Computer-based Assessments of High School Mathematics in Myanmar

Hnin Nwe Nwe Tun

a1679758

B. Ed (Specialized in Mathematics, Physics, Chemistry) (Yangon University of Education), M.Ed (Specialized in Mathematics Teaching) (Yangon University of Education), and M.Ed (Specialised in Testing and Measurement) (The University of Adelaide)

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

In the

School of Education, Faculty of Arts

The University of Adelaide

January 2022

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree. I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and through web search engines, unless permission has been granted by the University to restrict access for a period of time. I acknowledge the support I have received for my research through the provision of an Adelaide Graduate Research Scholarship.

Signed

Hnin

Monday, February 28, 2022

(Other name: Hnin Inioluwa Tun Agbaje)

(Student #: a1679758)

Dedication

To God Almighty, Jehovah. *"Thou art worthy, O Lord, to receive glory and hour and power for thou hast created all things and for thy pleasure they are and were created. Thou art worthy O Lord (Revelation 4: 11)".* Also, this work is dedicated to the brightest future that God has declared for me.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am greatly indebted to you, Almighty God, Jehovah-Jirah, for your grace, mercy, provision, and protection over my life, "New every morning, Great is thy faithfulness (Lamentations 3:23)". Thank you for shining light unto my path and for guiding my feet (Psalm 119:105).

My sincere thank you also go to Adelaide Graduate Research Centre, University of Adelaide, for receiving Adelaide Graduate Research Scholarship for the duration of my study, it has significantly aided the study and helped me to become who I am today.

I would like to express my profound gratitude to my supervisors Dr I Gusti Ngurah Darmawan and Associate Professor Nicholas Buchdahl for your gentle push, and encouragement which make me enjoy my PhD lifestyle. It has been a great privilege to be their PhD candidate. A special thank you to Dr I Gusti Ngurah Darmawan, for making my research journey enjoyable and learning new things about quantitative studies. Thank you for the opportunities that make me to learn the proper ways of conducting research and data analysis. My deepest appreciation goes to Associate Professor Nicholas Buchdahl, who devoted so much time guiding me to develop my thesis and improve my academic writing throughout study. Thank you for always being there, ready to listen, and provide constructive feedback for improvements.

My success as a graduate student would not have been possible without the support of Associate Professor Sivakumar Alagumalai, my former principal supervisor, who gave freely his valuable advice and support, especially on a proposal of my research study. Your insightful suggestions contributed a lot to the success of this research work. You have been a great source of inspiration.

Many thanks to the staff of the School of Education, University of Adelaide for their support and for making my research journey a comfortable one. Each of them contributed directly or indirectly to this success. Thank you to friends who are/were PhD candidates in Level 6, school of education, for sharing your experience, knowledge, laughter, and tears. It is a blessing to have all of you on board during this episode of life. A special thank you goes to Safitri Yosita Ratri, who made my PhD full of wonderful memories, and Mutinta Sifelani Musindo, who helped and shared experiences throughout PhD journey. May we always keep in contact and extend friendships forged here to span across the globe!

My deepest appreciation goes to the Department of Education, Ministry of Education, Myanmar, and principals from five participant schools. They willingly allowed me to have access to their students, teachers, classrooms, ICT facilitated classrooms, and information necessary for me to carry out my research. My thanks are also addressed to the students and teachers who participated in my research. Without them, I would not have completed this project. You are well-appreciated, Phyo Hein Kyaw, my former student and now working as a freelance accountant. Thank you for devoting your precious time and effort to assisting me during data collection. Thank you for always being there.

To all families in Deeper Life Ministry, Adelaide, especially Pastor Wale Tijani, Sister Pauline Tijani, and to my two sisters, Mojetijesu (Jeti) and Ifeoluwa, thank you very much for your love, friendship, and support and for making me feel to have a second home in Adelaide.

It was because of my teachers, people who saw my potential that I am more than just what my past might suggest. I am truly blessed, and for this, I am grateful to my primary and secondary school teachers and teachers from Yangon University of Education: Daw Mya San, Profs Dr Wai Wai Oo and Dr Naing Naing Thein, Myanmar.

Warm thank you go to my late father, U Tun Aung Kyaw, who began a dream of my PhD journey. A sincere thank you to my mother, Daw San Win Phyu, who is also my first mathematics teacher. Thank you, mom, for making me fond of numbers and arithmetic. You have influenced the journey of psychometric and quantitative studies. Thank you to my brother, Htun Lynn Kyaw, who made my life lovable.

I started this study as Hnin Nwe Nwe Tun and ended it as Hnin Inioluwa Tun Agbaje. Thank you to my husband, Dr Oluwatoosin B. A. Agbaje, my life journey partner and my love. It is impossible to accomplish my PhD study without your support, understanding, and encouragement both physically and spiritually. Thank you for making my dream come true. To my first-born son, TiOluwani A. O. Agbaje, who makes me understand the true definition of love, thank you for healthy growth that gives me the courage to complete my thesis successfully with no worries! You have always made me proud to be your mother.

Hnin Inioluwa Tun Agbaje

Abstract

Computer-based assessment (CBA) is a versatile educational tool in the twenty-first century. It offers many new opportunities for innovation in educational assessment through rich new assessment tasks and improves the learning progress of students. Educators have begun to benefit from CBA as it reduces the timing in reporting scores and increases assessment efficiency that enables immediate feedback. However, assessments in Myanmar high schools are mainly in paper-and-pencil test (PPT). Due to the large class and a limited number of teachers, this regular assessment causes them more workload in administering tests and providing scores and feedback. As a result, teachers spend most of their time assessing, scoring, and providing feedback. These activities negatively affect allocated hours of teaching and learning, which, in turn, are ineffective on the learning progress of students. The aims are: (1) to examine high schools in Myanmar whether computer-based assessment, this is, linear-online test (LOT) and computer-adaptive test (CAT) is more effective test mode than PPT as a formative assessment for the learning progress of students; (2) to identify contextual scales that influence students learning progress due to computer-based assessments and regular paper-and-pencil test.

Of intervention design with explanatory mix-method, this study applied counter-balanced quasi-experimental research to compare effects of computer-based and paper-based assessments in terms of the achievement improvement of students. This study conducted surveys among students and teachers, followed by semi-structured interviews from five high schools in Yangon Region, Myanmar. Students from these high schools took the computer-based test and paper-based format as formative assessments. For constructing an online formative assessment test, both the Concerto platform and online Monkey Survey were applied, and through the Rasch Dichotomous model, items are assembled in the item-banks of the computer-based assessment. The analysis of variance (ANOVA) was applied to examine the effect of the test modes. The results of the computer-based test mode showed that students who received their specific scores and feedback immediately improved their mathematics achievement significantly higher than those who received the delayed score and feedback from the paper-based test mode.

Structural equation modelling is used to analyse the structural relationship between measured variables. This model shows that positive attitude of students towards either computer-based or paper-based is the ultimate mechanism for more remarkable achievement. Although the two test delivery media may affect different groups of participant students in

different ways, this concerns equity issues. For example, findings showed that educational background of parents, students' gender, and attitude towards paper-based assessment or computer-based assessment could influence or affect the achievement of students. In addition, the specific practices of teachers towards formative assessment influence the attitude of students. The attitude of teachers concerning computers and technology affects the attitude of students towards innovative assessment formats. As shown by hierarchical linear modelling, the cross-level interaction effect from the teacher-level on the slope of the attitude of students towards paper-based assessment and their achievement improvement is specific practices of formative assessment. In addition, this study showed that the attitude of students to information and communications technology (ICT) and the attitude of teachers to formative assessment and ICT directly affect the achievement improvement of students.

This thesis reveals significant gaps in understanding concerning formative assessment in Myanmar and contributes to the theoretical, practical, and methodological implications in mathematics assessment and learning. In addition, the findings provide (albeit for Myanmar educational systems) a practical resource for assessment developers and a useful framework for the discussion of innovative assessment formats and use in computer-based settings.

List of Abbreviations

AMOS	Analysis of Moment Structures
AVE	Average variance extracted
CAT	Computer-adaptive test
CESR	Comprehensive Education Sector Review
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
FA	Formative assessment
HLM	Hierarchical linear modelling
ICT	Information and Communication Technology
INFIT MNSQ	Weighted mean square
LOT	Linear-online test
MCQ	Multiple-choice question
OUTFIT MNSQ	Unweighted mean square
PPT	Paper-and-pencil test
RMM	Rasch Measurement Model
RMSEA	Root Mean Square Error of Approximation
SEAMEO	Southeast Asian Ministers of Education Organisation
SEM	Structural Equation Modelling
SPSS	Statistical Package for the Social Sciences
SRMR	Standardised Root Mean Residual
TLI	Tucker-Lewis's index

Chapter 1

Introduction

1.1 Introduction

Mathematics education is the essential life skill for everyone, from daily basic to advanced professional level. Improvement of mathematics education for student is taken into account as one of key factors in the process of education innovation of every country. Myanmar, which is situated in southeast Asia, has started revolution in education system since 2012. From 2013 to 2018, mathematics achievement of high school students was below the average level (MOE, 2004; Nyunt, 2017). Report from Myanmar's National Education Strategic Plan (NESP) (2016) pinpointed the innovative paradigm shift from assessment systems focused on rote learning to a more balanced educational assessment system on learning progress as one of important shift in education revolution (MOE NESP, 2016). The integration of information communication technology (ICT) into the classroom assessment system is needed, based on the systematic review of NESP report (MOE NESP, 2016). Like other developing countries, class sizes of Myanmar high schools are large. Classroom teachers have more workload regarding with formative assessments of large classes. This issue reduces the effectiveness in administering formative assessment and providing scores and feedback. As a result, classroom formative assessment became to be ineffectiveness on the learning progress of students. The researcher conducted a study which will investigate whether the integration of ICT-based test modes can save these mentioned problems. The research documented in this thesis involves the motivation, conceptualisation, theory, development, and innovation of education assessment system of computer-based assessment methods or modes for developing countries. It is a study that is one of the pioneering studies for the development of educational assessment system in Myanmar. This study was conducted by the supervisors in education, mathematics and science, and the researcher. Conceptual modelling, mixed methods, data analyses in social sciences has been an instrumental ingredient underpinning the thesis. This introduction chapter is sequenced to highlight the importance of mathematics education, revolution of education system in Myanmar, innovative paradigm shift of assessment system in Myanmar, mathematics achievement of high school students in Myanmar, problem statement, aim of the study and research questions.

1.2 Importance of mathematics education

As the world's reliance on technology has grown, so too has the demand for individuals who can think in the abstract terms of mathematics and science. Today, technical jobs makeup nearly one-third of all employment opportunities, and to complement these, education systems have tried to keep pace with the demands of an increasingly competitive technological world by stiffening their mathematics requirements and invoking a system of high-stakes testing. Hence, the requirements in mathematics have a positive impact on the improvement in overall mathematics proficiency in their school (Shin, Sutherland, Norris, & Soloway, 2012).

Mathematics is essential in daily life-skills for counting, managing money, data-handling, and building things (Vorderman, Porkess, Budd, Dunne, & Rahman-hart, 2011; Shin, et.al, 2012). In addition, learning mathematics improves higher-order thinking skills such as logical thinking, critical thinking, creativity, and problem solving for our daily life. These higher-order thinking skills help an individual to have a successful life. Solving mathematical problems in school tends to improve the persistence and perseverance of individuals towards problem solving strategies and confidence (Rahman-hart, 2011; Shin, et.al., 2012). Further, solving mathematical problems for young children improved the ability to formulate, represent, and solve daily simple problems and reason and explain their mathematical activities (National Research Council, 2009).

Science needs mathematics, and it is vital to economics, finance (Shin, et.al., 2012) and related to higher education courses such as engineering, psychology, sciences, and social sciences (Shin, et.al., 2012). Current rapid technological advances demand mathematical skills for every citizen (Burghes, 2011; Vorderman et al., 2011; Shin, et.al, 2012). Mathematics has become a major course for the computing technology and software development underlying our technologically advanced and information-based world.

Mathematics education can prepare students for a better future life and work in the digital age (Gravemeijer, Stephan, Julie, Lin, & Ohtani, 2017). Ofsted (2011) claims that every citizen will be equipped with sound mathematics education for their future lives. In addition, many career fields require a strong mathematical foundation, such as engineering, architecture, accounting, banking, business, medicine, ecology, and aerospace. Consequently, mathematics education is foundational and important for an individual's everyday life.

Evidence of the importance of mathematics appeared between the 1950s and 1960s. After space race with the Soviet Union, the United States introduced better mathematics curricula with problem-solving skills (Public Broadcasting Service (PBS), 2019). In the early 1980s, schools raised graduation requirements for mathematics and introduced minimum competency testing in response to a government report on the state of education titled "*A Nation at Risk*" (Public Broadcasting Service (PBS), 2019). In the late 1980s, the National Council of Teachers of Mathematics revised the content and mode of mathematics teaching. At the same time, standards-based tests with rigorous mathematics sections were included as part of the graduation requirements in many schools.

More challenging graduation requirements in mathematics have had a generally positive effect — improving overall mathematics proficiency in the U.S., and mathematics became important and compulsory for graduation in college (Public Broadcasting Service (PBS), 2019). Consequently, mathematics plays an important role in achieving a better income compared to a typical high school graduate (Public Broadcasting Service (PBS), 2019). The key to this increment is that schools teach students practical job skills to which mathematics contributes immensely.

