DEVELOPING RASCH-CALIBRATED ITEM BANK USING COMPUTERIZED ADAPTIVE TEST (CAT) TO MEASURE SECONDARY SCHOOL STUDENTS' PERFORMANCE IN MATHEMATICS

SARAH EMILIA BINTI MERZA

UNIVERSITI SAINS MALAYSIA

2025

DEVELOPING RASCH-CALIBRATED ITEM BANK USING COMPUTERIZED ADAPTIVE TEST (CAT) TO MEASURE SECONDARY SCHOOL STUDENTS' PERFORMANCE IN MATHEMATICS

by

SARAH EMILIA BINTI MERZA

Thesis submitted in fulfillment of the requirements
for the Degree of
Doctor of Philosophy

January 2025

ACKNOWLEDGEMENT

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

Alhamdulillah. I am grateful for all the bounties that Allah has showered on me which enable me to complete this thesis. Alhamdulillah. Thank you Allah for providing me with supportive family members and friends during my study.

I would like to express my appreciation to all individuals without whom the completion of this thesis would not be possible. First of all, my heartfelt thanks go to my main supervisor, Professor Madya Dr Mohd Ali Bin Samsudin for his continual support, patience, and perseverance in guiding me through the entire process and for his invaluable academic advice. My deepest thanks also go to my co-supervisor Dr Muzirah Binti Musa for her invaluable assistance.

I would also like to express my particular thanks to all the headmasters and teachers from various schools that were directly or indirectly involved in this research. Many thanks also go to the faculty and administrative staff of the School of Educational Studies and the Institute of Postgraduate Studies for their assistance and support.

Last but not least, my affectionate thanks to my family members, especially my parents, for their continual understanding, prayers, and selfless support. Words are inadequate to express my gratitude for their sacrifice, support, and patience. I would also like to thank my cats for all the entertainment and emotional support. May Allah reward and bless all my family members and all personnel who helped me during this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
CAT	Computerized Adaptive Testing
CBT	Computer Based Testing
CTT	Classical Test Theory
CVI	Content Validity Index
IMMS	Instructional Materials Motivation Survey
IRT	Item Response Theory
MNSQ	Mean Square
OIB	Ordered Item Booklet
PBT	Paper-based esting
PCA	Principal Components Analysis
PISA	Programme for International Student Assessment
PLD	Performance Level Descriptors
PLS-SEM	Partial Least Squares Structural Equation Modeling
PT3	Pentaksiran Tingkatan Tiga
SBA	School Based Assessment
SPM	Sijil Pelajaran Malaysia
TIMSS	Trends in International Mathematics and Science Study
UPSR	Ujian Pencapaian Sekolah Rendah

LIST OF APPENDICES

MEMBANGUNKAN ITEM BANK BERKALIBRASI MENGGUNAKAN RASCH MODEL UNTUK UJIAN BERKOMPUTER SECARA ADAPTIF BAGI MENGUKUR PENCAPAIAN PELAJAR SEKOLAH MENENGAH DALAM MATEMATIK

ABSTRAK

Kajian ini bertujuan untuk membangunkan bank item berkalibrasi dalam ujian berkomputer secara adaptif bagi mengukur pencapaian pelajar sekolah menengah dalam matematik. Satu standard pencapaian untuk menilai pencapaian pelajar setelah menggunakan ujian berkomputer secara adaptif juga ditentukan dan ciri-ciri ujian berkomputer secara adaptif berdasarkan model ARCS diperiksa. Kajian ini dijalankan dalam empat peringkat merangkumi (1) pembangunan bank item, (2) pembangunan ujian adaptif berkomputer, (3) penetapan standard pencapaian pelajar dan (4) menganalisa ARCS model. Model pengukuran Rasch digunakan dalam pembangunan bank item dimana 600 soalan matematik dikalibrasi menggunakan kaedah penyamaaan dan penyataraan menggunakan item setara. Sebanyak 334 item didapati memenuhi ciri-ciri psikometrik dalam membangunkan bank item berkalibrasi bagi ujian berkomputer secara adaptif. Kesukaran item diperolehi akan digunakan sebagai parameter dalam membangunkan ujian berkomputer secara adaptif pada peringkat kedua. Ujian berkomputer secara adaptif dibangunkan mengunakan platform Concerto dan ditadbirkan kepada 382 pelajar sekolah menengah dari 6 buah sekolah. Pencapaian matematik dianalisa berdasarkan skor sempadan yang diperoleh pada peringkat ketiga. Skor sempadan ini diperolehi hasil daripada kajian tetapan standard menggunakan kaedah Bookmark. Empat aras pencapaian iaitu Bawah Asas, Asas, Mahir dan Lanjut telah dihasilkan berdasarkan skor sempadan. Pada pengingkat terakhir, model ARCS dianalisa menggunakan PLS-SEM untuk mengkaji ciri-ciri ujian berkomputer secara

adaptif dalam mempengaruhi motivasi pelajar. Dapatan kajian mendapati pelajar berpuas hati dengan penggunaan ujian berkomputer secara adaptif dan ujian berkomputer secara adaptif mampu meningkatkan motivasi pelajar untuk mempelajari matematik.

