

1.11.5 Item Analysis

Item analysis is an activity to evaluate the items in a test in which the activity could generate a form of a testing tool with minimal items but maximum reliability and validity (Kaplan & Saccuzzo, 2017). In contrast, item analysis is described by Salkind (2010) as the set of qualitative and quantitative techniques and processes used to measure the features of test items before and after the development and construction of the test. For educators and psychometricians, item analysis is referred to as an analysis of student responses towards each exam questions with the intention of

assessing the exam quality (Lee, 2019). Statistical methods are used for selecting items in item analysis. The process of item analysis is various as it depends on the measurement model used.

1.11.6 Item Fit and Person Fit

Item fit is a psychometric concept of infit and outfit statistics based on the Rasch modelling (Linacre, 2002). According to Reise (2016),item fit can be used to assess the test dimensionality that influences the validity of the test results and indicates errors that occur in the items calibrations. The acceptable range for MNSQ statistics by item dichotomous is 0.5 to 1.5 (Linacre, 2005). However, the proposed range is between 0.7 to 1.3 (Adams & Khoo, 1996; Bond & Fox, 2015).

Person fit is referred as the reproducibility of the sequence of order for every person when another set of items is given to assess the same construct (Wright & Masters, 1982). While Linacre (2012) referred to person fit as the assumption of individual abilities in the sample is consistent even with different sets of items but still measuring the same constructs.

1.11.7 Item Polarity

The validity of an instrument can be determined by reference to the analysis of the output program in accordance with the Rasch Model. The primary output to be referred to is the item polarity in order to find a measuring-point correlation coefficient which is recognized as point–measure correlation coefficient (PTMEA Corr.).

A high PTMEA Corr. indicates that the ability of respondents can be differentiated by an item. A negative value or null implies that the relation is conflicting with the variable or construct for the item response or respondent (Linacre,

2006). According to Allen and Yen (2001), if every PTMEA Corr. is ranged within 0.30 to 0.70, it indicates that the items might contribute to the respondents' measurement. This may distinguish between the different types of respondents' intelligence.

1.11.8 Local Independence

Local independence is the assumption that depends on the latent variable(s) on how the response of the items is not related to each other (Edwards et al., 2018). Literature refers to local independence as a person's response on an item is unaffected by his or her responses to the other items on the identical test. Precisely, this means that there will be insignificant covariance between any sets of items (Embretson & Reise, 2000; Sijtsma & Molenaar, 2002; Yen, 1993). According to Hambleton, Hambleton, and Swaminathan (1985), "the content of one item must not provide any clues to the answer to another item". This fundamental assumption asserts that the observed items are independent, therefore provided a person's score based on the latent construct(s) (Lewis-Beck, Bryman, & Liao, 2004).

In contrast, Finch and French (2018) defined local independence as no correlation among item responses once educators take into account the latent trait measured. As Cappelleri et al. (2014) point out that there is no or trivial correlation between any set of items for a fixed or specified level of the trait. According to Linacre (2010), both items are local dependence if the correlation value of any pair items exceeds 0.7, and thus only a single item remains for every pair.

Local item dependence could cause to inaccurate estimation of difficulties of the item, test statistics and person abilities, as well as overestimation of the function of reliability and information (Sireci, Thissen, & Wainer, 1991; G. T. Smith, 2005;

Thissen, Steinberg, & Mooney, 1989; Zenisky, Hambleton, & Sireci, 2001). Furthermore, local item dependence initiates unintended dimensions of the test at the expense of the construct of interest (Wainer & Thissen, 2005).

1.11.9 Logit

Logit is known as interval level units of measurement that match the total scores that have underwent an exponential conversion (Rasch, 1960). Yet, according to Bond and Fox (2015), logit is a unit derived from the transformation of ordinal data into an interval scale. In contrast, Ludlow and Haley (1995) defined logit as a natural log of an odds ratio. Royal (2010) explains in detail that logits are the measures produced from ordinal data that appear from the raw score (frequency) when computed via the Rasch model into odd probability and then converted into logarithm, thus, yields a measure that has interval properties which is interpretable.

1.11.10 Raw Score

The cumulative scores a test taker attains by responding to the questions correctly during a test (Tan & Michel, 2011). According to Urbina (2004), raw score is a number that summarizes or captures some aspect of a person's performance in the selected and observed behavior samples that constructs psychological tests.

1.11.11 Separation Index

An index that classifies individuals or items into multiple groups. Separation index determines reliability. Separation index is the square root value of the ratio between the true variance and the error variance in the data (Linacre, 2012). There are two facets of measurement interest which are person and item. Therefore, the internal

consistency reliability of both facets need to be analyzed. The Person Separation Index (PSI) is a measure of reliability similar to the Cronbach's alpha. However, it is calculated on the basis of a non-linear transformation of the raw scores (Tennant & Conaghan, 2007). PSI is a reliability statistic comparable to Cronbach's alpha which quantifying the error associated with the measurement of person in the sample. Andrich (1982) mentioned that higher value of PSI indicates a higher reliability. He also added that if the PSI value is greater than 0.7, the reliability is considered adequate.

In earlier studies, person separation reliability is known as the reliability index that corresponds to the traditional KR-20 or Cronbach's alpha which oftenly used in the classical test theory (Osborne, 2008; Wright & Masters, 1982). In contrast, item separation is the distance in logits between items of different difficulty levels (Draugalis & Jackson, 2004). The person and item reliability indices are calculated to guarantee consistency using two forms of reliability coefficients which are reliability (analogous to Cronbach's alpha) with the value between 0 and 1 and separation index (the number statistically different performance strata that can be identified in the sample by the test) (Fisher, 1992).

Person and item separation index and reliability of separation also measure the spread of the items throughout the continuum of the trait. For an instrument, separation index must greater than 1.0 as higher values of separation index indicate items and persons are widely spread along the continuum. Linacre (2005) points out that the value of separation index of person and items which is greater than 2 are regarded as good. Separation index may range from 0 to infinity and this index has no ceiling (Boone, Staver, & Yale, 2014).

Low separation index signify that several items are redundant and low variability of person on the trait. On the other hand, higher separation results in greater reliability due to it is in accordance with variance in the position of the person or item.

1.11.12 Unidimensionality

The existence of a dominant ability or trait which affects performance of the tests (Hambleton, Swaminathan, & Rogers, 1991). Unidimensionality is a basic assumption in the Rasch model which is fulfilled if the same trait is measured by all items in the instrument (Sijtsma & Molenaar, 2002; R. M. Smith, 2004; Wright & Masters, 1982). This does not mean that a single psychological process strictly affects the items' performance. The unidimensionality assumption is met even if the act of responding to items involves multiple psychological processes as long as they are affected by the same fundamental process (Bejar, 2016).

In assessing the unidimensional, Rasch (1980) stated that the raw variance explained by the measurement should be at minimum of 40%. While Linacre (2005) notes that if the residual factor has a strength of 5 items or more in Principal Component Analysis (PCA), this factor may provide useful information to ensure that a test is unidimensional which is measuring only a single construct.

1.12 Summary

This chapter examines the basis of this study such as the background problems related to item analysis of year-end Form Four Chemistry Paper 1. The objectives, the research questions, the delimitations and the limitations of the study including the definition of the terms used in this study are also provided in this chapter. Achievement test or examination is the most popular traditional tool that has been used widely to examine

the understanding of subject matter among students. The results obtained can help the teachers to improve their instructions in order to increase the student's ability.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter focuses on the review of relevant literature related to Rasch Model, validity and item analysis. In the first part of this chapter, researcher would review the international assessment, standardized test and multiple-choice question as Chemistry Paper 1 is a multiple-choice question format. Next part, researcher would describe the studies in term of item analysis and theoretical framework comprises classical test theory, item response theory (IRT) and Rasch Model. For the final part, researcher would explain about validity and reliability of the items and instrument.

2.2 Theoretical Framework of Study

2.2.1 Classical Test Theory (CTT)

The classical test model is a basic test performance model and a core of classical test theory. It is a traditional quantitative approach to test a scale's reliability and validity based on its items (Cappelleri et al., 2014). It is also recognized as a true score theory due to its relation with the analysis of the test results (M. Wu, Tam, & Jen, 2016) . Classical test theory assumes that observed score or raw score (X) is made up of true score (T) and measurement error (E) component, expressed in the form:

$$X = T + E$$

True score reflects the notion of ability. It can be defined as the expected value of the performance observed on the test (Hambleton et al., 1991). The measurement error is the result of any systematic or random factor that unrelated to the construct

being measured and represents the discrepancy between students' observed score and their corresponding true score (Cappelleri et al., 2014; Salkind & Rasmussen, 2008).

A significant fact in CTT is the measurement error or also known as magnitude of the error variance. The measurement error influenced the true score which is represented by the observed score. Therefore, the lower the measurement error, the more accurate the true score. A finding from an empirical study indicates that the most common reason for errors is associated with differences between items (Anthony, DiPerna, & Lei, 2016). It is possible to quantify and indirectly measure this type of error by ascertaining the internal consistency of a scale that shows how well items are linked to the scale (Barchard & Brouwers, 2016). Maximization of internal consistency is the main focus of scale measurement in CTT (Streiner, 2003).

The fundamental aspect of CTT is linearity which the observed scores are linearly regressed on the latent constructs and the interval scale assumptions. A major problem with the linear relationship assumption is that the observed scores or the summated scores are treated as if they were also interval scores when the latent attribute is assumed to be on an interval scale. CTT provides a rich framework for conducting analyses. When the two main hypotheses, either theoretically or empirically justified, the observed scores will lie on a metric scale and the functional relationship becomes linear (Hambleton, 1993; Rusch, Lowry, Mair, & Treiblmaier, 2017).

Although CTT is excellent in providing overall and comprehensive summaries of data, it is inadequate for truly objective measurement (Royal, 2010) but results obtained from CTT can be used as a basis to select best-fitted item response theory (IRT) model (Wiberg, 2004).

2.2.2 Item Response Theory (IRT)

Apart from CTT, an alternative approach of test development and item calibration that extensively used is item response theory (IRT) or latent trait theory which is employed to solve practical testing issues in a wide-scale cognitive achievement (Reise & Waller, 2009). IRT, often referred to as modern test theory has been the preferable statistical methodology for item analysis since it was developed in the 1950s and 1960s as it provides a statistical foundation that can be employed in a variety of contexts (eg. test development, item analysis, equating, item banking and computer adaptive testing). According to psychometricians (Crocker & Algina, 2008; Gorin & Embretson, 2007; Osteen, 2010), the rationale for the preference is due to IRT provides a range of statistical tools for assessing the measured characteristics and yields an overall picture of how an item functions.

IRT is derived from CTT with its purpose to overcome two major limitations: the test-dependent score and the presence of a single standard error of measurement for the population (Hambleton, 1994; Hambleton et al., 1991; Van der Linden & Hambleton, 1997). But, the main objective of IRT is to calculate the parameters of a mathematical function, normally a logistic function, that "models" the relationship between the latent trait and the item responses (Reise & Waller, 2009).

Cappelleri et al. (2014) described IRT as a collection of measurement models that is able to explain the connection between responses of observed items on a test and an underlying trait which is measured by the person's responses. Particularly, IRT models are mathematical equations that define the relationship between respondents' levels on a latent trait and the likelihood of a specific response to an item, using a non-linear monotonic function (Hays, Morales, & Reise, 2000). In addition, IRT is also known as a set of psychometric models for the development and refinement of

psychological measures, administering scales and measuring individual differences on psychological construct (Embretson & Reise, 2000; Reise & Haviland, 2005).

A recent study conducted by Sulis and Toland (2016) reported that IRT is a probabilistic framework used for the development and analysis of a multiple item instrument. It has the potential to elucidate both items and person characteristics of the scale by conjointly linking item parameters and latent trait values on the same scale. Principally, IRT is based on the basis that only two components are accountable for an individual's response on any given item: the response of the individual, and the features of the item (Bond & Fox, 2015).

Essentially, the concept of IRT is on the likelihood of a person achieving a certain score on a test as a result of that person's ability on the latent trait and the difficulties of the item (Reise & Haviland, 2005). Therefore, when the ability of a person changes, the likelihood of endorsing a correct response also changes (Hambleton et al., 1991). The ability of a person does not depend on the items. The precision of the measurement in IRT relies on the latent trait value, therefore the precision also depends on the items. Estimates of ability are more precise when they're based on items that are close to a person's ability level. If there is a discrepancy between the ability of the person and item difficulty, then the precision will be reduced.

The most significant difference between IRT and CTT is the shape of the relationship between the item score and the construct score. IRT models a curvilinear relationship between the two score, while a simple linear relationship between them is modelled by CTT models.

Table 2.1
Main Differences Between Classical and Item Response Theories

Area	Classical Test Theory	Item Response Theory
Model-main difference	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (easy to meet with test data)	Strong (more difficult to meet test data)
Item-ability relationship	Not specified	Item characteristics functions
Invariance of item and person statistics	No - item and person parameters are sample dependent	Yes - item and person parameters are sample independent, if model fits the test data
Item statistics	*p*, *r*	*b*, *a* and *c* (for three parameter model) plus corresponding item information functions
Sample size (for item parameter)	200 to 500 (in general)	Depends on the IRT model but larger samples, eg: over 500 in general

Source: Hambleton and Jones (1993)

A number of studies have reported that the selection of IRT model depicts the likelihood of a certain response to a particular item in agreement with the parameters of the item and the respondents' latent traits (Hambleton et al., 1991; Reise, Widaman, & Pugh, 1993; Tavares, Andrade, & Pereira, 2004; Van der Linden & Hambleton, 1997). However, the most popular IRT models are based on the unidimensionality assumption, according to which responses to a set of items can be explained only by a single underlying trait, although multidimensional IRT models have been also developed (Reckase, 2009).

Table 2.2
Type of IRT Models for Dichotomous Data

	Item Difficulty	Item Discrimination	Guessing Parameter
1-Parameter (Rasch)	✓		
2-Parameter	✓	✓	
3-Parameter	✓	✓	✓

Source: Hays et al. (2000)

Three common IRT models for dichotomous data are one, two and three parameter logistic models. These IRT models can be differentiated based on the number of parameters estimated. The one-parameter logistic model (1PLM) is used to estimate the probability that a person will answer the item (of difficulty, *b*) correctly. While, the two-parameter logistic model (2PLM) is not only aimed to estimate the probability of a correct answer but also allows estimating the discrimination of the item (a). On the other hand, the three-parameter logistic model (3PLM) estimates the other two parameters described (*a* and *b*) and the probability for guessing (the "*c*" parameter) (Hambleton et al., 1991).

An empirical study indicates that IRT can predict a student's scores through a function called the "item characteristics curve" (ICC) that is based on his or her capabilities or latent traits as well as substantiate a relationship between his/her item performance and the set of traits underlying item performance (Hambleton et al., 1991). ICC is the primary unit in IRT and incline to be S-shaped curves when the test data are dichotomous. This function also allows flexible specifications of the theoretical relationship between the latent underlying features and the answer format, contexts, or theoretical assumptions regarding the response process (Rusch et al.,

2017). The curves in ICC signify which items are harder and which items are more discriminating. With appropriate model fit, the ICC follows closely the actual test data. The IRT models with higher flexibility enables a close fit between a function and a data to be attained (Rusch et al., 2017).

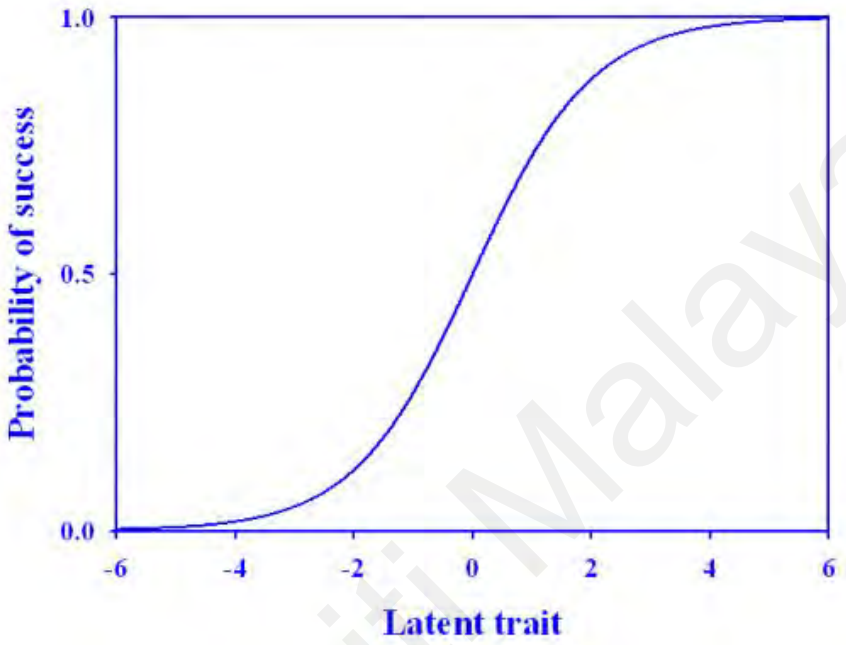

Figure 2.1 Item characteristic curve (ICC) showing the relationship between the location on the latent trait and the probability of answering the item correctly

Despite of IRT is a complex model due to its ability of modelling the outcomes at item level, it is very comprehensive in provides solutions in term of test performance. Many literature studies have reported similar findings regarding advantages of IRT over CTT (Courville, 2004; Embretson & Reise, 2000; Fan, 1998; Harvey & Hammer, 2016). The major advantage of IRT is that its facilitates the graphic evaluation of the item performance (Edelen & Reeve, 2007). Graphic presentation enables the test developer to understand the item properties immediately. Another advantage of IRT is its treatment of reliability and measurement error by item knowledge functions that are calculated for each item (Carlson & Von Davier, 2017;

Lord, 1980). The item knowledge feature takes into account all item parameters and demonstrates item measurement efficiency at various ability levels.

In general, IRT is able to provide information in-advance-of the classical test theory (CTT). IRT becomes a popular statistical approach as it can be employed for dichotomous or ordinal data due to the transformation of raw scores into an interval scale. This theory also a better strategy to be employed if the construct measured is presumed unidimensional due to the additional information provided in the item analysis. IRT analyses is able to generate reliable results with very few items, while CTT reliability varies depending on the number of items.

2.2.3 Rasch Model

The Rasch Model is a psychometric technique designed to improve the accuracy of a designed instrument, track the consistency of an instrument, and measure respondents' performance (Boone, 2016). The item complexity and person capability in the Rasch model are calculated in a logit scale (Runnels, 2012).

The Rasch model is an analytical measurement model that was developed by Georg Rasch in 1960s by taking into account the respondents' ability to answer questionnaires, assessments or instruments, and the complexity of each item (Rasch, 1980). Most experts classified the Rasch model as the simplest IRT model with strong measuring properties due to its one parameter (Afrassa, 2005). This conventional model is widely used because it is among one of the most efficient and suitable methods of assessing the abilities of students. In addition, the Rasch model is suitable for analyzing dichotomous scored data (Lake & Holster, 2016). An Eastern empirical study has shown that the Rasch Model is closely linked to Item Response Theory

(IRT), originating from a distinct set of fundamental postulates, and the most important element is objectivity (Sadia Mahzabin et al., 2015).

The Rasch model will establish and analyze measuring instruments (Rasch, 1980). Mostly, it has been used for educational testing. A unique aspect of the Rasch model is that it offers measurement that do not rely on the distribution of the person involved, as the data matches the Rasch model (Andrich, 1988). This causes invariant comparisons in both the latent and separate sample groups.

A feature of the Rasch model that is derived from the theory that a priori is developed based on the data (Andrich, 1985). In Rasch Measurement Theory, data are compared to the model, which is the opposite of conventional statistical modelling (classical test theory) to describe or explain data (Hagquist, 2001)

There are two main versions of the Rasch model: i) the dichotomous model and (ii) the polytomous model for more than two ordered categories. In the Rasch analysis, the data are examined against the Rasch model and the fit test which is a test of deviation from perfection. Thus, items can be useful for measuring even if they do not match the model perfectly (Bock & Jones, 1968).

The Rasch Model is used not only to assess students ' skills, but their attitudes and personal characteristics as well. Literature reported numerous studies in various fields have successfully developed and validate their instruments using the Rasch Model. For instance, Draugalis and Jackson (2004) used the Rasch Model to assess student and item performance and evaluate curriculum strengths and weakness using 65-item, multiple-choice examination while Azrilah Abd Aziz et al. (2008) used the Rasch Model to validate the construct of measurement instruments. In another significant studies, Nordin Abd. Razak, Ahmad Zamri Khairani, and Thien (2012) uses the Rasch Model to examine the quality of Mathematics test items and Siti Aminah

Osman, Syahdatul Isnain Naam, Othman Jaafar, and Wan Hamidon Wan Badaruzzaman (2012) uses the Rasch Model in measuring student performance in civil engineering design. In one hand, McCreary et al. (2013) used the Rasch model in psychometric analysis. While on the other hand, M. N. Rashidi, R. Ara Begum, Mokhtar, and Pereira (2014) used Rasch model to measure the weights of items.

Numerous measurement experts collectively agree that Rasch model is the only measurement model compatible with the 'objective measurement' idea due to its properties is necessary for objective measurement and its capability in generates a linear interval scale. This idea implies that a common metric is used to present the results regardless of what construct is being measured, or what measuring instrument is being used (Program Committee of the Institute for Objective Measurement, 2000). Unidimensionality is a prerequisite of objective measurement and linearity concept is one of the key ideas for understanding why Rasch is an essential tool for measurement experts. Rasch model only takes into account item difficulty as its parameter and assumes that all discrimination parameters are constant and equal to one (Rasch, 1960).

Rasch model is a probabilistic unidimensional model that measures a single latent trait, therefore two assumptions are required to be fulfilled beforehand; (a) the easier question the more likely for students to respond correctly and (b) the more capable of students, the more likely he will pass the questions compared to less able students (Hambleton, 1989; Henning, 2013; McNamara, 1996; Wright & Stone, 1979). In constructing tests using this model, items that do not follow these assumptions are often discarded. When designing tests using this model, objects that don't follow these assumptions are always discarded. The Rasch model is considered prescriptive because it prescribes the criteria and specifies that a test should follow to be considered

a good measuring tool (Runnels, 2012). One condition is testing one characteristic at a time (Bond & Fox, 2015). Theoretically, however, it is impractical to create a test evaluating only one trait (Wright & Stone, 1979). Rasch rewards this by using psychometric rather than psychological dimensional intervals (M. Wu & Adams, 2007). It is demonstrated in the data as a reflection of the underlying construct or dimension (McNamara, 1996).

In the survey context, due to Rasch models are psychometric models, it capable of guiding to prove the quality of items to enhance the validity of the survey instrument (Aziz, Masodi, & Zaharim, 2013). In Rasch models, the main latent trait measured for a person is usually the individuals' tendency to endorse given items. Likewise, a person with the main latent trait will always have a higher likelihood of endorsing any item than an individual a lesser amount of the latent trait. Estimated item's difficulties are also produced based on the tendency of each endorsed item. Items that are difficult to endorse will always have a lesser probability than the easy items.

Rasch model differs from another traditional approach because it offers a consistent and reputable repeatable measuring instrument instead of the 'best fit line' (Azrilah Abd Aziz et al., 2008). Rasch focuses on correctly designing the measuring instrument, rather than fitting the data to a measuring model with errors (Azrilah Abd Aziz et al., 2008; Mohd Saidfudin Masodi, Azrilah Abd Aziz, N. A. Rodzo'An, & Omar., 2010). This is well documented in the literature. An earlier study notes that Rasch's measurement is a solution to validity issues as the Rasch model provides very useful statistics analysis and offers great opportunities to test validity as well as facilitate and produce more efficient and reliable measures while enhancing confidence for users (Azrilah Abdul Aziz, Azlinah Mohamed, Noor Habibah Arshad, Sohaimi Zakaria, & Mohd Saidfudin Masodi, 2007; Bond & Fox, 2015). While

Zulkifli Mohd Nopiah et al. (2012) added that Rasch Model's predictive feature makes it capable of overcoming missing measurement data. As far as questionnaire is concerned, an earlier study has shown that studies to determine the validity and reliability of the instrument are very crucial in maintaining the instrument accuracy (Siti Rahayah Ariffin, Bishanani Omar, et al., 2010).

Numerous studies indicate that Rasch model was able to determine the relationship between a person's skill and an item difficulty on the same scale, where findings allow a high level of ability to answer questions with a lower level of difficulty (Bond & Fox, 2015; Rasch, 1980; Zamalia Mahmud, Nor Azura Md Ghani, & Rosli A. Rahim, 2013). For example, Zulkifli Mohd Nopiah, Mohd Haniff Osman, Noorhelyna Razali, and Izamarlina Asshaari (2010) have shown a significant finding when applied a dichotomous Rasch model using 0-1 scoring of multi-objective questions relevant to a linear algebra course at Universiti Kebangsaan Malaysia. Findings showed that the Rasch model is appropriate for assessing student ability and validity. In general, it can be concluded that the Rasch framework offers procedures for developing and revising social science measuring instruments and recording instrument properties (Boone, 2016).

The Rasch model locates the estimated values for the students' ability and item difficulty into the same interval, represented in units of logit (Θ) thus converting the result into a linear correlation (Rasch, 1960; Rozeha A.Rasyid, Azami Zaharim, & Mohd Saidfudin Masodi, 2007). The use of logit ruler is practical for evaluating specific results, such as students ' abilities and when validating a question, construct online (Zulkifli Mohd Nopiah et al., 2012). A number of empirical studies have reported a similar finding pertaining to item difficulty which is the degree of the item complexity is expressed by the distribution of the item over the logit scale; the higher

item is considered more difficult than the lower item (Azrilah Abd Aziz et al., 2008). The parameter of item difficulty is usually standardized to a mean. In addition, the objective of Rasch analysis is to test whether the data fit the model is assessed by whether the data response pattern matches the model's predicted theory model (Tennant & Conaghan, 2007).

Various studies stated that Rasch analysis is a statistical tool in assessing the psychometric properties of a questionnaire which are not evaluated through classical test theory techniques, e.g. how well an item performs in terms of its significance as a utility for measuring the underlying trait, the amount of the construct targeted by each question, the possible redundancy of an item as compared to other items in the scale, and the relevance of the response categories (Bond & Fox, 2015; Tesio, 2003; Wolfe & Smith, 2007) and can yield accurate findings even by using a small data set. The key reason for using Rasch 's analysis is that raw scores are non-linear, and variations between any two consecutive raw scores cannot be presumed to be equal (Wright, 1992). Rasch analysis was designed to improve the precision with which researchers develop instruments, track instrument quality, and compute student performance. Rasch analysis helps researchers to create alternate forms of measurement instruments, resulting in alterations in line with student growth and change. Azrilah Abdul Aziz et al. (2007) explained how Rasch analysis is more meaningful in supporting academic reports and enhancing student achievement the targeted outcomes by student classification hence better management of assessment.