1.3 Revolution of education system in Myanmar

The Republic of the Union of Myanmar (but mostly called Myanmar and formerly known as Burma) is situated in the mainland of Southeast Asia. Its landmass is the second largest in Southeast Asia.

Myanmar has been one of the world's poorest countries with a level of GDP per capita of approximately between US$800 and US$1,500, according to the World Bank's 'least developed nations' category (World Bank, 2020). Of note, the education system in Myanmar has been and is in a very weakened state (Hayden, & Martin, 2013). At present, its whole education system is in a prolonged recovery process with some challenges (Hayden, & Martin, 2013). The Ministry of Education (MOE) is one of the main change agents in national education innovation. After 2015, the Comprehensive Education Sector Review (CESR) and Myanmar's National Education Strategic Plan (NESP) (2016-2021) are working towards the upgrade of the quality of education.

Nowadays, the national education system in Myanmar is composed of a basic education system and a higher education system. The basic education system is composed of primary education, lower secondary education (middle-school level), and upper secondary education (high school level) (**Figure 1.1**).

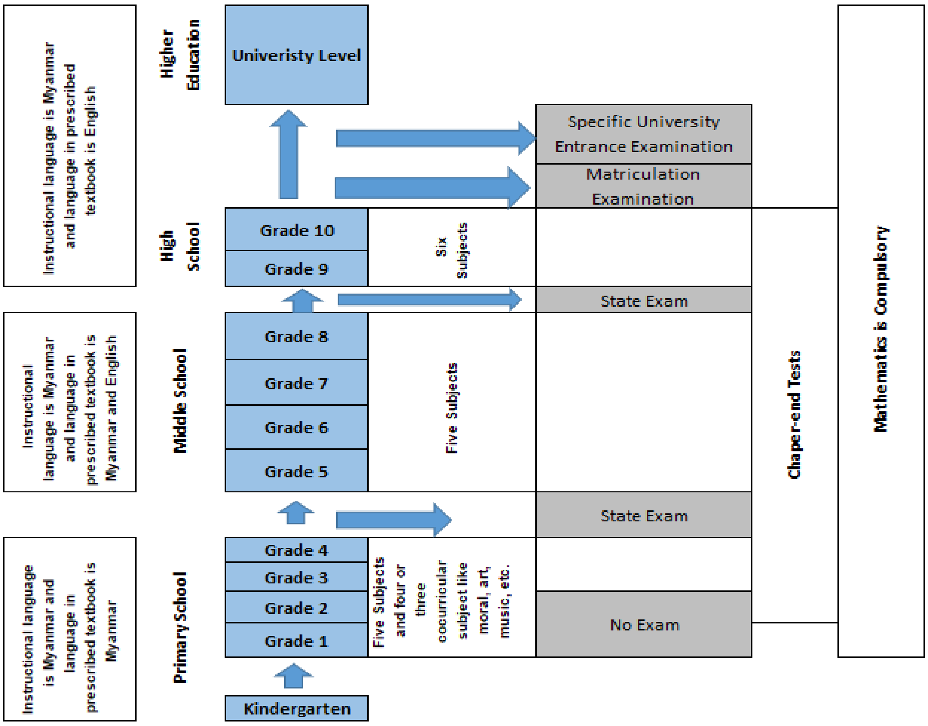

Figure 1.1 Education system in Myanmar

The basic education reform commenced in 2000 to enhance the quality of education. The *Thirty-year long-term basic education development plan* (2001/02 to 2030/31 Financial Year) was amended with ten basic education programmes. The programmes have planned to upgrade the quality of basic education, provide ICT-based facilities, and emerge with a modernized and developed education system. According to this plan, the Motto on Education (2000) is "*Building a modern developed nation through education*." Further, the motto leads to the Vision Statement on Education (2000) is "*To create an education system that can generate a learning society capable of facing the challenges of the Knowledge Age*". From this motto, quality enhancement is correlated with ICT integration.

Of education policy provided by Former President U Thein Sein at the first regular session of *Pyidaungsu Hluttaw* (Parliament) on 30 March 2011, the primary concern of the ten-point guideline for the quality enhancement of education is "*to nurture new generation as intellectuals and intelligentsia and to upgrade education standard to international level*". Consequently, emphasis is placed upon upgrading the quality of basic education to the international level in utilizing more effective teaching-learning materials and aids and integrating ICT into teaching-learning situations.

The CESR teams and eighteen working groups have cooperated with the Education Promotion Implementation Committee (EPIC) in the reforms of the basic education sector and the rapid assessment since October 2013 (MOE NESP, 2016). They suggested planning the restructure of the whole education system and guidelines for its establishment. That restructure involves changes in curriculum, classroom teaching and learning, classroom formative and summative assessment, technology-based assessment, and the evaluation system. Consequently, they emphasized upgrading the quality of basic education and providing facilities for e-Education ICT.

In addition, the NESP highly recommends that the quality of basic education is more related to upgrading facilities and curriculum and integrating ICT in basic education (MOE NESP, 2016). According to NESP (2016-2021), there is a need to upgrade technology-based assessment with the provision of reliable and valid student learning outcomes. Consequently, the upcoming technology-based assessment system and learning are necessary to modify the education system (MOE NESP, 2016).

1.4 Innovative paradigm shift of assessment system in Myanmar

One of the primary reforms concerning the quality of education is the shift of assessment systems focused on rote learning to a more balanced educational assessment system on learning progress against national learning standards (MOE NESP, 2016). The standards are targeted for educational developments of students and skills for their lifelong learning. In addition, the reform for the quality of education involves the integration of ICT into the whole system, especially into teaching and learning situations. Therefore, there is need for ICT-integrated classroom assessment for education reform in Myanmar.

Presently, there are three types of summative assessment in basic education schools: (1) Continuous Assessment and Progression System (CAPS), (2) Year-end examinations in Grade-5 and 9, and (3) Grade-10 Matriculation Examination (ME). Among these three, the practice of CAPS has commenced in all government schools, from Grade-3 to Grade-10, since 2000. CAPS takes place between 5 to 7 times a year under the guidance of the MOE. The primary aim of CAPS is to provide prompt feedback to all students and timely remedial teaching to students who fail tests, based on evaluating the performance of the students (MOE, 2004; Hayden, & Martin, 2013).

Teachers and students constantly struggle to finish every chapter in time due to the overuse of CAPS (Hardman, Stoff, Aung & Elliott, 2014). Again, another difficulty is the student-to-teacher ratio (45:1), which is high at the secondary level in Myanmar (Oxford Business Group, 2017). This high ratio causes teachers to be usually overworked and loaded with testing and scoring, and students are struggling with taking tests every month (Hardman, Stoff, Aung & Elliott, 2014). As a result, teachers cannot provide adequate feedback related to the strengths and weaknesses of students in CAPS. Moreover, teachers do not have enough time to raise questions on monthly CAPS to generate or improve higher-order thinking skills (MOE, 2004). The mindset of teachers is mostly dominated on how to teach, “what will be tested”, and students are also inclined to learn only “whatever might be on the exam” (Hayden, & Martin, 2013, p. 23).

Students and their parents often focus on how to judge the performance of teachers based on the success of their children in examinations, and school authorities spend much office time evaluating the performance of teachers on the same basis. In some instances, those facts lead to teachers being even more anxious than their students are about examination results (Hayden, & Martin, 2013). In addition, the overuse of the CAPS system has most likely contributed to poor learning outcomes in basic education, and the system mainly focuses on rote learning and exam-oriented learning (Hayden, & Martin, 2013). Consequently, the MOE has commenced the development of formative assessment, a new effective integrated-assessment policy, and framework.

In Myanmar, the broader use of formative assessment as the classroom-based evaluation has been commenced as long as the basic education national curriculum framework has been updated according to NESP since 2018 (MOE, 2016, NESP). As a result, the workload for teachers has been increased due to classroom formative assessments. In the successful implementation of the assessment reforms, the NESP report recommends that ICT systems play a vital role in reforms of assessment and examinations systems as the component-3 of quality of education (MOE, 2016, NESP). Effective ICT systems can produce statistical analyses of the assessment result and enable many administrative processes to work efficiently (MOE NESP, 2016). The reason is that ICT systems can store the data of questions and responses that teachers, parents, and students can access.

Previous studies mentioned that the computer-adaptive test (CAT) and linear-online test (LOT) help teachers reduce their workload related to test development and administration and improve learning with prompt feedback (Burghof, 2001; Eggen, 2007; Rashad, et.al, 2008). Therefore, the research will investigate the effectiveness of CAT and LOT for Grade-10 students’ mathematics achievement in comparing the Paper-and-pencil test (hereafter PPT), the current test mode in Myanmar.

1.5 Mathematics achievement of high school students in Myanmar

Like other countries, mathematics education in Myanmar is an essential gateway to tertiary education, especially technical vocational education and training or a better future (MOE, 2004; Nyunt, 2017). Mathematics is a compulsory subject for the Matriculation Exam (ME). Some subjects in high school, such as physics, chemistry and economics are mathematics-based subjects. However, from 2013 to 2018, less than fifty percent of students passed the mathematics requirement (**Figure 1.2**).

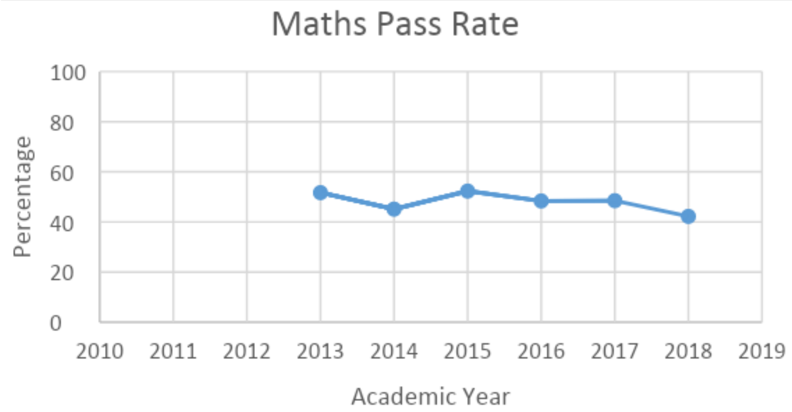

Figure 1.2 Mathematics pass rate of matriculation examination from 2013 to 2018

It is necessary to identify why students cannot perform well in mathematics. However, in Myanmar, there is no follow-up research or statistical analysis of mathematics achievement in the matriculation examination. Additionally, no investigation of the contextual background factors that affect/influence mathematics achievement. Consequently, studies are required to determine how to improve students' mathematics achievement.

1.6 Definition of key terms

Paper-and-pencil test mode (PPT) is the traditional fixed-length-test, applying paper and pencil or pen.

Linear-online test mode (LOT) is the computer-based fixed-length-test, applying the Web or online as a means of delivery (Weiss, 2004).

Computer-adaptive test mode (CAT) is a scale-length test, referring to a test in which new items are administered based on the responses of examinees to previous items (Kingston, 2010).

1.7 Problem statement

The above problems lead to the need to investigate the effectiveness of new computer-based test modes on student mathematics learning progress and teacher workload reduction for the classroom formative assessments, especially for large class sizes.

The Matriculation Exam (ME) in Myanmar is a high-stakes assessment at the end of Grade-10, in which mathematics is one of the compulsory subjects. However, the pass rate of mathematics in the ME is still lower than 50 percent from 2013 to 2018 (Department of Basic Education, 2018). Consequently, half of the ME examinees cannot access higher education. That pass rate indirectly shows that there is not enough mathematics learning progress throughout their previous grades. Hence, this leads to more studies integrating new interventions or assessments for their mathematics learning progress. Consequently, this study focuses on high school mathematics learning progress.

In Myanmar, teacher-made classroom formative assessments from Grade-3 to Grade-10 are predominantly PPT mode (MOE, 2004). They aim to enhance the learning progress of students by providing scores and feedback and timely remediation, but only for students who fail tests (MOE, 2004). The shortfall of this scope is because the large class size and the limited number of teachers in high schools hinder the fulfilment of this aim (Hardman, Stoff, Aung & Elliott, 2014; Oxford Business Group, 2017).

Further, teachers have an additional monthly workload for test development, administration, and scoring, thereby delaying the provision of immediate score and feedback and preparation of more engaging lessons (Hardman, Stoff, Aung & Elliott, 2014). Consequently, it can be said that such formative assessments in PPT mode are less likely to be practical classroom assessments for enhancing the learning progress of students.

Another reason for conducting this study is that it is the right time to integrate computer-based tests (CBTs) in Myanmar classroom assessments according to the Myanmar National Education Strategic Plan (2016-2021), in which traditional assessments will be upgraded as technology-based assessments to enhance and track the learning progress of students.

However, there are some technical and theoretical challenges. For instance, the availability and compatibility of hardware, software, item-banks, and technology-based test-delivery programs, which should have been initially installed for the accomplishment and/or for the integration of new computer-based test modes, were faced as challenges.

It is vital to examine whether the integration of LOT and CAT can solve the issues faced by classroom teachers and students (for example, ineffectiveness of classroom formative assessment in PPT mode to students learning progress and the workload of teachers for scoring and feedback) in Myanmar. Consequently, this study will need to examine this practical research question: which test modes (LOT, CAT, and PPT) when applied as classroom formative assessments, are more effective in their achievement improvement in Myanmar high school students? Its findings will provide a framework for the transition from conventional test mode to new computer-based ones, and it will also promote the quality of classroom formative assessments for the large class sizes.