DEVELOPING RASCH-CALIBRATED ITEM BANK USING COMPUTERIZED ADAPTIVE TEST (CAT) TO MEASURE SECONDARY SCHOOL STUDENTS' PERFORMANCE IN MATHEMATICS

ABSTRACT

The purpose of this study was to develop a calibrated item bank in computerized adaptive test for measuring students' performance in mathematics. The standard performnce for CAT was also determined and the characteristics of CAT based on ARCS model were examined. This study was conducted in four stages including (1) developing an item bank, (2) developing a CAT, (3) setting a performance standard and (4) analysing ARCS model. Rasch measurement model was used in the development of item bank where 600 mathematics item were calibrated using concurrent comment item equating method. A total of 334 items were found fulfilled the psychometrics properties in developing an item bank for CAT. The item difficulty for the items was used as item parameter in developing a CAT on platform in the second stage. The CAT was developed on Concerto platform and adminitered to 381 secondary students from 6 schools. The students' mathematics performance was analysed based on the recommended cutscores obtained in stage three. In stage three, the recommended cutscores for mathematics performance using CAT was acquired from a standard setting study using Bookmark method. Four performance levels, namely, Below Basic, Basic, Proficent and Advanced level was yielded from three recommended cutscores. In final stage, the ARCS model were analyzed using partial least square structural equation modeling (PLS-SEM) to examined the characteristisc of CAT. Findings found that students were satisfied after the used of CAT and CAT motivates secondary students to learn mathematics.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Assessment is the process of evaluating students' performance on certain knowledge or subject. Assessment is one of the most important methods that helps students in the process of learning and motivation. The quality of assessment can be elevated by integration of technology (Wall, 2004). The role of technology in education cannot be denied, especially incorporating the power of technology to promote students' learning (Al Musawi, 2011). For example, CTB/McGraw-Hill, Educational Testing Service (ETS) and many other companies have developed an online automated writing evaluation that employs artificial intelligence to evaluate essays and offer feedback. This integration and incorporation of technology and education has been widely used as a tool that helps students in assessment like GRE and TOEFL (Burstein et al., 2013) and in the classroom (Grimes & Warschauer, 2010). Studies have shown that students have a positive attitude toward the use of technology in education and it helps them build up their confidence (Tang et al., 2012; Grimes & Warschauer, 2010).

Further, modern computer technology has opened several new possibilities for conducting advanced psychometric analyses and optimizing the administration of educational and psychological tests. Technology-based assessment is the practice of using electronic devices like computers instead of paper to administer tests. Computer based testing (CBT) refers to administering tests or assessments with computer as an alternative medium instead of on paper.

For example, the Test of English for International Communication (TOEIC) is an international standardized English test for non-native speakers. In Malaysia, the TOIEC Listening & Reading Test is conducted using paper-based testing (PBT) method while the TOIEC Speaking & Writing Test is conducted using computer-based testing method where the test is pre-scheduled and online. The CBT is offered in larger cities while the PBT is more common in smaller cities and towns (TOEIC Malaysia, 2020). Computerized adaptive test (CAT) is an algorithm-based assessment that administers the most relevant test items based on the test takers' responses and psychometrics properties. CAT depends on the item parameter information obtained from the calibration process using item response theory (IRT) which is the most common psychometric approach in CAT (Linacre, 2000). A calibrated item bank is one of the core elements in the development of a CAT assessment where the calibrated item bank is comparable as the items are on the same scale (Li & Wu, 2022; Chuesathuchon & Waugh, 2010; Hambleton et al., 1991) and Rasch model measures item parameters and person ability simultaneously (Bond & Fox, 2015). Administrating assessment using CAT saves time and reduces the number of items needed to be administered without affecting the measurement quality (Weiss, 2004; Brown & Weiss, 1977). This approach has more advantages over the PBT approach, and it increases students' motivation to learn (Ross, et al, 2018; Wainer, 1990).

In education, performance is a common measurement tool used to assess the level of proficiency in certain knowledge. Mathematics performance can be defined as proficiency in mathematics or competency in mathematics. Mathematics is an important subject as the knowledge gained in mathematics is used in other fields like engineering, technology, science and even medicine.

Mathematics is an area of knowledge that includes the topics of arithmetic (numbers), algebra (formulas), geometry (shapes and space) and statistics (numerical data) (Ziegler & Loos, 2014). These topic areas are interrelated as seen from the curriculum. However, learning mathematics can be difficult as students perceived mathematics as a difficult and boring subject (Gafoor & Kurukkan, 2015; Brown et al, 2008) and this leads to an unenthusiastic attitude towards learning mathematics (Puteh & Khalin, 2016; Zakaria et al., 2012, Mohamed & Tarmizi, 2010). Most students perceived algebra as a difficult area in comparison to other topic areas (Gafoor & Kurukkan, 2015). This could be due to its abstract nature and the use of variables such as algebraic expressions. Students were found to have difficulties in visualizing algebraic forms (Muchoko et al., 2019) and understanding the meaning of variables (Pramesti & Retnawati, 2019).