The Rasch Model relies on two fundamental assumptions which is unidimensionality of the latent trait and local independence. Unidimensionality means that only one construct is assessed by the items in a measurement. The assumption of unidimensionality demands “the item function in unison and any non-random variance

in the data may be accounted for by the ability of the person and the difficulty of the item (Wale, 2013). Local independence means that the items are not interrelated when the latent characteristics or characteristics are regulated (McDonald, 1981). The presumption of local independence calls for the examinee's answer to an item not to affect his or her answer to some other item. The items should then provide an indication of the correct answer for another item (Alavi & Bordbar, 2017). Local independence is obtained by defining the complete latent trait space in the model.

The test measurement using Rasch model has various advantages (Wright, 1977). Firstly, the Rasch model will determine if the object is fit and recognize whether there is a bias. Second, its item calibration is not affected by sample ability, which means it is sample-free. Third, standard calibration error can be manipulated to test each item 's accuracy. Fourth, Rasch model will estimate and turn item difficulties from different samples into common scale. As a result, item banks can be automatically equated as a standard calibration shared by all items. Fifth, two people's ability may be correlated, even though they have no item in common by translating the ability figures into a common scale. This is called a test-free measure (Tinsley & Dawis, 1977). Sixthly, person-fit chi-square may be used to determine measurement accuracy. Finally, it encourages the creation and design of the best test and tailor-made test using the Rasch model.

In short, although it is easy to conclude from Rasch Model the success rate of researchers in analyzing the test items that depends on the level of ability and item difficulty level, there is a main limitation in its use (Engelhard, 2013). Rasch model is not a causal model but is a relational model therefore it is excellent at identifying relationships only (Royal, 2010). In general, the Rasch model provides valuable data analysis for the development, modification, evaluation and monitoring of valid

measurement instruments for industry, medicine and educational research communities. Furthermore, the Rasch model also offers invariant interval scales. Invariant in this sense means that the scale described by items can be controlled from time to time to define the latent trait in the same way (Boone & Scantlebury, 2006).

In the essence, Rasch analysis is the widely used technique to prevail over the deficiency of CTT due to its provide useful data that can be used with confidence in descriptive and parametric statistics as well as facilitating the development of instruments. In addition, Rasch analysis provides outcome indicators that provide researchers with meaningful guidance (Boone, 2016).

2.2.3.1 Item Difficulty and Person Estimate Ability

In Rasch Model, students' abilities and item difficulties are estimated together and placed on a numerical scale called logit (log odds ratio unit). The conceptualization of the ability (latent) continuum as a ruler is illustrated in Figure 2.2.

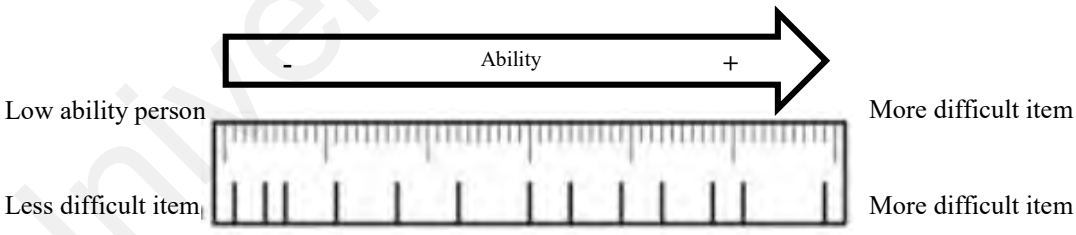

Figure 2.2 Conceptualization of the ability continuum
Source: Nevin et al. (2015)

Person-parameters and item-parameter are aligned at this scale where the probability of success is regularly determined at 0.5 which means the ability of a person to set at a point where he has a chance of 50 percent to succeed or fail (Bond & Fox, 2015; Hambleton et al., 1991). The logit scale is stated according to the scale of the interval where the mean and the standard deviation is arbitrary (Bond & Fox,

2015; Nunnally & Bernstein, 1994). Thus, the approximate item-parameter and person-parameter are about relative estimation, not an absolute measure.

Rasch analysis is comprises both mathematics and theory. Figure 2.3 shows a schematic diagram summarizing the Rasch model's fundamental mathematical and theoretical concepts. The single vertical line represents the construct to be measured by the test. Student and item measures placed along a vertical line. In this diagram (Figure 2.3), along the vertical line a notation regarding the level of ability of a student Anis is given along with the variable. There are three test items plotted along the variable. Each item is located in a position that indicates the difficulty or ease of each item with respect to the variable. The most importance is that each item along the variable has a probability of the respondent to answer correctly. However, item 3 exhibits the higher level of difficulty than the level of student ability.

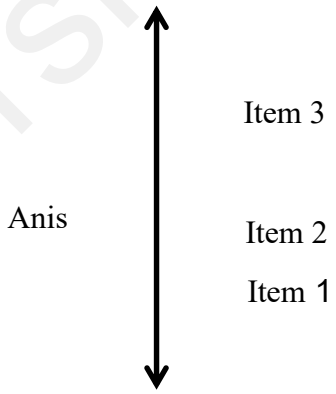

Figure 2.3 Rasch measurement schematic
Adapted: Boone (2016)

The mathematical model that describes the relationship between items and person is expressed by:

$$B_n - D_i = \ln(P_{ni}/(1-P_{ni}))$$

Figure 2.4 The dichotomous Rasch model
Source: Boone (2016)

Where:

B_n = the ability of a person along the variable

D_i = the difficulty of a test item

P_{ni} = the probability of the a person correctly answering a specific test item is solely related to his/her ability and the difficulty of the item being answered

$1-P_{ni}$ = the probability of a person incorrectly answering a test item

2.3 International Assessment

Since two decades ago, diagnostic assessment such as the Program for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS) have been employed as a direct comparison study of the quality of educational success across multiple systems by assessing various cognitive skills.

PISA is a worldwide assessment program and an international comparison study sponsored by the Organization for Economic Co-operation and Development (OECD) which is executed every three years and measures 15-year-olds student achievement in Mathematics, Science and Reading literacy. For the first time, Malaysia participated in PISA 2009 with the aim of evaluating the performance of the education system and quality of pupils developed by the system as well as comparing the existing educational system with other countries (Abdul Halim Abdullah, Johari Surif, & Ibrahim., 2014). In PISA 2015, 72 countries from around the world with approximately 540,000 students took part in studies that aim to determine the level of mastery and literacy in Mathematics, Science and Reading. Students are also assessed

in collaborative problem-solving skills and the ability of applying their knowledge in daily life. Malaysia took on the challenge by participating in computer-based assessments in PISA 2015 which is a daring shift from the traditional written assessment approaches.

TIMSS is conducted by the International Association for the Evaluation of Educational Achievement (IEA). The study offered by IEA is TIMSS Grade 4, TIMSS Grade 8 and TIMSS Advanced and administered with a four-year cycle. In addition to assessing the level of mathematical knowledge and science among students, this assessment also involves evaluating learning contexts. The assessment framework covers two aspects, namely content knowledge such as algebra and geometry, and cognitive skills such as knowledge-driven processes, applications and reasoning.

Students, teachers and schools are questioned regarding the environmental aspects of learning content that are taught, practiced and applied. The TIMSS result provides valuable information resources for each country on factors affecting the success of mathematics and science to enable the nation to improve their education system. However, Malaysia chose to join TIMSS Grade 8 only which is equivalent to Form 2 and the first study participated was on TIMSS 1999. The main objective of Malaysia's participation in TIMSS is to evaluate the effectiveness of mathematics and science learning among students compared to their peers from other countries. Research findings provide inputs for curriculum improvement, learning and teaching approaches, and school and national assessments. Malaysia has participated in the TIMSS 2015 cycle which began in 2013 and ends in December 2016.

2.4 Standardized Test

In Malaysia, there are various types of standardized assessment such as achievement assessments, scholastic aptitude and intelligence assessments, specific aptitude assessments and school readiness assessment. The standardized assessment allows for comparisons to be made among schools in regards to student achievement and ensures accountability for educators as well as informs instruction. In the essence, standardized assessment is mostly used to measure student achievement as well as predicting their achievements in actual exams. Although, many standardized test are used, but high-stakes achievement tests have caused many controversies among students, educators and the school administers whether this standardized test can really predict student achievement in schools.

Standardized test is a common feature of education system in many countries including Malaysia. The test has been used since decades ago to reports student achievement. A seminal study has defined the standardized test as a test administered under a standardized and controlled condition that determines where, how, and how long a student may responses to questions or "prompt" (Goodwin & Driscoll, 1980) whereas Popham (2020) stated standardized test as any administered test, scored and interpreted in a standard and predictable state (a pre-determined manner).

There are various types of tests and assessments that can be "standardized" but this term is more closely related to the administration of large scale tests given to large student populations. This standardized test is also known as the norm-referenced test because it is a type of formal test that has procedures with specific guidelines for administration, scores and interpretation of results.

The usage of standardized test has been identified in many discipline especially in education. For example, in the United States, standardized tests such as SAT are

used as an admission test for university student intake. Standardized assessments are used to make informed high-stakes decisions, including decisions that reduce access to education program, career pathways and other beneficial opportunities. Ideally, test scores collected complement other relevant information from sources such as interviews, feedback, findings, and job portfolios. Uniform test results are considered to be relevant in most cases as they meet the same assessment objectives in terms of information, achievement and skills. However, if the information given is incorrect or has an undue impact on this judgement, the standard test score may be violated. Some research and reports relate to deficiencies in the standardized test and undue dependence on scores in high -stakes decision-making (Santelices & Wilson, 2010). If the test is developed correctly and used responsibly, the standardized test will help students measure their progress and help people assess the effectiveness of a school. But if the test improperly designed, it may end up measuring wrong thing.

Table 2.3
Advantage and disadvantage of standardized testing

Advantage	**Disadvantage**
Has a positive impact on student achievement	Has a negative impact on student education
A reliable and objective measurement	Can be predictable
Allows for equal and equivalent content for all students	Assume that all students start from the same point of understanding
Teaches student prioritization	Only look at raw comprehension data Teacher evaluation have been tied to standardized test results Narrows the curriculum More time is spent on test preparation instead of actual learning

Source: https://vittana.org/12-advantages-and-disadvantages-of-standardized-testing

2.5 Chemistry

Chemistry curriculum usually incorporate many abstract concepts, which become the center of further learning in both chemistry and other sciences including Mathematics (Taber, 2002). This curriculum allows students to understand what is happening and how it contributes to the quality of life on the planet (Ware, 2001). Nevertheless, findings from a number of past studies have proved that students who have poor Mathematics skills and lack of basic knowledge as well as weak in mastering the concepts involved encountered difficulties in learning Chemistry (Aziz Nordin, 2007; M. J. Grove & Pugh, 2015; Ranga, 2018; Taber, 2019).

In addition, due to the vast amount of information in Chemistry, students lack the time to learn all the concepts (Edomwonyi-Otu & Abaraham, 2011; Jegede, 2007; Pollard & Triggs, 2000; Ward, Roden, Hewlett, & Foreman, 2005). These conceptual learning are essential in chemistry to understand the further concepts and theories as it forms the basis knowledge of Chemistry (Coll & Treagust, 2003; Nakhleh, 1992). A result from a study conducted in Kolej Matrikulasi Negeri Sembilan reports there are negative effects on the conceptual learning where students inclined to memorize what they have learned and emphasized in algorithm instead of exploring the topics and understand the concept that have been taught (Dani Asmadi Ibrahim et al., 2015).

The Chemistry curriculum in Malaysia aims to produce active students by giving them ample opportunities to participate in scientific studies through hands-on and experimental experiences. Due to its importance, the contents and contexts suggested are selected and students are encouraged to increase their interest in the subject. There are five themes with nine learning areas in the curriculum that need to be studied by the form four of the science stream students (Appendix XIV).

2.6 Multiple-Choice Item

The tests need to be systematically and clearly constructed so that a valid and reliable score can be obtained. The purpose of the test conducted is to measure the type of response from the students (Veloo, 2011). There are two types of test, objective test and subjective test. Multiple-choice questions which consist multiple-choice item are one of the popular objective test forms and widely used in assessing student achievement either at the school level or institutions of higher learning.

Teachers, schools, and assessment organizations typically use multiple-choice questions because they are quick, simple, and computer scorable. It can also be fairly graded and thus may give the test appearance of being more reliable than subjectively scored tests (Bailey, 1998). This test form are very suitable assessment tools to determine the level of knowledge of many students at various academic levels in the different subjects (Burton, Sudweeks, Merril, & Wood, 1991). Multiple choices often enable students to discover their misconceptions by using inaccuracies in the choices (Treagust, 1988). Hence, this test form is one of the form that been used by Malaysian Examination Syndicate (MES) to construct SPM Chemistry paper besides subjective test.

Many assessment experts agreed multiple-choices are one of the conventional evaluation instruments most common and widely used besides true or false tests, short answer and essays for measuring simple and complex learning outcomes which is knowledge, understanding and application (Ben-Simon, Budescu, & Nevo, 2016; Waugh & Gronlund, 2013). In addition, it enables the practitioner to evaluate the achievement thoroughly and easily by providing multiple questions in a short period of time (Burton et al., 1991; Treagust, 1988). This type of item is used extensively in achievement testing due to its flexibility and high quality items that adaptable to most

of subject-matter content. In term of assessment method, Scouller (1998) discovered that students were more likely, within a multi-choice test setting, to use surface learning methods and perceived multi-choice tests as measurement of knowledge based or lower intellectual processing levels.

A typical multiple-choice item is a type of item which consists of two distinct parts: a stem that poses a situation of problem and a sequence of responses that, in turn, consists of several possible solutions to the problem. The stem can be a question or an incomplete statement. The alternatives include the correct answer which is called key or key alternative and several plausible responses called distractor.

Distractors are structured to distinguish students who have studied from those who have not. Useful distractors will usually cover documented misunderstandings encountered by previous students and factual errors common to the instructor and should have a student response value for each distractor at least 20-30 percent. It is a waste of time for academics and students to apply distractors to an object with very limited response values for students. Often these types of items are called selection items since people need to assess each choice and choose the best response. Technically speaking, matching items, true-to-false items and a number of other specific item styles with correct answers and students selecting the correct answer are all multiple questions. Thus, in this report the researcher addresses only those questions that involve a stem followed by a sequence of answers to this stem. This is generally the format considered to be a traditional question of multiple choice and often known as an analytical question.

In multiple-choice questions, a student's test score is normally calculated by counting the number of correct answers in the test and used for assessing the knowledge of the person on the materials and contents covered by the test.

Nonetheless, the overall total test score achieved which consists of two numbers: the number of questions the student who knows the answer and the number of questions in which the student guesses the answer correctly due to the partial knowledge or uncertainty (Ng & Chan, 2009). A previous research has reported that objective questions are called objectives because the marking are objective and even checking works and scoring are also easy to implement (Mohamad Fauzi Yunus, 1996). In addition to that, the test has a consistent score fidelity and implausible answer to be checked. This means that although a test is examined by many teachers, the monitoring consistency can be maintained and secured (Bhasah Abu Bakar, 2003).

Table 2.4
The Advantages and Disadvantages of Multiple-Choice Questions

Advantages of Multiple-Choice Items	**Disadvantages of Multiple-Choice Items**
Can be used to measure learning outcomes at almost any level	Writing a good item is difficult and time-consuming especially on higher order thinking skills
Easy to understand (if well written)	Amount of reading is a barrier
Writing is not a barrier	Limit creativity
Minimize guessing	May have more than one correct answer
Easy to score	Does not evaluate performance
Marking is objective	
Well-constructed items can be used for determining misconceptions	
Quick and easy to administer	
Assessment cover a broad range of concepts	
Inter-marker reliability is maximised	
Can be easily analyzed for its effectiveness	

Source: http://in.sagepub.com/upm-data/45668_8.pdf

The empirical study conducted showed that scholars are less concerned with achievement test items especially for multiple-choice items (Cronbach & Furby, 1970; T. M. Haladyna & Downing, 1989; T. M. Haladyna et al., 2002). Multiple choice items are often criticized for relying on what student recall and for not assessing the ability of students to apply and evaluate course knowledge (Seldomridge & Walsh, 2006). The format of multiple choice items makes it possible for students to guess the answers even if they do not have substantial knowledge of the subject under consideration (Biggs, 2006). However, blind guessing for a well-written test is very rare and this provides a valid measure of student achievement (S. M. Downing, 2003).

Multiple-choice items allow for profound analysis of the item whether the item could discriminate between persons with the high ability and low ability. This can be achieved by completing an item review that involves creating a difficulty index and an index of discrimination for each item. The rationale of an item analysis is to determine how well an item differentiates between a high ability and a low ability person. However, the most common mistakes that occurred in the analysis of items is when the individual evaluates the whole test and not every item. The aim of this item review is to gain input on how well each item is updated to make it a better item in the future.

2.7 Item Analysis

In the field of measurement and psychology, item analysis is often underestimated and sometimes neglected although it is a very important process in determining the quality of an item or test which can be used in the future. Item analysis is an important method in education assessment. Earlier studies have reported that item analysis as the systematic process of examining students' response to make decision about each item

by using statistical techniques for consideration of the quality of the items, focusing on the objective items (Brookhart & Nitko, 2019; Lange et al., 2005).

An empirical study defined item analysis as a process of identifying and selecting items that function after items are pre-tested to find out the statistics for each item. Selected items are collected and stored in the item bank (Bhasah Abu Bakar, 2003). In essence, item analysis is the method by which student's answers to individual test items (qualities) are evaluated to determine the content of these items and the whole test by their internal accuracy or validity of an internal structure, concentrated on verifying a single or one-trait test.

Item analysis uses statistics and expert evaluations to measure tests based on the quality of individual items, the sets of items and the entire set of items, along with the relationship of each item to another item. Regarding the value of each item, it is assessed by comparing student's item responses to the overall test scores. A previous seminal study has indicated that item analysis "researches and uses this knowledge to enhance item quality and measure the output of the items considered individually either in relation to certain external parameters or in relation to other test items" (Thompson & Levitov, 1985). For norm-referenced and criterion-referenced tests, the principles of item analysis are identical, but vary in particular ways. The item analysis may be used for scored (right or wrong) items dichotomously and for scored polytomous (more than two score categories) items. Item analysis is not restricted to quantitative but also qualitative analysis (Popham, 2020).

Besides improving item and test quality, item analysis also used to remove vague or confusing elements in a single test management framework and improve the expertise of the instructor in test construction and to recognize particular areas of contents that need greater focus or clarification.

2.8 Past Studies on Item Analysis Using Rasch Model

Measurement experts have used the Rasch Model in their studies to analyze the psychometric properties of the instrument due to its objectivity and comprehensive analysis output. Arnold, Boone, Kremer, and Mayer (2018) used an open-ended response format Scientific Inquiry Competence (SIC) instrument to demonstrate the strength of using Rasch approach in conducting psychometric analysis of an instrument. The authors stated that, the raw data is not assumed as linear in Rasch computation compared to traditional analysis. Therefore, the data wouldn't be flawed. The Rasch analysis assists researchers in improving the quality of the instrument by allowing them to optimize the instrument.

The majority of chemistry educators' instruments are relatively new (Arjoon, Xu, & Lewis, 2013). Hence, gathering and reporting the validity and reliability are crucial due to the interpretations that are made from the raw score. It is very important for the chemistry educators to comprehend the significance of the quality of the assessment tools employed to produce a comprehensive report.

Rasch analysis is used for examining the unidimensionality, item functioning, item difficulty and person reliability which assists in validating a Chemistry multiple-choice question in order to measure students' ability (Yee, Fah, & Ling, 2018). A comparable study (Winarti & Mubarak, 2019) used Rasch analysis to review the student's progress in the learning process and as a guideline for designing the chemical strategies. They concluded that dichotomous items are worthy of being used as assessment tool of cognitive ability.

The Rasch model has been utilized to evaluate content knowledge and the topic specific pedagogical content knowledge and quantify the relationship for both variables in organic chemistry by providing empirical evidence. It was reported that

both tests met the performance of the good test design and the correlation between both tests is strong (Davidowitz & Potgieter, 2016). Furthermore, the Rasch analysis is able to establish the psychometric properties of the concepts inventories (Nedungadi, Paek, & Brown, 2019). The authors note that Rasch analysis can also be used as an information source in the creation of concept inventories. These studies concluded that Rasch provides robust information on the instrument and its the functionality.

The Rasch model nonetheless has been used as a diagnostic tool to investigate the misconception about chemistry by analyzing the pattern of responses for each answer choice (Herrmann-Abell & DeBoer, 2011). Researchers used the probability curves for each answer option to determine the validity and reliability of the items and analyze the probability of students that have a certain level of comprehension. This information about misconceptions of the items are most beneficial and can be very helpful for teachers in informing and improving instruction so that the instruction is more effective.

These studies have shown the value of using Rasch analysis to assess the efficiency of chemistry education assessment tools. Rasch analysis provided a lot of information about the feasibility of an item in measuring the level of student understanding (Winarti & Mubarak, 2019). This study therefore uses the Rasch model to test psychometric properties and to generate valuable results that reflect on the validity and reliability of the Chemistry test that comprises multiple-choice questions. The results of this study signified the unidimensionality of the test paper, local independence, validity, and reliability of the person, item difficulty, mapping person-item, and distractor analysis. An additional analysis result such as DIF able to strengthen the significant finding in determining the item criterion.

2.8.1 Validity on Chemistry Achievement Test (CAT)

A few studies conducted on Chemistry subject proved that it is a difficult subject compared to other Science subject as it incorporates many abstract concepts (Harris, 2003; Sirhan, 2007). Essentially, Chemistry is the centre of science because it connects physics with other natural sciences such as biology and geology (Veloo et al., 2015). Therefore, many researchers attracted to conduct studies on the Chemistry subject which is focused on the achievement test as the main question in chemistry subject is how to trigger students into what will be a new way of seeing and thinking. Achievement test is one of the measuring tool for teachers and students to analyze the success rate of the learning process. The analysis is indispensable to strengthen the process of teaching and learning. The performance test should also be accurate and reliable in order to assess the skill of the students.

Validity refers to the degree of proof as well as theory in the measurement sense that supports the interpretation of test results as the intended use of the test (AERA, APA, & NCME, 2014). Validation is simply a compilation of facts to provide the objective foundation for both the analysis of test scores and the ability of an instrument to measure a concept or construct matches for its proposed use. It is important to analyze how the instrument measures what it is meant to evaluate (validity) and the accuracy (reliability) of the test (Pae, 2011) when measuring Chemistry ability in the high-stakes achievement test.

Siti Salbiah Omar et al. (2016) in their studies mentioned that the content validity gave more focus on the suitability of the contents of the Chemistry curriculum of Malaysia. However, the authors focused on the content and linguistic validity only. In line with Siti Salbiah Omar et al. (2016), Espinosa (2014) added achievement test for Chemistry should be face and content validated as it affects the performance of the

students. Nevertheless, construct validity is disregarded in both studies even though this type validity has equivalent importance to content validity.

In the current study, the content validity is referred to Table of Specification (TOS) which is drafted by the experienced Chemistry teachers in order to make sure the items comply with the learning constructs and in line with the Form Four Syllabus. In contrast, construct validity is evaluated using Rasch model as it endorsed the probability to answer an item correctly by determining the difference between the latent trait of the individual and the difficulty of the item (Wei, Liu, Wang, & Wang, 2012). The author added that Rasch model principles are related to the Messickian construct-validity issues.

2.9 Conceptual Framework

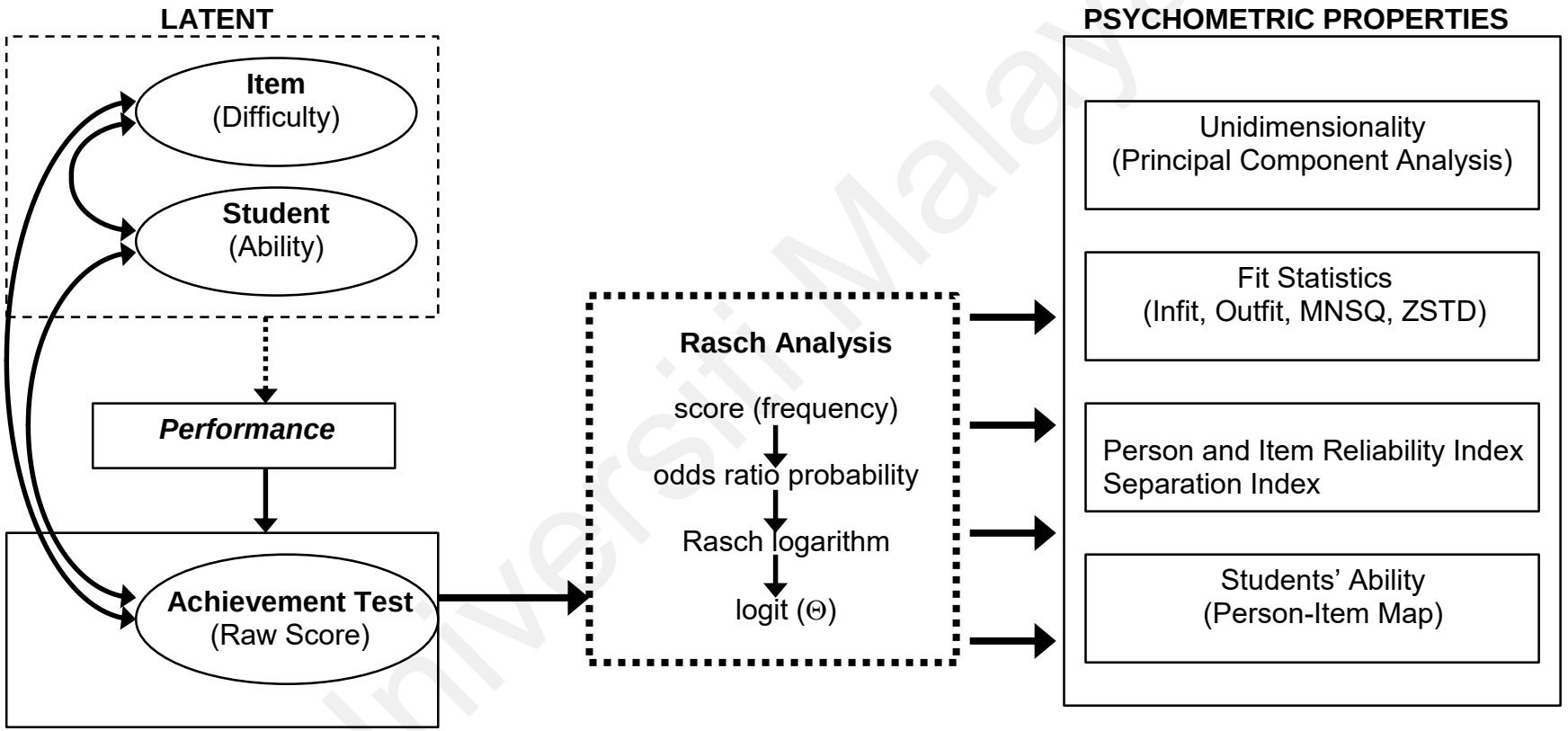

Figure 2.5 Conceptual-psychometric framework of achievement test
Source: Adapted from Eckes (2015)

Basically, item analysis was the main focus of the current study. The Rasch analysis of the student performance data is lay on a conceptual model of factors that influence the achievement test. Figure 2.5 depicts these factors and their mutual relationships. The left-hand side of Figure 2.5 comprised factors that have an immediate impact on the score awarded to the student. It shows two categories of variables that influence the student performance. Some of these factors may interact with one another.