1.8 Aims of the study

Classroom formative assessments of large classes causes more workload for teachers and reduce the effectiveness in administering tests and providing scores and feedback. These circumstances lead to ineffectiveness on the learning progress of students. This study will investigate whether the integration of LOT and CAT can save time and immediately provide scores and feedback that will solve the abovementioned problems.

To address the research problem, the study aims:

1. to explore the integration of LOT and CAT in high schools of Myanmar,
2. to examine whether LOT or CAT is a more effective test mode than PPT for classroom formative assessment to enhance Grade-10 students' mathematics achievement and reduce their teachers' workload to provide score and feedback, and
3. to identify the contextual scale which influences students' achievement improvement due to LOT and CAT.

1.9 Research questions

According to the overarching aim of the project, there are two general research questions and three specific questions associated with the second research question. These questions are the combination of other specific questions related to technical challenges and psychometric properties of test modes, contextual factors of students (gender, school geolocation, parental education and occupation, attitude towards mathematics, attitude towards a computer, computer familiarity, and their attitude towards the three test modes and score reporting and elaborated feedback) and teachers' perception towards these test modes.

The two general research questions are as follows:

Research Question 1: *Which test mode (PPT, LOT, or CAT) makes students' achievement improve more when the test mode is applied as classroom formative assessments in Myanmar high schools?*

Research Question 2: *Which contextual factors from student-level and teacher-level significantly influence the achievement improvement due to test modes (PPT, LOT or CAT)?*

The three specific questions associated with Research Question 2 are as follow:

2.1 *What are students' contextual and attitudinal factors, which significantly influence their achievement improvement due to the different test mode (PPT, LOT and CAT)?*

2.2 *What are the contextual and attitudinal factors of teachers, which significantly influence the achievement improvement of students due to the different test mode (PPT, LOT and CAT)?*

2.3 *What are the perceptions of students and teachers towards PPT, LOT and CAT as classroom formative assessment?*

1.10 Summary

This chapter provides an introduction and sets up the context of this thesis. It presents the issues that the study investigates. This chapter highlights the source and the context of the issues and provides the background by introducing the education system, the state of basic education, and the recent education reforms in Myanmar. The chapter briefly discusses the innovative shift of assessment and evaluation, the context of the education reforms, and the innovation of assessment reform towards the quality of education in Myanmar. In addition, this chapter presents the statement of the problems, and the objectives and research questions. Further, the Literature Review (Chapter 2) discussed in detail the points presented in this chapter.

Chapter 2

Literature Review

2.1 Importance of assessing classroom mathematics achievement

Assessing mathematics is an important component of the education system in all nations. If citizens are to play a role in today's technological world, there is a need to have strong mathematical skills (Burghes, 2011; Vorderman, Porkess, Budd, Dunne, & Rahman-hart, 2011). For this reason, mathematics is compulsory for higher education courses such as engineering, economics, sciences, and social sciences. Importantly, learning mathematics is one of the sources for improving higher-order thinking skills such as logical thinking, critical thinking, creativity, and problem solving, for our daily lives (Vorderman et al., 2011). Consequently, students need to be equipped with sound mathematics for their future lives.

It is important to assess students' mathematics and enhance their mathematics learning. There are various ways to achieve this aim. One of the ways is classroom formative assessment (Quellmalz & Pellegrino, 2009). Information collected from a classroom formative assessment could provide prompt formative feedback for subsequent teaching and learning as it provides the teachers with knowledge of strengths and weaknesses of a group of students for developing the next instruction set (Suurtamm, Thompson, Kim, Moreno, Sayac, Schukajlow, Silver, Ufer, Vos, 2016).

The desirable attribute of feedback is that it is timely and specific and helps to suggest ways to progress further learning. Also, more importantly, a timely score which is a type of feedback from classroom formative assessment provides the bridge between assessment and classroom teaching and learning, as it enhances future instruction and aids the individual learning progress of students (Van Lent, 2009, p.83). Given this, score and feedback are vital as they improve mathematics achievement.

2.2 Importance of formative assessment in education

Assessment is one of the important terms in education for providing evaluative information to make good decisions that lead to better education. Assessment is an essential tool in the teaching and learning process in the classroom. There are two types of assessment, namely summative assessment, and formative assessment. Summative assessment provides the high-stakes decisions about grades and changes in placement of students at the end of the school year to summarize what they have already learned (Thorndike & Thorndike-Christ, 2010). It aims to rank-order students and schools. Guskey (2003) pinpoints that the result from summative assessment, a score, or grade, cannot directly inform better teaching and learning. On the other hand, formative assessment is a type of teacher-made assessment that improves learning (Quellmalz & Pellegrino, 2009). This is because classroom formative assessment aims to gather the information that helps determine what students have learned and consequently adjusts their learning environment and instruction (Quellmalz & Pellegrino, 2009).

Guskey (2003) ascertains that formative assessment is the best tool for learning progress in the classroom. Classroom teachers apply formative assessment with some purposes: (1) to ensure students achieve instructional objectives; (2) to monitor the progress of the whole class and individual students; (3) to make decisions for preparing their next instruction; (4) to identify the necessary remedial teaching for students to achieve mastery level (Thorndike & Thorndike-Christ, 2010). Consequently, classroom formative assessments are applied as tools: (1) to accurately diagnose current ability of students, (2) trace their learning progress and (3) provide timely feedback. If a teacher applies the information from formative assessment, this assists their instruction to be effective and their students' learning to be better.

If the teaching and learning process is to be effective, there is a need to have assessment tools that are efficient and provide accurate scores and timely feedback. However, formative assessment comes with challenges, especially when classes are large. Teachers spend most of their time in assessment, which negatively affects the allocated teaching and learning hours. Stiggins (1999) highlights the importance of teacher assessment practices that directly affect their classroom formative assessment. They merely rely on the question collection book to develop tests when they do not have such practices. This indirectly reveals teachers' burden for test development. If such resources are not available, it is not assured that their questions may be valid or well-organized. Guskey (2003) suggests that the information from invalid and unreliable tests leads to the misleading classroom teaching and learning, thereby hindering the learning progress of students. The technology-based item-bank, from LOT and CAT, (that is, the collection of item or achievement questions), can solve such problem (Terzis, & Economides, 2011a; Bennett, 2011). This is because it involves a variety of high-quality questions.

2.3 Importance of classroom formative assessment with scoring and feedback

Classrooms commonly practice formative assessments as diagnostic tests to seek evidence of learning progress for students. Bejamin Bloom (1968) asserted that classroom formative assessment plays a key role in enhancing teaching-learning processes. Guskey

(2003) states that classroom formative tests can analyze scores to identify whether students achieve the targeted learning outcome and prepare the next instruction based on the evidence on the tests. However, in practice, there is no sophisticated analysis of student results for regular classroom tests, and the tally method on that result is enough (Guskey, 2003). This will be true and useful for the small class size (Guskey, 2003). Nevertheless, this will be troublesome for large class sizes because teachers cannot easily observe the strengths and weaknesses of their students (Guskey, 2003).

In addition, feedback from classroom formative assessment is a powerful tool to inform students what and how to progress their learning (Kluger, & DeNisi, 1996; Hattie, & Timperley, 2007; Hattie, 2008). The essential aim of providing feedback is to minimise the differences between students' current achievement and their intended achievement (Stobart, 2008). In addition, receiving timely feedback make the progress of active learning and the more-engaged instruction to the current achievement of students, according to Stobart (2008).

There are two tasks that teachers should provide to students in their classroom formative assessments: (1) the diagnostic evidence (which means ‘score’ or ‘right or wrong information’ on each item or the whole test); and (2) the information and way to progress their learning (which means elaborated feedback) (Thorndike, & Thorndike-Christ, 2010). In PPT modes, the manual feedback of a teacher is provided, and it may be delayed if the class size is large. CBT modes can solve this problem (Hattie & Timperley, 2007; Lopez, 2009).

Some previous studies underline the importance of time allowed to receive and apply feedback, the content in the feedback, and the willingness of students towards feedback (Hattie & Timperley, 2007; Stobart, 2008). Clariana, Wagner, and Murphy (2000) emphasise that if students can receive immediate feedback on each item as soon as they have responded to that item during the test, this provides students another chance to manage to answer or solve a problem. The authors conclude that such feedback cannot help to enhance learning progress well.

Importantly, feedback should involve clear information presentation (Mory, 2004) and available reference materials (Stobart, 2008). The feedback that explains the correct answer and refers to relevant learning materials is more effective for learning progress (Lee, Lim, & Grabowski, 2010). Additionally, feedback on learning progress is related to the willingness of students to receive feedback and apply it for them to improve learning (Hattie & Timperley, 2007; Stobart, 2008). Timmers and Veldkamp (2011) show that students pay more attention to feedback for an incorrect answer to a question than to correct answers to that question.

Nowadays, some classrooms have already applied LOT and CAT in place of PPT. Such test modes can provide prompt feedback to enhance learning progress, supply remediation, and adapt to all student abilities (Hartnell-Young, Harrison, Crook, Davies, Fisher, Pemberton, & Smallwood, 2007; DEECD, 2010). However, some studies show contradicting findings that LOT, and CAT have no effect on students' learning progress (Nicol & Macfarlane-Dick, 2007; Kluger & DeNisi, 1996; Hattie & Timperley, 2007; Shute, 2008). Further, other studies still show that there is no positive impact of feedback from LOT and CAT on learning progress rather than feedback from PPT (Hattie & Timperley, 2007; Shute, 2008; Stobart, 2008). Consequently, more studies need to find out the impact of LOT and CAT and their feedback on learning progress of students.This study examines the effects of feedback from classroom formative assessment in the different test modes on students' mathematics achievement. In addition, it explores the attitudes of students towards feedback from the different test modes and compares the effect of timely feedback from LOT and CAT with that of manual teacher feedback from PPT.

2.4 Integration of ICT into assessment as innovative paradigm shift

Typically, educational systems use assessment for accountability across all levels of the national education system, and for the improvement of the school system and for the support of student learning improvement (Kellaghan, & Greaney, 2001). The ultimate purpose of assessment is to improve education quality, which leads to improvement in student learning outcomes (Kellaghan, & Greaney, 2001). Effective assessment enables teachers to track learning progress of students and identify areas for improvement. It also enables students to demonstrate their achievement through qualifications they gain from assessments (Kellaghan, & Greaney, 2001).

In addition, assessment is the cornerstone of a high-performing education system (MOE NESP, 2016). For instance, lessons from the assessment system of the nations in the Asia Pacific Region have led policy makers and change agents to reform their assessment systems from content-based assessment systems to skilled-based ones (MOE NESP, 2016). The reform prioritises the acquisition of skills that their educational system expects students to know and be able to do in their lives because of their time spent in school (MOE NESP, 2016).

Consequently, assessment reform is a key part of basic education reforms, and it can help to improve classroom teaching and student learning achievement.

As one of the key assessments, classroom-based assessment involves formative assessment and summative assessment. Formative assessment can enable teachers to adapt their teaching to students' needs, and support students to expand their own learning while summative assessment can provide a snapshot in time about student learning (MOE NESP, 2016). In addition, formative assessment can provide teachers and students with vital information about student learning progress (MOE NESP, 2016). The information includes the formal grading and the more informal observation of student work.

The use of information and communication technologies (ICT) in the twenty-first century is becoming more and more important in all aspects of life. This is because ICT permeates all aspects of life, providing newer, better, and quicker ways for individuals to learn, network, and access information across the globe. Its importance will continue to grow and develop in education (Hernandez, 2017). ICT systems are integrated to support education innovation more and help to shift towards student-centred learning settings (Oliver, 2002). In addition, it can give a better solution to overcome the tensions between teachers and students (Oliver, 2002). The assessment strategies also need to be better harmonised with 21st century learning approaches by developing ICT-based assessments and re-focusing on the importance of providing timely and meaningful feedback to both learners and teachers (Redecker, & Johannessen, 2013).

The ICT integration in assessment and test systems had already commenced before 1990. Bunderson et al. (1989) forecast generations of computerised educational measurement, namely: computer-based assessment (CBA) and computer-adaptive Test (CAT) that are feasible in the classroom as an integral part of regular teaching and learning. Martin (2008) pointed out that both CBA and CAT have led to more effective and efficient delivery of traditional assessments. Today, CBA and CAT applications are now seen in mainstream assessment systems. Some developed nations have already applied computer-based tests. The computer-based tests are already being administered widely for various educational purposes, especially in the United States (Csapó et al, 2010), but increasingly also in Europe (Moe, 2009). CATs are commonly applied in every sector of education in the US, at primary and secondary school level (Bennett, 2010; Bridgeman, 2009; Csapó et al., 2010), and at higher education level (Bridgeman, 2009). Some European countries, such as the Netherlands (Eggen & Straetmans, 2009) and Denmark (Wandall, 2009), have already applied CAT.

It is proposed that computer-based assessment can be used to support formative and diagnostic testing (Redecker & Johannessen, 2013). For instance, in Norway, a linear-online test (LOT) of the formative assessment for digital literacy commenced in 2012. Hungary developed a networked platform for diagnostic assessment from an online assessment system developed for reading, mathematics, and science in grades 1 through 6. In the north of England and Wales, linear-online test is used to formative assessment of grammar and writing tasks to improve middle- and low-performing students in 2012 (Redecker & Johannessen, 2013).