There are many kinds of educational assessment tools such as observations, interviews, and assessments. Formative assessment focuses on the mastery of content that empowers students to learn (Wang et al., 2006) and digital formative assessment used in classrooms increases student performance, motivation, and interest (Morriem, 2016). Over the years, researchers have studied the relationship between mathematics performance and motivation towards learning mathematics. Studies have shown that mathematics performance and motivation have a positive relationship (Özen, 2017; Mahyuddin et al., 2009). The implementation of technology in learning would improve students' performance and motivation (Song & Keller, 2001; Abu Bakar et al., 2010; Aliasgari et al., 2010; Bai et al., 2012). Education assessment tools that incorporate technology such as digital assessments and adaptive quizzes improved students'

performance and motivation (Morriem, 2016; Ross et al., 2018; Faber et al., 2017). Ross et al., (2018) suggested more research is needed in motivation and engagement using adaptive testing technologies among students.

1.2 Research Background

Performance is a measurement for determining a person's level of knowledge proficiency. Mathematics performance can be defined as proficiency in mathematics or a measure of how well the students perform in mathematics. In Malaysia, secondary school students' mathematics performance is assessed through school assessments and national public assessments such as PT3 and SPM. At the school level, mathematics performance is usually assessed through formative assessment such as short tests or quizzes and summative assessment such as end of the year tests. At the international level, mathematics performance is assessed through international assessments such as Mathematics TIMSS Grade 8 and PISA (Kelly, et al, 2020; Mullis & Martin, 2017).

Trends in International Mathematics and Science Study (TIMSS) is an ongoing 4-yearly international assessment that provides comprehensive information regarding students' performance in mathematics and science to assist policymakers, educators, and researchers (Neidorf et al., 2004). Malaysia participated in TIMSS for the first time in 1999. The international benchmark tests conducted over the last twenty years revealed that the performance of Malaysian students in these subjects was promising in the early stages but has been on the decline with the students achieving means below the global averages (Mullis et al., 2020). TIMSS 1999 reported the average score of Malaysian

students was higher than the international average for mathematics with 519. However, the Malaysian students average score has continued to report a steady decline, especially in mathematics for four consecutive cycles of TIMSS, where the average score declined drastically to 440 in 2011. Nevertheless, Malaysian students' performance in mathematics has a slight improvement, scoring 25 points higher from 440 to 465 in 2015, but the score still fell behind the international benchmark average. In 2019, the average score decreased slightly from 465 to 461 and Malaysia remained far below the global average compared to top countries like Singapore that ranked 1st place with 611 in mathematics.

Furthermore, the decline in mathematics is very dramatic with Malaysia being ranked below Thailand and Vietnam (PISA 2012). Programme for International Student Assessment (PISA) is another international assessment that measures how students utilize knowledge and skills in real life challenges in reading, mathematics, and science (OECD, 2019). The assessments have been conducted since 2000, with Malaysia taking part for the first time in 2009 and continue to participate for the next three cycles. According to the results released by the OECD, Malaysia scored 404, 421, 446 and 440 respectively to PISA 2009, PISA 2012, PISA 2015, and PISA 2018 for mathematics literacy. Although the PISA 2015 results were quite an improvement over previous PISA results, the PISA 2018 results showed a deterioration. Furthermore, the global average score is 489 in Mathematics and Malaysia's score remained lower than the international benchmark. Singapore maintained as the top performer in the international assessments despite having similarity with Malaysia in term of geographical location and ethnicity (OECD, 2019; TIMSS 2019). The TIMSS and PISA results were a major setback to the Malaysian Education Blueprint

2013-2025 target in achieving higher than global average in international assessments by 2025 (Malaysian Ministry of Education, 2013).

In Malaysia, students must go through at least 11 years of schooling, six years in primary level and five years in secondary level. Students have to sit for three national assessments, namely, UPSR, PT3 and SPM throughout the 11 years of schooling in Malaysian formal education system. Students sit for UPSR at the end of Standard Six (12 years old), PT3 at the end of Form Three (15 years old) and SPM at the end of Form Five (17 years old). The UPSR and PT3 has been abolished effectively in 2021 and 2022 respectively due to Covid-19. Meanwhile, SPM is a centralized high-stakes assessment taken by most secondary school students at the end of schooling. For the SPM Mathematics results, the scores had declined from 4.94 in 2020 to 5.57 in 2021. This score was calculated based on the *Gred Purata Mata Pelajaran* (GPMP) where the score indicates that the lower the score, the higher the grade (Ministry of Education, 2022). Compared to the score in 2021, the GPMP for SPM Mathematics has improved with a score of 5.43 in 2022 and 5.28 in 2023. However, the score was considered low compared to the score in 2020. Meanwhile, SPM Additional Mathematics results also shown a deterioration where in 2020, the GPMP score was 5.64 and in SPM 2021 the score increased to 5.84 and the score increased even further to 5.99 in SPM 2022. These results from the SPM assessments revealed a concerning trend in Malaysian student performance and indicated that Malaysian students are struggling with mathematics subjects. These issues need to be addressed in order to improve students' performance in mathematics.