The middle part of the diagram shows the transformation process of the raw score from the achievement test into the log-odds units or logits that occurs while running Rasch analysis. On the right-hand side of Figure 2.5, the diagram lists major types of output from a Rasch analysis of the achievement test. From Rasch's analysis, the unidimensionality of the test can be determined to warrant the test only measuring a single latent trait by using principal component analysis (PCA). In addition, person and item reliability index, separation index, fit statistics, the person-item map can be obtained for interpretation of each item and each student. Reliability of test items and students influence the validity and reliability of the achievement test.

Ideally, the Rasch model provides detailed insight into the functioning of each factor that is deemed relevant in the particular assessment context. Ultimately, the Rasch analysis is employed to determine the students' reliability and item reliability as well as the validity of the achievement test besides identifying the items difficulty and the students' ability (Bond & Fox, 2015).

2.10 Reliability

Reliability is a prerequisite for validity however it is not sufficient condition for validity. Typically, reliability can be described as to how consistent the instrument measures what it is meant to measure (Kimberlin & Winterstein, 2008). An earlier

study states reliability is the consistency, accuracy and stability of a test in measuring what is being tested. Accuracy and stability are the scores produced in a test obtained by uninterrupted students. Consistency means if students take the same test for the second time without any change in behavior, they will get the same score or almost the same as the first test (Azman Wan Chik, 1994). In order to scrutinize the validity and reliability of instruments, techniques such as factor analysis, item analysis and reliability analysis can be used (McCoach, Gable, & Madura, 2013) depending on the objective for which the test is used.

The reliability of a measurement scale linked with the consistency of the assessment results from time to time. Psychometricians have found that reliability of an instrument is refers to the consistency of the measurement; the extent to which the instrument can give the same score when used on the subject several times (Anastasi, 1988; Mehrens & Lehmann, 1984). However, on the other hand, reliability is defined as a measurement of an instrument's precision or accuracy (Radhakrishna, 2007). Nevertheless, in the evaluation sense, reliability refers to the degree to which test scores are error-free (Muijs, 2011).

Test scores are data sources that usually associate with achievement and psychometric tests. Some of the measures taken from data sources are more objectives such as achievement tests because the reliability and validity of the indicators are known, with error margins and results reporting usually following stringent standards. Yet, most sources of data require a greater degree of subjectivity in assessment or other possible error sources. Adequate reliability depends upon a low magnitude of errors of measurement. Therefore, it is imperative to control the known sources of error and to report the reliability and validity of measurement used.

In contrast to Muijs (2011), the Rasch model's reliability is determined by its ability to identify the trait level (Bond & Fox, 2015). Furthermore, the reliability of the test is divided: a) the reliability of the test person and b) the reliability of the test item (Khine, 2020). Computer software for Rasch analysis, such as WINSTEPS, provided three fit indices for the reliability of the Rasch model: Cronbach's alpha, item, and person reliability and item and person separation (Sumintono & Widhiarso, 2015). Hence, the reliability of the Rasch model is capable of providing exhaustive test information.

However, due to CTT focuses on the overall test score, the reliability only captures the consistency of the scores obtained from the application of the instruments (Alagumalai et al., 2006). Generally speaking, the reliability of CTT only concerns the consistency of the measurement (Andrich & Marais, 2019).

In the present study, researcher is more focusing on internal consistency rather than stability. Internal consistency also referred to as the extent the items make the scale 'hang together' (Pallant, 2007). The internal consistency coefficient offers an approximation of measurement reliability, based on the assumption that items measuring the same construct can correspond.

Cronbach's alpha is a popular method used for measuring internal consistency that was developed by Lee Cronbach in 1951 using a scale of zero to one (Kimberlin & Winterstein, 2008; Tavakol & Dennick, 2011). It is an indicator of an instrument 's internal consistency in testing certain concepts (Jackson, 2006). The alpha value "describes to what level all items in a test measure the same principle or construct" and may help determine the amount of measurement error in the testing instrument. Cronbach's alpha is a function of the average inter correlations between items and the amount of elements used for summarized scales because of its high sensitivity to the

number of items. A high alpha value demonstrates a high degree of internal consistency between test items and allows the researcher to determine whether all the questions answer the same construct. However, a relatively low alpha Cronbach value is common when using a short scale. Ideally for the reliability of a test, most researchers have agreed that an alpha greater than 0.8 indicates a reliable instrument (DeVellis, 2003).

Person reliability is described as the likelihood of the respondent's consistency with respect to the correct response to each item as measured by its difficulties (Linacre, 2012). While, reliability of items is described as the effectiveness of the item sample sizes in a precise position on the latent variable (Linacre, 2012). Both person and item reliabilities for an instrument is measured and interpreted using the alpha benchmark (Fisher, 2007). In the more recent study, Cordier et al. (2018) clearly mentioned that person reliability is commensurate with the standard Cronbach's alpha and shows the internal accuracy of the measurement.

As for the reliability of item separation and reliability of entity separation can be measured by using Rasch model (Mappiasse, 2006; Wright & Masters, 1982). Estimating a test 's internal consistency reliability is dependent on reliability of the individual separation. Each person's logit scale estimate is used to measure reliability (Bhakta, Tennant, Horton, Lawton, & Andrich, 2005). The reliability of the item separation shows how well the reacting sample will distinguish the instrument items. Although reliability in separation offers information about how well individual items can differentiate subgroups within a sample.

Table 2.5
Reliability in Rasch Analysis

Statistics	Fit Indices	Interpretation
Cronbach's alpha (KR-20)	< 0.5	Low
	0.5 - 0.6	Moderate
	0.6 - 0.7	Good
	0.7 - 0.8	High
	> 0.8	Very High
Item and Person Reliability	< 0.67	Low
	0.67 – 0.80	Sufficient
	0.81 – 0.90	Good
	0.91 – 0.94	Very Good
	> 0.94	Excellent
Person Separation	>2	High separation value
Item Separation	>3	indicates that the instrument has a good quality since it can identify the group of persons and items

Source: Sumintono and Widhiarso (2015) and (Boone et al., 2014)

2.11 Validity

The most significant concept in assessment is validity. According to Baghaei (2008), the evaluation must be accurate and valid to make sure that correct assessment are conducted (Banta, 2007; Figlio & Lucas, 2004; Fuchs et al., 1999) and informative interpretation can be made (Wright & Mok, 2004). For educators, validity and scientific measurement are critical as they are seeking the valid output from assessment (Bond & Fox, 2015). A previous study has shown that validity is not influenced by reliability (Roseni Din et al., 2009). However, reliability is required for validity of an instrument. Ebel and Frisbie (1991) stated that there are two aspects need to be considered in term of validity which is what is measured and how consistently it is measured.

Traditionally, validity of the instrument is regarded as test characteristic yet validity is not an instrument property. In term of definition, numerous empirical studies have shown that validity is the the degree of the test's ability to assemble the information about the quality of the instrument (Mohamad Majid Konting, 2005; Muijs, 2011; Tavakol & Dennick, 2011).

Nevertheless, Standards for Educational and Psychological Testing (AERA et al., 2014) defined validity as "the degree to which evidence and theory support interpretations of test scores resulting from proposed testing uses" (Plake & Wise, 2014). While this definition seems transparent, it is rather broad, comprising a variety of evidence and theory. In line with AERA et al. (2014), Wright and Stone (1999) explained in general, validity is a declaration of test conformity and its elements, truth of test results and interpretation. While measurement experts believed that validity is an evaluation of the adequacy, accuracy and appropriateness of the interpretations and assessment result usage that suit the purpose of the particular test (Miller et al., 2013; Reynolds, Livingston, & Willson, 2009).

Nuttall (1987) described validity as the accumulation of proof warranting clear analysis of test findings. Several psychometricians extend the definition by focusing on the school assessment to which the results of the assessment that been carried out by the teacher can be used as inferences to predict the student learning process (D. Cohen & Rhydderch, 2010; McMillan, 2011). Essentially, in measurement, validity is refers to as the validity of a score-based inference and not the test-based inferences as the test does not possess validity (Popham, 2020). Achievement tests are used to make inferences about the student's status which lead to one inference only. Technically, validity relies on the inferences from educators Therefore, it is inaccurate to infer about the validity of a test. If the test does not measure what it should measure, then the test

cannot be used because the interpretation made on it is invalid or irrelevant. As for a well-constructed test, if administered with the wrong group of students or under unsuitable circumstances can also lead to the invalid inferences.

Messick (1995) in his study stated that validation is a process of evaluation of meaning empirically and measurement ramification. This validation process represents a predominant stage due to the strong justifications as it uses score interpretation as a reference to make decisions (Bachman & Palmer, 1996).

According to Frankel, Wallen, and Hyun (2011), all the interpretations and findings gathered in order to support an assessment objective and validity proof is typically divided into three major validity areas according to their techniques; content, criterion and construct validity. In general, content validity emphasizes on the research material and testing procedures. Criterion validity depends on the same target 's external measures. In contrast with content and criterion validity, construct validity concentrates on the underlying function involves relationships with other measures. Though, in the present study, validity of a year-end Chemistry Paper 1 only focuses on content validity and construct validity.

2.11.1 Content Validity

Creswell and Guetterman (2018) described validity of content as the degree to which the items on the instrument and the scores from those items reflect all possible items that may be asked about content or ability that is being measured. The validity of content is a key component of construct validity because it decides the suitability of the items for construct operationalization and provide appropriate and representative sample of all things capable of calculating the construct in the instrument. Nunnally and Bernstein (1994) stated that content validity is assessed by subject matter experts

that are able to differentiate between a theoretical definition and an empirical measurement. The objective of the content validity is to ensure that the information contained within the measure is reflective in the content of the assessment.

According to Runnels (2012), the evaluation items associated with the responses to these items should be appropriate and descriptive of the domain to be assessed. Kimberlin and Winterstein (2008) added that there is no statistical test required as content validity typically depends on expert judgement in the relevant area. The material validation of an evaluation method inevitably requires validation and often refinement of the targeted construct (G. T. Smith & McCarthy, 1995). The objective of validating content is to ensure the constructed items are capable of capturing the measure of specific learning appropriately (Anastasi, 1988; Noraini Idris, 2010; Sidek Mohd Noah, 2003).

In the education field, Table of Specification (TOS) is widely used for content validation because it helps teachers frame the decision-making process of test creation and strengthen the validity of teacher assessments based on tests constructed (Fives & DiDonato-Barnes, 2013). Literature describe TOS as a table that assists teachers align objectives, instruction, and assessment (Notar, Zuelke, Wilson, & Yunker, 2004). Furthermore, TOS serves as a tool to ensure that testing or evaluation tests the material and thinking skills the test aims to assess (Fives & Barnes, 2018).

2.11.2 Construct Validity

Construct validity can be defined as the extent to which the measures used in the study assesses the construct in instruments. According to Picardi and Masick (2014), this type of validity ensures that the construct measured is appropriately operationalized to assess the construct in the instrument. However, in an earlier study,

Messick (1989) referred construct validity as the evaluation of the appropriateness of interpretations and usage of the results of the assessment, based on the empirical evidence and theory. Crocker and Algina (2008) emphasize the requirement of the relationship of the items that being evaluated to be examined with the construct measured by the instrument while examining the construct validity.

The Rasch model proposes a robust analysis method of the internal construct validity of the result of assessment (Hendriks, Fyfe, Styles, Skinner, & Merriman, 2012; Tennant & Conaghan, 2007). Therefore, the Rasch model was employed to pre-validate the construct of the instrument. The output analysis provided by Rasch model are statistics which is very useful at item level such as a person-item map, fit statistics and factor analysis residual (Bond & Fox, 2015; Rasch, 1980). Furthermore, the validity of the instrument using the Rasch Model could be established on the basis of an analysis from the misfit order of the items.

According to Sumintono and Widhiarso (2015), item fit allows the researcher to decide whether the item usually functions in performing the alleged measurements and to evaluate the suitability of the item. Misfit items indicate that students have a misconception regarding the items. Boone et al. (2014) and Bond and Fox (2015) unanimously agreed that three criteria can be used to assessing the item fit which are Outfit Mean Square Values (MNSQ), Outfit Z-Standardized Values (ZSTD) and Point Measure Correlation (PTMEA-Corr.). Sumintono and Widhiarso (2015) asserts that items that fail to satisfy these three requirements must be enhanced or changed to ensure the quality and suitability of the item.

Table 2.6
Fit Indices for Item Fit

Statistics	**Fit Indices**
Outfit mean square values (MNSQ)	0.50 – 1.50
Outfit z-standardized values (ZSTD)	-2.00 – 2.00
Point Measure Correlation (PTMEA – CORR)	0.40 – 0.85

Source: Boone et al. (2014)

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter delineates the various aspects of the research methodology involving the design of the research, the selection of the sample, instrumentation, procedures of the data collection and statistical analysis performed to answer the research questions. The reliability and validity of the research instrument is addressed.

3.2 Research Design

Many earlier studies have shown that quantitative research began from physical sciences, primarily chemistry and physics (Creswell & Guetterman, 2018). In this study, the researcher chooses a quantitative research with an instrument validation method. This design focuses on describing and explaining (Creswell & Guetterman, 2018) the present situations as well as establishing the relationship between the abilities of students and item difficulties. According to Leedy and Ormrod (2015), quantitative research is very specific in seeking explanations, predictions and develop generalizations that contribute to the existing theories. In quantitative research, a theoretical framework is usually presented as a model that incorporates the variables and the relationships between these variables. The model determines the variables that are evaluated through an empirical study. The standard format in the quantitative research design is for each student to be asked the same questions, ensuring that the overall data sample able to be analyzed fairly. The focus point on objectivity is referred to what enables the researcher to generalize the findings of a research study beyond the specific situation (eg. students) involved.

The numerical data obtained is used to objectively measure reality and draw logical conclusions (Williams, 2007). In this study, the reality is referred to the ability of the students. Furthermore, data can also be analyzed in a quantifiable manner using statistical techniques in order to support or rebut "alternate knowledge claims" (Creswell & Creswell, 2017). The researcher therefore employs mathematical models such as the Rasch Model, a data analysis method that is used to predict, explain and confirm the findings. In the essence, quantitative research rests on the numerical data collection and analysis for describing, explaining, predicting or controlling the variables and phenomena of interest (Gay, Mills, & Airasian, 2009).

3.3 Population

A population consists of all the subjects who are being studied which researcher seeks to draw an inference (Salkind, 2010). However, a review of the literature define population as the entire collection of all objects, subjects or members that comply with the specifications set of the studied group (Mertens, 2019). In more recent studies, Rohana (2004) added that every individual or object in a population may vary in many ways. As for the current study, the target population are the Form Four Pure Science students who will be sitting for the Sijil Pelajaran Malaysia examination (SPM) in 2021. This target population consists of 1552 Pure Science students from 25 secondary schools in the Petaling Utama district and were accessible.

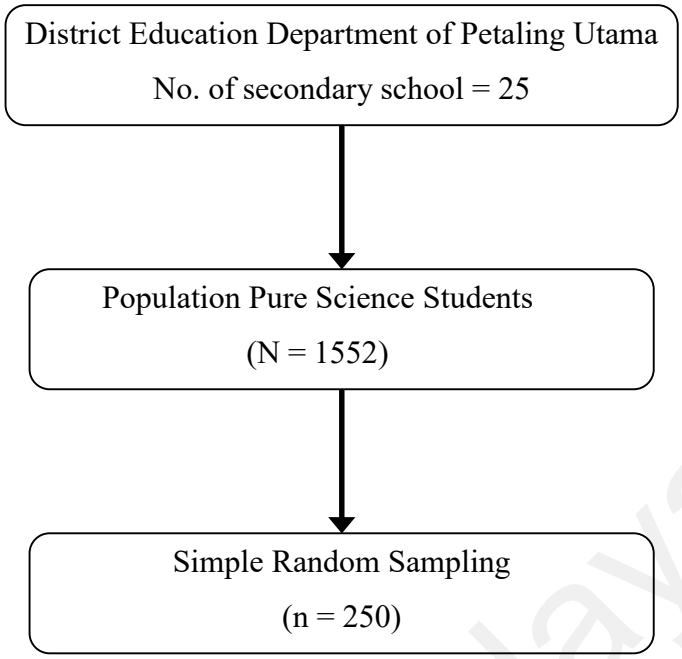

Figure 3.1 Population and Sampling of Pure Science Students in Petaling Utama District (2019)
Source: District Education Department of Petaling Utama (2019)

3.4 Sampling

In any study, the correct sampling selection is very important because it can affect the validity of the study (King, Rosopa, & Minium, 2018). In the present study, the researcher chooses simple random sampling as it can best address the issue under investigation. In addition, all secondary schools have same probability of being selected.

Simple random sampling is the most intuitive sampling approach due to its straightforward sampling strategy. It is the most popular technique to draw samples from the population. In simple random sampling, the probability of being selected as part of the sample is equal for every member of the population. There are two type of simple random sampling which are sampling with replacement and sampling without replacement. The selection is said to be non-replaced if a member of the population

cannot be selected more than once in the same sample. The members of the sample are usually selected consecutively. Each selection shall be made from a population other than those already selected. Therefore, the samples are not independent of each other. Simple random sampling is considered to be substituted if each selected sample member is replaced in the available population pool for subsequent sampling. In practice, replacement sampling is not as common as non-replacement sampling.

Various methods are employed to create a simple random sample, for example, a lottery technique, using a random number table or by computer. Researcher choose computer-aided random selection application which is more convenient compared to the traditional method due to the large population in the Petaling Utama district. In the present study, the researcher forms a sampling frame list that comprised of 25 secondary schools and uses a computer to generate random numbers. Spreadsheet packages such as Microsoft Excel is used to select the sample in the randomly ordered list from the sampling frame list and report the results (Lind, Marchal, & Wathen, 2018). The researcher uses the Excel's RAND () function to generate random numbers for each secondary school of the 25 schools in the sampling frame list. Then, the researcher sorts the list in increasing order of their corresponding random number, and selects the first 4 schools on that sorted list.

Gravetter and Forzano (2018) have stated that samples of random sampling are unbiased and representative if the sampling frame list is not subdivided or partitioned. Therefore, the inferences derived are most generalizable. Ideally, sample size of more than a few hundred is needed to be able to apply simple random sampling in an appropriate way (Saunders, Lewis, & Thornhill, 2019).

The sample size of the study plays an important role in producing meaningful results. According to (Uttley, 2019), the sample size would affect the sensitivity of the

study and its ability to reveal the reality about the sampled population. Normally, item response theory (IRT) models need broad samples to gain accurate and stable parameters, but the one-parameter (Rasch) model can be estimated with more moderate samples. Cappelleri et al. (2014) note that there are several factors that need to be considered in sample size estimation even if there is no ultimate answer given. The type of response options is one of the factors that will affected the required sample size which is of a dichotomous type (right or wrong) in this study. In general, as the number of response classes increases, a huge sample size is needed.

Literature studies suggests sample sizes of at least 200 with a minimum of 20 items are needed for the one-parameter (Rasch) model for dichotomous items if the standard errors of item difficulty are proved to be in the range of 0.14 - 0.21 for this sample size (Reeve & Fayers, 2005; Wright & Stone, 1979). Yet, an empirical study has reported that the numbers of person and items are similar because Rasch Model is a blind model (Linacre, 1994). For example, if 30 items are administered to 30 persons, a stable measure is produced. Linacre (1994) asserted that the expected sample size as small as 30 respondents would be sufficient for a Rasch model with dichotomous items in terms of item difficulty calibration to be within one logit of a stable value with 95% confidence. Nevertheless, if sample were consisting only 2 or 3 respondents, results could be very unstable. Evidence from an earlier study indicates a negative finding on small sample that could lead to the deprivation of potentially valuable results and are comparable to a loss of power in the test used for analysis (Pituch & Stevens, 2015).

In the present study, the researcher uses the sample size of 250 as suggested by Linacre (1994) and Boone et al. (2014) that the estimated difficulty of the item is within a definitive range of its stable value which lead to 99% confidence level. Thus, researcher can make analytical generalizations about the population being studied. It

is crucial for a researcher to contemplate whether the size of the sample is adequate to ensure sufficient accuracy in making decisions on the findings with confidence. Nonetheless, large samples are expensive and time-consuming (Linacre, 1994).

3.5 Instrument of the Study

A year-end Chemistry Paper 1 examination was used to obtain the data for this study. This instrument was prepared in collaboration with Majlis Pengetua Malaysia (MPM) Semenanjung Malaysia and District Education Department (DEP) Petaling Utama. A group of test developers that comprised of 10 experienced Chemistry teachers who were appointed by the District Education Department Petaling Utama, constructed the Chemistry items according to the Test Specification Table (TOS) in the actual SPM Chemistry Paper 1 (2018) format. There were 50 multiple-choice questions from different topics with varying degrees of difficulty in this instrument that needed to be answered within one hour and fifteen minutes. The instrument that had been prepared was distributed in softcopy form to the schools for print out.

3.6 Data Collection

First of all, a written permission to conduct a study in the selected schools was obtained from the Education Planning and Research Division (EPRD) of the Ministry of Education. The researcher also needed to acquire permission from the State Education Department before conducting the study.

Next, the researcher would visit the selected schools with the permission of EPRD and state education department to conduct the study. Then, she would brief the school principals about the purpose of the study, the number of students who will be involved and the instrument used. At the end of the day, the researcher would meet the

Chemistry teachers who have been appointed by the school principals to collect all the answered question papers that contained the multiple choice items that have been answered by students before transferring it to the answer sheet. The administration of the Chemistry examination took place in May 2019. The researcher will return all the question papers of the students to schools for revision purpose in a week.

3.7 Data Analysis

There are various computer programs for conducting Rasch analysis such as WINSTEPS (Linacre, 2009), RUM2030 (Andrich, Lyne, Sheridan, & Luo, 2003), FACETS (Linacre, 2006), QUEST (Adams & Khoo, 1996) and ConQuest (Adams, Wu, Cloney, & Wilson, 2020) depending on the type of Rasch model used. However, the most extensively used software is WINSTEPS (Linacre, 2009), which is a Windows-based application. It applies the joint maximum likelihood (JML) method of parameters estimation to generate a respective interval scale logit scaled score. Apart from the standard Rasch person-item maps, persons and item measures and fit statistics, this software also offers a number of sophisticated analytical statistics, for instance, distractor analysis, principal component analysis of Rasch residuals and analysis of differential item functioning.

In contrast to WINSTEPS, the RUM2030 is a highly interactive program that offers comprehensive diagnostics using familiar point and click set-up features in both tabular and graphical forms. This program can be used for large scale assessments including vertical equating. The estimation in RUM2030 is on the basis of pairwise conditional maximum likelihood. Thus, the data is more complex.

As for FACETS which is a Windows-based software, it is suitable for data such as essay grading, portfolio assessment and other types of rated performances or paired

comparisons. It offers calibrations of response format structures, for example, rating scales or partial credit model. Results are displayed in tables as well as graphically. Although, FACETS is robust against many type of misfits and missing data, it is unsuitable to be used in this study as the type of data is dichotomous.

QUEST is the first program that is implemented for Rasch analysis. It offers a comprehensive analysis by incorporating Rasch measurement and a range of traditional analysis procedures. The results can be accessed through a variety of flexible and informative output. This program can be run in batch mode, interactive mode or a combination. Nevertheless, the data structure is quite complicated and researchers who intend to use this software need a license agreement.

ConQuest is the latest computer software for Rasch analysis and has been developed as an enchanced QUEST update. It is an effective and versatile program because its output provides a broad range of informative graphs, charts and variable maps. However, due to its complexity, many researchers are reluctant to use this software.

In comparison with all other software programs that are applicable for Rasch analysis, the researcher chooses the WINSTEPS due to the dichotomous data and the easiness of analysis handling. WINSTEPS is able to scrutinize the model requirements for unidimensionality and offers an impartial estimation of reliability. In addition, this software allows researchers to configure the analysis in a more familiar graphical design such as tables, files, plots and graphs. WINSTEPS offers a Wright map (person-item map) that locates the items and persons along a continuum.

For conducting Rasch analysis using the WINSTEPS software, firstly, the researcher needs to store the data to be analyzed in an Excel spreadsheet. Next, the researcher has to create a control file that uses a special Winsteps control language to

specify the model parameter, data structure, and output format. The data stored in the Excel file is annexed to the end of the control file. This control language is subsequently saved as a text file and then executed by the Winsteps program.

The unidimensionality of the year-end examination of the Chemistry Paper 1 can only be determined by conducting item analysis even though the Table of Specification (TOS) is used in constructing the items. Rasch analysis is performed as unidimensionality is a fundamental assumption underpinning Rasch model. It can be measured by point-biserial measure correlation, Rasch fit indicators and Rasch factor analysis. Principal component analysis (PCA) of residuals from Rasch analysis may also be applied to examine if there is only one dimension captured by the model (Boone et al., 2014; Linacre, 2015). The following conditions are suggested to establish whether the assumption of unidimensionality holds: (a) the variance explained by items should be higher by four times the first principal component in the residuals (Linacre, 2009); (b) the total variance explained by the first measures should exceed 50% (Bond & Fox, 2015; Linacre, 2015); (c) the eigenvalues of the residuals should not exceed 3 (Reckase, 2016); and (d) the unexplained variance by the first principal component in the residuals should be lower than 5% (Linacre, 2006). For a good calibration, Reckase (2016) suggested that the total percentage of variance explained by the first component must be at least 20% to claim the unidimensional assumption. However, in contrast with Reckase (2016), measurement experts argued that 40% or more of the total variance should be accounted for the first component (Bond & Fox, 2015; Carmines & Zeller, 1979; Linacre, 2003).

Next, the item fit statistics are examined. These item fit statistics are represented by non-weighted (outfit) and weighted (infit) mean square errors (MNSQ). The outfit MNSQ directly squares and averages standardized residuals; whereas the

infit MNSQ averages standardized residuals with weights. When items comply the expectations of the model, their outfit or infit MNSQ would have an expected value of 1. All items that comply with the assumption of model fit data should have the values of infit and outfit MNSQ within an acceptable range between 0.70 to 1.30 (Bond & Fox, 2001; Pesudovs, Garamendi, Keeves, & Elliott, 2003; Wright & Masters, 1982).