Due to the advanced technology in computers and networking, the earliest computer-based or linear-online tests quickly jumped on CAT (Ejim, 2018). CAT can fulfil the promise of more accuracy with shorter tests than paper-and-pencil test (Ejim, 2018). The earliest computerized-adaptive testing programs for the large-scale study was the College Board ACCUPLACER® testing program in 1985, the certified Novell Corporation network engineer (CNE) examination in 1990, Education Testing Service (ETS) Graduate Record Examination (GRE) in 1992, the Graduate Management Admission in 1997, the United States Medical Licensing Examination (USMLE) in 1999 (Ejim, 2018). More recently, many organizations are considering a new CAT in the multistage testing mode to administer licensure and certification. For example, the multistage-adaptive Massachusetts Adult Proficiency Tests administered in 2006 (Sireci et al., 2006, cited in Ejim, 2018) the Uniform CPA Examination as the computer-adaptive multistage testing frameworks for large-scale applications (Breithaupt, Ariel, & Hare, 2010, cited in Ejim, 2018).

Southeast Asia Ministries of Education Organization (SEAMEO) member countries increasingly focused on the assessment reforms to embrace the paradigm shifts from summative assessment to formative assessment. This is because of the 'overly relying on summative assessment'; 'neglecting formative assessment', overemphasis on national examinations and certification tests (SEAMEO INNOTECH, 2015). Formative assessment and its feedback can ease exam pressure and stress and deeply emphasize passing the test rather than learning to pass high-stakes national examinations. However, provision of feedback is time-consuming, particularly for large classes, which is a weakness of formative assessment (SEAMEO INNOTECH, 2015). Hence, SEAMEO member countries introduced ICT integration in formative assessment as innovative practices and improvement strategies in the education system (SEAMEO INNOTECH, 2015). Some countries, such as Malaysia and Singapore, involved in SEAMEO, focus on introducing ICT in assessment, such as linear-online test and computer-adaptive test (SEAMEO INNOTECH, 2015).

2.5 Different test modes for classroom formative assessment

Various test modes of formative assessment can be used to assess mathematics in the classroom, such as paper-and-pencil test (hereafter PPT), and computer-based test (hereafter CBT). Of these modes, the traditional PPT has been the most common and dominant method of assessment delivery for many years (Weiss, 2011). PPT refers to a test or a set of questions in paper form and students write down answers in separate answer sheets within the assigned amount of time (Weiss, 2004). For PPT, teachers can manually provide the score and feedback to each student. Consequently, teachers generally report the overall score of a test for the individual (Ricketts, Filmore, Lowry, & Wilks, 2003). Although this assessment method takes a large number of teachers or large amount of time and effort of teachers to report the specific scores of each content area and cognitive level, students still prefer manual score and feedback from their teacher (Ricketts, Filmore, Lowry, & Wilks, 2003).

Additionally, the task for scoring, providing feedback, and adjusting instruction using score and feedback would be intimidating in PPT for large class sizes (Rudner, 1998; Ricketts, Filmore, Lowry, & Wilks, 2003). Consequently, the provision of immediate score and prompt feedback to individual students can be impossible for a teacher with a large class, preventing students from knowing their weaknesses in a timely manner. This, in turn, can negatively affect their learning progress.

The introduction of computers into the education system has led to the computer is one of the possible testing delivery devices. The 1970s saw the development of the CBT (Bunderson, Inouye, & Olsen, 1988; Drasgow, 2002). CBT refers to a test delivered in an electronic form and students respond electronically (Bugbee, 1996). At the end of the test, the computer can provide automatic scoring and feedback for their achievement (Bugbee, 1996). Applying CBT to educational measurement promises the opportunity for classroom assessment with fewer demands on teachers (Greenwood, Cole, McBride, Morrison, Cowan, & Lee, 2000). Moreover, classroom assessment in CBT mode, provides more opportunities for streamlined data collection, immediate scoring, and quick feedback to both students and teachers (Ripley, 2009). In a study by Deno (1985), CBT for classroom assessment has developed a bridge between instruction and assessment, providing teachers with a precise and straightforward way to monitor student progress in mathematics.

The integration of the Web into CBT designs is a logical next step from CBT to linear-online test (hereafter LOT) (van der Linden & Glas, 2000; Wainer et al., 2000; Luecht &

Clauser, 2002). In LOT, a set of questions is given on student computer screen via the internet and student answers by clicking on an option or typing their response in the given answer box on their computer, which is sent back to the main server or website within the assigned amount of time. Another technology-based assessment, computer-adaptive test (hereafter CAT), on the other hand, is the most sophisticated type of CBT delivered over the internet. In CAT, new questions are administered based on the responses of examinees to previous questions (Kingston, 2010). For example, if a test taker correctly responds to an item of middle difficulty level, they will be given a relatively more difficult item. Otherwise, a relatively easier/more straightforward item would be given. CAT can tailor the ability of students and automatically provide the score at the end of the test (Kingston, 2010).

Assessment which provides frequent and reliable feedback for students' learning progress is an important aspect of secondary school teaching and learning (Csapó, Molnár, & Tóth, 2009, p.120). Feedback and CBT scores help students assess their individual progress (Csapó, Molnár, & Tóth, 2009, p.120) and reveal which content domains and cognitive or proficiency levels are weaker (Meijer, 2009, p.104). Also, teachers with large class sizes can solve their problem of classroom formative assessment by applying LOT or CAT, which can automatically provide scoring and feedback to each student instead of manual scoring. Consequently, LOT and CAT become alternative test modes of classroom formative assessment instead of PPT.

2.6 Benefits and challenges of PPT, LOT, and CAT

Weiss (2004) identifies that PPT, and LOT are types of *fixed-length test*. *Fixed-length test* refers to a test that teachers or test developers create with a fixed number of items. The advantage of fixed-length tests is that they are useful for better measurement of students with average ability, not for students with a high ability or poor ability (Weiss, 2004). However, such tests cannot sufficiently cover the whole range of ability (Weiss, 2004). Therefore, the information or score from the test cannot be sufficiently relevant for the learning progress of students whose ability is either higher or lower than the average (Weiss, 2004). PPT and LOT can be the best measure for a student with average ability, but it is a substandard measure for intelligent performers and poor performers (Weiss, 2011).

A teacher can easily develop PPT mode without technological support or advanced psychometric knowledge. However, teachers who teach large classes still require more effort to develop items, assemble them in a test, print and deliver tests and score. They also pay more attention to exam invigilation and the security of question-and-answer sheets because all examinees take the same test at the same time.

Another mode of *fixed-length test*, LOT, is the upgraded one of computer-based *fixed-length-test*, applying the Web or online as a means of delivery (Weiss, 2004). Unlike PPT, the teacher needs less effort for test security in LOT because it can randomly assign items delivered to the computers of students and protect questions and answers with the password (Hensley, 2015). Sometimes, teachers do not need any effort for item development because LOT can draw and assign items from item-banks developed by experts. With less effort by teachers, LOT can provide feedback and specific scores in each content and competency level to students (Rashad, Youssif, Abdel-Ghafar, & Labib, 2008; Hensley, 2015).

Nowadays, classroom teachers apply Google Forms, Learning Management Systems (LMS), and Student Response Systems (SRS) to develop LOT (Waters, 2012). However, teachers will need computer accessibility, computer literacy, and computer familiarity for applying their classroom formative assessment in LOT mode. CAT is a scale-length test, referring to a test in which new items are administered based on the responses of examinees to previous items (Kingston, 2010). CAT can overcome the deficiency associated with fixed length tests because CAT measures the continuum from lower achievers to higher achievers more accurately and precisely within a short testing time (Özyurt, & Özyurt, 2015; Lazendic, 2015; Ling, Attali, Finn, & Stone, 2017). This raises the motivation of students and reduces their test anxiety (Rezaie & Golshan, 2015). Olson (2004) also ascertains that CAT assures better-prepared students, more substantive learning, and more effective teaching because it can identify and meet the needs of all students within the shortest possible test-taking time. Consequently, CAT could be a best measure for all students with different ability.

CAT is the best diagnostic tool, especially as the classroom formative assessment because of its immediate scoring and feedback that provide information that is more meaningful to teachers in designing the subsequent learning activities for poor performers especially and to all students by identifying their mastery level in the specific content area (Weiss, & Kingsbury, 1984; Eggen, 2007). CAT can assign items automatically from its item-bank, enabling teachers to save time. Like LOT, CAT can also provide automatic scoring and providing feedback. Test security in CAT is better than LOT and PPT because students do not take the same items according to their different abilities. Teachers need time for test administration in CAT.

However, classroom teachers will face many challenges related to technological issues and psychometric issues if they develop CAT. Some considerations to develop CAT are item-bank, item selection algorithms, scoring algorithms, and test delivery system (see Weiss, 2004; Way, Davis, & Fitzpatrick, 2006; Thompson & Weiss, 2011; Jacobsen, Ackermann, Egüez, Ganguli, Rickard, & Taylor, 2011). Consequently, classroom formative assessment in CAT mode will essentially require well-developed CAT systems, at the national level, designed by a team of psychometricians, technicians, and subject-specialists. According to the studies of Csapó, Molnár, and Tóth, (2009) and Meijer (2009), LOT and CAT make students improve their achievement rather than PPT. They proved that computer-based test modes could provide immediate feedback and scoring. Specifically, while LOT and CAT substitute in PPT, LOT can enhance the learning achievement because of prompt feedback (Hartnell-Young, Harrison, Crook, Davies, Fisher, Pemberton, & Smallwood, 2007). However, some studies by Hattie and Timperley, (2007), Shute (2008) and Stobart (2008) proved that there is no significant impact of test mode on learning progress in LOT and CAT modes. Further, researchers such as Nicol and Macfarlane-Dick (2007) indicated that LOT and CAT cannot improve student achievement.

2.7 Advantages and challenges of multiple-choice questions

There are commonly two types of mathematics questions: multiple-choice questions (hereafter MCQs) and constructed-response questions (hereafter CRQ). MCQ is composed of a stem (or questions) and more than two alternatives, a key answer and more than two distractors. Students can choose an alternative and describe it in their answer sheet, but they do not need to prove why they choose their answer. In CRQ, students need to construct and prove their answers on their answer sheet. MCQ is the more suitable and better question type for large class size than CRQ and ECRQs (Torres, Lopes, Azevedo, & Babo, 2009). This is because MCQs need less time for grading, with high scoring accuracy and objectivity (Torres, Lopes, Azevedo, & Babo, 2009). MCQs efficiently cover large numbers of content domains (Torres, Lopes, Azevedo, & Babo, 2009) and could be applied to measure from knowledge-recalling to more complex meta-cognitive (Torres, Lopes, Azevedo, & Babo, 2009). However, MCQs cannot measure the ability of information communication, explanations, organization, and the capability to produce original ideas as well as the writing ability of mathematics symbols which CRQs or ECRQs can measure (Torres, Lopes, Azevedo, & Babo, 2009). CRQs and ECRQs can also measure partial understanding of students, while MCQ with dichotomous scoring cannot. However, CRQs need more time for scoring than MCQs, especially for large classes.

Roediger and Marsh (2005) ascertain that student essentially need immediate scoring and feedback on MCQs for reviewing their incorrect responses or misconceptions to those questions. However, teachers spend a huge amount of time providing scores and feedback on MCQs in PPT mode for large classes (Rickards & Friedman, 1978). MCQs in CBT are more effective than in PPT because CBT can provide automatic scoring and prompt feedback for incorrect answers or misconception with less time and effort (Torres, Lopes, Azevedo, & Babo, 2009). CBT can record only the final responses of students, but it cannot determine their partial understanding. In PPT, teachers can see how students solve MCQs in their answer sheets (Rickards & Friedman, 1978).

MCQ is more difficult to develop than CRQ or ECRQs (Oermann, & Gaberson, 2006). Especially, the development of MCQs for higher-order thinking skills require expertise, time, and funding (Oermann, & Gaberson, 2006). Good MCQs should involve unambiguous language and well-functioning distractors (i.e., alternatives describing incorrect answers in MCQ). Ambiguous language leads learners to respond wrongly because of incorrect interpretation (Roediger & Marsh, 2005). Students are more likely to get the right answer by guessing MCQs with distractors that cannot function well. Therefore, such MCQs cannot be an effective measure to check the misconceptions of students. However, Roediger and Marsh (2005) ascertain that developing an item-bank can solve the problem of developing high-quality MCQs. In the bank, MCQs are kept securely, and they can be reused, refined, and reviewed. Consequently, teachers save time for item development (Roediger, & Marsh, 2005).

2.8 Assessment of mathematics and its non-cognitive predictors

The realization of the importance of mathematics education leads to the assessment of mathematics knowledge in cognitive and content domains and the non-cognitive predictor. Such assessment is found in international, regional, and large-scale national studies as well as state assessments and classroom assessments (Anderson, Lin, Treagust, Ross, & Yore, 2007, Education Policy and Data Center, 2012). Through the studies and assessments, the areas in which students are good at; and in which they need to improve learning, as well as factors that predict the achievement of students are revealed. These studies integrate student achievement and contextual predictors to disclose current achievement levels and the national education system. The contextual predicators have associated with the factors of students, teachers, and school principals (Education Policy and Data Center, 2012).