School-based assessment (SBA) was introduced in 2011 as a component of assessment in Malaysian school and the implementation of SBA demonstrates the ministry's effort to decentralize the national assessment system and improve educational system. The abolition of UPSR and PT3 was to ensure that teachers could implement teaching and learning in a more creative and innovative way to encourage an enjoyable learning environment instead of exam-oriented practice and to ensure that SBA can be implemented effectively (Bernama, 2022). There has been a major change in assessing students since SBA has been introduced. From a centralized high-stakes assessment to a classrooms assessment. The transition has changed in terms of the test developers where in SBA, teachers are authorized to carry out assessment tasks that suit students' competency and needs. However, several drawbacks hinder the transformation of the national assessment as teachers are less ready to implement SBA (Ghazali, 2016; Singh et al., 2017), lack of supporting materials for implementation of SBA (Ghazali, 2016; Kenayathulla & Ibrahim, 2016) and the implementation of SBA seems to add workloads and burdensome to the teachers (Ghazali, 2016; Mohd Noh & Mohd Matore, 2020). Furthermore, the Covid-19 pandemic has given a lesson that we can no longer depend on the current assessment method where most of school assessments and national assessments are still using fixed item tests to assess students. An alternative method of assessment that is more appropriate with time and the future is needed in Malaysian education system.

An adaptive test is a test where the items or questions administered to each test taker are selected based on the responses throughout the test (Weiss, 2015). Computerized adaptive tests (CAT) extend the basic ideas of adaptive testing and combine them for maximum efficiency and effectiveness. A calibrated item bank

is one of the core elements in the development of a CAT assessments. Item bank also known as item pool is a collection of calibrated items that were derived from measurement model theory called item response theory (IRT) (Weiss & Kingsbury, 1984). The one parameter logistic (1PL) IRT also known of Rasch model measures item parameters and person ability simultaneously (Bond & Fox, 2015). The Rasch model assumes unidimensionality as it measures a single latent trait such as mathematical ability and enables CAT systems to assess the ability that focused on one particular trait. The use of Rasch model in item banking would result in an improvement in learning and reporting of performance (Umar, 1999). The results of tests that used items in calibrated item bank are comparable as the items are on the same scale (Chuesathuchon & Waugh, 2010; Hambleton et al., 1991). Furthermore, the Rasch model was chosen because Rasch model has rigid standards in controlling the data and has strong assumptions that make it appropriate to use in calibrating the item bank. Thus, a Rasch calibrated item bank produces more flexible and appropriate tests. Rasch calibrated item bank coupled with sophisticated computer software such as CAT makes the application of CAT possible to be administered at any levels such as at school level (Hambleton et al., 1991).

In addition, computerized adaptive test (CAT) is instant and personalized (Wainer, 2000). In CAT, a student will be given an item at the medium level of difficulty, and he/she answers the item correctly the subsequent item presented to the student will be of a higher difficulty level (Weiss, 2011). This process will continue until the student achieves the levels of difficulty of test items that are appropriate to his/her ability level (Lyman, 1998). Further, CAT always yielded accurate estimates in assessing students' performance (Weiss, 2011). In addition,

CAT is always accompanied by an automatic report of performance at the end of each test. Student does not need to wait for teachers' feedback as the score is given instantly when using CAT. Without CAT, the students are fully dependent on the teacher. CAT also can be an excellent complement to the SBA by providing more precise and efficient assessments as CAT is a dynamic and personalized approach to evaluating students' knowledge and abilities without affecting the measurement quality (Wainer, 2000; Weiss, 2004). In classroom, teachers can use CAT in administrating formative assessments such as quizzes or topical tests and with CAT, teachers can continuously monitor their students' performance levels and identify students' strengths and weaknesses (Zhang et al., 2013, Weiss & von Manden, 2011). Applying CAT in SBA also reduced cheating among students as administrating tests using CAT, all students will be getting different set of questions since CAT is tailored to each students' abilities. This showed that CAT is more secure and fairer to all students as well as helping teachers in observing students' growth. Furthermore, applying CAT in the SBA can reduce teachers' workloads as the CAT system helped teachers in marking the assessments and gave immediate feedback to the students (Aggrey, et al., 2017).

Moreover, CAT has been implemented globally. CAT is used in educational settings to test basic arithmetical skills and language skills in Netherland and as aid in teaching and learning process in Alberta (IACAT, 2016). Most Asian countries implemented CAT as a method of testing in language test. In Japan, there are three language tests that use CAT method to evaluate the proficiency of the language, namely, J-CAT to evaluate Japanese language, C-CAT to evaluate Chinese language and BJ-CAT to evaluate Business Japanese

language skills. Meanwhile, in China, CAT is conducted via internet and has been used in assessing non-native Chinese speakers' abilities to use the language in real business environment. The United States has been implemented CAT method for decades in GRE and GMAT examinations. This indicates that implementation of CAT in educational setting whether at school level or in high-stakes examination is an effective approach in assessing students' performance. In Malaysia, the use of CAT could be implemented as a formative test in school like the Netherland did or implementing in lower secondary examination such as the new PT3 like the United States of America did with GRE examination. The implementation of CAT in educational setting as schools can conduct assessments more efficiently because the CAT assessments are personalized for each student (Wainer, 2000) and the time taken to complete the test using CAT is shorter without compromising the quality of the outcomes (Weiss, 2004). Additionally, CAT provides immediate feedback to the students that would help students and teachers to identify students' strengths and allows teachers to adjust their lesson plans accordingly (Choi & McClenen, 2020; Faber, et al., 2017; Jung, et al., 2015). Thus, CAT should be explored further in order to discover alternative method of assessment in Malaysia.