On the other hand, MNSQ values that are very close to zero indicate redundancy of items, therefore it is considered as overfitting while MNSQ values greater than 1 indicates too much ransom noise and are termed as 'mis-fitting' (Bond & Fox, 2015). A further range of values between 0.5 and 1.5, as advocated by Linacre (2005) for the MNSQ statistics, is usually suggested as a critical range for a productive of measurement. Items with an outfit or infit MNSQ that out of this range are deemed as misfit. These misfit items need to be set aside for modification or repair before being discarded (Linacre, 2005). Conversely, Ahmad Zamri Khairani and Nordin Abd. Razak (2015) state that these items need to be discarded from further analysis because they are only measuring 'noise' without any meaningful contribution of the intended construct. Unidimensionality is ensured and interval measures could be generated by removing misfit items.

Rasch analysis should be carried out continuously until all items comply with the model fit requirements. In terms of confirming a factor structure, Rasch model functions excellently for factor analysis. After 'mis-fitting' items are identified and eliminated from the instrument, the researcher has to re-run the analysis. This process is performed iteratively until no misfit item is observed. When deciding on a 'mis-fitting' item, every attempts should be made to ensure a person separation value exceed 2.00 even if one has to retain a misffiting item (Garamendi, Pesudovs, Stevens, & Elliott, 2006; Mallinson, Stelmack, & Velozo, 2004).

A person-item map (Wright Map) is used to illustrate the items arrangement according to the levels of difficulty and the location of the ability of the students as well as assess the strength and weakness of the instrument. In addition, it assists the researcher to compare the theory with what is observed in the data set.

Person separation index, item separation index, person reliability index, and item reliability index are employed to evaluate the reliability. For item reliability, the low reliability value indicates that the sample size is small to precisely locate the items on the latent variable (Linacre, 2012). While person reliability indicates whether the test is able to separate the students into different levels.

Separation is the signal-to-noise ratio in the data. Person separation is used to differentiate students into certain groups. Low person separation with the related sample indicates that the instrument is less sensitive to differentiate between high and low ability students. This exhibits the power of the items to distinguish between students. Literature studies suggest that the value for person separation index which is an index of 1.50 is acceptable, 2.00 is good and 3.00 is excellent (Boone & Noltemeyer, 2017; Duncan, Bode, Min Lai, & Perera, 2003).

In contrast with person separation, item separation is utilized to verify the item ranking. Low separation of the item indicates that the sample is inadequate to validate the item difficulty of the instrument. An item separation index value of 1.5 is needed for analysis at the person level and 2.5 is needed for groups analysis (Tennant & Conaghan, 2007).

Point-measure correlation coefficient (PTMEA Corr.) for each item is carried out to determine if these items move in the same direction with the construct. The positive range index shows that the items measured are comparable to the construct (Siti Rahayah Ariffin, Bishanani Omar, et al., 2010). While, a high PTMEA Corr.

implies that student's ability can be differentiated by an item. In contrast, a negative value or null suggests that the connection is disputed with the variable or construct for the item response or student (Linacre, 2003).

3.8 Summary

This chapter outlined the quantitative research method used in the present study. Besides that, the description of the population, sample, instrumentation as well as the method of collecting data are discussed in detail. The procedures for the statistical analysis of the data are also explained.

CHAPTER 4

RESULTS

4.1 Introduction

The primary objective of this study is to determine the psychometric properties of the year-end examination of the Form Four Chemistry Paper 1 of the selected schools in Petaling Utama district. Four research questions have been formulated that have been addressed in Chapter 1. The Winsteps® version 3.73 of the computer program (Linacre, 2012, 2015) has been employed to execute the Rasch analysis.

Two stages of Rasch analysis were carried out. The first stage of analysis was to examine the Rasch assumption on unidimensionality by using Principal Component Analysis (PCA). In this stage, the statistical results of PCA reveal that unidimensionality is held across the Chemistry test. However, this dimension is only reasonably acceptable. These findings depict that perhaps most of the students have a similar ability or the level of difficulty for certain items are equal.

While in the second stage of analysis, the appropriateness of Chemistry test items was measured against Rasch standard analysis. Numerous Rasch statistical output (e.g., fit statistics, person and item reliability indices, person and item separation indices, person-item map, and point-measure correlation index) were analyzed and interpreted. The thorough results of these statistical analyses are presented in several sub-sections.

Generally, in the second stage, the findings indicate that the Chemistry test had high reliability and was able to discriminate students into three groups according to their abilities. Based on the PTMEA Corr. result, all items were positively correlated among them and moving in one direction. Furthermore, the Wright map showed that

the Chemistry test items were well-distributed among pure science students. Nevertheless, item difficulty result indicated that majority of the items measured the students' abilities only at a certain range. As for distractor analysis, particularly all the distractors were working properly. A DIF analysis was performed to examine the student's responses by gender and the results designated that six items were identified as difficult to be answered by male students compared to female students.

4.2 Sampling Profile

The sample consisted of 435 Form Four Pure Science students who were retrieved from four randomly selected secondary schools in Petaling Utama district. All students sat for the final examination of Chemistry Paper 1 in the year 2019 as it is compulsory. This is because Chemistry is one of the elective subjects offered in Science Stream.

Table 4.1
Sampling Profile According to Gender

Gender	Total	Percentage (%)
Male	245	56.32
Female	190	43.68

4.3 First Stage of Analysis - Unidimensionality

In the present study, the researcher employed Rasch residual analysis to test the dimension of the instrument used. The first stage of analysis was based on Research Question 1 which involved the analyzing process of the final examination of Chemistry Paper 1. J. D. Brown (2000) in his study stated that construct validity was conventionally characterized as an experimental demonstration in which the test

measures the construct it asserts to be measuring. The test result is said to provide evidence of construct validity when it can differentiate the group with constructs and groups without the constructs. Construct validation can be addressed either by several perspectives (logical thinking) or an accumulation of pieces of evidence (empirical). In agreement with J. D. Brown (2000), Messick (1989) added that construct validity is an evaluation of the suitability of the interpretations and use of the assessment results based on a posteriori evidence and theories.

During the item development, constructs for all Chemistry test items were validated by the experts using logic thinking. However, Othman Talib (2013) in his study notes that content validation of test items, solely by a panel of experts is insufficient and needs to be supported with statistical analysis such as construct validity. Bond and Fox (2015) stressed that statistical analysis is vital in providing the information primarily on item fitness or item suitability against the Rasch model standard. In general, a summary of the fit statistics from the Rasch residual analysis offered additional evidence to support construct validity of the test items.

Rasch residual analysis is designed to satisfy the fundamental postulates of the Rasch Model which are unidimensionality and local independence. Sick (2011) that unidimensionality is a collection of items which is designed to measure a single construct. The instrument items should target only one dimension or trait at one time (Bond & Fox, 2015). In this study, the students' answers of the Chemistry Paper 1 test had to solely reflect the students' ability. This basic assumption of unidimensionality needs to be ascertained to ensure the data set is in one direction to form one pattern (Wright & Masters, 1982). It is crucial not to violate this assumption in order to benefit from the Rasch Model.

4.3.1 Research Question 1: To what extent does the data fit the Rasch Model?

Based on this research question, the researcher compared and mapped the Chemistry data set with the Rasch model linearly using a variety of residual analysis such as Principal Component Analysis (PCA) correlation matrix, factor loading, fit statistics and local independence. PCA correlation matrix is the most frequently used Rasch statistical analysis in identifying the presence of secondary dimensions of an instrument due to its robust explanations. There are few aspects in PCA that can be employed to elucidate the unidimensionality.

4.3.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) correlation matrix is an extension of Rasch fit analysis conducted by Sick (2011) to measure the dimensionality of the data set which in this study is to identify the ability of students that influences the pattern of response of students. When the data conforms to the Rasch model, statistical results indicate that items measure the intended unidimensional construct. Furthermore, the PCA is used to confirm whether a single dimension difficulty and ability is adequately accounted for in all the non-random variances in the data (Sick, 2011). The PCA is also use for defining the common variance which is unexplained by the Rasch model. When a predominant measure is found among the items which is unexplained by the Rasch model, it means that a second dimension has affected the test data. Unidimensionality can be established by several statistical analyses.

Reckase (2016) in his study recommended that an observed raw variance measure should be able to explain more than 20% of the variance to substantiate the unidimensionality assumption. In contrast with Reckase (2016), Rasch (1980) sets a

percentage of 40% or more for observed raw variance measure to be the indicator of a strong unidimensionality while 30% of variance as a moderate dimension.

The findings of this study showed that the observed raw variance measures were 27.7% that were more than 20% of the variance for all item (Table 4.2). This data indicated that the Rasch dimension for this instrument only explains 27.7% of the variance. Comparison between observed raw variance measure for the data and the raw variance of expectation model signifies that both variances were quite equivalent which was 27.8%. Thus, it could be concluded that the test dimension of the Chemistry test was moderate and acceptable. The low percentage of the raw variance explained may be due to small range of students' abilities or some items' difficulty level. In other words, similar abilities between the Pure Science students and equal difficulty level of the Chemistry test items could have produced a low percentage of explained raw variance.

According to the guidelines by measurement experts, a minimum of 3:1 ratio was suggested from the variance explained by a Rasch measure against the variance of the first principal components of residuals (Conrad et al., 2012; Embretson & Reise, 2000). The result of the Rasch analysis exhibits the first contrast in the residuals which explains 27.7% of the variance meets the criteria of unidimensionality (Reckase, 2016). Therefore, the ratio for the variance explained by measurement dimension compared to the variance of the first principal components of residuals was 9.23:1. This finding proved that the Chemistry test paper met the requirement of unidimensionality.

Unidimensionality was claimed when the eigenvalue of unexplained variance by the first contrast was less than or equivalent to 2.0 and the total variance is less than 5% and less than 10% of the unexplained variance (Fisher, 2007; Linacre, 2006; E. V.

Smith, Jr., 2002). An eigenvalue of 2.0 denotes that the residual factor had a strength of approximately two items, the lowest value deemed to be the existence of a second dimension (Linacre, 2006). A finding of a study by Raîche (2005) showed that items must be reviewed to ascertain the presence of the second dimension if the eigenvalue exceeded 2.0.

Linacre and Tennant (2009) however revised the critical value of the eigenvalue of 2.0 that had been used widely by many previous researchers. He argued convincingly that the eigenvalue below 2.0 indicated that the dimension of the data was less than the two items' strength. This indication exhibited that no matter how powerful the dimension might be diagnostically, it had little data strength. Linacre then asserted that the significance had to be lied on the strength of the factors and not on the magnitude of their eigenvalues. Therefore, he concluded by the general rule of thumb that the eigenvalues in the unexplained variance of a secondary dimension should have the strength of at least 3 items. If a factor had an eigenvalue of less than 3 (with a reasonable length test) then the test is unidimensional. These analyses supported the Rasch assumption of unidimensionality (Reeve et al., 2007).

The unexplained variance by the first contrast is also known as a level of noise. Fisher (2007) and Linacre (2006) recommended that the acceptance value of unexplained variance by the first contrast should not exceed 5% and less than 15% of the regulatory limit. Furthermore, Aziz et al. (2013) asserted that the value of unexplained variance by the first contrast that accounted for more than 15% indicated that there was too much noise for the instrument. In the present study, Rasch analysis demonstrated the level of noise is 3% compared to the variance. Hence, this statistical result is considered very good because it approximately shows that there are no residual factors to measure students' abilities.

Table 4.2 shows that the eigenvalue of the first contrast had a strength of 2 items out of 50 items. Linacre (2005) stated that a residual factor warrants 3 items or more to provide useful information to guarantee a test had distinctive dimensions. Therefore, this finding indicated that the existence of a second dimension was not evident. All statistical analysis of PCA signifies that the residual is random and sufficient unidimensionality is likely held (Linacre & Tennant, 2009; Raîche, 2005), across the entire Chemistry Paper 1 test instrument that is consistent with the assessment design.

Table 4.2
Principal Component Analysis (PCA) of Chemistry Data Set

Standardized residual variance (in Eigenvalue unit)		Observed (%)		Expected (%)
Total raw variance in observations	69.2	100		100
Raw variance explained by measures	19.2	27.7		27.8
Raw variance explained by persons	6.9	9.9		9.9
Raw variance explained by items	12.3	17.8		17.8
Raw unexplained variance (total)	50	72.3	100	
Unexplained variance in first contrast	2.1	3.0	4.2	
Unexplained variance in second contrast	1.8	2.6	3.6	
Unexplained variance in third contrast	1.6	2.4	3.3	

Factor loadings of all Chemistry items ranged from -0.01 to 0.45. Table 4.3 exhibits that items 40 and 27 had positive factor loadings that exceeded the factor loading benchmark of 0.40. The grouping of these two items is significant because it recommends that both items have a common meaning that differs from the Rasch measurement standard (Bond & Fox, 2015). Therefore, it can be summarized that a secondary dimension was exists in this instrument with only a small influence. Bond

and Fox (2015) stressed out that any item with factor loading ≥0.40 should be examined.

Table 4.3
Factor loading of Chemistry test items that signify multidimensionality

Contrast	Loading	Measure	Infit MnSq	Outfit MnSq	Item
1	**0.45**	0.55	1.24	1.24	**40**
1	**0.41**	− 0.39	1.2	1.30	**27**
1	0.39	1.57	1.25	1.36	45
1	0.35	0.68	1.09	1.11	20
1	0.32	0.35	1.24	1.29	18
1	0.30	0.27	1.22	1.36	39
1	0.29	0.25	1.15	1.17	28
1	0.19	− 0.81	1.05	1.03	2
1	0.19	0.85	1.12	1.19	4
1	0.17	1.78	1.17	1.40	16
1	0.15	1.33	1.03	1.07	44
1	0.12	2.12	1.09	1.31	38

Notes: Factor loading ≥ 0.40 are in boldface. The information presented is an excerpt from the complete table in Table 4.5.

4.3.1.2 Fit Statistics

Fit statistics is an alternative analytic method employed instead of Principal Component Analysis (PCA) for assessing psychometric unidimensionality. This method can determine the extent of the empirical data meets the Rasch model requirements (Bond & Fox, 2015; Boone & Scantlebury, 2006)With regards to the construct validity, the fit statistics result is able to show the fitness of the data to the intended construct measured (Bond & Fox, 2015).

Wright et al. (1994) and Bond and Fox (2015) recommended that the critical value of the mean square statistic for the measurement be 1.4 because it could indicate 40% more variation than the Rasch model predicted. Test items with a mean square statistic value ranging from 1.4 to 2.0 are possibly unproductive but are not degrading

to measure (Linacre, 2012). However, a mean square statistic value of items that exceeds 2.0 may distort or degrade the scale resulting in an inaccurate measurement (Wright et al., 1994).

In the present study, a summary of the fit statistics shown in Table 4.4 designates that the Chemistry Paper 1 test data were conformed to the Rasch model. The result of both infit and outfit statistics result was evidence of an ideal fit of the items in the Rasch model as the data set of items ranged from 0.70 to 1.30 (Wright et al., 1994).

4.3.1.3 Local Independence

Despite the unidimensionality, the second assumption of the Rasch model that needs to be fulfilled in Rasch statistical analysis is local independence. If this assumption is violated, any statistical analysis based on it would be misleading and flawed especially on the validity of the estimates that would lead to incorrect decisions (Baghaei, 2007). Researchers have agreed that local independence is automatically proven with evidence of unidimensionality assumption (Embretson & Reise, 2000; McDonald, 1981). In this study, local independence pertained to the students' answers of the test items. However, in theory, local independence is related to the correlations among the test items which entails the test items to be statistically independent (Zunita Maskor & Harun Baharudin, 2019) and the item responses are not swayed by one another (Gnaldi, 2013).

Linacre (2015) clearly has stated that local independence value is established by MNSQ value that is less than 0.7. He adds that when the residual correlation value of two items exceeded 0.7, only one item has to remain while the others should be eliminated. Nevertheless, the process of eliminating the item must be thorough to

ensure the content and construct are secured. In line with Linacre (2015), Sharifah Nurulhuda Tuan Mohd Yasin, Mohd Fauzi Mohd Yunus, and Izwah Ismail (2018) add that the relatively large residual correlation value either suggests that the pair of items has something more in common characteristics among each other or both combining several other dimensions that are shared and confusing. Briefly, the high residual correlation values between items indicate a breach of the local independence assumption which means that the items are dependent on each other. Hence, for local independence to be ideal, the value of an item response should not affect the response of another item when the underlying data has been taken into account (Boone, 2016).

The local independence result of 50 items of the Chemistry Paper 1 test was established by analyzing the standardized residual correlations (Appendix I). The finding of the present study indicated that the residual correlations for every pair of test items was less than 0.7 and this means that there was no violation of the local independence principal (Linacre, 2015). Therefore, it can be concluded that the Chemistry Paper 1 test items were locally independence and there is no redundancy of items in this test. In other words, local independence is held in the entire Chemistry Paper 1 test instrument.

4.4 Second Stage of Analysis - Appropriateness of the Chemistry Paper 1 Test

For any certifying examination or test, reliability is an essential prerequisite because an error of measurement and smaller reliability will affect the accuracy of discrimination. In the second stage of the Rasch analysis, the researcher used several meaningful indices such as fit statistics and reliability analysis, separation indices, item polarity, and person-item map to evaluate test reliability (E. V. Smith, Jr., 2001). The analysis interpretation from these indices are able to accommodate Research Questions

2, 3, 4, and 5. In this study, Rasch statistical analysis acts as a measurement tool in assessing the ability of students and item difficulty. Furthermore, Rasch analysis is able to highlight any problematic items, and by modification, these items will increase the precision of the measurement.

Despite reliability, validity is considered the utmost importance for any effective assessment. The suitability of the tests, the rationality and adequacy of the test scores can be explicate by utilizing validity (AERA et al., 2014; Messick, 1989; Zou, 2005). Briefly, the validity of a measurement can be described as how well it measures what it intends to measure (Azizi Yahaya, Peter Voo, Ismail Maakip, & Mohd Dahlan A Malek, 2017)

Besides reliability and validity, quality control is another aspect of test development that needs to be taken into account, especially for certifying examinations that are used to distinguish students. In Rasch statistical analysis, fit statistics could be regarded as a quality control tool to assess the validity of student's response patterns with those expected by the Rasch analysis (Boone & Noltemeyer, 2017; Houston, Kearney, & Savoldelli, 2006; Norhayati Mohd Noor, Fatin Imtithal Adnan, & Nor Akma Mat Junoh, 2020; Wright & Stone, 1999).

4.4.1 Research Question 2: What are the students' reliability and item reliability of the year-end examination of the Form Four Chemistry Paper 1?

4.4.1.1 Fit statistics and reliability analysis

Rasch Model offers two statistical indicators with standardized residuals for evaluating the fitness of the model which are infit and outfit mean square indices. Infit and outfit mean square fit statistics present a summary of Rasch residuals

for each item and person where their responses are different from those expected in the Rasch model. The infit statistic is more about the overall pattern, whether persons and items comply with the expected model of higher ability means answering more difficult items correctly, and lower ability makes this less likely. Furthermore, infit statistics are less sensitive to outlier effects (Linacre, 2002) but, is sensitive to unforeseen responses to items close with the ability level of the person (McCreary et al., 2013).

In contrast, the presence of outlier is sensitive to the outfit statistic (Boone et al., 2014; Brinthaupt & Kang, 2014). The outfit also acts as an indicator of outlier whether items may too difficult or too easy (Linacre, 2002). Outfit statistic also discusses the difference between observed and expected responses irrespective of the extent of the endorsability of the item from the ability of a person (McCreary et al., 2013). The expected fit values indicating the ideal fit is 1.0 while fit values above 1.5 indicate poor fit (Linacre, 1998). Fit statistics that range from 0.5 to 1.5 indicate a good fit and productive for measurement (Linacre, 2002; Wright et al., 1994). However, fit values that are below 0.5 may produce misleading of good reliabilities and separation (Wright et al., 1994).

Table 4.4 represents that the overall mean infit and outfit was 1.00 and 0.98, accordingly. The mean standardized infit and outfit is 0.0, respectively. When the mean square fit statistics are high, a significant number of unforeseen responses are indicated. This maybe because of the flawed design of items such as obscure wording, have more than one key answer etc. or it could imply that a different construct is measured by the items.

The reliability of the Chemistry Paper 1 test was 0.87 which was near to 1.0. This denotes a high level of confidence in the placement of person (Bond & Fox,

2015). Furthermore, this data exhibits the notion of replicating persons along with the construct within measurement error when similar items are given to measure the same abilities. This means that the Chemistry test had sufficient reliability of the test score and the students answered the Chemistry test earnestly. The acceptable reliability index shows the stability and internal consistency of the Chemistry Paper 1 test. It also shows the pattern of the responses of students. The high person means square values indicate the students who have randomly filled in the responses or have unusual knowledge gaps. They also may belong to a demographic group that systematically responds differently to certain items.

The separation index supporting the notion of a logit interval scale in separating items and persons. Yet, the item separation index can be used as a construct validity index while the person separation index represents criterion validity. A high separation index indicates that an item or person is subject to adequate discrimination. Green and Frantom (2002) and Krishnan and Noraini Idris (2014) agreed that data are widely spread in terms of range if the separation index is greater than 1.

The results from the analysis denotes person separation index was 2.54 (Table 4.4). This index implies that the abilities of student varied well and the Chemistry Paper 1 test reliably separated Pure Science students into at least 3 statistically different ability groups (Bond & Fox, 2015; Sumintono & Widhiarso, 2014). In line with De Ayala (2013), the separation index represents how good the test can discriminate students according to their ability. In addition, Linacre (2005) and Siti Rahayah Ariffin (2008)) note that the separation is considered to be well spread and the item's position has high reliability when the separation index exceeds the minimum value of 2. Sumintono and Widhiarso (2014) reiterated that a higher separation index produces a higher quality instrument. Thus, it is concluded that discrimination of

students can determine the level of ability with the item difficulties (Bond & Fox, 2015; Wright & Masters, 1982).

The item separation index was 8.95, indicating very reliable item difficulty estimation and good variability (Table 4.4). This index also denotes that the items in the Chemistry Paper 1 could be separated into 9 groups according to the answers of the students. Table 4.4 demonstrates the item reliability index was 0.99, indicating that the Chemistry test items were fairly well-distributed across the logit interval scale, indicating an adequate breath of position on the linear continuum from students with insufficient knowledge to students with sufficient knowledge in the Chemistry. The high-reliability index signifies a high level of confidence in the replication of items' placement within the measurement error. A reliability index greater than 0.94 is considered excellent (Fisher, 2007). Hence, it can be concluded that all the items in Chemistry Paper 1 were in the acceptable range of between 0.6 and 1.4 as suggested by Bond and Fox (2015) and were excellent as well (Fisher, 2007).

Table 4.4
Analysis of Reliability and Separation Index

	Person		Item
N	435		50
Measures			
Mean	0.35		0.00
SD, standard deviation	0.97		1.07
SE, standard error	0.05		0.15
Outfit Mean Square			
Mean	0.98		0.98
SD	0.25		0.19
Separation	2.54		8.95
Reliability	0.87		0.99
Cronbach's alpha		0.87	
Chi-square (χ^2)		23075.57	
Unidimensionality		19.20%	

Rasch person reliability analysis can be interpreted by Cronbach's alpha that is used in Classical Test Theory (CTT) (Fisher, 2007). The reliability of the person is based on the estimated locations of persons along the logit interval scale. The low person reliability indicated that the sample range of student abilities is small (Linacre, 2012). However, reliability of the person is often lower because extreme scores are excluded in the computation.

A statistical analysis result demonstrates that the reliability of the Chemistry Paper 1 test using Cronbach's alpha coefficient is 0.87 which is strong and within the high range of 0.71 to 0.99 as mentioned by Bond and Fox (2015). In line with Bond and Fox (2015), DeVellis (2003) added that an alpha above 0.8 indicates a reliable instrument. Therefore, the interpretation result of the Rasch's analysis shows that the Chemistry test has a high reliability of internal consistency. Cronbach's alpha represents a measure of the relationship between test items. A high Cronbach's alpha implies that the items have a close relationship (Barker, Donovan, Schubert, & Walker, 2017) and the instrument is in good condition and acceptable. Hence, it can be deduced that the mean value of the instrument is very good with a high level of consistency (Bond & Fox, 2015).

4.4.2 Research Question 3: What are the items validity of the year-end Form Four Chemistry Paper 1?

4.4.2.1 Item validity

The validity of the items is examined through various statistical analysis provided by Rasch analysis (Linacre, 2003). The validity of the various items in the Chemistry Paper 1 test was established based on the misfit order of the items. Fitness of items is able to influence the reliability and validity of an instrument. When the item

is fit, the item is well-functioned to perform the intended measurements and to assess the item's suitability. On the other hand, misfit items indicate the students had misconception regarding the items.

Item fit is able to assess the constructs through three criteria which are outfit Mean Square values (MNSQ), outfit Z-Standardized values, and PTMEA Corr. (Bond & Fox, 2015; Boone et al., 2014). When these three criteria are within the fit range, the item measured is considered as fit (Azrilah Abd Aziz, M. S. Jusoh, A.R. Omar, Harith Amlus, & Salleh, 2014). However, if all the three criteria are not within the range, then the item is deemed as misfit (Azrilah Abd Aziz et al., 2014). Outfit MNSQ values need to be considered before infit values in determining the suitability of the items that measure the constructs (Aziz et al., 2013). Z-Standardized values can be ignored if the values of outfit and infit are acceptable (Linacre, 2007).

The infit MNSQ values represent the abnormality of the responses to items according to the students' ability. The outfit MNSQ values represent the abnormality of the responses to items beyond the students' ability. The range for the outfit and infit MNSQ values should be between 0.50 logits and 1.50 logits to ensure the appropriateness of the items (Bond & Fox, 2015; Boone et al., 2014; Linacre, 2003). According to Bond and Fox (2015), if an MNSQ value exceeds 1.7, the item is considered to be misfit, and indicates that the item does not reflect the construct. (Linacre, 2009) has added that items are confusing if the value of MNSQ exceeds 1.5 logits. If the MNSQ value is below 0.5, this depicts an inadmissiable overfit item and is easily predictable. It also suggests there is a high likelihood that the item is a replicate of other items (Bond & Fox, 2015; Linacre, 2007). Misfit or outfit items have to be assessed, rectified and eliminated from the scale. In other words, the item fit is deemed as unacceptable (Bond & Fox, 2015). As for multiple-choice tests which are

low stakes, Wright et al. (1994) have suggested that MNSQ values below 1.3 are permissible.