The assessment of mathematics achievement and numeracy plays a significant role in large-scale studies. Among them, the most prominent studies for mathematics achievement are *Trends in International Mathematics and Science Study* (TIMSS); and *Programme for International Student Assessment* (PISA). In those studies, along with the mathematics achievement of students, information is collected on student attitudes and perceptions related to schooling, home background, and school information (Anderson, Lin, Treagust, Ross, & Yore, 2007; Mullis, & Martin, 2017; Gustafsson, 2018). A regional study, *Latin-American Laboratory for Assessment of the Quality of Education* (LLECE), measures mathematics in logical reasoning, problem-solving, and accuracy and collects information about the context of learning by administering questionnaires to students, their parents, teachers, and school principals (Education Policy and Data Center, 2012). Another regional assessment for countries in anglophone East Africa is the *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SACMEQ), which examines Grade 6 students' mathematics achievement and its related factors (Education Policy and Data Center, 2012).

Another regional mathematics assessments in the *Southeast Asian countries, Southeast Asia Primary Learning Metrics* (SEA-PLM), assesses mathematics learning outcomes of Grade-5 students and collects key factors from students, parents, teachers, and schools to monitor the progress of students (Belisle, Cassity, Kacilala, Seniloli, & Taoi, 2016). The factors of students, families, teachers, and schools directly influence students' achievement improvement according to a study of ASEAN countries by Maamin, Maat, and Iksan, (2021). The results of the study by Akyüz (2014) showed that students' attitudinal factors towards mathematics learning, social economic status, and home and school educational resources positively affect the TIMSS 2011 mathematics scores of eighth-grade students in Singapore. There is also impact of student-, teacher-and school-level factors on mathematics achievement in Singapore (Ker, 2016). Students' mathematics achievement in Indonesian senior high school is affected by their socioeconomic status and parental involvement in learning (Kusaeri, et.al, 2018). Some studies conducted in Myanmar reveals that there is positive relationship between attitudes, motivation, and self-efficacy toward mathematics with mathematics achievement of grade 10 students (Thein & Thein, 2018; June, & Eamoraphan, 2019).

Consequently, it is paramount to mention that every education system considers the assessment of mathematics education as one of the key players in education innovation. In addition, the assessment of mathematics achievement is associated with examining contextual factors such as attitude and background information.

2.9 Theoretical background of extraneous scales on achievement improvement due to test modes

In social research, where multistage sampling is employed to reduce cost and collect more data, it is essential to consider a multilevel analysis (Raudenbush & Bryk, 1992). Most scales from social sciences studies are intertwined in a hierarchical structure (Hox, 1995; O'Connel & McCoach, 2008; Teddlie, & Tashakkori, 2009). In education, students are nested in classrooms, and classrooms are nested in schools. Student scales such as achievement, attitude, ability, and proficiency, can vary according to factors from their classrooms and their schools (Hox, 1995). Such hierarchical data analysis can provide meaningful results (Hox, 1995; Hox, 2002; O'Connell & McCoach, 2008). However, some studies have failed to address issues of hierarchical structure related to their datasets. For example, while investigating students' attitudes, motivation, and anxiety towards assessment at the student-level, it needs to consider that students are nested in the classroom and classes are nested in school. Scales in education are nested in hierarchical structures. Therefore, examination of single level analysis will not provide meaningful explanations for the introduction of new assessment modes. One of the studies conducted by Duque and Schools (2016) compares differences in the scores of students in the Partnership for Assessment of Readiness for College and Careers (PARCC), based on two test modes, which are PPT and CBT. The research methodology of their study is two-level hierarchical linear modelling (HLM) because of students nested in schools.

In most previous studies investigating the effectiveness of LOT or CAT, the researchers identified a direct relationship of the scales on student achievement in each level. However, very few studies focus on the relationship of some scales on student achievement in the hierarchical structure. For example, Backes and Cowan (2018) suggest that, due to CBTs, the demographic background scales of teachers could directly affect the achievement improvement of students. In addition, background scales of students could directly affect their achievement improvement. This study will investigate which student, and class scales have a relationship with student achievement improvement, and which have a cross-sectional relationship with student achievement.

In this study, the characteristics and attitude attributes of the principals, teachers, and students are related in the multilevel relationships. In other words, there are multilevel relationships among teacher and student characteristics and educational outcomes at both teachers- and students-levels in hierarchical structures (Kozlowski & Klein, 2000). This concept aligns with multilevel organisation theory (MOT) (Kozlowski & Klein, 2000). This theory explains that the interaction process can co-occur at the lower (or student-level) and higher levels (or teacher-level). The theory specifies the relationships between the lower and higher levels by highlighting the top-down processes, referring to the direct effects from a higher-level unit (teacher characteristics) to a lower-level unit (such as student performance). In the teacher-student relationship context, the higher-level unit refers to teacher factors such as teachers' professional knowledge, whereas the lower-level unit refers to an educational outcome such as student performance. In this study, student-level scales refer to the effective scales, whereas the teacher-level scales refer to the aggregated student-level scales at the teacher-level.

Regarding literature review in order to examine the different mode effects, there are three issues: psychometric properties of the modes, technology issues and participant issues. Technology issues related to the modes are overcome by the application of CBT. Test developers must examine the psychometric properties of items in PPT, LOT, and CAT for examining their validation. Factors related to participants need to be investigate because they affect students' achievement either directly or indirectly. According to social science studies, students are nested in classrooms that are nested in schools. In examining the effect of assessment mode on achievement, scales in three levels which are student-level, class-level and school level cannot be left.

2.9.1 Classroom-related scales

Teachers play a key role in classroom instruction, test administration, and the achievement of students (Thorndike, & Thorndike-Christ, 2010). Some major scales related to teachers' background information affect students' achievement due to the effectiveness of CBTs. Those scales are the qualifications of teachers, teaching experience, computer experiences and CBT experiences (Jamil, Topping & Tariq, 2012) and their attitude towards CBT (Khoshsima, et.al., 2015). This study investigates whether above scales affect the achievement of students due to PPT, LOT, and CAT.

There are relationships between teachers' background information and perceptions towards CBT and PPT. Jamil, Topping & Tariq, (2012) identified some related major scales, such as, gender, qualifications, teaching experiences, computer-training certifications and CBT examination experiences, that are peculiar to teachers' perceptions towards CBT and PPT. Most teachers prefer CBT to PPT. Particularly, female teachers, highly qualified, less teaching experienced, less computer familiarity. Some teachers with the CBT examination experiences were more positive towards CBT. Furthermore, teachers' self-efficacy, enjoyment, usefulness, behavioural intention to use CBTs, system satisfaction, system challenges are examined in a study which is conducted by Fageeh (2011). This study points out that examining such scales is important for creating CBT and they are required to identify the effectiveness and success of the adoption, implementation, and diffusion of CBT. Teachers are ones of the key performers in this integration of new test modes. There is a positive effect of teachers' attitude towards CBTs, on their students' learning progress due to these test modes (see Hu, Chau, Sheng, & Tam, 1999; Venkatesh, 1999; Ong & Lai, 2006; Lee, 2006; Van Raaij & Schepers, 2008, Khoshsima, et.al., 2019). According to these previous studies, there is a relationship between the attitude of teachers to test modes and their students' achievement improvement.

The highest education levels of participant teachers and their highest qualification level in education are extrernal factors that can affect the achievment improvement of students due to the effect of PPT and CBTs.Therefore, their highest education levels and their highest qualification level in education positively influence the achievment improvement of their students due to the effect of PPT and CBTs (Brown, 2004, Jamil, Topping, & Tariq, 2012). The workload, especially in class size and multi-subject teaching, negatively affects students' learning because work loading can decrease the quality of test development, administration, teaching and learning (Burghof, 2001; Eggen, 2007; Rashad, et.al, 2008). Teachers' formative assessments are also important for their students' achievement improvement due to PPT and CBTs. These practices of teachers help students to achieve better informs them on how well they are doing, helps them to improve specific points, encourages students to work hard and guides them to what they need to focus on when they are having difficulty (Ellis, Loewen & Erlam, 2006, Quyen, & Khairani, 2017).

Another scale related to formative assessment, teachers' attitude towards formative assessment can help their students to progress in their learning, and feedback helps them improve specific points, or to help plan their learning. There is a direct effect of the attitude of teachers towards formative assessment and their students' achievement progress in CBTs. The attitude of teachers to formative assessment positively influences the achievement improvement of their students in any test modes (Pinchok & Brandt, 2009; Opre, 2010, Quyen, & Khairani, 2017). However, Nesa (2014) mentioned that there is no relationship between the attitude of teachers towards formative assessment and their students' achievement improvement in any test modes.

Teachers' ICT-related scale needs to be considered in studies that investigated the effectiveness of CBTs. ICT familiarity and the attitude teachers towards ICT improve the performance of students in CBTs. Quyen, and Khairani (2017) found that that teachers' ICT familiarity and their attitude towards ICT are the factor, which have positive effect on the achievement of students in CBT. Hence, this study will examine to what extent the gender of teachers, highest education levels, and their highest qualification level in education affect their students. In addition, this study will consider the class size, multi-subject teaching, ICT familiarity, attitude to ICT, practices of formative assessments, attitude towards formative assessment and attitude towards test modes (PPT, LOT and CAT) of teachers that have impacts on their students' achievement improvement due to the effect of PPT, LOT and CAT.

2.9.2 Student-related scales

Generally, the background information of students, such as gender, expected education, parental education levels, directly impact their achievement improvement due to different test modes. Some contextual background information such as gender influences cognitive achievement between PPT and CBT (Leeson, 2006; Hensley, 2015). However, there is no score difference for gender in other studies due to the test mode effect (Clariana, & Wallace, 2002; Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Terzis, & Economides, 2011b; Poggio, & McJunkin, 2012). Students' parent education level and their expected education level are also the key factors in their learning progress. Parent education level significantly influences more on PPT achievement than CBT (Clariana, & Wallace, 2002; Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Acharya & Joshi, 2009). In addition, their expected education level or their inspiration for their further education are directly related to their learning progress due to different test modes (Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Poggio, & McJunkin, 2012; Hensley, 2015).

Motivation is considered as a critical determinant of behaviour and its level can determine the persistency of a test-taker for providing the stability of performance speed (Cohen, Sparling-Cohen, & O'Donnell, 1993)). The motivation of students towards their learning helps their achievement improvement due to PPT and CBT (Acharya & Joshi, 2009; Hensley, 2015; Alaga, 2016). In addition, their self-efficacy and attitude towards mathematics are key factors for improving achievement due to encountering CBTs (Chua, 2012; Timmers, et.al, 2013; Hensley, 2015; Chua & Don, 2013; Alaga, 2016). Terzis, and Economides (2011b) also find out motivation, self-efficacy and attitude towards mathematics have impact on their achievement improvement due to PPT and CBTs. The studies by Ak and Sayil, (2006), Newton and Mwisukha (2009) and Geddes, Murrell, and Bauguss (2010) indicated that there is a significant relationship between the concept of attitude towards learning mathematics and academic achievement improvement due to test modes. Candeias, Rebelo and Oliveira (2010), and Verešová and Malá (2016) mentioned that higher academic achievement is associated with a positive attitude of a student towards learning due to any test modes. Schunk (1995) has highlighted that motivation and self-efficacy are key factors in students; achievement improvement because their level of motivation and self-efficacy is enhanced when students achieve more due to CBTs. The study by Liu and Koriala (2009) also shows that mathematics self-efficacy is a significant positive predictor of mathematics achievement because of the effect of CBTs.

ICT familiarity and attitude towards ICT are important factors affecting students' achievement improvement due to ICT-related test modes or CBTs. Among them, amount of ICT familiarity and attitude towards ICT help some examinees to achieve more in CBT and improve their motivation in their learning (Parshall & Kromrey, 1993; Clariana & Wallace, 2002; Weaver and Raptis, 2001; Goldberg, & Pedulla, 2002; Leeson; 2006; Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Thorndike, & Thorndike-Christ, 2010; Duque & Schools, 2016). However, some studies reveal that the amount of ICT familiarity and attitude towards ICT are not related to performance improvement due to different assessment modes (Waston, 2001; Wang, Jiao, Young, Brooks, & Olson, 2008).

In this study, attitude towards formative assessment is an important factor because achievement improvement is mainly related to the effect of formative assessment in different test modes. The studies by Bennett (2011), Ajogbeje, (2014) and Quyen and Khairani (2017) reveal the significant relationships between the attitude of students towards formative assessment and their improvement in achievement in PPT and CBTs. The more students have a positive attitude towards formative assessment in any test modes, the greater their achievement improvement. Attitude towards test modes is considered as a key scale that will affect the improvement of students due to the effect of test modes. Several studies reveal that the attitude of students towards CBTs positively impacts their achievement improvement due to those test modes (see Hu, Chau, Sheng, & Tam, 1999; Venkatesh, 1999; Ong & Lai, 2006; Lee, 2006; Van Raaij & Schepers, 2008; Way, Davis, and Fitzpatrick, 2006; Penuel, 2006). Further, these studies mention that students who believe that computer-based test modes will improve their achievement for the course will achieve more in their learning.

According to previous studies, this study will examine to what extent students' gender, parental education, expected education, motivation, self-efficacy, attitude towards learning mathematics, ICT familiarity, attitude towards ICT, attitude towards formative assessment, and attitude towards test modes (PPT, LOT and CAT) impacts on their achievement improvement due to the effect of PPT, LOT and CAT.