Motivation is the driving force that stimulates and sustains learning behaviour. Administrating assessment using CAT has more advantages over the PBT approach where many studies found that CAT increases students' performance and motivation to learn (Ross, et al, 2018; Ling, et al, 2017). Keller (2010) developed a model focuses on motivations which clustered motivational concepts into four constructs, namely, attention (A), relevance (R), confidence (C) and satisfaction (S). The ARCS model has been widely used and validated

as a systematic approach to improve learner motivation and performance in educational settings (Song & Keller, 2001; Leong, 2015; Wu et al., 2012; Kim & Keller, 2008; Zhang, 2017). The ARCS model is applicable to all learning environments as it consists of general principles for motivation and learning. Attention is a condition in the learning process that refers to the act of gaining interest, simulating curiosity, and sustaining attention of the learners (Keller, 2010). Relevance refers to individual opinions of importance and value of learning as well as the usefulness of the material (Keller, 2010). Confidence refers to the feeling of self-belief in succeeding in the learning task (Keller, 2010). Satisfaction refers to the good feeling of accomplishment in the learning process where learners should be satisfied with what they achieved during the learning process (Keller, 2010). Incorporating the ARCS model into CAT in this study can enhance both the development of the CAT system and students' motivation and engagement in learning mathematics. The use of technology in educational settings especially in school has proven to attain students' attentions and increase learning engagement among students (Timotheou, et al, 2023; Yeh, et al., 2019). Previous studies also found that there were improvements in students' confidence and satisfaction in learning mathematics when using adaptive assessments (Faber, et al., 2017). Thus, the characteristics of CAT in term of ARCS elements should be explored in order to understand and enhance the development CAT in this study.

Moreover, many studies have been conducted in assessing students' motivation in learning, however there is a limited amount of research on the motivation in mathematics education (Wæge, 2009; Hannula, 2006) and the application of motivation theory in the assessment setting has not been studied.

Studies have shown that the use of technologies such as computer and tablet in teaching and learning contributes to the increase of students' performance as well as students' motivation (Song & Keller, 2001; Abu Bakar et al., 2010; Aliasgari et al., 2010, Bai et al., 2012; Ling et al., 2017). In the study in the United States of America, the effectiveness of using technology in learning mathematics was assessed, and findings suggested that the implementation of technology in mathematics education can increased mathematical knowledge and maintained students' motivation to learn (Bai, et al., 2012). In computerized adaptive test (CAT), a student will be given questions that are appropriate to his/her ability level. A high-performance student would continue to get difficult questions as he/she continues to answer correctly, and this process could increase motivation as the test is getting interesting and challenging. Meanwhile, low performance students will receive questions that are appropriate their levels and they might feel more confidence as the tests tailor to their abilities. Ling et al. (2017) reported adaptive test resulted higher engagement in a study on middle school students who completed a mathematics assessment. Students were found very favourable toward the adaptive test method and perceived to increase their motivation as adaptive test method help them in learning (Ross et al., 2018; Rossano et al., 2017; House et al., 2016). Therefore, motivation plays an important role in the students' learning process and the use of technology in assessment helps to motivates students to learn. Thus, research in student motivation using adaptive testing technologies is needed.

1.3 Problem Statement

Mathematics is a compulsory subject for all levels in Malaysian schools. Learning mathematics strengthens the skills of thinking critically, solving problems, and seeking evidence, conclusions, and proof. For such reasons, mathematics is a highly foundation-based subject as it is important to learn at early stage of education. Various methods and strategies have been used by the teachers overcome the lack of mastery in mathematics among students. However, the level of mastery in mathematics among Malaysian students are still at level below global average (TIMSS 2019, PISA 2018). Even the results of national assessments such as SPM have shown a drastic deterioration in recent year when compared to the previous five years. The recent results in Malaysian national assessments aligned with international assessments' results where Malaysian students showed their mathematics performance were getting worse, demonstrating the teachers' attempts to enhance learning process and students' mathematics performance were ineffective. This trend may also raise the possibility that the current assessments method might not be effectively assisting students' mathematical learning and new approaches are needed to improve both understanding and performance in mathematics.

Studies showed that students perceived mathematics as a difficult subject and they found mathematics as uninteresting subject (Gafoor & Kurukkan, 2015; Brown et al., 2008). Studies also shown that Malaysian students struggled to perform in mathematics especially in problem-solving that required mathematical reasoning skills (Hasan, 2019; Tambychik & Meerah; 2010, Tarmizah, 2005). Students perceived algebra as difficult topic in mathematics (Gafoor & Kurukkan, 2015; Muchoko et al., 2019; Pramesti & Retnawati, 2019).