Table 4.5 demonstrates the item mean square outfit ranging from 0.62 to 1.40. While, the minimum value and maximum value of infit are 0.85 and 1.25, respectively and within the range of productive items for measurement which is 0.5 logit to 1.5 logits (Linacre, 2007). All the items in the Chemistry Paper 1 test fulfilled all the criteria proposed by Boone et al. (2014). Therefore, no item was modified or discarded from the instrument

4.4.2.2 Item Polarity

Point-measure correlations (PTMEA Corr.) in the Rasch analysis are equivalent to point-biserial correlation in CTT. It explains the contribution of each item to the total test scores as well as shows whether all items have empirically equal item discrimination as demanded by Rasch Model. This statistical analysis is a basic stage in determining the validity of the constructs and the instrument. Item measure correlations are affected by data predictability, the targeted item on the test-takers sample and the distribution of the person measures (Linacre 2004). The analysis from the present study indicated all the items in the Chemistry Paper 1 were positively correlated with the construct to be measured and moved in the same direction as well as able to differentiate the students' ability (Table 4.5). The PTMEA Corr. values were within the range of 0.16 to 0.50. Bond and Fox (2015) asserted that the value of the PTMEA Corr. must be positive and high in order for the item to distinguish the abilities of the students. In addition, Nunnally and Bernstein (1994) and Linacre (2006) advocated that the value of the PTMEA Corr. which were less than 0.30 signify the items are sags.

On the contrary, negative or zero value of item measure correlations indicate united responses to the items or the students which is contradicted to the constructs (Linacre, 2003) as well as contrasts the direction of the measurement (Runnels, 2012). In other words, the items are misfit according to the Rasch standard and are unable to evaluate the constructs that are intended to be measured and need to be discarded as they may be difficult or misleading the questions (Bond & Fox, 2015; Linacre, 1998). Discarding the misfit items from the measurement increases the PTMEA Corr. value.

In addition, item correlation measure also used in identifying item difficulties. Item correlations that are near to zero designates either the item is very easy, incredibly difficult or suggests that the item may measures the construct in a different direction from the other items (Wolfe & Smith, 2007). Interpretations from statistical analysis of the current research exhibits that the hardest item is item 35 while the easiest item is item 34.

Table 4.5
Item Measure for Fit Statistics

No.	Item	Logit	Standard Error	Outfit		Point Measure Corr.
				MNSQ	ZSTD	
1	I35	2.21	0.14	0.78	-1.8	0.48
2	I38	2.12	0.13	1.31	2.3	**0.25**
3	I16	1.78	0.12	1.4	3.5	**0.20**
4	I45	1.57	0.12	1.36	3.7	**0.16**
5	I48	1.45	0.12	0.95	-0.6	0.43
6	I30	1.37	0.12	1.00	0.0	0.41
7	I44	1.33	0.11	1.07	0.9	0.37
8	I9	1.22	0.11	0.89	-1.5	0.49
9	I11	0.91	0.11	0.94	-0.9	0.44
10	I25	0.87	0.11	0.96	-0.7	0.44
11	I47	0.87	0.11	1.02	0.3	0.40
12	I4	0.85	0.11	1.19	3.0	**0.28**
13	I10	0.82	0.11	0.96	-0.7	0.45
14	I22	0.70	0.11	0.93	-1.3	0.50
15	I20	0.68	0.11	1.11	2.0	0.32

Table 4.5
Item Measure for Fit Statistics (continued)

No.	Item	Logit	Standard Error	Outfit		Point Measure Corr.
				MNSQ	ZSTD	
16	I46	0.65	0.11	1.01	0.3	0.39
17	I40	0.55	0.11	1.34	5.7	**0.16**
18	I42	0.44	0.10	0.88	-2.3	0.48
19	I18	0.35	0.10	1.29	4.9	**0.17**
20	I39	0.27	0.10	1.36	5.9	**0.16**
21	I28	0.25	0.10	1.17	3.0	**0.25**
22	I24	0.18	0.10	0.96	-0.7	0.40
23	I26	0.16	0.10	1.09	1.6	0.34
24	I50	0.15	0.10	0.91	-1.6	0.44
25	I29	0.09	0.10	0.83	-3.1	0.51
26	I8	-0.01	0.11	0.88	-2.1	0.46
27	I23	-0.10	0.11	1.15	2.3	0.30
28	I12	-0.12	0.11	0.88	-2.0	0.46
29	I33	-0.16	0.11	0.87	-2.0	0.45
30	I1	-0.28	0.11	0.87	-2.0	0.46
31	I36	-0.28	0.11	1.02	0.3	0.41
32	I27	-0.39	0.11	1.3	3.7	**0.16**
33	I13	-0.45	0.11	0.77	-3.3	0.51
34	I37	-0.46	0.11	0.81	-2.6	0.46
35	I41	-0.56	0.11	0.97	-0.3	0.34
36	I14	-0.61	0.11	0.74	-3.3	0.50
37	I6	-0.77	0.11	0.94	-0.6	0.39
38	I31	-0.80	0.11	0.98	-0.2	0.39
39	I2	-0.81	0.11	1.03	0.4	**0.28**
40	I3	-0.82	0.11	0.86	-1.5	0.39
42	I43	-0.96	0.12	1.13	1.2	**0.26**
43	I21	-1.22	0.12	0.78	-1.9	0.40
44	I32	-1.23	0.12	1.05	0.4	**0.29**
45	I17	-1.28	0.13	0.86	-1.1	0.34
46	I19	-1.54	0.13	0.76	-1.7	0.36
47	I15	-1.67	0.14	1.02	0.2	**0.29**
48	I7	-1.90	0.15	0.75	-1.4	0.34
49	I5	-2.10	0.16	0.62	-2.1	0.35
50	I34	-2.41	0.18	0.68	-1.4	**0.28**

Note: PTMEA Corr < 0.30 are in boldface.

4.4.2.3 Distractor Analysis

It is a difficult task to design options with equal plausibility in constructing multiple-choice questions. Distractors functionality, item writing defects, and the optimum number of options are interconnected and should be given full

attention as they may influence the quality of the item, the performance of the item and the results of the test. Tarrant, Ware, and Mohammed (2009) summarized that items with two functional distractors were more difficult than item with three functional distractors.

In Rasch analysis, the Winsteps computer program (Linacre, 2015) was used to acquire the frequency of students selecting each answer (A, B, C or D) along with the range of students's ability estimates on the logit scale at each point. Koizumi et al. (2011) in their study emphasized on the significance of good and quality distractors in imparting information on error pattern profiles. As an item in the test is comprised of stem and distractor, therefore it can reflect on how well the test items are developed. The distractors functionality acts as an autonomous indicator of the functioning of the item.

An effective distractor is referred to as the option that attracts students with misconceptions or errors in thinking and reasoning, normally students with low abilities (Rodriguez, Kettler, & Elliott, 2014). These distractors are also called functional distractors if they are being selected by one or more students. However, not all distractors work equally (Andrich & Styles, 2011). Some distractors may draw away students more than any other distractor. Students with moderate ability might not select an implausible distractor, but it may be selected or not selected at all by lower ability students. Hughes (2008) claimed that certain defects distractor for items were problematic because they had more than one correct answer or had no right answer or had any clues to the right answer option and the options of ineffective responses.

Nonetheless, Linacre (2012) notes that the acceptance of items with a problem of distractors should meet certain conditions. Items which have good fit values and the

average measure of incorrect options less than the average measure of correct options may be accepted and retained for further use. Thus, the items that are 'mis-fitting' and the average measure of the incorrect options higher than the average measure of the correct option must be reviewed or eliminated.

Distractor analysis results from the present study show that distractors for all Chemistry Paper 1 items are good distractors due to the logit values increase systematically according to the increasing trend (Appendix VIII). This finding indicates that all distractors are effective and able to discriminate the students according to their ability. The selection of distractor is crucial for teaching and learning as distractors choice may offer details on the misunderstandings and misconceptions of low ability students as well as address the possible reasons for their low achievement (Asril & Marais, 2011).

4.4.3 Research Question 4: What are the appropriateness between item difficulty and students' ability?

4.4.3.1 Item Difficulty

An achievement test is considered ideal when the difficulty level is set up in accordance with the abilities of students (Susongko, 2016). In other words, the test represents the full range of the abilities of all students. Sick (2011) described that difficult items are expected to be answered only by high ability students while easy items are expected to be answered by students with low ability.

A standardized achievement test such as SPM Chemistry Paper 1 that consisted of 50 multiple-choice question was constructed according to the item ratio principle of 5:3:2 that referred to the different constructs in the Bloom's Taxonomy. In the SPM Chemistry Paper 1, this ratio represents 25 items on the knowledge construct, 15 items

on the understanding construct and 10 items on the application construct (Lembaga Peperiksaan, 2002).

Table 4.6
Item Difficulty Level

Item Difficulty Level			
	Difficult (1.22 -2.21)	Moderate (0.91 – (-0.96))	Easy (-1.22 - (-2.41))
Item no.	35, 38, 16, 45, 48, 30, 44, 9	11, 25, 47, 4, 10, 22, 20, 46, 40, 42, 18, 39, 28, 24, 26, 50, 29, 8, 23, 12,33, 1, 36 ,27, 13, 37, 41, 14, 6, 31, 2, 3, 49, 43	21, 32, 17, 19, 15, 7, 5, 34
Total (Percentage)	8 (16%)	34 (68%)	8 (16%)
Ratio	2	8.5	2

Table 4.6 shows the number of Chemistry test items that was developed according to the different construct, 8 items (16%) are on knowledge, 34 items (68%) are on understanding and 8 items (16%) are on application. Ratio comparison with the actual SPM Chemistry Paper 1 indicates that the ratio for the Chemistry test was 2:8.5:2. Therefore, from this finding it can be concluded that the Chemistry test measures the abilities of students in certain constructs only which is mostly on their understanding.

4.4.3.2 Mapping of Student and Item

The person-item map (Wright map) depicted in Figure 4.1 is a significant feature of the Rasch model. This map is a graphical representation that capable of illustrating the dispersal of persons estimates and the items difficulty on a

common logit scale (Zunita Maskor & Harun Baharudin, 2019) and represents the relationship of person-item (Boone & Noltemeyer, 2017). Theoretically, this map can explain the extent of the item coverage or comprehensiveness, the amount of redundancy and the extent of the latent trait in the test-takers (Cappelleri et al., 2014). The information attained from person-item map is vital for test developers to construct a high quality and valuable instrument.

The person-item map yielded from the data set exhibited a significant picture of the linear continuum of the Science student's abilities compared to the Chemistry test items. The distribution of the student ability level, signified by "#s", is shown from the highest to the lowest and from top to the bottom of the scale on the left hand side of the map. The higher logit values of the person measure indicate a higher degree of the ability of students in Chemistry test and a better test performance. The lower logit values of the person measure signify the low abilities of students in their test performance. The upper left quadrant depicts students who has knowledge in Chemistry while the lower left quadrant signifies students with insufficient knowledge. On the right hand side of the map, the difficulty level of the items is dispersed from the hardest item to the easiest item in descending order. More difficult items are placed at the top of the person-item map and easier items are placed at the bottom of the map.

The letter M denotes mean on students' ability ("M" on the left hand side of the map) and level of difficulty of the item ("M" on the right hand side of the map). The mean difficulty of item is normally set to 0 logits (Iramaneerat, Smith, & Smith, 2008). If a student is plotted at the same level as an item, this means that a student has a 50% probability of answering that item correctly. As the items tend to be difficult, the odds of success is reduced, means that less chance to answer correctly (Rasch,

1960, 1961). These items are estimated beyond the students' ability. The mean of student's ability is compared with the mean of item difficulty in establishing how well the Chemistry test items are spread according by the level of students' abilities. For this data set, mean items is 0 while mean for students is 0.35 logit which are very close to each other. This signifies that the test items for the students are well-targeted without ceiling and floor effect (Boone, 2016). It also means that the level of difficulty of the test items is appropriate for the Science students. In short, the Chemistry test items are moderately difficult for the students.

Measurement experts clearly advocate that the instrument should be able to evaluate students with high and low abilities. Each item has different difficulty level to discriminate the abilities of students according to their levels. The item difficulty measures range from logit -2.41 to logit 2.21 while student abilities estimates range from -2.01 to 3.66 which is slightly higher than the item difficulty measurement. The wide distribution of student abilities compared to item difficulty dispersal exhibits that only certain items are able to cover the range of the measurement traits (Green & Frantom, 2002).

The maximum level of item measurement was 2.21 logit (SE: 0.15) while the maximum measure of a student was 3.66 logit (SE: 0.05). Most of the Chemistry Paper 1 test items have difficulties level near the mid-range of the logit scale which is within one standard deviation. Linacre (2009) stated that students with high ability are able to answer difficult items while students with high and low ability can easily answer the easy items. The most difficult item answered by students is I35 with 74 correct responses out of 435. This item is about molar mass of a compound. On the contrary, item I34 is the easiest with 400 correct responses out of 435. The easy item is on napthalene graph.

The person-item map posits a normal distribution of items and students in the logit interval scale continuum and falls into the mid-range zone of the scale. Linacre (2005) notes that in which norm reference interpretations are required, the distribution of students' ability should be synchronized with the distribution of item difficulty. Nonetheless, there are some gaps exist in the item location distribution (I9 and I10, I38 and I16) in the map that indicate students in the middle and upper levels of the map were not sensibly aimed by the Chemistry test items due to the content aspects of the constructs under study, lacking of some representations and compromise on the validity of the test (Messick, 1989). The existence of gap between two consecutive items in the mid-range of the person-item map and the lack of appropriate items with the higher ability students at the upper level of logit scale indicate that some important aspects have not been measured by the instrument (Huey Fern & Hooi Lian, 2017).

There are eight items (I17, I21, I32, I19, I15, I7, I5, I34) that fall below the ability of students. Although these items fit the model but they do not contribute to the measurement precision. Hence, these items can be discarded from the instrument. On the contrary, a few students with high abilities are located above the logit 2.21. Stelmack et al. (2004) emphasized that if there were more students at the high end of the difficulty range, then more item could be needed to guarantee that all abilities were measured. However, unlike the aforesaid items, the difficult items should not be eliminated from the instrument to avoid the ceiling effect. The precision of measurement would be useless if the abilities of students is beyond the demand of the test.

Person-item map also exhibits the redundancy or overlapping of items in the Chemistry test and assesses the same level of difficulty of the construct (Boone & Noltemeyer, 2017). In line with the previous study, Bond and Fox (2015) added that

the overlapping items at different levels of difficulty are considered analogous in terms of measuring the same construct. Items with the same measurement are viewed as a "cutting thermometer" at the same location (Boone & Noltemeyer, 2017). Based on the analysis from the person-item map, I25 and I47 (both at logit 0.87) and I1 and I36 (both at logit -0.28) all measure at the same level of difficulty. Therefore, these redundant items are suggested to be discarded to maintain the integrity of the test (Boone & Scantlebury, 2006) in distinguishing the students.

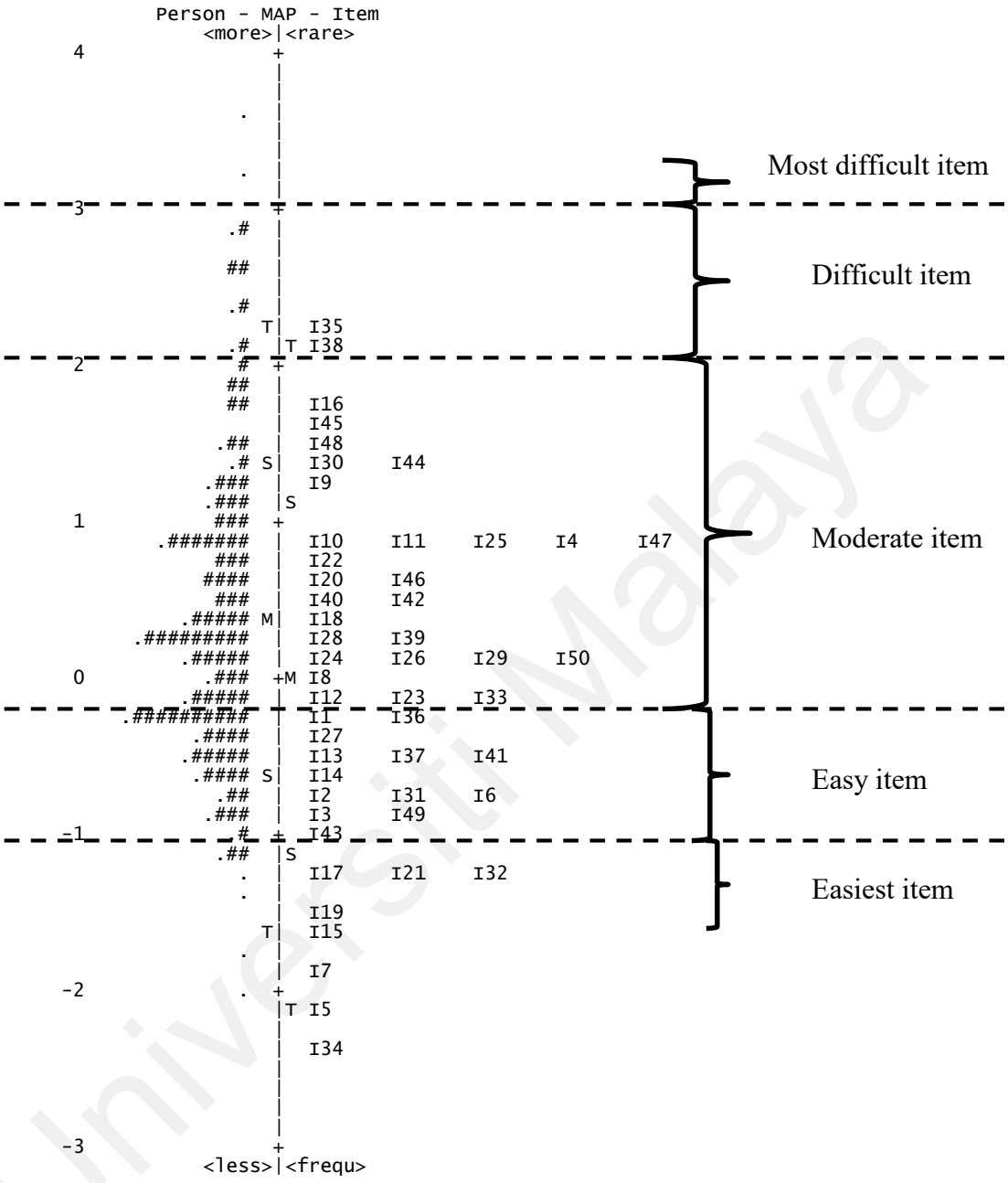

Figure 4.1 Person-item map of Chemistry test. Each "#" represents 4 participants, whereas "." indicates 3 participants.`

4.4.3.3 DIF Analysis

A crucial phase in the validation of the test is the identification of items which demonstrate differential item functioning (DIF) for various groups of students. DIF items could result to decisions on bias testing and threaten the validity of the test (A. S. Cohen, Kim, & Wollack, 1996; Magis & De Boeck, 2011). At the same time, it is possible to dismiss its accession by the test takers and policy makers.

DIF analysis was conducted at the item level to ascertain whether any irrelevant factor intervened with the construct that is being measured. Crocker and Algina (2006) clarified that bias is present when test results represent unrelated factors or characteristic that do not reflect the construct of interest (e.g., demographic variables). By this definition, the construct validity would be impaired by bias through the interpretation of the test score. Hence, a Rasch analysis of the uniform differential test and the functioning of the items were implemented to the Chemistry data set. This is to identify specific items that exhibit gender bias (e.g., male versus female).

Two criteria to be taken into account in the DIF analysis were recommended by Linacre (2012). For the first criteria, the likelihood of the item DIF has to be small which is the likelihood of the item DIF has to be statistically significant different with $p \leq 0.05$. While, the second criteria referred to the DIF contrast in which it has to be at minimum of 0.5 logit to establish a significant DIF difference.

Table 4.7
Differential Item Functioning (DIF)

Male	DIF Measure	DIF S.E.	Female	DIF Measure	DIF S.E.	DIF Contrast	Rasch-Welch's t	p	Item
L	-1.58	0.18	P	-2.61	0.31	1.03	2.84	0.0188	I7
L	-0.34	0.14	P	-1.01	0.19	0.67	2.85	0.0240	I14
L	-1.4	0.17	P	-2.19	0.27	0.78	2.48	0.0408	I15
L	-1.35	0.17	P	-1.87	0.24	0.52	1.79	0.0677	I19
L	0.05	0.14	P	-0.46	0.17	0.51	2.33	0.1393	I33
L	-0.03	0.14	P	-0.63	0.17	0.60	2.72	0.0543	I36

The results of the DIF analysis display a significantly apparent DIF for the Chemistry items. Six items (I7, I14, I15, I19, I33 and I36) were statistically significant at DIF contrasts ranged between 0.51 and 1.03 (Table 4.7). The answers of students based on gender indicated these items were difficult for male students than female students. In general, the results of the uniform DIF analysis were aligned with the Rasch model which supported to the structural aspects of validity of the Chemistry data set from a gender perspective.

4.5 Psychometric Analysis for the Items in the Chemistry Paper 1

Four indicators have been considered to determine the quality of each item:

a) index value of infit MNSQ

b) item measure in logit unit

c) item polarity index

d) distractor analysis

The accepted infit MNSQ index ranges from 0.50 to 1.50 and the point-biserial measure correlation index value ranges between 0.30 and 0.70. Item measure exhibits the accurate difficulty level of the test item.

Table 4.8
Point Biserial Measure Coefficients for Distractors Analysis

Scale Range	Indication
0.30 or above	very good test distractor
0.20 to 0.29	reasonably good test distractor
0.09 to 0.19	needs improvement
< 0.09	poor test distractor

Source: 'Using Assessment Data' (2015), retrieved 3 September 2015, https://www.unthsc.edu/center-for-innovative/using-assessment-data/

The distractor of each item is considered to be functioning if it is selected by at least 5% of the students (T. M. Haladyna & Downing, 2016) and has negative value of point-biserial measure correlation (Boone & Staver, 2020a). If a distractor has positive point-biserial measure correlation values where the values are close to or greater than the point-biserial correlation value of the correct answer, the distractor has a potential to be the correct answer. In this case, the distractor is considered to be not working properly. As for the correct answer, it should have a positive value of point-biserial measure correlation and the highest average ability. Table 4.8 reflects the ranges of point-biserial measure correlation and its indication in the distractor analysis interpretation.

Table 4.9
Point Biserial Measure Coefficients Indication for an Item

Scale Range	Indication
< 0.00 (Negative)	Unacceptable/need item examination
0.00 to 0.24	Room for improvement
0.25 to 0.39	Good item
0.40 to 1.00	Excellent item

Source: 'Using Assessment Data' (2015), retrieved 3 September 2015, https://www.unthsc.edu/center-for-innovative/using-assessment-data/

For Rasch analysis, point-measure biserial correlation can serve as item discrimination (S. Brown et al., 2005; Linacre, 2012). Besides differentiating students, item discrimination also measures the effectiveness of an item whether it is low, medium or high (Table 4.9). Lake and Holster (2016) stated in their study that item discrimination allows a researcher to identify items that behave unexpectedly.

The following are examples of comprehensive analysis result of a few Chemistry test items using Winsteps software. The complete analysis result of each the Chemistry test item is presented in Appendix IX.

Table 4.10
Analysis of Item 9

Infit MNSQ	0.90			
Item Measure	1.22			
Option	Data		Average Ability	PTMea Corr
	Count	Percentage		
A	46	11	0.21	-0.05
*B	139	32	1.05	0.49
C	201	46	0.06	-0.28
D	43	10	-0.21	-0.19

* correct answer

The infit MNSQ value for item 9 is 0.90 which is within the acceptable range. The point-biserial measure correlation of item 9 is 0.49, that indicates this item has a good discrimination index and the item's difficulty level is hard. Option B is the key answer thus it has a positive point-biserial correlation. Distractor C is not functioning well in discriminating students although the point-biserial measure correlation is negative. This analysis clearly showed that many students including high ability students also chosen distractor C as their answer instead of the key answer. On the other hand, distractors A and D are working properly where the frequency of distractor selection has exceeded 5% respectively. As a conclusion, item 9 is an excellent item but needs some modification whether in terms of distractor or in the stem of the item.

Table 4.11
Analysis of Item 14

Infit MNSQ	0.86			
Item Measure	-0.61			
Option	Data		Average Ability	PTMea Corr
	Count	Percentage		
A	29	7	-0.38	-0.20
*B	299	69	0.68	0.50
C	39	9	-0.32	-0.22
D	56	13	-0.31	-0.26

* correct answer

The infit MNSQ value for item 14 is 0.86 which is within the acceptable range. The point-biserial measure correlation of item 14 is 0.50, that indicates this item has an excellent discrimination index and is able to discriminate students into 3 groups namely high ability, moderate ability and low ability. The difficulty level of the item is moderate. Option B is the key answer thus it has a positive point-biserial measure correlation. All distractors, namely D, C and A are working properly because they

have negative point-biserial measure correlation values and are able to attract more than 5% of the students. Therefore, it can be concluded that item 14 is an excellent item.

Table 4.12
Analysis of Item 25

Infit MNSQ	0.97			
Item Measure	0.87			
Option	Data		Average Ability	PTMea Corr
	Count	Percentage		
A	86	20	0.01	-0.17
B	57	13	0.18	-0.07
*C	168	39	0.88	0.44
D	114	26	0.00	-0.21

* correct answer

The infit MNSQ value for item 25 is 0.97 which is within the acceptable range. This item has a point-biserial measure correlation value of 0.44 that indicates it is able to discriminate students properly. The difficulty level of item 25 is moderate. Option C is the key answer due to a positive point-biserial measure correlation value and the highest percentage of students that have selected this option. However, a majority of the high ability students have also chosen distractor D as their right answer. This indicates that the distractor D has a potential to be the key answer. Distractors A and B are good distractors because both have negative point-biserial measure correlation values and the frequency of distractor selection has exceeded the common benchmark of distractor functionality. Hence, it can be concluded that item 25 is an excellent item. Nevertheless, it is suggested that this item should be reviewed and the focus is particularly given to distractor D so that the quality of item 25 can be increased.

The exhaustive analysis result of each Chemistry items showed that not all distractors were performing properly. The non-functioning distractors need to be removed or replaced with more plausible distractors as the existence of non-functioning distractors can affect the quality of the test items. These non-functioning distractors were added as "fillers" to complete the requisite options. T. M. Haladyna and Downing (2016) reported that more than 38% of the test distractors were discarded as less than 5% of the students selected them. Generally, the results of this psychometric analysis clearly revealed that it is difficult to develop equally plausible distractors (T. M. Haladyna & Downing, 1989).

4.6 Summary

The findings of the study based on the statistical analysis of student scores in Chemistry Paper 1 were presented in this chapter using Winsteps. The analysis results clearly exhibit the data set conforms the Rasch Model. This means unidimensionality is held across the Chemistry test paper. In fact, there is no violation of the local independence assumption. Particularly, the comprehensive analysis performed indicates that all the Chemistry items are well-targeted and suitable to measure the level of the knowledge and understanding of the students. Overall, the results of the Rasch analysis have shown that the Chemistry Paper 1 has good psychometric properties.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Introduction

This chapter summarizes the results of the analysis and discusses the psychometric properties of the Chemistry items. The quality of the items and the sources of error affecting the test scores are also reviewed. The findings of this study are based on the objectives and research questions addressed in Chapter One. This chapter also includes discussions, implications of the study, and recommendations for future research in addition to the summary of the findings.