2.10 Theoretical Framework

The basic concepts of the theoretical framework in this study lie in Biggs's 3P model (Presage-Process-Product) (Biggs, 1989) (see **Figure 2.1**). There are three stages (Presage-Process-Product) in Biggs's 3P model. The factors in the presage stage provide the contextual background, which impacts the teaching-learning process in the classroom and the learning outcomes. The factors in the process stage facilitate the effect on the outcome. Finally, the resulting factor, learning outcome, in the product stage, is influenced by the factors from the presage and process stages.

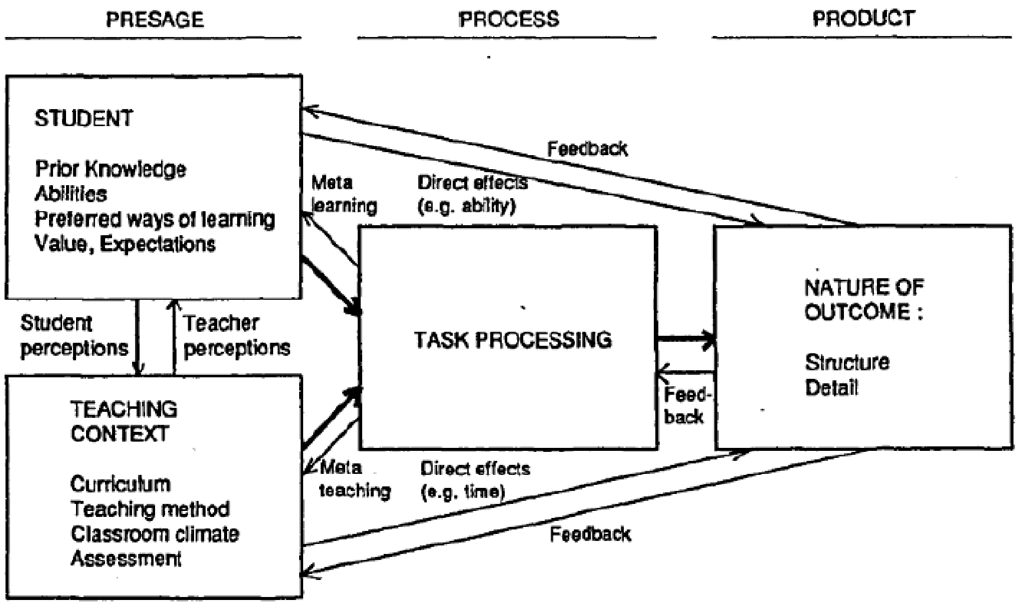

Figure 2.1 Biggs' 3P Model of classroom learning (adopted from Biggs, 1989)

In the theoretical framework of this study, the presage stage is defined as that students, teachers, and schools bring into the classrooms. This is because the model identifies that the demographic characteristics of students and teachers and their perceived attitudes and motivation, anxiety, self-efficacy, and concepts of learning mainly influence the achievement of students. In addition, school characteristics and ICT support the impact on the student- and teacher-level scales. The process stage is defined as the teaching, learning, and assessment practices in which learning approaches of students, the methods of teaching by the teachers, and their assessment practices are involved. These scales impact the mathematics achievement of students, which is from the production stage. This framework will analyse the interaction of students' and teachers' characteristics and teaching, learning, and assessment approach to evaluate the mathematics achievement of students. **Table 2.1** shows the focus of each stage for this research, and **Figure 2.2** shows the theoretical framework, which describes the interactive relationship of factors from the different levels for all test modes.

The study aims to compare the impacts on the achievement improvement of students by using three test modes, namely: PPT, LOT, and CAT. Thus, there is a graphical figure (**Figure 2.2**) for the theoretical framework, which represents PPT, LOT, and CAT. In conclusion, this study will investigate the mode effect while administering both online and paper versions of assessments. The scales at the student-level focus on the characteristics and achievement of students. They include gender, parent education level, expected education level, motivation, self-efficacy, attitude towards mathematics learning, attitude towards formative assessment, attitude to ICT, IT familiarity, and attitude towards test mode. The scales at the class-level to the characteristics of teachers, involving highest education levels, highest qualification level in education, class size, multi-subject teaching, assessment practices of formative assessment, attitude towards formative assessment, attitude towards ICT, ICT familiarity, and attitude towards test modes. Based on these scales of the two levels, two contextual questionnaires (for students and teachers) are developed for the evaluation of different mode effects.

Table 2.1
Focused Factors of the Presage, Process and Product Phases

Stages	Focused Factors
Presage	1. students' gender, parent education level, expected education levels
	2. teachers' highest education levels, highest qualification level in education, class size, multi-subject teaching
Process	1. students' motivation, self-efficacy and attitudes towards learning mathematics, attitude towards formative assessment, ICT familiarity and attitude towards ICT, attitude towards test modes (PPT, LOT, and CAT)
	2. teachers' formative assessment practices, attitude towards formative assessment, workload, ICT familiarity and attitude towards ICT, attitude towards test modes (PPT, LOT, and CAT)
Product	1. students' mathematics achievement improvement due to test modes (PPT, LOT, and CAT)

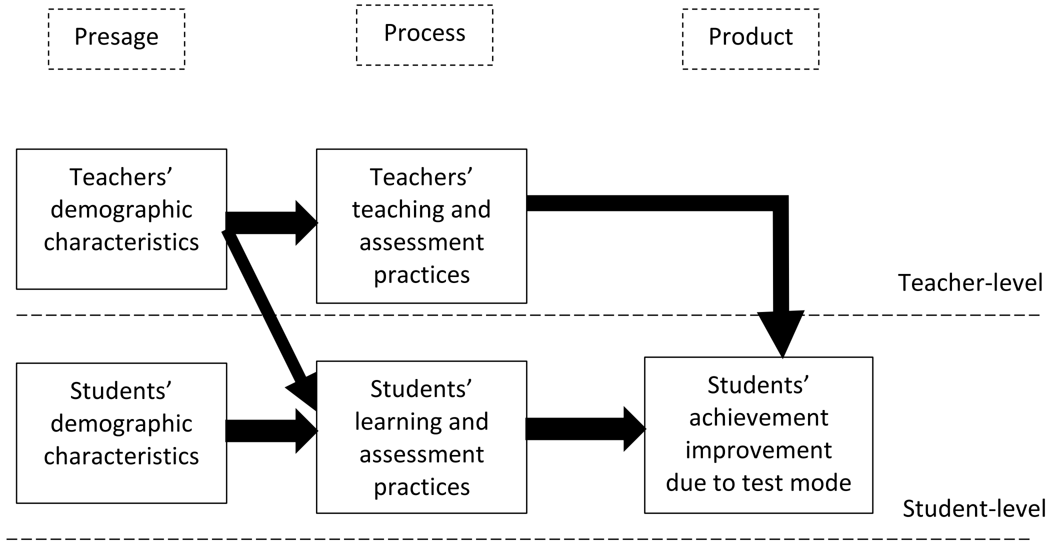

Figure 2.2 Theoretical framework for the study (adopted from Biggs, 1989)

2.11 Summary

This chapter has presented the literature review and theoretical framework. The chapter described the importance of formative assessment in education, the use of different test mode for the classroom formative assessment, the benefits, and challenges of PPT, LOT, and CAT, and extraneous scales on achievement improvement due to test modes. The literature review is used to determine the related factors from students and teachers directly affecting the achievement improvement due to PPT, LOT and CAT and their interrelationship. Then, teacher and student factors that are related to contextual and psychological background indirectly influence the achievement improvement of students due to PPT, LOT, and CAT. The Biggs' 3P model of classroom learning is adopted as a theoretical framework. This model includes the presage stage, process stage, and product stage. Concerning the factors mentioned in the literature review, the related factors are mentioned in each stage. In addition, this study constructed the proposed theoretical framework. The next chapter, Chapter 3, describes the research methodology and the data collection process for this thesis.

Chapter 3

Research Methodology

3.1 Introduction

This study is the very first of its kind to be administered in the Yangon Region, Myanmar. The overall aim of this study is to investigate which test modes (paper-and-pencil test (PPT), linear-online test (LOT), computer-adaptive test (CAT)) leads to the better improvement of students in mathematics learning from the Myanmar context as the research questions in Chapter 1 and literature evidence in Chapter 2. This chapter discusses the suitable research design and methods used to gather, analyse, and interpret the data to secure reliable data for analysis and for appropriate findings, conclusions, and recommendations. The chapter proceeds with a discussion of sampling methods, including the population, sample, and research locale. Then, it follows the procedure or steps of data collection in constructing the experiment, how the gathered data were treated and interpreted. In addition, a section at the end of this chapter highlights the key points.

3.2 Intervention design with explanatory sequential mix-method

Intervention study, which is also called experimental study is a kind of quantitative research in which the investigator or researcher determines whether an activity or material or treatment or intervention makes a difference in results for participants (Creswell, 2014; Creswell, 2015). Through the intervention study, the researcher assesses this impact by giving one group one set of interventions and withholding the set from another group (Creswell, 2014; Creswell, 2015). To get best understanding and explanation a research problem, it is necessary to combine quantitative and qualitative data which is called mixed method (Creswell, 2014; Creswell, 2015). Thus, intervention design with explanatory mixed method is the most appropriate research methodology because the design can explore the problem with a new experiment or intervention into the current educational context by combining data of quantitative and qualitative approaches (Creswell, 2014; Creswell, 2015). The design was applied to conduct three experiments, and a survey and add the qualitative data after the experiments and survey (Creswell, 2015). The experiments in this study examine which test mode is the most effective and efficient for a large classroom to improve the learning progress of students. The survey explores the perceptions of teachers and students in different test modes. This study applies qualitative data through interview to explain the statistical data from the experiment in detail. This design can be called an intervention design with explanatory mixed method (Creswell, 2014; Creswell, 2015). Consequently, the design predominantly quantitative study with limited follow-up qualitative aspect to provide a better understanding of the research problem.

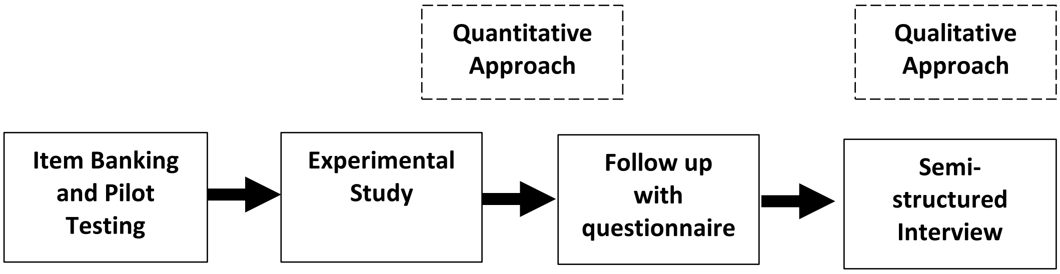

Figure 3.1 Intervention design with explanatory mix-method

3.3 Counter-balanced quasi-experimental research approach

Experiment research is the most powerful method for investigating the cause and effect (Fraenkel, Wallen, & Hyun, 2011, p.266). Experimental research can investigate the effect of a new intervention and how it happens (Fraenkel, Wallen, & Hyun, 2011). There are three main types of experiment research approaches: (1) pre-experiment, (2) true-experiment, and (3) quasi-experiment (Creswell, 2009; Fraenkel, Wallen, & Hyun, 2011). Pre-experimental research is a study in which a single group will be provided with the interaction during the experiment. In quasi-experimental research, the participants are not randomly assigned to control or experimental groups.

True-experimental research that applies random assignment of participants is the best because random assignment helps overcome threats to internal validity and provide optimal findings (Creswell, 2009: Fraenkel, Wallen, & Hyun, 2011). However, that research is often associated with practical problems such as ‘compliance, cost-effectiveness and the ethics of withholding full participation’ (Koedinger, Mclaughim & Heffernan, 2010, p. 497). Moreover, random assignment cannot be possible for all experimental research in social studies because the participants are already in intact groups, naturally formed groups (Fraenkel, Wallen, & Hyun, 2011; Creswell, 2009). If students already in a class are randomly assigned into groups for an experimental study, this will need more classes and teachers and reduce their willingness for full participation in the study. Consequently, quasi-experimental research is more favoured in social studies. Slavin (2008) also suggests that findings from quasi-experiment are more likely to be more similar to those of true experiments.

Counter-balanced research is quasi-experimental research. All experimental groups are exposed to all experiments to equate them (Fraenkel, Wallen, & Hyun, 2011). The random orders of the experiments reduce order effects and participant characteristics threat to internal validity, which most quasi-experimental research commonly face. It helps to get unbiased estimates of the effects of experiments. However, its weakness is multi-experiment interference because the performance during a particular experiment may be affected by one or more of previous experiments (Fraenkel, Wallen, & Hyun, 2011).

Counter-balanced quasi-experimental research can cover both between-subject design and the within-group design, which most experimental design cannot cover (Creswell, 2009). Between-subject design can compare the outcome from two or more individual groups affected by different experiments. Within-group design can study only one group that are affected by different experiments at different times within the experimental period. Pollatsek and Well (1995) suggest that counter-balanced design is a powerful experimental design for repeated measurement because data can be specifically and thoroughly analysed by between-subject design and within-group design.