This could be due to its abstract nature and requirement of mathematical reasoning skills in algebraic problem solving (Tajudin, 2015; Rubin & Rajakaruna, 2015). Studies has shown that implementation of technology in learning could improve students' performance and motivation (Bai et al., 2012; Morriem, 2016). Assessment using technology has a positive impact on mathematics performance and motivation of students (Morriem, 2016; Faber et al., 2017). To address the issues of learning mathematics, this study intends to integrate technology in assessment by developing a computerized adaptive test (CAT) as an effort to find reliable and engaging assessment towards learning mathematics and indirectly improves students' mathematics performance. Studies found that CAT helps in increasing students' performance and motivation (Wainer, 1990; Ross et al., 2018).

In traditional testing, Classical Test Theory (CTT) is a common and widely used theory in evaluating an assessment especially in paper-based assessment. A disadvantage associated with CTT is the statistics associated with the test are dependent upon the items of the test where each test takers has a true score on a test (Cohen & Swedlik, 2016). This disadvantage is a downfall for some test takers. The test takers might have a particular score on one test, but it might not reflect or represent his/her true capabilities. In addition, the test items could be selected and arranged by the test developer based on theory only without any statistical evidence. The Rasch model able to measure the ability of the test taker and level of difficulty of an item simultaneously (Bond & Fox, 2015). The results from tests of calibrated item bank are comparable as the items are on the same metric (Chuesathuchon & Waugh, 2010). This means that with calibrated item bank, different tests can be administered to different groups of test takers, yet the

results can be compared. Thus, testing become more flexible with Rasch-calibrated item bank. The CAT assessment constructed based on the Rasch-calibrated item bank that can be used to develop effective and efficient systems of tests (Zhang et al., 2013; Reckase, 2010). Thus, use of Rasch model in the development of item bank for CAT provides a solution the problems of scoring in an assessment. Further, the use of item bank can provide more detailed information on the students' understanding levels of the relevant topics.

Paper based testing (PBT) is the traditional mode of testing and commonly used in schools. In fact, in Malaysia, the national assessment like SPM is still administered using PBT method. However, PBT has few drawbacks and one of major disadvantages of PBT is that it requires a long time to check and score the tests and feedback cannot be given immediately after the test. For example, SPM results took about two to three months until the students received their results. Even a simple formative assessment in school could take a minimum of a day for the teachers to give the feedback. This shortcoming has been overcome by using computers and online technology testing called computer-based testing (CBT) (Piaw, 2012). However, both PBT and CBT employed conventional fixed-form tests, differing only on the platforms used. Further, CBT too has serious drawbacks, such as execution in rigid linear sequences and having weak security on the items (Zhang & Chang, 2005). The limitations of both PBT and CBT can be overcome by using CAT as the use of CAT is always accompanied by a report of performance at the end of each test that can act as learning feedback (Lilley & Baker, 2007). In addition, application of CAT often reduces time taken to complete the test and CAT administers a shorter test compared to PBT and CBT, as CAT can result in reduction of the number of items

administered by 50%-90% without affecting the quality of measurement (Brown & Weiss, 1977; Weiss, 2004). Further the issues of item security in CBT could be overcome by application of CAT as CAT applied item selection algorithm that could potentially create millions of permutations of ordering items (Thompson, 2011; Lilley et al., 2015). Besides, CAT always produced more accurate measures in estimating ability compared to PBT or CBT (Weiss, 2011). Therefore, the use of CAT in administrating assessment would be beneficial to both students and teachers. Consequently, the development of CAT would overcome the issues of current assessment method.

Moreover, school-based assessment has been implementing in Malaysia education system since 2011. With the implementation of SBA, teachers' assessment quality needs to be reliable and valid in assessing students' performance. Fook & Sidhu (2011) found out teachers did not concern about the validity and reliability of the tests constructed and Ing et al., (2015) revealed that teachers did not refer to the table of test specification while building the items for assessment and teachers are lack of awareness on the importance of the table of specification. These behaviours could affect the quality of the school-based assessment. Mohd Noh & Mohd Matore (2020) suggested potential solutions to the challenges facing by teachers in implementation of school-based assessment in Malaysia such as utilizing a computer program to automatically mark and grade students' answer (Shermis & Burstein, 2013). The adaptive features in CAT seem like a way to overcome the problem in assessment quality and the use of CAT assessment would provide limitless number of assessments with no decrease in the quality of the measurement.

Additionally, CAT tests are more efficient than PBT because test length can be minimized without affecting precision in measurement (Weiss, 2004; Fenwick, et al., 2020). It is important to have a good stopping criterion to ensure accuracy, efficiency, and fairness of testing process. There is no consensus in the literature on which stopping criteria is the best for CAT. Chuesathuchon (2008) developed CAT to measure mathematical ability among primary students and used variable length criteria as stopping rule criteria. The study examined the differences in test length and found that there was significant different in the mean number of items administered. Findings showed that CAT with stopping criteria of standard error of measurement less than 0.2 administered only mean of 4.34 items and the CAT stopped at the average of 8.14 items when the stopping criteria of standard error of measurement less than 0.4. These findings show that lack of consistency and lack of predictability in variable length criteria compared to fixed length criteria. Fixed length criteria relatively easy to implement and consistent and predictable as test takers know exactly how long the test will take to complete. This can be useful in situations when time is a constraint. Babcock & Weiss (2009) suggested in order to ensure the stability of measurement when using CAT, a minimum number of items should always be used. The lack of a clear and reliable stopping criteria in CAT can lead to inconsistent and unreliable test scores, potentially affecting the validity and fairness of the assessment. Therefore, investigation on stopping rules criteria in term of fixed length by measuring its impact on the test scores should be conducted to enhance the development of CAT.