5.2 Summary of the Study

In this study, the psychometric properties of the instrument are explored by applying the Rasch model. Through the implementation of a series of item analysis tests, all items of the Chemistry test were found to conform with the Rasch model. The findings are discussed under four subsections in accordance with research questions:

i. Data fit the Rasch Model
ii. Student reliability and item reliability of the year-end Form Four Chemistry Paper 1
iii. Item validity of the year-end examination
iv. The appropriateness between item difficulty and the ability of students

5.2.1 Data Fit the Rasch Model

The unidimensionality of an instrument plays an important role in determining of its validity. The researcher was able to evaluate, with the Rasch analysis, if all items

worked together in measuring a single variable (Bond & Fox, 2015). A Principal Component Analysis (PCA) and various statistical analyses were performed using 50 Chemistry items to assess their dimensionality. Practically, all statistical analysis demonstrated that the Chemistry test items met the unidimensionality assumption, however the raw variance explained by measures of all items was only 27.7%. (Linacre, 2006) pointed out that a measurement higher than 40% is considered as a strong dimension, higher than 30% as moderately strong dimension, and those higher than 20% as moderate dimension. Based on this indicator, the Chemistry test dimension is concluded as moderate and acceptable. Indirectly, it means all Chemistry items are clear and not confusing. Nonetheless, the unidimensionality of this test has to be examined and needs to be improved.

Besides PCA, an additional criterion that was considered for unidimensionality was the item fit statistics. The analysis revealed that the Chemistry test data fitted the Rasch standard reasonably well. It was found that the item fit was found to be within a productive range (Bond & Fox, 2015) These items were contributing significantly to the measurement of the construct (Linacre, 2012). The lack of 'mis-fitting' items show that the Chemistry items for the test define a unidimensionality characteristic.

One of the cornerstones of the Rasch model is local independence. But researchers (Winarti & Mubarak, 2019; Yee et al., 2018) who have conducted a similar study have not discussed local independence in their findings. Usually, once unidimensionality is proven, indirectly local independence assumption is also fulfilled. Violations of local independence can increase reliability estimates and problems with construct validity. Dependency among items can portray a fake impression of the precision and quality of the test (Christensen, Makransky, & Horton, 2017). There were no significant correlations among test items when analyzing the local

independence of the Chemistry test data, since the residual correlations for each pair of test items was less than 0.7 as suggested by (Linacre, 2015). Furthermore, it also demonstrates that there is no redundancy of items in this test.

5.2.2 Student Reliability and Item Reliability of The Year-End Examination Form Four Chemistry Paper 1

For Rasch analysis, reliability is taken into account both from the standpoint of the persons and the items. The statistical analysis of the study shows that the Chemistry test has a strong internal consistency and very high item reliability. In other words, the test is sufficiently reliable for measuring the abilities of students. Besides that, the Chemistry test has a very high value of item separation which signifies that the instrument has very well distribution of items (Klooster, Taal, & Van de Laar, 2008). A separation item value that exceeds 2.00 is strongly accepted and suggests that the actual difference related to the student ability is easy to distinguish for the items (Jailani Yunos et al., 2017). On the contrary, a finding by Krishnan and Noraini Idris (2014) denotes that the items are well dispersed if the item separation value exceeds 1.00.

The analysis has also revealed that the instrument has high person reliability and person separation index. A high person separation that exceeds 2.00 implies that the instrument is able to distinguish students efficiently (Gracia, 2005). Statistically speaking, the person separation index of 2.54 designates three strata of the students (high, moderate, low). On the other hand, Jailani Yunos et al. (2017) and Khamis and Che Yahya (2015) have agreed that the separation value of less than 2.00 is considered low. Therefore, removing the problematic items will improve the reliability and the

separation index as well as enhance the quality of the instrument (Siew & Mohammad Syafiq Abd Rahman, 2019).

5.2.3 Item Validity of The Year-End Examination

The misfit order of the items in the Rasch analysis informs about the appropriateness of the item measure the constructs of the Chemistry test. The item mean square outfit of the 50 items is ranged between 0.62 to 1.40 which is within an acceptable range (Bond & Fox, 2015). Thus, there is no misfit item in the Chemistry test. In general, all the Chemistry items have contributed to the measurement. This valuable information is vital for teachers as a reference for improving the quality of their instruction. A similar study by Winarti and Mubarak (2019) found that the information yielded could prevent and deal with attacks of the misconception that could arise in the future. A shred of alternative evidence that examined the psychometric properties of items graphically is the item characteristic curve (ICC). It provides comprehensive information about the test items as well.

Analysis of item polarity is an essential step for measuring construct validity. Item polarity or point-measure correlation (PTMEA Corr.) is the initial recognition of construct validity (Bond & Fox, 2015) to examine the connection among the items in assessing the required constructs. PTMEA Corr. not only benefits in measuring item fitness but also on item checking to find out if the items move in the same direction with the constructs (Linacre, 2015). The positive values were indications that the items were parallel to the construct to be measured by the researcher (Linacre, 2015). The acceptable PTMEA Corr. value suggested by Linacre (2012) and Bond and Fox (2015) is in range between 0.3 and 0.7.

Based on the statistical analysis, PTMEA Corr. values were all positive with less acceptable items correlation strength to the constructs of the model (Bond & Fox, 2015) which ranged between 0.16 and 0.51. However, although PTMEA Corr. were all positive, there are 14 items that have PTMEA Corr. value less than 0.30. Tran, Dorofeeva, and Loskutova (2018) recommended that an item with PTMEA Corr. value that is out of the acceptable range needs to be removed. Nevertheless, this analysis result indicates the items worked together efficiently and moved in same direction in measuring the proposed construct (Bond & Fox, 2015). It can be summarized that the item discrimination is very good due to its discrimination power of the abilities of students.

For multiple-choice items, it is critical to examine responses of the distractors. An analysis of the distractors for assessing students' achievement offers relevant information on their understanding of the measured variable in the classroom (Asril & Marais, 2011). This analysis is able to determine the student performance in various ranges of ability on different distractors. A distractor is defined as functional when it is aimed to be plausible for low-performance students (Testa, Toscano, & Rosato, 2018). On the other hand, implausible distractors only extend the test duration without improving the test accuracy (DiBattista & Kurzawa, 2011). In addition, the information in a distractor implies that a person who chooses that distractor has greater ability than the person who chooses another distractor with no information (Andrich & Styles, 2011).

In this study, the increasing trend of distractor analysis depicts the distractors for Chemistry test items are well-functioned based on the logit values of PTMEA Corr. Thus, all the test items are able to distinguish the students efficiently. Despite of PTMEA Corr. value, distractor analysis plot can also be used to illustrate the link

between the estimates of the achievement of the students and the proportion of students who select a specific answer (A, B, C or D) in a graphical form.

However, in terms of the comprehensive analysis of the distractors for items, I16, I17, and I34 possess a distractor that has only been selected by either 1% or 0% of the students. By referring to the table of specification (TOS) (Appendix XI), items I16 and I34 are based on the topic Structure of the Atom while item I17 is on the topic of Chemical Bond. As for items I38, I44, and I48, the analysis showed the existence of non-effective distractors as many students were attracted to choose the distractor as the correct answer. On the other hand, for item I45, students tend to select the distractors and correct answers equally. Based on TOS (Appendix XI), the topic presented in item I38 was on the Periodic Table of Elements, items I44, and I48 were on Acid and Base while item I45 was on Salts. Ultimately, it can be summarized that students have difficulty in these topics because they require them to have conceptual understanding and application as well as memorization skills.

5.2.4 The Appropriateness Between Item Difficulty and Ability Of Students

In item analysis, it is important to establish the item difficulty because it reveals whether the item would be too easy or extremely difficult for the students. Results from the present study indicates that 8 items (I35, I38, I16, I45, I48, I30, I44, I44, I9) are within the range of 1.22 to 2.21 in linear continuum that indicates that these items are difficult. While, 8 items (I21, I32, I17, I19, I15, I7, I5, I34) are easy as their locations are between logit -1.22 and -2.41. Moderate items are ranged at logit 0.91 to -0.96.

According to the format of the assessment of Chemistry paper for SPM 2003 issued by Malaysian Examination Syndicate (Lembaga Peperiksaan, 2002), the constructs measured in the Chemistry Paper 1 included 80 % questions on knowledge and understanding while 20% of the questions were based on application of the subject matter. The measurable construct, the coverage of the context and the distribution of items difficulty level must comply with the table of specification standards. In addition, all the items developed need to be pretested and have empirical evidence that meets the standards before setting an instrument. Popham (2020) and Doody and Doody (2015) pointed out that a pilot test was carried out to increase the quality of the items and the confidence level in data interpretation. Furthermore, besides detecting the weaknesses of the instrument, Sharifah Nurulhuda Tuan Mohd Yasin et al. (2018) mentioned that the reliability of the instruments was also tested using a pilot test.

Based on the assessment standard, the item difficulty level is in the ratio (low: medium: high) of 5: 3: 2. However, the Chemistry test paper used in this study was found to have only a ratio of 2:8:2 and did not meet MES standard. This finding was consistent with the findings by Yee et al. (2018). According to their research, when the ratio of difficulty does not meet the table of specification, it indicates Chemistry items are not in accordance with the test developer's expectation. Ultimately, it can be summarized that this instrument assesses mostly moderate ability students due to the high ratio on a moderate level of item difficulty.

The construct hierarchy or person-item map (Wright map) is the key source of data and acts as the heart of the analysis in the Rasch model (Nazlinda Abdullah, Shereen Noranee, & Mohd Khamis, 2017). This map illustrates both persons and items located on the psychometric ruler and visualizes how the parameters interact (Boone & Staver, 2020b). From the Wright map (Figure 4.1), the mean item difficulty was set

to 0.0 by default while students' ability mean was 0.35 logit that was very close to each other. This suggested that the Chemistry test items were on average, slightly easier for the students and quite well targeted. Despite this, there were few students whose ability estimates far exceeded the item difficulty estimates.

Although the Wright map shows the spread of the majority of the ability of students (-2.01 to 3.66) fall within the range of the item difficulty distribution (-2.41 to 2.21) but in general the ability of the students' distribution is slightly higher than the item distribution. This finding indicates some of the items solely measure within a particular range of the students' ability. Ultimately, this Chemistry test enables the teachers to relate the students' ability to the difficulty of the test items. As a result, McCamey (2014) stressed that teachers or test developers could refine the test by eliminating items with low difficulty, lessen the number of items with the same difficulty and adding items with a higher level of difficulty to create a better instrument.

The significance test of DIF was carried out to ensure the Chemistry test was fair for every student who took it. DIF study is the primary method in instrument development specifically on evaluation in education. This is because it focused on the identification of the differences. The test could show either no similarity or almost the same function when administered to a group of students with similar abilities (Siti Rahayah Ariffin, Rodiah Idris, & Noriah Mohd Ishak, 2010).

The result of the uniform DIF analysis indicated that six items had a statistically significant DIF. These identified items should be examined because they contain construct-irrelevant variance that could change the measurement precision and affect the structural aspect of the construct validity (Al-Owidha, 2018). This finding is important for test developers and content experts to decide whether the item is biased

or should be eliminated but this substantive difference has to be supported by statistical significance as cautioned by (Linacre, 2004).

Generally speaking, the findings of this study show that the Chemistry Paper 1 possesses good psychometric properties. However, there are still a few limitations that have to be considered. As the data collected was limited to only four schools, the findings of this study cannot be generalized to the population. In addition, the sample size of 435 is practically insufficient to conduct the DIF method. R. M. Smith (2004) stated that the appropriate sample size for the Rasch-DIF method should exceed 500 in each subpopulation. He added that this method was not capable of detecting biased items below 0.5 logits for sample size that was less than 500.

5.3 Implications

The present study is conducted to determine the psychometric properties of the standardized Chemistry Paper 1 instrument using the Rasch Model. The result of this study would benefit Chemistry teachers and test developers, especially in the process of item development. There are practical and methodological implications gained from this study.

5.3.1 Practical Implications

This study has significant effects on test developers and teachers in reviewing the multiple-choice items and the assessment standards primarily on item development. The data analysis from this study reveals that the standardized Chemistry test instrument is unidimensional and has good psychometric properties. Nedungadi et al. (2019) even advocate that psychometric data are important because it shows that the instrument only measures the intended construct. While, Arjoon et al.

(2013) have claimed that the validity of interpretation from the test scores is dictated by the psychometric evidence, therefore collection and reporting validity and reliability of the evidence is the important aspect.

According to Danielson (2001), teachers should be able to construct valid and reliable assessment instruments during instruction. Therefore, teachers can be test developers in terms of assessment at the school level. Furthermore, teachers can secure content validity as they are content experts in their subjects. A valid instrument can increase the confidence level of teachers in using the instrument for measuring the knowledge of students and their level of understanding. The findings of this study serve as an indicator of the state of Chemistry measurement. This is supported by Nor Hasnida Che Md Ghazali (2016) in her study on the psychometric properties where she has emphasized that teachers must be able to utilize validated instruments as self-assessment tools in identifying their strengths and weaknesses. In addition, Winarti and Mubarak (2019) also agreed with Nor Hasnida Che Md Ghazali (2016) that the assessment of learning such as multiple-choice items are able to evaluate students' progress in the learning process, and provide guidance for creating chemical learning strategies and recognizing students' understanding of chemical material.

Test development is a standardized process that needs iterative refinement (Irwing & Hughes, 2018). This process is usually standardized through the use of test development as it is compulsory to produce fair and equitable assessment tasks (T. Haladyna & Rodriguez, 2013). In educational tests especially in schools, multiple-choice items are commonly used because they directly measure knowledge, skills, and competencies (Gierl, Bulut, Guo & Zhang, 2017; Shin, Guo & Girl, 2019). Some researchers have agreed that multiple-choice items with distractors are able to estimate the students' ability in terms of understanding the subject and its becomes a strategy

in preventing potential student misconception (Herrmann-Abell & DeBoer, 2011; Yee et al., 2018).

The findings of this study can guide chemistry teachers in constructing quality items that met the psychometric properties in accordance with the multiple-choice items standards as proposed by T. Haladyna and Rodriguez (2013). In items development, the use of a table of specification is of utmost importance because it's able to warrant the content validity of the instrument including the thinking skills that the teachers intended to measure. As pointed out by Fives and DiDonato-Barnes (2013) in their study, the table of specification assists teachers in developing multiple-choice items so that the items are well-aligned with the topics studied and the cognitive process during instruction.

Teachers can make valid interpretations of the total test scores on their students' ability and comprehension when the items difficulty level match with the thinking level of instruction. Arjoon et al. (2013) in their study asserted that chemistry teachers should comprehend the quality of the psychometric evidence associated with the instruments they would like to use for valid interpretations. Items with good psychometric properties are useful for diagnosing the thinking of students and distinguishing them according to their actual level of ability. This valuable information can assist teachers to target their instructions more effectively (Herrmann-Abell & DeBoer, 2011).

The multiple-choice items provide better psychometric properties such as reliability and validity compared to another forms of tests like open-ended questions (Wells & Wollack, 2003). In addition, Miller et al. (2013) have mentioned that the higher reliability of multiple-choice items is due to objective scoring. Popham (2020)

even added that the multiple-choice items show high evidence of content validity because it can sample a wide range of content domains.

A multiple-choice item is consisting of stem, distractors, and auxiliary information. Findings from the previous studies showed that distractors were the important part of a multiple-choice item as it could influence the quality of the items and learning outcomes (T. M. Haladyna & Downing, 1989; Hansen & Dexter, 1997; Thissen, Steinberg, & Fitzpatrick, 1989). The plausibility of each distractor could adversely impact the psychometric properties of the right and wrong options.

The result of the item analysis is able to establish the functionality of the Chemistry test items as the test items determine the quality of the instrument and influence the scores of students. In general, the scores of students depend not only on their ability but also on the item difficulty level. Yee et al. (2018) claimed that teachers should analyze the items since item analysis is able to determine whether the item difficulty set by teachers is in accordance with the table of specification and matched perfectly with the ability of students that is measured by their answers.

The most common statistic reported in item analysis are item difficulty and item discrimination. Despite item analysis, distractor analyses are another critical analysis used to measure how well each of the incorrect option contributes to the quality of a multiple-choice item. A previous study by Gierl et al. (2017) has proved that distractor analysis is able to assist the test developers and teachers comprehend why students make errors and thus guide the diagnostic conclusions about test performance. He added that this analysis could identify the areas of content that require instructional improvement and provide remedial instruction to students in those areas.

Al-Owidha (2018) finding demonstrated that test developers may be driven by item analysis data to enhance the assessment tool's effectiveness whereby the items

might be removed or modified. Ultimately, the Rasch analysis result would benefit test developers in evaluating the psychometric properties of the test to assess the construct-related validity. The assessment of this construct fortifies the interpretation of the test scores. Items that are functioning well can be stored into an item bank. By having an item bank that is made up of these tested and analyzed items, it helps teachers to prepare for the test paper efficiently. Nevertheless, teachers can provide exercises that are equivalent to the actual test paper that has been standardized by Malaysian Examination Syndicate.

5.3.2 Methodological Implications

Numerous measurement experts have proclaimed the Rasch model to be the "gold standard" approach for psychometric studies, as it is solely measurement model which have the properties of invariance for objective measurement that overcomes the limitation of the traditional statistical models (Bond & Fox, 2015; Royal, 2010; Royal & Gonzalez, 2016; Salzberger, 2015; Wright, 2005a, 2005b). The present study offers a comprehensive psychometric validation using this state-of-the-art measurement model since the psychometric evidence gathered is viewed as a collective activity and reported at the level of detail as revealed by Arjoon et al. (2013). When attempting to discern validity, the Messick's framework is used to assess validity evidence and make a collective judgement of the construct validity which might be beneficial to other researchers (Royal & Flammer, 2015; Royal & Gonzalez, 2016) In line with Royal and Flammer (2015) and Royal and Gonzalez (2016), McCreary et al. (2013) added that the utilization of Rasch analysis enabled a crucial psychometric analysis beyond what was possible with only classical test theory. As supported by Tennant and

Conaghan (2007), Rasch analysis provides a comprehensive evaluation of the item and scale performance than a classical test theory alone.

Practically, this study provides a useful tool for test developers and teachers to measure how well students understand what has been taught in the classrooms. It is crucial to have insights into what exactly students understand, as conflicting views with teachers may potentially result in inaccurate methods of instruction and interpretation.

An earlier study by McCamey (2014) asserted that the Rasch Model assists test developers to diagnose instruments by calibrating the difficulty and stability of the items to a common scale that is independent of the norm reference group. He also stated that test developers and teachers are able to develop better instruments to optimize the number of items, and eliminate items of the same difficulty, as well as most closely matches the level of difficulty of the items to students' ability with this calibration function. Ideally, it is impossible to construct a truly fair and equitable set of items for all students with different levels of ability.

Particularly, Rasch analysis has created a paradigm shift in assessment and item development due to the valuable data yielded. A quality instrument can be constructed by examining individual test items. The statistical evidence provided by Rasch analysis for instruments can be used in the future.

5.4 Recommendation

The present study was conducted with psychometric analysis in four selected schools using only multiple-choice question paper. Although the results suggested that the Chemistry test possesses high validity and reliability, there may be alternative ways in which the Chemistry test can be improved concerning some items especially, the

distractors and the key answer. As for future studies, the researcher recommends that comparison studies are conducted using demographic factors such as ethnicity, type of school, school location, and state. The collective data can provide a broader analysis of the student's achievements and reinforce the Chemistry's test reliability and validity.

Apart from the psychometric properties, key balancing also contributes to the quality of the test. Key balancing is referred to as the position of the correct answer in a reasonable distribution between all possible options (Towns, 2014). Test developers can revise the test items by performing a frequency count of the correct answers and key balancing. The unbalanced distribution of correct answers can inflate the test scores of students (Bar-Hillel & Attali, 2002) and results in an imprecise measurement of the ability of students. For future studies, the researcher suggests that key balancing is taken into account when conducting a study on the validity and reliability of an instrument.

The Chemistry test comprises of 3 papers namely Paper 1 (multiple-choice question), Paper 2 (subjective question) and Paper 3 (essay question) with different formats. For a comprehensive evaluation of the Chemistry test paper, future studies can be done on Chemistry test papers that consist of the subjective and essay questions by using different methods of the Rasch Model. These methods such as multi-facet, partial-credit model, rating scale model or graded response theory suit the types of data collection obtained. The subjective and essay questions need to be analyzed to obtain the empirical data that can determine the item characteristics and the suitability of the items in accordance with the abilities of students. On top of that, test scores for Papers 2 and Paper 3 should also be analyzed to identify the source of error that can contribute

to the variance and to determine the appropriate number of subjective and essay items in the Chemistry test at the SPM level.

Furthermore, the psychometric analysis conducted was only on the Chemistry subject. Researcher suggests that psychometric analyses should be extended to other subjects, so that the reliability of the scores of students can also be examined for those subjects

For school-level assessment improvement, the researcher recommends several suggestions that need to be taken into account:

a) Establishment of an item bank at school and district levels

The item bank is a system that stores various test items that are encoded based on the subject, level of teaching, measured teaching objectives, and other features of the related item. Each item in the item bank has been tested and has statistical values such as item difficulty value, index discrimination, and item fit. Items are organized by specific topics, types of items such as objective items, structures, and essays. The item bank is run by each subject committee in the school. A test developer only needs to specify the characteristics of the items on the basis of the table of specifications and then select the items that are needed to prepare a test paper. Therefore, by having an item bank for schools and at district level, a set of quality test papers for each subject can be prepared easily.

b) Item Development Course

The school administration organizes courses or workshops on item development so that all teachers understand and have knowledge of assessment theory, concepts, and principles. These courses can guide the subject teachers to develop the assessment skills and evaluate the items sufficiently and finally, produce the instruments that have high quality and are in accordance with the assessment standard.

Teachers should never copy questions from workbooks or practice books for examination or test papers without verifying their psychometric properties.

c) Skill to analyze items based on IRT or CTT

In terms of item analysis, Malaysia Examination Syndicate (MES) and schools are still adopting the Classical Test Theory (CTT). Many literature studies have explained the disadvantages of CTT-based analysis of items. Schools conduct CTT-based item analysis as it is easy to understand and practical. However, the item statistics produced depend on the characteristics of the student group used in the analysis process. The item statistical value yielded will be biased if the sample used is incorrect. Therefore, the skills of item analysis that are based on modern test theory need to be expanded so that relevant information on the psychometric properties of the items can be used to produce a set of quality test papers and reliable scores of students.

5.5 Conclusion

This study was aimed to assess the psychometric properties of the Chemistry test. The psychometric properties of the test items are essential elements for assessing the quality of the test items. In line with He, Liu, Zheng, and Jia (2016), the present study shows how to use the Rasch model for validating the assessment instruments in science education. A comprehensive analysis of items using the Rasch model provides accurate empirical information on the psychometric properties of items rather than raw scores that determines the quality of items. This valuable information benefits the teachers and test developers to determine the functional items and non-functional items in developing a well-constructed test. Karlin and Karlin (2018) found that the only reason for not using Rasch analysis is due to its complicated process.

According to measurement theory, the presence of even a few flawed items can reduce the reliability and validity of the entire test. The non-reliable and non-valid test could not measure the students' understanding and their ability in the content of the subject. It is therefore essential that these flawed items are identified to ensure the tests result are meaningful. Flawed items not only reduce the reliability of the test but also confuse students during the test-taking process.

The psychometric interpretation enables teachers to improvise and modify their instructions according to the abilities of students besides ensuring the appropriate use of a test as a tool of assessment. The result from the thorough analysis designates the Chemistry test possesses good psychometric properties and is capable of yielding valid and reliable scores in measuring the cognitive domain of students. Despite the target of the item particularly well on students' abilities, there are no suitable items to assess students with the highest ability. The Chemistry test only measures mostly on the understanding of students. The quality of the Chemistry items can be enhanced by replacing less functional items or modify the items that have less functional distractors. Besides, more difficult items should be added to the instrument to measure the students with the highest ability.

In particular, the year-end examination of the Chemistry Paper 1 test constructed by the Principals Council of Peninsular Malaysia has almost the same standard as the actual Chemistry Paper 1 test of the Malaysian Certificate of Education produced by the Malaysian Examination Syndicate. However, a few modifications are needed to improve the quality of the test paper. The researcher concludes that an accurate and reliable test result provides valuable information on the progress of students as well as a valid prediction of their achievement and the effectiveness of the pedagogical method.

REFERENCE

Abdul Halim Abdullah, Johari Surif, & Ibrahim., N. H. (2014). *PISA 2012: DI MANA KEDUDUKAN MALAYSIA UNTUK SUBJEK MATEMATIK?* Paper presented at the Prosiding Seminar Antarabangsa Kelestarian Insan 2014 (INSAN2014), Universiti Tun Hussein Onn Malaysia, Johor.

Adams, R., & Khoo, S. (1996). *Quest: The Interactive Test Analysis System*: Australian Council for Educational Research.

Adams, R., Wu, M., Cloney, D., & Wilson, M. (2020). Acer conquest: Generalised item response modelling software [computer software]. Version 5. Camberwell, Victoria: Australian Council for Educational Research.

Adibah Abdul Latif, Ibnatul Jalilah, Nor Amin, Wilfredo Libunao, & Yusri., S. (2016). Multiple-choice items analysis using classical test theory and rasch measurement model. *Man in India, 96*(1-2), 173-181.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, USA: Joint Committee on Standards for Educational Psychological Testing.

Afrassa, T. M. (2005). Monitoring mathematics achievement over time. In? (Ed.), *Applied Rasch Measurement: A Book of Exemplars* (pp. 61-77): Springer.

Ahmad Zamri Khairani, & Nordin Abd. Razak. (2015). Modeling a multiple choice mathematics test with the Rasch model. *Indian Journal of Science and Technology, 8*(12), 1. doi:10.17485/ijst/2015/v8i12/70650

Ainol Mardziah Zubairi, & Noor Lide Abu Kassim. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research (MaJER), 1*(1).

Al-Owidha, A. (2018). Investigating the psychometric properties of the Qiyas for L1 Arabic language test using a Rasch measurement framework. *Language Testing in Asia, 8*. doi:10.1186/s40468-018-0064-5

Alagumalai, S., Curtis, D. D., & Hungi, N. (2006). *Applied Rasch measurement: A book of exemplars: Papers in honour of John P. Keeves*. Netherlands: Springer.

Alavi, S. M., & Bordbar, S. (2017). Differential item functioning analysis of high-stakes test in terms of gender: A Rasch model approach. *Malaysian Online Journal of Educational Sciences, 5*(1), 10-24.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. IL, USA: Waveland Press.