This study applied counter-balanced quasi-experimental research because this thesis aims to compare the effect of the different test modes in terms of the achievement progress of students during the experiment, their perspective towards these modes and their teachers' perspective. Consequently, all participants need to experience all the test modes (see **Table 3.1**).

Table 3.1

Sample counter-balanced quasi-experimental research

Groups	Pre	Expt 1	Obs	Pre	Expt 2	Obs	Pre	Expt 3	Obs
Group-1 (Test mode-order-1)	Pre	PPT	O	Pre	LOT	O	Pre	CAT	O
Group-2 (Test mode-order-2)	Pre	LOT	O	Pre	CAT	O	Pre	PPT	O
Group-3 (Test mode-order-3)	Pre	CAT	O	Pre	PPT	O	Pre	LOT	O

Note: Pre – Pre-test; Expt – Experiment; Obs – Observation; O – Post-test; PPT – Classroom formative assessment in PPT mode; LOT – Classroom formative assessment in LOT mode; CAT – Classroom formative assessment in CAT mode

(Adapted from Fraenkel, Wallen, & Hyun, 2003)

Table 3.1 shows the sample of the research. Group-1 (Test mode-order-1) receives PPT and is observed by post-test (summative assessment), then receives LOT, observed by post-test, and lastly receives CAT, observed by post-test (See Group-2 (Test mode-order-2) and Group-3 (Test mode-order-3) in **Table 3.1**). The researcher will examine which mode is more effective in enhancing the mathematics learning progress of students by comparing the average score between different groups and within group. This will help to examine the score comparability across modes to identify tests' psychometric properties.

3.4 Purposive stratified random sampling

The population of a study is defined as the larger group to which the findings of the study will be applied (Fraenkel, Wallen, & Hyun, 2011). Due to the financial and time constraint, researchers collect the information from a sample (Fraenkel, Wallen, & Hyun, 2011). The smaller sample helps researchers save time, effort, and finance, but this limits the generalizability of the finding (Fraenkel, Wallen, & Hyun, 2011).

This study selects a representative sample of schools by stratified random sampling. In the method, sub-groups or strata are selected for the sample in the same proportion out of the population (Fraenkel, Wallen, & Hyun, 2011). In Yangon Region, there are four districts: North Yangon, East Yangon, South Yangon, and West Yangon (Department of Basic Education, 2016). **Table 3.2** shows a number of schools in districts. In the district of the east, there are more populated than other districts. Consequently, the number of schools is twice than another district. The 11 schools out of the total can access the internet and multimedia classroom, and a number of schools out of the total area from the very remote area. By applying a table of random numbers, 50% from each district high-schools with ICT access were randomly selected. There are five schools out of eleven schools with ICT access that participated in the main experiment. In addition, a classroom from another school with ICT access participated in the pilot study. The six classrooms from the other two schools paricipated in the data collection of the item bank.

Figure 3.2 Four districts in Yangon Region, Myanmar (source from https://th.bing.com/th/id/OIP.bDmIsrLo337KXuLMx-_3pAHaKP?pid=ImgDet&rs=1)

Table 3.2

Number of schools in population and in selected sample

District	Total	ICT access	Selected
North	47	2	1
East	102	5	2
South	54	2	1
West	49	2	1
Total	252	11	5

3.5 Data collection

The main research approach of this study is counter-balanced quasi-experimental research. As LOT and CAT are newly integrated test modes in Myanmar high schools, item-bank need to be constructed. Therefore, there are five phases of data collections for: (1) item-bank (2) construction of CAT, LOT and PPT, (3) pilot testing (4) counter-balanced quasi-experimental research (5) semi-structured interview. Before the experiment, this study conducted a pilot study to confirm the feasibility of research conduction.

3.5.1 Item-bank

An item-bank refers to a source of items, which are the valid, reliable, and defensible test items for any high-quality computer-based test (CBT) (Velasco, 2017). Although item-bank can be applied to assign items for any form of test modes (for example, PPT, LOT, and

CAT), the item-bank is a key performer for administering CAT. There are some considerations for assembling items in an item-bank: (1) planning items, (2) assembling items, (3) calibrating items, and (4) entering items into a database for an item-bank.

(1) The initial step in item-bank is used to draw a blueprint in order to identify each item for content areas and cognitive domains. Then, Bloom's Taxonomy can be used to categorise items into different cognitive domains that assess different thinking skills (Granello, 2001).

    Items are selected into three sub-topics (*function*, *remainder,* and *factor*) from the algebra content of the Grade-10 textbook. The items covered three cognitive domains (knowing, reasoning, and applying). For this study, items were identified by creating a blueprint and consequently used an equal number of items in each content and cognitive domain.

(2) The next stage is item assembling. The present study estimated the total number of items in the item-bank. Wood and Skurnik (1969) recommended a total of 200 or more items. However, this study applied the method of Sands and coworkers (1997). An item-bank contains a ratio of items to the number of test-takers. For instance, if there are 30 students (i.e., the average number of students in each class) (x=30) who will take a test on computer and the minimum number of items to the maximum number of items is 5 to 10, 150 to 300 items are needed for 40 test-takers. The minimum number of items is 150, and the maximum of items is 300.

    Hence, 165 items of Grade-10 are selected from the previous matriculation examinations (from 2011 to 2017), which have already standardised (Myanmar Examination Board, 2018). This number helps to collect the targeted number of items, and the number collected is of MCQ items with one key answer and four distractors.

(3) The third step is calibrating items for the item-bank. For 165 items, seven test booklets were selected. In each test-booklet, there are 25 items for 25 minutes to answer. There are two hundred and fifty student participants from two schools. After examining the psychometric properties of items to fit the Rasch measurement model and on the same Rasch logit-scale, about 150 items with good quality were assembled in the item-bank.

(4) The last step is to enter items into the item-bank. The selected items were assembled to the item-bank. The item-bank helped test developers review and manage items easily and store items with high security.

3.5.2 Construction of CAT, LOT, PPT and elaborated feedback

For this thesis, the experiment instruments that are the different assessment modes and feedbacks are developed. Concerto platform (v.5) by the psychometric centre of the University of Cambridge, and online Monkey Survey, were applied for developing CAT and LOT respectively. Concerto is free for academic use, but the developer or programmer is needed to be paid for uploading Concerto platform in Internet server. Concerto platforms allow anyone create online assessments, from simple surveys to sophisticated IRT-based CAT with textual and graphical feedback. In the creating CAT, item banking, item selection, stopping criteria and ability estimation method are essential features. Item bank (CIB) refers to a group of efficient items which are constructed according to the proper principles of testing construction, and the constructed items are to be systematically kept and administered by using a computer. In this study, the items in the bank were Grade-10 mathematics items, each item was calibrated by using Rasch Measurement Model through the ACER ConQuest and all items fitted the measurement model. Its process will be described in Chapter 5. Item selection refers to the method of selection of each item from the item pool of the item bank by matching the item difficulty and mathematics ability of the examinee which is estimated by Concerto platform of CAT program. Stopping criteria (SC) refers to the values specified for stopping the testing of each examinee. The standard error of estimation (SEE) of an examinee's ability will be used. Stopping criteria, SEE ≤ 0.5, which is the best for getting appropriate test length, was used through observation. Ability estimation method refers to the method used to estimate an examinee ability using the Computerized Adaptive Testing process. In the present research, an Updating Bayesian method, which is default of the Concerto platform was used to calculate because an examinee ability estimated using this method is more stable, least biased, and more accurate than the others, when there are less than 500 examinees in testing. All test items are set up in each test form and link with feedback loop in the Concerto platform. The platform sends out the test via weblink.

Monkey Survey also allow anyone create online assessments with textual and graphical feedback. Monkey SurveyMonkey is an online service that allows users to create web browser -based surveys. This service offers both a free account and a paid account that includes enhanced features. LOT test was created through paid account from online platform of Monkey Survey for adding the feedback loop. Firstly, log into Survey Monkey platform, choose online test templates, add MCQ questions into the test template, enable MCQ test mode by checking score the question for each question, assign an answer options points with the plus or minus signs, add feedback in each question through feedback loop, send out the test via weblink.

These online platforms are valid and reliable online comprehensive testing programs designed by Ph.D. psychometricians and be accessible both for examiners and for examinees anywhere. Test developers can conduct item development, review, and modification with the full security; design their own examinee score reports and feedback, analyze data by advanced psychometrics models (IRT or Rasch); and access data anytime and anywhere. For each item, the researcher constructed elaborate feedback by applying Grade-10 textbook and other relevant supplementary resources. The researcher took face validity and content validity of all feedbacks from the experts who are either mathematics educators or senior mathematics teachers.

3.5.3 Pilot testing

A pilot study was conducted to examine the accessibility of computers, software, and the website for administering the tests. In addition, pilot study examined the functionality of each instrument, the conformity of the technical feasibility of the experiment instruments. The study examined the validity, reliability, and accuracy of the experiment instruments (for example, PPT-1, LOT-1, CAT-1) as to fix any problems that may arise. A class from a school that has already participated for data collection of item calibrating participated in the pilot study. After the pilot study, the researcher asked some participants whether they face any problem or over any mistakes during the test administration. The researcher of this study made some adaptions after conducting the pilot study and also visited all five participating schools to confirm any problems during the experiments.

3.5.4 Counter-balanced quasi-experimental research

3.5.4.1 Participants

Most high-school classes have student characteristics of mixed-sex (the ratio of boys and girls is nearly 1:1). According to the Department of Basic Education (2016), the students are in the age range of 15.5 to 16.5 years. An approximate number of students in a class is around 45 in each school. Five schools participated in the main experiment in this study. A minimum number of classes in each school is three; consequently, three classes were selected for the study.

About 675 students and 5 mathematics teachers participated in this study. The participated teachers were voluntarily involved in a semi-structured interview. For semi-structed interview, two students were selected from each school according to intensity sampling method of qualitative approach (Gay, et.al., 2012). The researcher divided two groups: males and females, for gender balance. She selected one student from each group by drawing lots. Classes from the selected school participated as an intact group and were assigned as experimental groups after confirming the agreement between the school principals, teachers, and students.

3.5.4.2 Experiments

This study investigated the effectiveness of classroom formative assessments in three test modes, including LOT, CAT, and PPT, on the mathematics learning progress of students. The experimental study took six weeks. Each experiment covered a sub-topic and there are three concepts, namely, Function, Remainder, Factor. Three experimental groups provided six weekly classrooms formative assessments for three concepts in the three test modes randomly (see **Table 3.3**). For example, Group-1 (Test mode-order-1) took two weekly classroom formative assessments for Function concept in PPT mode (F/PPT-1, F/PPT-2), followed for Remainder concept by LOT mode (F/LOT-3, F/LOT-4), and for Factor concept by CAT mode (F/CAT-5, F/CAT-6) (See Group-2 (Test mode-order-2) and Group-3 (Test mode-order-3) in **Table 3.3**). A classroom formative assessment contains 20 MCQ items and takes about 25 minutes to answer.

In PPT mode, teachers manually scored and provided feedback to an individual student. However, a specific score report and elaborated feedback on each item were delayed for some days. Elaborated feedback refers to an explanation with additional resources to get the correct answers on each item. In LOT and CAT modes, specific score report and elaborated feedback on each item were received immediately from LOT computer programme after submitting the assessment. In CAT, individual students took different items according to their ability with varied test-taking time, finding out their weakness. In addition, CAT recorded a test-taking time of individuals.

Table 3.3

Counter-balanced experimental research

Experimental Groups	Pre-test	Expt - 1	Post-test	Pre-test	Expt - 2	Post-test	Pre-test	Expt - 3	Post-test
Experimental Group-1 (Test mode-order-1)	Pre-test 1	F/PPT-1 with Score report and Feedback (*Function*)	Post-test-1	Pre-test 2	F/LOT-2 with Score report and Feedback (*Remainder*)	Post-test 2	Pre-test 3	F/CAT-3 with Score report and Feedback (*Factor*)	Post-test 3
Experimental Group-2 (Test mode-order-2)	Pre-test 1	F/LOT-1 with Score report and Feedback (*Remainder*)	Post-test-1	Pre-test 2	F/CAT-2 with Score report and Feedback (*Factor*)	Post-test 2	Pre-test 3	F/PPT-3 with Score report and Feedback (*Function*)	Post-test 3
Experimental Group-3 (Test mode-order-3)	Pre-test 1	F/CAT-1 with Score report and Feedback (*Factor*)	Post-test-1	Pre-test 2	F/PPT-2 with Score report and Feedback (*Function*)	Post-test 2	Pre-test 3	F/LOT-3 with Score report and Feedback (*Remainder*)	Post-test 3

Note**:** Expt – Experiment; F/ refers to a formative assessment during the experiment (For example, F/PPT-1 means formative assessment in PPT mode in the first experiment and provide the score report and feedback after assessment)

3.5.4.3 Measures

(a) Pre-test and post-tests

Pre-test and post-tests are applied in this experiment. A pre-test is conducted before each experiment, and post-tests are conducted after each experiment. The two tests were conducted to measure the mathematics progress of students due to the experiment. Pre-tests contain all algebra contents for examining their prior knowledge. It was possible to compare three different average scores resulting from post-tests after different experiments within a group and experiments between different groups (see **Table 3.3**). A post-test covers a sub-topic, which classroom formative assessments in a test mode have already experienced. Both pre-test and post-tests, including 20 items, took about 25-minutes in PPT. Further, the score of pre-test and post-test are securely stored in the main dataset in the computer and hard drive.