Further, standards are used to ensure the level of quality or performance of products or services. Performance standards describe the quality of students'

performance whether students are mastering the subject or not (Hambleton, 2001). Feedback on students' performance is an essential aspect in any assessment and influenced students' learning process (Mamoon-Al-Bashir et al., 2016). Cutscores refer to passing points on a test scale to differentiate students into different performance levels (Cizek, 2001). Standard setting is a systematic approach to develop a performance standard for assessment (Cizek & Bunch, 2007). The use of standard performance standard provides more accurate measure in differentiating students' performance (Cizek, 2001; National Research Council, 1999) and studies have shown that performance standards provide guidance for curriculum development, instruction, and assessment (Bejar, 2008; National Research Council, 2002; Cizek, 2001).

In Malaysia, most schools use students' raw scores in the evaluation of students' performance. For example, one school implies that those who get lower than 45% is considered as fail or grade E. Others may use different passing point like 40% and below marks as grade E that indicates failure. This unsystematic practice could be because there is no theoretical foundation employed when establishing the passing scores. In addition, comparison of students' performance by arranging the test scores in rank order also a usual practice in school. By looking at the ranking, categorization of their performance can be determined. However, the ranking of 10 is good for a larger number of students such as 50 but not in a smaller group of 20. This practice might be good for selecting the best student, but the ranking does not represent the difficulties experienced by students. For example, one student scored 50 in mathematics test and ranked 10 in his/her classroom of 30 students. The ranking looked good, and the student felt that he/she was good in mathematics. These norm-referenced scoring

practices gave a misleading interpretation. Having a standard cutscore in assessing the performance levels is crucial in CATs. Standard setting is one of the components in managing CAT where it defines the minimal borderline of students' competency to distinguish between a competent student and an important student (Seo, 2017). Standard setting provides consistency across administrations of the tests (Bejar, 2008) and provides a reliable and valid way to measure student learning and progress to ensure the tests administered were fair and accurate (Plake et al., 2000; Gonullu et al., 2023). Therefore, a standard performance levels for mathematics test should be set because a mere scale like the average score of the tests or a grade or even ranking would not be sufficient to really measure or determine the performance or ability of the students and the performance standard should be set to complement and enhance the development of CAT.

Furthermore, motivation is one of the factors that influences the students' learning process. Studies has shown that there is correlation between motivation and performance (Özen, 2017). According to Özen (2017), there is a positive correlation between motivation and student performance. Motivation towards learning mathematics can be increased by changing the way of teaching mathematics and with the implementation of technologies in teaching and learning mathematics (Abu Bakar et al., 2010; Aliasgari et al., 2010; Bai et al., 2012). Adaptive testing in formative assessments such as quizzes has been shown to have positive learning benefits and improved students' performance (Barla et al., 2010; Ross, et al., 2018). The use of CAT is likely to improve student enthusiasm and increase student motivation as CAT is adapted to their abilities. According to Wainer (1990), CAT leads in increasing student's

motivation as CAT is a test that provides an appropriate level of challenge for each student. This shows that CAT is personalized to each student, thus its results will determine each student's growth and potential that can be used for prescribing remedial or accelerated programs to further enhance their abilities. The modern features available in CAT is seen as a way to engage students' interest in learning. However, some studies found that students were not satisfied and reluctant to engage in learning with technology (Thompson, 2013; Kundu & Bej, 2021). Kundu & Bej (2021) found that students were lacking in concentration and stressed in using computer during the assessments. Therefore, the characteristics of the CAT should be examined further to find whether CAT is a good assessment tool in engaging students to learn.

Thus, this study intends to address these issues by developing Rasch-calibrated item bank in computerized adaptive test (CAT) with established performance standards to provide a reliable and engaging assessment experience. Additionally, this study aims to assess students' performance in mathematics when using CAT and examine the motivational impact of CAT. This study also focuses on the CAT's features and examine stopping criteria as an effort in developing a fair, effective and accurate assessment tool that supports and motivates students as well as providing teachers with insightful information about students' performance.

1.4 Research Objectives

The main objective of this study is to develop and investigate the effectiveness of computerized adaptive test (CAT) in assessing mathematics performance of Malaysian secondary school students by using calibrated item bank. Thus, the objectives of this study are:

1.4.1 To identify the psychometrics properties for computerized adaptive test (CAT) items.

1.4.2 To set the cutscores of mathematics performance for the computerized adaptive test (CAT) item bank for secondary school students.

1.4.3 To examine the stopping rules criteria of computerized adaptive test (CAT) on secondary school students' performance in mathematics.

1.4.4 To investigate the computerized adaptive test (CAT) characteristics on secondary school students' motivation needs.

1.5 Research Questions

The following research questions are based on the research objectives of this study:

RQ1- What are the psychometrics properties for developing item for computerized adaptive test (CAT)?

1.5.1 Is concurrent common item equating method able to produce calibrated items that met the item fit analysis requirements?