Anastasi, A. (1988). *Psychological testing, 6th ed*. New York, NY, England: Macmillan Publishing Co, Inc.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Rasch Models for Measurement in Educational and Psychological Research Education Research and Perspectives, 9*(1), 95-104.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In? (Ed.), *Sociological methodology 1985.* (pp. 33-80). San Francisco, CA, US: Jossey-Bass.

Andrich, D. (1988). *Rasch models for measurement*. Thousand Oaks, CA, US: Sage Publications, Inc.

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). RUMM 2020 Perth, Australia: RUMM Laboratory.

Andrich, D., & Marais, I. (2019). *A course in rasch measurement theory*. Singapore: Springer.

Andrich, D., & Styles, I. (2011). Distractors with information in multiple choice items: a rationale based on the Rasch model. *J Appl Meas, 12*(1), 67-95.

Anthony, C. J., DiPerna, J. C., & Lei, P. W. (2016). Maximizing measurement efficiency of behavior rating scales using Item Response Theory: An example with the Social Skills Improvement System - Teacher Rating Scale. *Journal of School Psychology, 55*, 57-69. doi:10.1016/j.jsp.2015.12.005

Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education, 90*(5), 536-545. doi:10.1021/ed3002013

Arnold, J. C., Boone, W. J., Kremer, K., & Mayer, J. (2018). Assessment of competencies in scientific inquiry through the application of Rasch measurement techniques. *Education Sciences, 8*(4), 184. doi:10.3390/educsci8040184

Asril, A., & Marais, I. (2011). Applying a Rasch model distractor analysis. In? (Ed.), *Applications of Rasch measurement in learning environments research* (pp. 77-100).

Aziz, A. A., Masodi, M. S., & Zaharim, A. (2013). *Asas model pengukuran Rasch: Pembentukan skala dan struktur pengukuran*. Bangi, Malaysia: Penerbit Universiti Kebangsaan Malaysia.

Aziz Nordin. (2007). *Diagnosis dan pemulihan dalam proses pengajaran dan pembelajaran kimia*. Skudai, Johor: Penerbit Universiti Teknologi Malaysia.

Azizi Yahaya, Peter Voo, Ismail Maakip, & Mohd Dahlan A Malek. (2017). *Kaedah penyelidikan dalam pendidikan*. Tanjung Malim, Malaysia: UPSI.

Azman Wan Chik. (1994). *Pengujian bahasa: Kes Bahasa Melayu*. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Azrilah Abd Aziz, Azlinah Mohamed, Azami Zaharim, Sohaimi Zakaria, Hamzah Ahmad Ghulman, & Mohd Saidfudin Masodi. (2008). *Evaluation of information professionals competency face validity test using Rasch model.* Paper presented at the 5th WSEAS/IASME international conference on Engineering education, Heraklion, Greece.

Azrilah Abd Aziz, M. S. Jusoh, A.R. Omar, Harith Amlus, & Salleh, T. S. A. (2014). Construct validity: A Rasch measurement model approaches. *Journal of Applied Science and Agriculture, 9*(12), 7-12.

Azrilah Abdul Aziz, Azlinah Mohamed, Noor Habibah Arshad, Sohaimi Zakaria, & Mohd Saidfudin Masodi. (2007). *Appraisal of course learning outcomes using Rasch measurement: A case study in information technology education*. Paper presented at the 7th WSEAS InternationalConference on Artificial Intelligence, Knowledge Engineering and Data Bases, University of Cambridge, UK.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford,UK: Oxford University Press

Baghaei, P. (2007). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21*(3), 1105-1106.

Baghaei, P. (2008). The Rasch model as a construct validity tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

Bailey, K. M. (1998). *Learning about Language Assessment: Dilemmas, Decisions, and Directions*: Heinle & Heinle Publishers.

Banta, T. W. (2007). A warning on measuring learning outcomes. *Inside Higher Education*.

Bar-Hillel, M., & Attali, Y. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *The American Statistician, 56*(4), 299-303. doi:10.1198/000313002623

Barchard, K. A., & Brouwers, V. (2016). Internal consistency and power when comparing total scores from two groups. *Multivariate Behavioral Research, 51*(4), 482-494. doi:10.1080/00273171.2016.1166422

Barker, B. A., Donovan, N. J., Schubert, A. D., & Walker, E. A. (2017). Using Rasch analysis to examine the item-level psychometrics of the infant-toddler meaningful auditory integration scales. *Speech Lang Hear, 20*(3), 130-143. doi:10.1080/2050571x.2016.1243747

Bejar, I. I. (2016). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7*(3), 303-310. doi:10.1177/014662168300700306

Ben-Simon, A., Budescu, D. V., & Nevo, B. (2016). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*(1), 65-88. doi:10.1177/0146621697211006

Bereiter, C. (2013). *The Psychology of Written Composition*. New York, USA.

Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Med Educ, 5*(1), 9. doi:10.1186/1472-6920-5-9

Bhasah Abu Bakar. (2003). *Asas pengukuran bilik darjah*. Tanjong Malim: Quantum Books.

Biggs, J. (2006). What the student does: Teaching for enhanced learning. *Higher Education Research & Development, 18*(1), 57-75. doi:10.1080/0729436990180105

Bock, R. D., & Jones, J. V. (1968). *The measurement and prediction of judgment and choice*. Oxford, England: Holden-Day.

Bond, T., & Fox, C. (2001). *Applying the rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Bond, T., & Fox, C. (2015). *Applying the rasch model; fundamental measurement in the human sciences*. New York, USA: Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE life sciences education, 15*(4). doi:10.1187/cbe.16-04-0148

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education, 4*(1), 1416898. doi:10.1080/2331186x.2017.1416898

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*(2), 253-269. doi:10.1002/sce.20106

Boone, W. J., & Staver, J. R. (2020a). Point Measure Correlation. In? (Ed.), *Advances in Rasch Analyses in the Human Sciences* (pp. 25-38). Cham: Springer International Publishing.

Boone, W. J., & Staver, J. R. (2020b). Wright Maps (Part 3 and Counting...). In *Advances in rasch analyses in the human sciences*. Cham, Switzerland: Springer.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. New York, USA: Springer.

Borghans, L., Golsteyn, B. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences of the United States of America, 113*(47), 13354-13359. doi:10.1073/pnas.1601135113

Brinthaupt, T. M., & Kang, M. (2014). Many-faceted Rasch calibration:An example using the self-talk scale. *Assesment, 21*(2), 241-249. doi:10.1177/1073191112446653

Brookhart, S. M., & Nitko, A. J. (2019). *Educational assessment of students*. New York,USA: Pearson.

Brown, J. D. (2000). What is construct validity. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 4,* 8-12.

Brown, S., Race, P., & Smith, B. (2005). *500 Tips on assessment*. Abingdon,UK: Routledge.

Brualdi, A. C. (1998). Implementing performance assessment in the classroom. *Practical Assessment, Research, and Evaluation, 6*. doi:10.7275/kgwx-6q70

Bucat, B., & Mocerino, M. (2009). Learning at the sub-micro level: Structural representations. In J. K. Gilbert & D. F. Treagust (Eds.), *Multiple representations in chemical education* (pp. 11-29). Dordrecht, Netherlands: Springer.

Burton, S. J., Sudweeks, R., Merril, P., & Wood, B. (1991). *How to prepare better multiple choice tests: Guidelines for university faculty*. Provo, Utah: Brigham Young University Testing Services and The Department of Instructional Science.

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics, 36*(5), 648-662. doi:10.1016/j.clinthera.2014.04.006

Cardellini, L. (2012). Chemistry: Why the subject is difficult? *Educación Química, 23*(2), 305-310. doi:10.1016/s0187-893x(17)30158-1

Carlson, J. E., & Von Davier, M. (2017). Item response theory. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 133-178). NY,USA: Springer International Publishing.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Vol. 17). Thousand Oaks, CA: SAGE Publications, Inc. .

Chan, S. W., Zaleha Ismail, & Sumintono, B. (2014). A Rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics. *Procedia - Social and Behavioral Sciences, 129*, 133-139. doi:10.1016/j.sbspro.2014.03.658

Childs, P. E., Hayes, S. M., & O'dwyer, A. (2015). Chemistry and Everyday Life: Relating Secondary School Chemistry to the Current and Future Lives of Students. In? (Ed.), *Relevant Chemistry Education* (pp. 33-54).

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178-194. doi:10.1177/0146621616677520

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26. doi:10.1177/014662169602000102

Cohen, D., & Rhydderch, M. (2010). Making an objective assessment of a colleague's performance. *The Clinical Teacher, 7*(3), 171-174. doi:https://doi.org/10.1111/j.1743-498X.2010.00362.x

Cohen, R., Swerdlik, M. E., & Sturman, E. D. (2013). *Psychological testing and assessment : An introduction to tests and measurement*. New York: McGraw Hill.

Coll, R. K., & Treagust, D. F. (2003). Investigation of secondary school, undergraduate, and graduate learners' mental models of ionic bonding. *Journal of Research in Science Teaching, 40*(5), 464-486. doi:10.1002/tea.10085

Conrad, K. J., Conrad, K. M., Mazza, J., Riley, B. B., Funk, R., Stein, M. A., & Dennis, M. L. (2012). Dimensionality, hierarchical structure, age generalizability, and criterion validity of the GAIN's Behavioral Complexity Scale. *Psychological Assessment, 24*(4), 913-924. doi:10.1037/a0028196

Cordier, R., Speyer, R., Schindler, A., Michou, E., Heijnen, B. J., Baijens, L., . . . Joosten, A. (2018). Using Rasch analysis to evaluate the reliability and validity of the swallowing quality of life questionnaire: An Item Response Theory approach. *Dysphagia, 33*(4), 441-456. doi:10.1007/s00455-017-9873-4

Courville, T. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics.* Georgia Institute of Technology.

Creswell, J. W., & Creswell, D. J. (2017). *Research design:Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: SAGE Publications, Inc.

Creswell, J. W., & Guetterman, T. C. (2018). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, NJ: Pearson.

Crocker, L. M., & Algina, J. (2006). *Introduction to classical and modern test theory*. Boston, MA: Cengage Learning.

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68-80. doi:10.1037/h0029382

Curriculum Development Centre. (2005). *Curriculum specification chemistry form four*. Putrajaya: Ministry of Education Malaysia.

Dani Asmadi Ibrahim, Azraai Othman, & OthmanTalib. (2015). Pandangan pelajar dan guru terhadap tahap kesukaran tajuk-tajuk kimia. *Jurnal Kepimpinan Pendidikan, Vol. 2*, 32-46.

Danielson, C. (2001). New trends in teacher evaluation. *Educational Leadership, 58*(5), 12-15.

Davidowitz, B., & Potgieter, M. (2016). Use of the Rasch measurement model to explore the relationship between content knowledge and topic-specific pedagogical content knowledge for organic chemistry. *International Journal of Science Education, 38*(9), 1483-1503. doi:10.1080/09500693.2016.1196843

De Ayala, R. J. (2013). *The theory and practice of item response theory*. NY,USA: Guilford Publications.

DeMars, C. (2010). *Item response theory*. Oxford, UK: Oxford University Press.

Denny, P., Luxton-Reilly, A., & Simon, B. (2008, 01/01). *Evaluating a new exam question: Parsons problems.* Paper presented at the International Computing Education Research (ICER'08), Sydney, Australia.

DeVellis, R. (2003). *Scale Development: Theory and Applications (Ed.p.1-113)* (Vol. 26).

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning, 2*(2), 4. doi:10.5206/cjsotl-rcacea.2011.2.4

Doody, O., & Doody, C. M. (2015). Conducting a pilot study: case study of a novice researcher. *British Journal of Nursing, 24*(21), 1074-1078. doi:10.12968/bjon.2015.24.21.1074

Downing, S., & Haladyna, T. (2006). *Handbook of test development*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837. doi:10.1046/j.1365-2923.2003.01594.x

Draugalis, J., & Jackson, T. (2004). Objective curricular evaluation: Applying the Rasch model to a cumulative examination. *American Journal of Pharmaceutical Education, 68*(2), 1-12.

Duncan, P. W., Bode, R. K., Min Lai, S., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. *Archives of Physical Medicine and Rehabilitation, 84*(7), 950-963. doi:10.1016/s0003-9993(03)00035-2

Ebel, R., & Frisbie, D. (1991). *Essentials of Educational Measurement*. NJ,US: Prentice Hall.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. Frankfurt, Germany: Peter Lang Publishing Group.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16 Suppl 1*, 5-18. doi:10.1007/s11136-007-9198-0

Edomwonyi-Otu, L., & Abaraham, A. (2011). The challenge of effective teaching of chemistry: A case study. *Leonardo Electronic Journal of Practices and Technologies, 10*(18), 1-8.

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods, 23*(1), 138-149. doi:10.1037/met0000121

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341-349. doi:10.1037/1040-3590.8.4.341

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Engelhard, G. (2013). *Invariant measurement: Using rasch models in the social, behavioral, and health sciences*. New York, USA: Routledge.

Espinosa, A. A. (2014). Analysis of achievement tests in secondary chemistry and biology. *International Journal of Learning, Teaching and Educational Research, 4*(1).

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 1-17. doi:10.1177/0013164498058003001

Fensham, P. J. (1988). *Development and dilemmas in science education*. London, UK: Falmer Press.

Figlio, D. N., & Lucas, M. E. (2004). Do high grading standards affect student performance? *Journal of Public Economics, 88*(9-10), 1815-1834. doi:10.1016/s0047-2727(03)00039-2

Finch, W. H., & French, B. F. (2018). *Educational and psychological measurement*. New York, USA: Routledge.

Fisher, W. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions, 6*(3), 238.

Fisher, W. (2007). Rating scale instrument quality criteria. Retrieved from http://www.rasch.org/rmt/rmt211a.htm

Fives, H., & Barnes, N. (2018). Table of Specifications. In? (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1655-1657). Thousand Oaks,CA: SAGE Publications, Inc.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research, and Evaluation, 18*(3), 1-7. doi:10.7275/CZTT-7109

Frankel, J., Wallen, N., & Hyun, H. (2011). *How to design and evaluate research in education*. New York: McGraw-Hill Education.

Fuchs, L. S., Fuchs, D., & Kazdan, S. (1999). Effects of peer-assisted learning strategies on high school students with serious reading problems. *Remedial and Special Education, 20*(5), 309-318. doi:10.1177/074193259902000507

Ganglmair-Wooliscroft, A., & Lawson, R. (2003). Advantages of Rasch modelling for the development of a scale to measure affective response to consumption. *European Advances in Consumer Research, 6*, 162-167.

Garamendi, E., Pesudovs, K., Stevens, M. J., & Elliott, D. B. (2006). The Refractive Status and Vision Profile: evaluation of psychometric properties and comparison of Rasch and summated Likert-scaling. *Vision Research, 46*(8-9), 1375-1383. doi:10.1016/j.visres.2005.07.007

Gay, L. R., Mills, G. E., & Airasian, P. W. (2009). *Educational research: Competencies for analysis and applications*. New York, USA: Pearson.

Ginty, A. T. (2013). Psychometric properties. In? (Ed.), *Encyclopedia of Behavioral Medicine* (pp. 1563-1564). ?: ?

Gnaldi, M. (2013). Methods of item analysis in standardized student assessment: An application to an Italian case study. *The International Journal of Educational and Psychological Assessment*.

Goodwin, W., & Driscoll, L. A. (1980). *Handbook for measurement and evaluation in early childhood education: Issues, measures, and methods*. San Francisco, CA: Jossey-Bass Publishers.

Gorin, J. S., & Embretson, S. E. (2007). Item response theory and Rasch models. In? (Ed.), *Handbook of Research Methods in Abnormal and Clinical Psychology* (pp. 271-292). New Castle, England: SAGE.

Gracia, S. (2005). Analyzing CSR implementation with the Rasch model. *Faculty Publications*(271).

Gravetter, F. J., & Forzano, L.-A. B. (2018). *Research Methods for the Behavioral Sciences*. Boston, MA: Cengage Learning.

Green, K., & Frantom, C. (2002). *Survey development and validation with the Rasch model*. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston,SC.

Grove, M. J., & Pugh, S. (2015). Is a conceptual understanding of maths vital for chemistry? *Education in Chemistry, 52*(1), 26-29.

Grove, N., & Bretz, S. L. (2012). A continuum of learning: from rote memorization to meaningful learning in organic chemistry. *Chemistry Education Research and Practice, 13*(3), 201-208. doi:10.1039/c1rp90069b

Guskey, T. (2003). How classroom assessments improve learning. *Educational Leadership, 60*(5), 6-11.

Hagquist, C. (2001). Evaluating composite health measures using Rasch modelling: an illustrative example. *Social and Preventive Medicine, 46*(6), 369-378. doi:10.1007/bf01321663

Haladyna, T., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, USA: Routledge.

Haladyna, T. M., & Downing, S. M. (1989). A Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education, 2*(1), 37-50. doi:10.1207/s15324818ame0201_3

Haladyna, T. M., & Downing, S. M. (2016). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999-1010. doi:10.1177/0013164493053004013

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. doi:10.1207/S15324818AME1503_5

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In? (Ed.), *Educational measurement, 3rd ed.* (pp. 147-200). Washington, USA: American Council on Education.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9*(1), 57-68.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229-244.

Hambleton, R. K., Hambleton, T., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*: Springer Netherlands.

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

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

Hamilton, L., & Klein, S. (1999). *Large-Scale Testing: Current Practices and New Directions*. Santa Monica,CA: RAND Corporation.

Hansen, J. D., & Dexter, L. (1997). Quality Multiple-Choice Test Questions: Item-Writing Guidelines and an Analysis of Auditing Testbanks. *Journal of Education for Business, 73*(2), 94-97. doi:10.1080/08832329709601623

Harris, H. H. (2003). Chemical misconceptions-prevention, diagnosis and cure; volume i: Theoretical background; volume II: Classroom resources (Taber, Keith). *Journal of Chemical Education, 80*(5), 491. doi:10.1021/ed080p491.1

Harvey, R. J., & Hammer, A. L. (2016). Item Response Theory. *The Counseling Psychologist, 27*(3), 353-383. doi:10.1177/0011000099273004

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item Response Theory And health outcomes measurement in the 21st century. *Med Care, 38*(9 Suppl), II28-42. doi:10.1097/00005650-200009002-00007

He, P., Liu, X., Zheng, C., & Jia, M. (2016). Using Rasch measurement to validate an instrument for measuring the quality of classroom teaching in secondary chemistry lessons. *Chemistry Education Research and Practice, 17*(2), 381-393. doi:10.1039/c6rp00004e

Heckman, J. J., Humphries, J. E., & Kautz, T. (2014). *The myth of achievement tests: The ged and the role of character in american life*. Chicago, IL: University of Chicago Press.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19*(4), 451-464. doi:10.1016/j.labeco.2012.05.014

Hendriks, J., Fyfe, S., Styles, I., Skinner, S. R., & Merriman, G. (2012). Scale construction utilising the Rasch unidimensional measurement model: A measurement of adolescent attitudes towards abortion. *The Australasian medical journal, 5*(5), 251-261. doi:10.4066/AMJ.2012.952

Henning, G. (2013). *A guide to language testing: Development, evaluation, research*. Scotts Valley, CA: Createspace Independent Pub.

Henning, G. (2016). Dimensionality and construct validity of language tests. *Language Testing, 9*(1), 1-11. doi:10.1177/026553229200900102

Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice, 12*(2), 184-192. doi:10.1039/c1rp90023d

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *Journal Of Pakistan Medical Association, 62*(2), 142-147.

Houston, P., Kearney, R. A., & Savoldelli, G. (2006). The oral examination process - gold standard or fool's gold. *Canadian Journal of Anesthesia, 53*(7), 639-642. doi:10.1007/bf03021620

Howie, S., Long, C., Sherman, V., & Venter, E. (2009). *The role of IRT in selected examination systems*. Pretoria, South Africa: Umalasi Council of Quality Assurance In General aand Further Education and Training.

Huey Fern, L., & Hooi Lian, L. (2017). Pengesahan instrumen sikap terhadap matematik dalam kalangan murid tingkatan empat di Kedah. *Jurnal Kurikulum & Pengajaran Asia Pasifik, 4*(1), 1-13.

Iramaneerat, C., Smith, E. V., & Smith, R. M. (2008). An introduction to rasch measurement. In J. Osborne (Ed.), Best Practices in Quantitative Methods. Thousand Oaks, California: SAGE Publications, Inc. doi:10.4135/9781412995627

Irwing, P., & Hughes, D. J. (2018). Test Development. In? (Ed.), *The Wiley Handbook of Psychometric Testing* (pp. 1-47).

Jackson, S. L. (2006). *Research methods and statistics: A critical thinking approach*. Belmont, CA: Thomson Wadsworth.

Jailani Yunos, Marina Ibrahim Mukhtar, Maizam Alias, Ming Foong Lee, Tze Tee, Siti Nur Kamariah Rubani, . . . Sri Sumarwati. (2017). Validity of vocational pedagogy constructs using the Rasch measurement model. *Journal of Technical Education and Training, 9*(2), 35-45.

Jegede, S. (2007). Students' anxiety towards the learning of Chemistry in some Nigerian secondary schools. *Educational Research and Review, 2*(7), 193-197.

Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer Assisted Learning, 7*(2), 75-83. doi:10.1111/j.1365-2729.1991.tb00230.x

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues*. Toronto, CA: Nelson Education.

Karlin, O., & Karlin, S. (2018). Making better tests with the Rasch measurement model. *Insight: A Journal of Scholarly Teaching, 13*, 76-100. doi:10.46504/14201805ka

Kellaghan, T., & Greaney, V. (2001). *Using assessment to improve the quality of education:Fundamentals of educational planning.* Paris,France: United Nations Educational, Scientific, and Cultural Organization, Paris (France). International Inst. for Educational Planning.

Kementerian Pendidikan Malaysia. (2013). *Malaysia education blueprint, 2013-2025: Preschool to post-secondary education*. Putrajaya: Kementerian Pendidikan Malaysia.

Khamis, M. R., & Che Yahya, N. (2015). Does law enforcement influence compliance behaviour of business zakat among smes?: An evidence via Rasch measurement mode. *Global Journal Al Thaqafah, 5*(1), 19-32. doi:10.7187/gjat752015.05.01

Khine, M. S. (2020). *Rasch Measurement*. Singapore: Springer

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276-2284. doi:10.2146/ajhp070364

King, B. M., Rosopa, P. J., & Minium, E. W. (2018). *Statistical reasoning in the behavioral sciences*. Hoboken, NJ: John Wiley & Sons.

Klooster, P. M., Taal, E., & Van de Laar, M. A. (2008). Rasch analysis of the dutch health assessment questionnaire disability index and the health assessment questionnaire ii in patients with rheumatoid arthritis. *Arthritis & Rheumatism, 59*(12), 1721-1728. doi:10.1002/art.24065

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

Krause, S., & Kelly, J. (2011). Teaching, learning, and assessment resources for introductory materials science and engineering courses. *MRS Online Proceedings Library, 1364*, 409. doi:10.1557/opl.2011.1184

Krishnan, S., & Noraini Idris. (2014). Investigating reliability and validity for the construct of inferential statistics. *International Journal of Learning, Teaching and Educational Research, 4*(1), 51-60.

Lake, J., & Holster, T. (2016). Rasch analysis for dichotomous items. *Journal of Linear and Topological Algebra, 19*, 201-206.

Lange, A., Lehmann, I., & Mehrens, W. (2005). Using item analysis to improve tests. *Journal of Educational Measurement, 4*(2), 65-68. doi:10.1111/j.1745-3984.1967.tb00572.x

Lee, C. (2019). What is item analysis? And other important exam design principles:How item analysis can increase teaching efficacy and assessment accuracy. Retrieved from https://cutt.ly/phVhuy4

Leedy, P. D., & Ormrod, J. E. (2015). *Practical research : planning and design*. London, UK: Pearson.

Lembaga Peperiksaan. (2002). *Format of Chemistry paper for SPM 2003* Putrajaya, Malaysia: Ministry of Education.

Lembaga Peperiksaan. (2009). *Sistem Pentaksiran Pendidikan Kebangsaan*. Putrajaya: Kementerian Pendidikan Malaysia.

Lembaga Peperiksaan. (2019). Pengumuman analisis keputusan Sijil Pelajaran Malaysia 2018 [Press release]

Lewis-Beck, M., Bryman, A., & Liao, T. (2004). *The sage encyclopedia social science research methods*. Thousand Oaks,CA: SAGE Publications,Inc.

Linacre, J. M. (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions, 7*, 328.

Linacre, J. M. (1998). Book Review: The new measurement. *Language Testing, 15*(1), 114-117. doi:10.1177/026553229801500107

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2003). Size vs. Significance: Standardized Chi-Square Fit Statistic. *Rasch Measurement Transactions, 17*(1), 918.

Linacre, J. M. (2004). Assessing statistically and clinically meaningful construct defiency/saturation: Recommended criteria for content coverage and item writing. *Rasch Measurement Transactions, 17*(4), 954-955.

Linacre , J. M. (2004). Test validity and Rasch measurement: Construct, content, etc. *Rasch Measurement Transactions, 18*(1), 970-971.

Linacre, J. M. (2005). *A user's guide to Winsteps/Ministeps Rasch model programs*. Chicago,IL: MESA Press.

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Rasch Measurement Transactions, 20*, 1045-1054.

Linacre, J. M. (2007). Reliability and Separations. A Users Guide to Winsteps/Ministep Rasch–Model Computer Programs In. Chicago: Winsteps. Com.

Linacre, J. M. (2009). Winsteps (Version 3.68). Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2010). Predicting responses from Rasch measures. *J Appl Meas, 11*(1), 1-10.

Linacre, J. M. (2012). A user guide to Winsteps Ministep Rasch model computer programs: Program manual 3.75.0. Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Linacre, J. M. (2015). *Winsteps Rasch Measurement Computer Program User's Guide* (Vol. 29). Beaverton, OR: Winsteps.com.

Linacre, J. M., & Tennant, A. (2009). More about critical eigenvalue sizes in standardized-residual principal components analysis (PCA). *Rasch Measurement Transactions, 23*(3), 1228.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2018). *Statistical techniques in business & economics*. New York, USA: McGraw Hill.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. NJ, USA: Lawrence Erlbaum Associates, Inc.

Ludlow, L. H., & Haley, S. M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement, 55*(6), 967-975. doi:10.1177/0013164495055006005

M. N. Rashidi, R. Ara Begum, Mokhtar, M., & Pereira, J. J. (2014). Pelaksanaan analisis data menggunakan model pengukuran Rasch bagi menentukan wajaran item. *Journal of Advanced Research Design, 2*(1), 1-9.

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: A multivariate outlier detection approach. *Multivariate Behavioral Research, 46*(5), 733-755. doi:10.1080/00273171.2011.606757

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *CSN: General Cognitive Social Science (Topic), 1*(1), 1-11.

Maksy, M. M., & Zheng, L. (2008). Factors associated with student performance in advanced accounting and auditing. *Accounting Research Journal, 21*(1), 16-32. doi:10.1108/10309610810891328

Mallinson, T., Stelmack, J., & Velozo, C. (2004). A comparison of the separation ratio and coefficient alpha in the creation of minimum item sets. *Med Care, 42*(1), I17-24. doi:10.1097/01.mlr.0000103522.78233.c3

Mappiasse, S. (2006). Developing and validating instruments for measuring democratic climate of the civic education classroom and student engagement in North Sulawesi, …. *International Education Journal, 7*, 580-597.

Martin, M. O., Mullis, I. V., Foy, P., & Stanco, G. M. (2012). *Timss 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McCamey, R. (2014). A primer on the one-parameter Rasch Model. *American Journal of Economics and Business Administration, 6*(4), 159-163. doi:10.3844/ajebasp.2014.159.163

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. New York,US: Springer-Verlag.

McCreary, L. L., Conrad, K. M., Conrad, K. J., Scott, C. K., Funk, R. R., & Dennis, M. L. (2013). Using the Rasch measurement model in psychometric analysis of the Family Effectiveness Measure. *Nursing Research, 62*(3), 149-159. doi:10.1097/NNR.0b013e31828eafe6

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*(1), 100-117. doi:10.1111/j.2044-8317.1981.tb00621.x

McMillan, J. H. (2011). *Classroom assessment: Principles and practice for effective standards-based instruction*. London, UK: Pearson.

McNamara, T. (1996). *Measuring Second Language Performance (Applied Linguistics and Language Study)*. London, UK: Longman Pub Group.

Mead, R. (2008). *A Rasch Primer: The Measurement Theory of Georg Rasch*. Maple Grove, MN: Data Recognition Corporation.

Mehrens, W. A., & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart, and Winston.

Mertens, D. (2019). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods* Thousand Oak, CA: SAGE Publications, Inc.

Messick, S. (1989). Validity. In *Educational measurement, 3rd ed.* (pp. 13-103): American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:10.1037/0003-066X.50.9.741

Miller, M. D., Linn, R. L., & Gronlund, N. (2013). *Measurement and assessment in teaching* (Vol. 11th). Upper Saddle River, NJ: Pearson.

Mocerino, M., Chandrasegaran, A. L., & Treagust, D. F. (2009). Emphasizing multiple levels of representation to enhance students' understandings of the changes occurring during chemical reactions. *Journal of Chemical Education, 86*(12), 1433. doi:10.1021/ed086p1433

Mohamad Fauzi Yunus. (1996). *Pembinaan soalan dan analisis item. Kursus peningkatan ilmu dalam pengajaran sejarah*. Kuala Lumpur: Unit Kurikulum, Jabatan Pendidikan Wilayah Persekutuan.

Mohamad Majid Konting. (2005). *Kaedah penyelidikan pendidikan*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Mohd Saidfudin Masodi, Azrilah Abd Aziz, N. A. Rodzo'An, & Omar., M. Z. (2010). *Use of Rasch analysis to measure students performance in engineering education*. Paper presented at the Proceedings of the 7th WSEAS international conference on Engineering education, Corfu Island, Greece.

Muijs, D. (2011). *Doing quantitative research in education with SPSS*. London, GB: SAGE Publications Ltd

Nakhleh, M. B. (1992). Why some students don't learn chemistry: Chemical misconceptions. *Journal of Chemical Education, 69*(3), 191. doi:10.1021/ed069p191

Nazlinda Abdullah, Shereen Noranee, & Mohd Khamis. (2017). The use of Rasch Wright map in assessing conceptual understanding of electricity. *Pertanika Journal of Social Science and Humanities, 25*, 81-88.

Nedungadi, S., Paek, S. H., & Brown, C. E. (2019). Utilizing Rasch analysis to establish the psychometric properties of a concept inventory on concepts important for developing proficiency in organic reaction mechanisms. *Chemistry Teacher International, 2*(2). doi:10.1515/cti-2019-0004

Nevin, E., Behan, A., Duffy, G., Farrell, S., Harding, R., Howard, R., . . . Bowe, B. (2015). *Assessing the validity and reliability of dichotomous test results using Item Response Theory on a group of first year engineering students*. Paper presented at the The 6th Research in Engineering Education Symposium (REES 2015), Dublin, Ireland.

Ng, A., & Chan, A. (2009). Different methods of multiple-choice test: Implications and design for further research. *Lecture Notes in Engineering and Computer Science, 2175*.

Nor Hasnida Che Md Ghazali. (2016). The implementation of school-based assessment system in Malaysia: A study of teacher perceptions. *Geografia: Malaysian journal of society and space, 12*(9).

Noraini Idris. (2010). *Penyelidikan dalam pendidikan*: McGraw Hill (Malaysia).

Nordin Abd. Razak, Ahmad Zamri Khairani, & Thien, L. M. (2012). Examining quality of mathemtics test items using Rasch model: Preminarily analysis. *Procedia - Social and Behavioral Sciences, 69*, 2205-2214. doi:10.1016/j.sbspro.2012.12.187

Norhayati Mohd Noor, Fatin Imtithal Adnan, & Nor Akma Mat Junoh. (2020). Psychometric properties of the Malay version of the Women's Views of Birth Labour Satisfaction Questionnaire using the Rasch measurement model: a cross sectional study. *BMC Pregnancy and Childbirth, 20*(1), 295. doi:10.1186/s12884-020-02975-z

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, 31*(2), 115-129.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Nurul Awanis Abdul Wahid, Hazlina Abdul Hamid, Stephanie Low, & Zariyawati Mohd Ashhari. (2011, 14-15 January 2011). *Malaysian education system reform: Educationists perspectives.* Paper presented at the International Conference on Social Science, Economics and Art 2011, Hotel Equatorial Bangi-Putrajaya, Malaysia.

Nuttall, D. L. (1987). The validity of assessments. *European Journal of Psychology of Education, 2*(2), 109-118. doi:10.1007/bf03172641

Organization for Economic Cooperation and Development Global Science Forum. (2006). *Evolution of student interest in science and technology studies policy report*. Retrieved from http://www.oecd.org/science/inno/36645825.pdf:

Osborne, J. (2008). *Best Practices in Quantitative Methods*. In. Retrieved from https://methods.sagepub.com/book/best-practices-in-quantitative-methods doi:10.4135/9781412995627

Osteen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research, 1*(2), 66-82. doi:10.5243/jsswr.2010.6

Othman Talib. (2013). *Asas penulisan: tesis penyelidikan & statistik*. Serdang, Selangor: Penerbit Universiti Putra Malaysia.

Pada, A. U. T., Kartowagiran, B., & Subali, B. (2016). Separation index and fit items of creative thinking skills assessment. *Research and Evaluation in Education, 2*(1), 1-12. doi:10.21831/reid.v2i1.8260

Pae, H. K. (2011). *Research note: Differential item functioning and unidimensionality in the Pearson test of English academic*. New York,USA: Pearson Education Ltd.

Pallant, J. (2007). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows Version 15*. Maidenhead,UK: Open University Press.

Pang, V., & Lajium, D. (2008). Pengukuran dan penilaian dalam latihan mengajar. In? (Ed.), *Pengetahuan pedagogi guru* (pp. 161-174). Sabah, Malaysia: Penerbit UMS.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65-77. doi:10.1016/j.pse.2014.11.002

Pesudovs, K., Garamendi, E., Keeves, J. P., & Elliott, D. B. (2003). The Activities of Daily Vision Scale for cataract surgery outcomes: re-evaluating validity with Rasch analysis. *Investigative Ophthalmology & Visual Science, 44*(7), 2892-2899. doi:10.1167/iovs.02-1075

Picardi, C. A., & Masick, K. D. (2014). *Research Methods: Designing and Conducting Research With a Real-World Focus*. Thousand Oaks,CA: SAGE Publication,Inc.

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. New York, USA: Routledge.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement: Issues and Practice, 33*(4), 4-12. doi:10.1111/emip.12045

Pollard, A., & Triggs, P. (2000). *What pupils say: Changing policy and practice in primary education*. New York,USA: Routledge.

Popham, W. J. (2020). *Classroom assessment : What teachers need to know* (9th ed.). New York, USA: Pearson.

Portney, L. G., & Watkins, M. P. (2013). *Foundations of clinical research: Pearson new international edition: Applications to practice*. New York, USA: Pearson Education Limited.

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and quality of life outcomes, 1*(1), 27. doi:10.1186/1477-7525-1-27

Program Committee of the Institute for Objective Measurement. (2000). Definition of Objective Measurement. Retrieved from https://www.rasch.org/define.htm

Quaigrain, K., Arhin, A. K., & King Fai Hui, S. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1). doi:10.1080/2331186x.2017.1301013

Radhakrishna, R. B. (2007). Tips for developing and testing questionnaires/instruments. *Journal of Extension, 45*(1).

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*.

Ranga, J. S. (2018). ConfChem Conference on Mathematics in undergraduate Chemistry instruction: Impact of quick review of Math concepts. *Journal of Chemical Education, 95*(8), 1430-1431. doi:10.1021/acs.jchemed.8b00070

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Rasch, G. (1961). *On General Laws and the Meaning of Measurement in Psychology.* Paper presented at the IV Berkeley Symposium on Mathematical Statistics and Probability.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago,IL: University of Chicago Press.

Reckase, M. D. (2009). Multidimensional item response theory models. In? (Ed.), *Multidimensional item response theory* (pp. 79-112). ?: Springer.

Reckase, M. D. (2016). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230. doi:10.3102/10769986004003207

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. In? (Ed.), *Assessing Quality of Life in Clinical Trials: Methods and Practice* (pp. 55-73). Oxford: Oxford University Press.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care, 45*(5 Suppl 1), S22-31. doi:10.1097/01.mlr.0000250483.85507.04

Reise, S. P. (2016). A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT. *Applied Psychological Measurement, 14*(2), 127-137. doi:10.1177/014662169001400202

Reise, S. P., & Haviland, M. G. (2005). Item Response Theory and the measurement of clinical change. *Journal of Personality Assessment, 84*(3), 228-238. doi:10.1207/s15327752jpa8403_02

Reise, S. P., & Waller, N. G. (2009). Item Response Theory and clinical measurement. *Annual review of clinical psychology, 5*, 27-48. doi:10.1146/annurev.clinpsy.032408.153553

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566. doi:10.1037/0033-2909.114.3.552

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2th ed.). New Jersey, NJ: Allyn & Bacon/Pearson Education.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13. doi:10.1111/j.1745-3992.2005.00006.x

Rodriguez, M. C., Kettler, R., & Elliott, S. (2014). Distractor Functioning in Modified Items for Test Accessibility. *SAGE Open, 4*. doi:10.1177/2158244014553586

Rohana, Y. (2004). *Penyelidikan sains sosial*. Bentong, Pahang: PTS Publications & Distributors.

Roseni Din, Ahmad, M., Zaman, M. F., Sidek, N. M., Karim, A. A., Johar, N. A., . . . Ariffin, S. R. (2009). Kesahan dan kebolehpercayaan soal selidik gaya e-Pembelajaran (eLSE) versi 8.1 menggunakan model pengukuran Rasch. *Journal of Quality Measurement and Analysis, 5*(2), 15-27.

Royal, K. (2010). Making Meaningful Measurement in Survey Research: A Demonstration of the Utility of the Rasch Model. *IR Applications, 28*, 1-16.

Royal, K., & Flammer, K. (2015). Measuring Academic Misconduct: Evaluating the Construct Validity of the Exams and Assignments Scale. *American Journal of Applied Psychology, 4*, 58-64. doi:10.11648/j.ajap.s.2015040301.20

Royal, K., & Gonzalez, L. (2016). An evaluation of the psychometric properties of an advising survey for medical and professional program students. *Journal of Educational and Developmental Psychology, 6*(1). doi:10.5539/jedp.v6n1p195

Rozeha A.Rasyid, Azami Zaharim, & Mohd Saidfudin Masodi. (2007). *Application of Rasch measurement in evaluation of learning outcomes: A case study in electrical engineering.* Paper presented at the Regional Conference on Engineering Mathematics, Mechanics, Manufacturing & Architecture (EM3ARC).

Rubiah Sidin, Juriah Long, Khalid Abdullah, & Puteh Mohamed. (2001). Pembudayaan sains dan teknologi: Kesan pendidikan dan latihan di kalangan belia di malaysia. *Jurnal Pendidikan, 27*, 35-45.

Ruhaiza Rusmin. (2015, 27 April 2015). Tarik pelajar minat Sains, Matematik. *Harian Metro*.

Runnels, J. (2012). Using the Rasch model to validate a multiple choice English achievement test. *International Journal of Language Studies, 6*(4), 141-153.

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management, 54*(2), 189-203. doi:10.1016/j.im.2016.06.005

Sadia Mahzabin, Roszilah Hamid, & Shahrizan Baharom. (2015). Rasch model approach for final examination questions construct validity of two successive cohorts. *Journal of Engineering Science and Technology*, 42-52.

Salkind, N. (2010). *Encyclopedia of Research Design* (Vol. 1-10). Thousand Oaks, California.

Salkind, N., & Rasmussen, K. (2008). *Encyclopedia of educational psychology*. Thousand Oaks, CA: SAGE Publications.

Salzberger, T. (2015). Rasch Model. In *Wiley Encyclopedia of Management* (pp. 1-2).

Santelices, M. V., & Wilson, M. (2010). Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning. *Harvard Educational Review, 80*(1), 106-134. doi:10.17763/haer.80.1.j94675w001329270

Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students*. London, UK: Pearson Education.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*(4), 453-472. doi:10.1023/a:1003196224280

Seldomridge, L. A., & Walsh, C. M. (2006). Evaluating student performance in undergraduate preceptorships. *Journal of Nursing Education, 45*(5), 169-176. doi:10.3928/01484834-20060501-06

Sharifah Nurulhuda Tuan Mohd Yasin, Mohd Fauzi Mohd Yunus, & Izwah Ismail. (2018). The use of rasch measurement model for the validity and reliability. *Journal of Counseling and Educational Technology, 1*(2), 22-27. doi:10.32698/0111

Sick, J. (2010). Assumptions and requirements of Rasch measurement. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 14*(2), 23-29.

Sick, J. (2011). Rasch measurement and factor analysis. *SHIKEN: JALT Testing & Evaluation SIG Newsletter, 15*(1), 15-17.

Sidek Mohd Noah. (2003). *Reka bentuk penyelidikan : Falsafah, teori dan praktis : Sebuah buku mesra pengguna*. Serdang, Selangor: Penerbit Universiti Putra Malaysia.

Siew, N. M., & Mohammad Syafiq Abd Rahman. (2019). Assessing the validity and reliability of the future thinking test using rasch measurement model. *International Journal of Environmental and Science Education, 14*(4), 139-149

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: SAGE Publications.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. doi:10.1111/j.1745-3984.1991.tb00356.x

Sirhan, G. (2007). Learning difficulties in chemistry: An overview. *Journal of Turkish Science Education, 4*(2).

Siti Aminah Osman, Syahdatul Isnain Naam, Othman Jaafar, & Wan Hamidon Wan Badaruzzaman. (2012). Application of Rasch model in measuring students' performance in civil engineering design II course. *Procedia - Social and Behavioral Sciences, 56*, 59-66. doi:10.1016/j.sbspro.2012.09.632

Siti Rahayah Ariffin. (2008). *Inovasi dalam pengukuran dan penilaian pendidikan*. Bangi, Malaysia: Fakulti Pendidikan Universiti Kebangsaan Malaysia.

Siti Rahayah Ariffin, Bishanani Omar, Anita Isa, & Sharida Sharif. (2010). *Validity and reliability multiple intelligent item using rasch measurement model.* Paper presented at the World Conference On Learning, Teaching and Administration, The American University Cairo Egypt.

Siti Rahayah Ariffin, Rodiah Idris, & Noriah Mohd Ishak. (2010). Differential item functioning in Malaysian generic skills instrument (MyGSI). *Journal of Education, 32*(1), 1-10.

Siti Salbiah Omar, Jamalludin Harun, Johari Surif, Noor Dayana Abd Halim, & Suraiya Muhamad. (2016). A pilot study on chemistry achievement test. *Journal of Global Business and Social Entrepreneurship, 2*(3), 119-129.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*(1), 33. doi:10.1186/1471-2288-8-33

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *J Appl Meas, 2*(3), 281-311. doi:10.1680/cehlattv.28760.0010

Smith, E. V., Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas, 3*(2), 205-231.

Smith, G. T. (2005). On construct validity: issues of method and measurement. *Psychological Assessment, 17*(4), 396-408. doi:10.1037/1040-3590.17.4.396

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 300-308. doi:10.1037/1040-3590.7.3.300

Smith, R. M. (2004). Fit analysis in latent trait measurement models. *J Appl Meas, 1*, 73-92.

Stelmack, J., Szlyk, J. P., Stelmack, T., Babcock-Parziale, J., Demers-Turco, P., Williams, R. T., & Massof, R. W. (2004). Use of Rasch person-item map in exploratory data analysis: a clinical perspective. *Journal of Rehabilitation Research & Development, 41*(2), 233-241. doi:10.1682/jrrd.2004.02.0233

Streiner, D. L. (2003). Being inconsistent about consistency: when coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*(3), 217-222. doi:10.1207/S15327752JPA8003_01

Sulis, I., & Toland, M. D. (2016). Introduction to Multilevel Item Response Theory Analysis. *The Journal of Early Adolescence, 37*(1), 85-128. doi:10.1177/0272431616642328

Sumintono, B. (2018). *Rasch Model Measurements as Tools in Assesment for Learning.* Paper presented at the 1st International Conference on Education Innovation (ICEI 2017).

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model rasch untuk penelitian ilmu-ilmu sosial (edisi revisi)*. Bandung: Trim Komunikata Publishing House.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Bandung: Trim Komunikata Publishing House.

Sun, Y., & Chen, W. (2009). The relationship between teaching comprehensibility, and instructional time vs. Students' achievement in rational numbers. *The Journal of Human Resource and Adult Learning, 5*(2), 99-107.

Susongko, P. (2016). Validation of science achievement test with the Rasch model. *Indonesian Journal of Science Education, 5*(2), 268-277.

Taber, K. S. (2002). *Alternative conceptions in chemistry: Prevention, diagnosis and cure?* London,UK: The Royal Society of Chemistry.

Taber, K. S. (2013). Revisiting the chemistry triplet: drawing upon the nature of chemical knowledge and the psychology of learning to inform chemistry education. *Chemistry Education Research and Practice, 14*(2), 156-168. doi:10.1039/c3rp00012e

Taber, K. S. (2019). Chapter 1 The Challenge of Teaching and Learning Chemical Concepts. In *The Nature of the Chemical Concept: Re-constructing Chemical Knowledge in Teaching and Learning* (pp. 1-13): The Royal Society of Chemistry.

Talanquer, V. (2011). Macro, Submicro, and Symbolic: The many faces of the chemistry "triplet". *International Journal of Science Education, 33*, 179-195. doi:10.1080/09500690903386435

Tan, X., & Michel, R. (2011). *Why do standardized testing programs report scaled scores? Why not just report the raw or percent-correct scores?* Retrieved from https://bit.ly/2KmOXlL:

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Med Educ, 9*(1), 40. doi:10.1186/1472-6920-9-40

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. doi:10.5116/ijme.4dfb.8dfd

Tavares, H., Andrade, D., & Pereira, C. (2004). Detection of determinant genes and diagnostic via Item Response Theory. *Genetics and Molecular Biology - GENET MOL BIOL, 27*. doi:10.1590/S1415-47572004000400033

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism, 57*(8), 1358-1362. doi:10.1002/art.23108

Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. *J Rehabil Med, 35*(3), 105-115. doi:10.1080/16501970310010448

Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: Assessing its relation with item cognitive level classified according to Bloom's Taxonomy. *Frontiers in psychology, 9*, 1585. doi:10.3389/fpsyg.2018.01585

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-Choice Models: The Distractors Are Also Part of the Item. *Journal of Educational Measurement, 26*(2), 161-176. doi:10.1111/j.1745-3984.1989.tb00326.x

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260. doi:10.1111/j.1745-3984.1989.tb00331.x

Thompson, B., & Levitov, J. E. (1985). Using microcomputers to score and evaluate items. *Collegiate Microcomputer, 3*(2), 163-168.

Tinsley, H. E., & Dawis, R. V. (1977). Test-free person measurement with the Rasch simple logistic model. *Applied Psychological Measurement, 1*(4), 483-487. doi:10.1177/014662167700100404

Towns, M. H. (2014). Guide To Developing High-Quality, Reliable, and Valid Multiple-Choice Assessments. *Journal of Chemical Education, 91*(9), 1426-1431. doi:10.1021/ed500076x

Tran, V. D., Dorofeeva, V. V., & Loskutova, E. E. (2018). Development and validation of a scale to measure the quality of patient medication counseling using Rasch model. *Pharmacy Practice (Granada), 16*(4), 1327. doi:10.18549/PharmPract.2018.04.1327

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159-169. doi:10.1080/0950069880100204

Treagust, D. F., Duit, R., & Nieswandt, M. (2000). Sources of students difficulties in learning Chemistry. *Educación Química, 11*, 228-235. doi:10.22201/fq.18708404e.2000.2.66458

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ, US: John Wiley & Sons Inc.

Uttley, J. (2019). Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research. *LEUKOS, 15*(2-3), 143-162. doi:10.1080/15502724.2018.1533851

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York,USA: Springer.

Van Driel, J. H., Jong, O. D., & Verloop, N. (2002). The development of preservice chemistry teachers' pedagogical content knowledge. *Science Education, 86*(4), 572-590. doi:10.1002/sce.10010

Veloo, A. (2011). Keupayaan teori dan pelaksanaan pentaksiran dalam pembelajaran. *Journal of Governance and Development, 7*, 8-15.

Veloo, A., Rahimah Nor, & Rozalina Khalid. (2015). Attitude towards physics and additional mathematics achievement towards physics achievement. *International Education Studies, 8*(3), 35-43. doi:10.5539/ies.v8n3p35

Wainer, H., & Thissen, D. (2005). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*(1), 22-29. doi:10.1111/j.1745-3992.1996.tb00803.x

Wale, C. M. (2013). *Evaluation of the effect of a digital mathematics game on academic achievement.* University Of Northern Colorado, Greeley, Colorado.

Wang, W. C. (2008). Assessment of differential item functioning. *J Appl Meas, 9*(4), 387-408.

Ward, H., Roden, J., Hewlett, C., & Foreman, J. (2005). *Teaching Science in the Primary Classroom*. Thousand Oaks,CA: SAGE Publications Ltd.

Ware, S. A. (2001). Teaching chemistry from a societal perspective. *Pure and Applied Chemistry, 73*(7), 1209-1214. doi:10.1351/pac200173071209

Waugh, C. K., & Gronlund, N. E. (2013). *Assessment of student achievement*. New York,USA: Pearson.

Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using Rasch Measurement To Develop a Computer Modeling-Based Instrument To Assess Students' Conceptual Understanding of Matter. *Journal of Chemical Education, 89*(3), 335-345. doi:10.1021/ed100852t

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. Retrieved from https://cutt.ly/ihX93xE:

Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test*. Umeå, Sweden: Department of Educational Measurement, Umeå University.

Williams, C. (2007). Research Methods. *Journal of Business & Economic Research – March, 5*. doi:10.19030/jber.v5i3.2532

Winarti, A., & Mubarak, A. (2019). Rasch Modeling: A Multiple Choice Chemistry Test. *Indonesian Journal on Learning and Advanced Education (IJOLAE), 2*(1), 1-9. doi:10.23917/ijolae.v2i1.8985

Wolfe, E., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I-Instrument development tools. *J Appl Meas, 8*, 97-123.

Wright, B. D. (1977). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement, 14*(2), 97-116. doi:10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D. (1992). Raw scores are not linear measures: Rasch vs. classical test theory CTT comparison. *Rasch Measurement Transactions, 6*(1), 208.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*(3), 509-511.

Wright, B. D. (2005a). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45. doi:10.1111/j.1745-3992.1997.tb00606.x

Wright, B. D. (2005b). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116. doi:10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(370).

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL.: MESA press.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In? (Ed.), *Introduction to rasch measurement: Theory, models and applications* (pp. 1-24). Maple Grove, MN.

Wright, B. D., & Stone, M. (1999). *Measurement essentials*. Wilmington, Delaware: Wide Range, Inc.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wu, C., & Foos, J. (2010). Making Chemistry Fun to Learn. *Literature Information Computer Education Journal, 1*(1), 3-7. doi:10.20533/licej.2040.2589.2010.0001

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne,Australia: Educational Measurement Solutions.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Classical test theory. In? (Ed.), *Educational measurement for applied researchers* (pp. 73-90). Singapore: Springer.

Xu, Y., & Liu, Y. (2009). Teacher Assessment Knowledge and Practice: A Narrative Inquiry of a Chinese College EFL Teacher's Experience. *TESOL Quarterly, 43*(3), 493-513.

Yee, A. L. S., Fah, L. Y., & Ling, M.-T. (2018). *Rasch Analysis Properties of a Chemistry Test for Form Four Students.* Paper presented at the Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings.

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement, 30*(3), 187-213. doi:10.1111/j.1745-3984.1993.tb00423.x

Zamalia Mahmud, Nor Azura Md Ghani, & Rosli A. Rahim. (2013). Assessing Students' Learning Ability in a Postgraduate Statistical Course: A Rasch Analysis. *Procedia - Social and Behavioral Sciences, 89*, 890-894. doi:10.1016/j.sbspro.2013.08.951

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica, 29*(1), 18. doi:10.1186/s41155-016-0040-x

Zenisky, A., Hambleton, R. K., & Sireci, S. (2001). Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics. *MCAT Monograph, 5*.

Zoller, U. (1990). Students' misunderstandings and misconceptions in college freshman chemistry (general and organic). *Journal of Research in Science Teaching, 27*(10), 1053-1065. doi:10.1002/tea.3660271011

Zou, S. (2005). *Language Testing*. Shanghai, CH: Shanghai Foreign Language Education Press.

Zulkifli Mohd Nopiah, Mohd Haniff Osman, Noorhelyna Razali, & Izamarlina Asshaari. (2010). *How good was the test set up? From Rasch Analysis Perspective.* Paper presented at the Regional Conference on Engineering Education & Research in Higher Education, Kuching, Malaysia.

Zulkifli Mohd Nopiah, Mohd Helmi Jamalluddin, Nur Arzilah Ismail, Haliza Othman, Izamarlina Asshaari, & Mohd Hanif Osman. (2012). Reliability analisis of examination questions in a mathematics course using Rasch measurement model. *Sains Malaysiana, 41*(9), 1171-1176.

Zunita Maskor, & Harun Baharudin. (2019). Assessing psychometric properties of malaysian secondary school students' arabic vocabulary knowledge inventory. *Global Journal Al-Thaqafah*, 147.