1.5.2 Is concurrent common item equating method able to produce calibrated items that fulfilled the item polarity analysis?

1.5.3 What are the fit statistics used as the measure value for each item in the calibrated item bank based on the Rasch model analysis?

RQ2- What are the cutscores of mathematics performance for the computerized adaptive test (CAT) for secondary school students?

RQ3- What are the distribution of students' performance based on the stopping rules criteria of computerized adaptive test (CAT)?

RQ4- How does computerized adaptive test (CAT) characteristics affect the students' motivation needs in term of (a) Attention, (b) Relevance, (c) Confidence and (d) Satisfaction?

1.6 Research Hypotheses

There are two hypotheses in this study based on research question 3 (RQ3) and research question 4 (RQ4). There are sub hypotheses for two of the main hypotheses. The first hypothesis is for RQ3, and the hypotheses are as follows:

H_1: There are no significant differences in the students' mathematics performance between all the stopping rules criteria by using computerized adaptive test (CAT).

H_{1a}: There are no significant differences between the 40-Questions test and the 30-Questions test by using CAT.

H_{1b}: There are no significant differences between the 40-Questions test and the 25-Questions test by using CAT.

H_{1c}: There are no significant differences between the 40-Questions test and the 20-Questions test by using CAT.

H_{1d}: There are no significant differences between the 30-Questions test and the 25-Questions test by using CAT.

H_{1e}: There are no significant differences between the 30-Questions test and the 20-Questions test by using CAT.

H_{1f}: There are no significant differences between the 25-Questions test and the 20-Questions test by using CAT.

The second hypothesis is for RQ4, and the hypothesis is as follow:

H_2: There are positive effects in the students' motivation needs in term of attention, relevance, confidence, and satisfaction after the use of computerized adaptive test (CAT).

Based on the literature review (Chapter 2), a hypothesized model was developed to test the relationship between attention, relevance, confidence, and satisfaction. The following sub hypotheses are based on the hypothesized model (Figure 1.1):

H_{2a}: Attention has a positive effect on relevance.

H_{2b}: Attention has a positive effect on confidence.

H_{2c}: Attention has a positive effect on satisfaction.

H_{2d}: Relevance has a positive effect on confidence.

H_{2e}: Relevance has a positive effect on satisfaction.

H_{2f}: Confidence has a positive effect on satisfaction.

H_{2g}: Relevance mediates the relationship between attention and satisfaction positively.

H_{2h}: Confidence mediates the relationship between attention and satisfaction positively.

H_{2i}: Attention has a positive indirect effect on satisfaction through relevance and confidence.

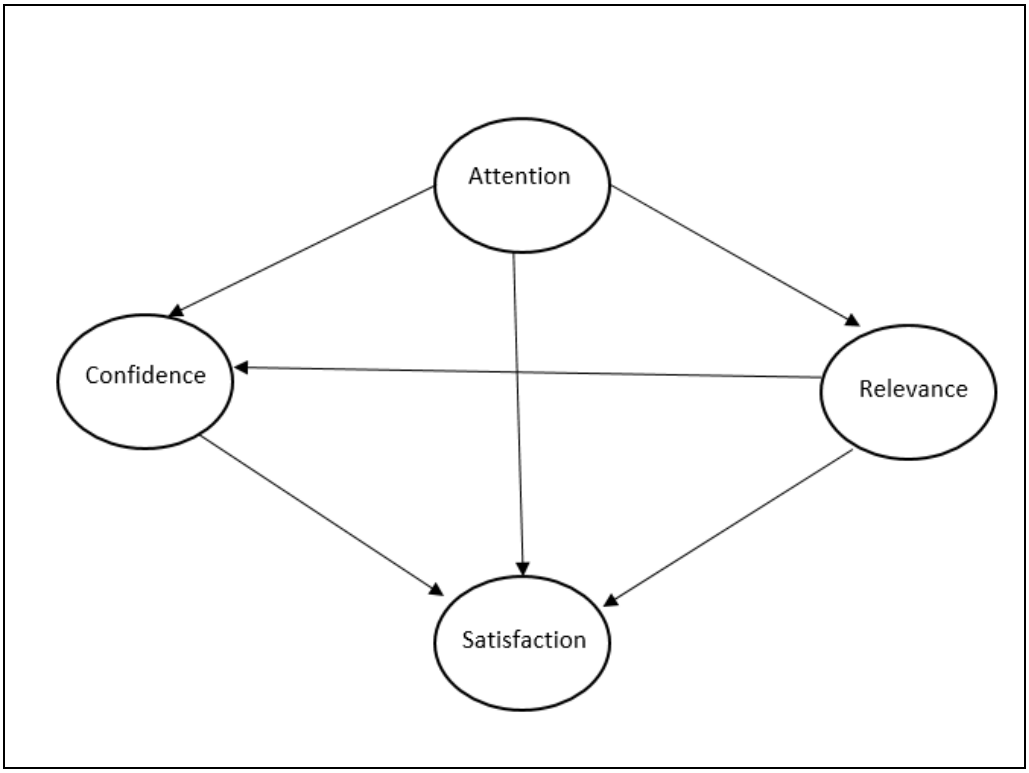

Figure 1.1: Hypothesized Model