(b) Follow-up questionaries

The study examined the relationship between teachers' and students' contextual factors and mathematics achievement in different test modes. There are two types of follow-up questionnaires: teachers' and students' questionnaires.

(i) Teachers' questionnaire

This thesis investigated the efficiency of classroom formative assessment in different test modes (LOT, CAT, and PPT) from the perspectives of participating teachers. The adapted questionnaire was conducted for teachers to determine their practices of classroom formative assessments. The questionnaire consisted of experiences and challenges of classroom formative assessments and perception and opinions on computer-based classroom formative assessments (Broughton, 2017). The items in the questionnaire include the attitude of teachers towards enhancing instruction and student learning and about the formative assessment (Karim, 2015). Further, the questionnaire was used to ascertain the attitude of teachers to formative assessment and feedback in teaching English for specific purposes (Al-Shebri, 2008).

The questionnaire involved teachers' contextual information and measures their practices and the challenges of ordinary classroom formative assessment; their attitude towards formative assessment and feedback for the learning progress of students; and their perceptions and opinions of computer-based classroom formative assessment in different test modes. In addition, teacher questionnaire involved 4-point Likert-scale items (varying from 1= strongly disagree to 4= strongly agree) and open-ended questions (for giving comments about different test modes for classroom formative assessments and providing feedback manually or automatically).

(ii) Students' questionnaire

The questionnaire in the literature was adapted for students to determine test-relevant motivation and engagement and subjective test experience (Martin & Lazendic, 2017). The items in the questionnaire include students' attitudes and motivation towards feedback (Miller, 2009). Further, this questionnaire consisted perceived usefulness of students to feedback (Harks, Rakoczy, Hattie, Besser, & Klieme, 2014). In addition, the questionnaire involved questions related to students' contextual information, computer familiarity, attitude towards the computer, attitude towards mathematics, attitude towards different test modes, attitude, and motivation towards feedback after experiencing the three different testing modes. The questionnaire consisted of items with a four-point Likert scale (varying from 1 = strongly disagree to 4 = strongly agree) and open-ended items. Students were able to comments on different test modes for classroom formative assessments and feedback in the open-ended items.

3.5.5 Semi-structured interview

The questions and topics in the semi-structured interview contained the challenges of ordinary classroom formative assessment, their attitude towards formative assessment and feedback for students' learning progress, and their perceptions and opinions of computer-based classroom formative assessment in different test modes. Questions for semi-structured interviews contained the opinion of students and view towards classroom formative assessment, feedback, and three different types of assessment modes. Data from the semi-structured interview was applied to explain their perception towards the effectiveness and challenges of new assessment modes (LOT and CAT) for mathematics learning progress and their challenges for applying LOT and CAT.

3.6 Summary

This chapter has presented the detail of the research methodology used for this study. The main methodology is the intervention design with an explanatory mix-method. Specifically, counter-balanced quasi-experimental research is applied. This study compares different test modes in terms of the achievement improvement of students during experiments, their perspective, and their teachers' perspective towards these modes. Consequently, all participants need to experience all of the test modes.

This study experimented with a pre-test and post-test to investigate which test mode is more effective for the improvement of students. Then, the surveys were conducted for students and teachers to explore the contextual factors from students and teachers according to the theoretical framework in Chapter 3. Finally, semi-structured interviews followed to explain their perception of the effectiveness and challenges of new assessment modes such as LOT and CAT.

For this thesis, purposive stratified random sampling was applied to select a representative sample of schools. There are four phases for these experiments. Phase 1 is for item-bank, while phase 2 revealed the construction of experiment instruments, including the technology test modes and their feedback. These involve the computer-adaptive test (CAT) and Linear-online test (LOT). CAT is based on Concerto and LOT based on Survey Monkey. Phase 3 is for piloting, and Phase 4 is for conducting counter-balanced quasi-experimental research in Myanmar. The included measures are pre-test and post-test for measuring the improved achievement and the follow-up questionnaires for contextual and psychological information from the teacher and student questionnaires. Further, semi-structured interview of students and teachers followed. The next chapter will explain the method and procedure of the data analysis for this thesis.

Chapter 4

Measurement for the Study: Methods of Data Analysis

4.1 Introduction

This study aims to explore the effectiveness of PPT, LOT, and CAT on improving mathematics achievement. As a social study, the effect of test modes' is not the sole factor, which influences the improvement of mathematics achievement. Hence, the study explores the factors that directly or indirectly affect the improvement of mathematics achievement. Further, as the study involves multiple scales derived from multiple levels that form a hierarchical structure, it is necessary to consider several statistical analytical techniques to cover the wide range of issues to be examined.

This chapter begins with discussing survey instrument development, modification, and adaptation. It is followed by a description of data preparation and then several analytical techniques and associated software used for this study. Further, the reliability and validity tests used for this study are described. The instrument validation, verification, calibration processes, and scoring of data are also illustrated. Several statistical software packages are used in this study, namely, Microsoft Excel, IBM SPSS 26, IBM AMOS (Analysis of Moment Structures) (v.26), ACER ConQuest (v.4), MPlus (v.8), and Hierarchical Linear Modelling (v.6). Finally, a summary of this chapter is provided.

4.2 Survey instrument development, modification, and adaption

The survey instruments employed in this study are designed to determine what factors influence the achievement improvement of students due to the different test modes. The survey instruments are adapted with modifications or developed, using some existing scales as a guide and in consultation with the supervisors. In adopting or developing the methodologies and software, the objectives and research questions of the study are used as the primary basis. In addition, the applicability of the items in the research context is considered to make the tools suitable and useful. Hence, items that are found irrelevant are discarded in the final tools. The scales are either adapted from the existing instruments or developed using the existing scales or literature as a guide. In adapting and developing the scales, some steps are taken to ensure that the participants answered the items with minimal disruption to their usual activities. Specifically, the study observed the clarity of language, brevity, clear format or structure, single cognitive load per item, clear directions, and applicability of all items to teacher and student respondents. Survey questionnaires for students and teachers are used in the study.

Student questionnaire involves five sections: (1) demographic information, (2) their attitude towards learning and assessment, (3) attitude to formative assessments, (4) their ICT familiarity and their attitude towards ICT, and (5) their attitude towards test modes. In addition, the scale of students' demographic information, *gender status, parental educational levels, and expected education* are included. The study involves scales for *self-efficacy, motivation, and attitude towards mathematics learning* for their attitude towards learning mathematics. The scale of *attitude towards formative assessment* is involved in examining their attitude towards assessment. The scales related to ICT are *attitude towards ICT and ICT familiarity*. Finally, the scale of the *attitudes towards paper-and-pencil test (PPT), linear-online test (LOT) and computer-adaptive test (CAT)* are examined for the attitude towards different test modes. The targeted scale is the *improvement of their achievement or their achievement improvement* due to the subtraction from the post-test scores to the pre-test score in each test modes experiment.

The questions concerning the scales of *self-efficacy, motivation, and attitude towards mathematics learning* are adapted and modified from student questionnaire from the Third International Mathematics and Science Study (TIMSS) background questionnaire (International Association for the Education of Educational achievement – IEA, 2015). The questions concerning the *ICT accessibility and use* scales are from the Programme for International Student Assessment (PISA)'s student questionnaire (Organisation for Economic Co-operation and Development (OECD), 2016). The use of ICT and ICT accessibility is formed into *ICT familiarity*. The scale *attitude towards ICT devices* is developed using a computer attitude scale for secondary students by Jones and Clarke (1994). The scale is composed of three parts: (1) effective component, (2) behavioural component, and (3) cognitive component. The total number of items in this scale for the student questionnaire is 16; 6 items for the affective component; 6 for the behavioural component, and 4 for a cognitive component. The attitude of students towards PPTs, attitude towards LOTs, and attitude towards CATs are developed using the scale from student computer-aided assessment survey by Broughton (2017). The *attitude towards formative assessment and feedback* is adapted from the scale of students' attitude towards formative assessment and corrective feedback by Fakeye (2016). Nine of the twelve items are adapted for this study. Students' *attitude towards PPT, attitudes towards LOT, and attitudes towards CAT* are developed using the scales from the survey of student computer-aided assessment by Broughton (2017).

The teacher questionnaire includes four sections: (1) demographic characteristics, (2) their assessment practices, (3) their ICT accessibility and familiarity and their attitude towards

ICT, and (4) their attitude towards new test modes. The designed questionnaire for teachers determined what factors influence their perceived ICT skills and use of ICT. All the items and scales are carefully selected from the available literature that was appropriate of this study.

There are two scales related to formative assessment practices: *general and specific*. The teachers' general formative assessment practices scale is adapted from a computer-aided assessment survey for lecturers (Broughton, 2017), and specific practices are adapted from a national survey examining teachers' formative assessment practices (Fishman, Riconscente, Snider, Tsai, & Plass, 2014).

There are two scales of *attitudes towards formative assessment*. The scales of *attitude towards formative assessment* are adapted by using the study of Karim (2015). The scale of *attitude towards ICT devices and internet surfing is* in the questionnaire developed for teachers, using a computer attitude scale designed by Jones and Clarke (1994).

For the teacher questionnaire, this scale is composed of 12 items: 5 items for the affective component, 4 for the behavioural component, and 3 for a cognitive component. The *attitude of* teachers *towards PPT, attitudes towards LOT, and attitudes towards CAT*, are developed, using the scale from the survey of lecturers' computer-aided assessment by Broughton (2017).

The items in questionnaires or surveys for this study are modified and/or used to construct similar but suitable items with respect to the context. Specifically, the questions are changed to generic items instead of subject-specific and are made to represent assignments and tests as the major constructs, as the purpose of this research instrument is to capture the general perceptions of students on assignments and tests in all learning areas. Some items are re-worded or restructured to suit the participants, and those that are believed to be irrelevant are excluded in the final form. The conceptual paradigm for the development of formative assessment practices and the perception of technology-based test modes is changed, and some items are revised, regrouped, and deleted based on the results of the expert, judgement, and face validation. Hence, this study validated, verified, and calibrated modified scales (see Chapters 6 and 7). The responses for all the items in scales are designed in a four-point Likert scale format.

This study also aims to explore the effects of demographic scales on factors at the teacher- and student- levels and on the outcome scales. For the student questionnaire, items related to gender status, highest educational level of fathers and mothers and their expected education are in MCQ format. For the teacher questionnaire, personal and professional background information obtained from teachers in the university include age, gender, and years of teaching experience through MCQ and open-ended item formats. Subsequently, items covering demographic scales are developed by the researcher and are included in the teacher and student questionnaires.

Items for the achievement test are adapted from the item collection of the Myanmar National Examination according to targeted contents such as *function, remainder,* and *factor*. Items are of MCQs that are commonly used in mathematics examinations of Myanmar. MCQ items can measure the ability of students in a wide range of higher-order thinking skills. The items can cover a number of content areas on a single exam, and students can answer the MCQ items in a class period. Items can be scored quickly and easily by hand or electronically. Each item in these achievement tests has four distractors and one correct answer. The Rasch Dichotomous Models verify these items, and the logit score of student participants are calibrated.

4.3 Ethics approval

Before data collection, the researcher of this thesis got ethics approval from the Human Research Ethics Committee (HREC) of the University of Adelaide. In addition, the Ministry of Education for high schools in Myanmar approved this study for data collection and analyses. Consequently, obtaining ethics approval for participation involved a considerable amount of paperwork and time in informing the educational authorities about the rationale and the methodology for the study.

This study was administered in Myanmar. Therefore, the ethics approval was collected from the Ministry of Education (MOE) in Myanmar. The approval from MOE was obtained before submitting the ethics approval. The permission to conduct the study in Myanmar was granted on 1st April by the Ministry of Education through email (see Appendix). Further, the education authorities in Myanmar requested a report of the data analysis as part of their condition for the approval of this study. In addition, gaining ethics approval also involved the necessity of ensuring that informed consent would be obtained from all participants in the study and that confidentiality would always be maintained. Then, in conducting the achievement tests, survey questionnaires, and interviews, the identity of every participant was to be kept confidential.

The relevant information sheet about the study was to be provided to every participant, involving school principals, classroom teachers, students, and their parents or guardians. In those information sheets, there are some statements or declarations mentioned. These include: (a) instructional and learning time were not to be infringed; (b) the employment of teachers and the academic standing of students was not to be affected in any way, and (c) participation was to be made voluntary and that every respondent was free to discontinue at any time during the conduct of the study. Where appropriate, some of these conditions were made clearer to the participants just before administering the survey to answer doubts and ensure that the research was living up to the ethical requirements.

Since this study included students under 18 years old, the Human Research Ethics Committee (HREC) requested that the researcher make sure that the study complies with child-related policies (which is AusAID Child Protection Code of Conduct) before the data collection could proceed. Subsequently, the researcher of this study and the supervisor panel has taken the necessary steps to ensure that this research is conducted in compliance with the Australian Code for the Responsible Conduct of Research, National Statement on Ethical Conduct in Human Research, the University's Child-Safe Environment Policy, and the AusAID Child Protection Policy. The researcher has also signed the AusAID Child Protection Code of Conduct under the guidance of the Head of the School of Education from the University of Adelaide. For those participants, the consent of their parents or guardians was obtained. These conditions were observed in the administration of the study, and they followed the ethics committee's approval. The HREC approved this study to proceed on the 1st of May 2019 (ethics approval number H-2019-039). The study was undertaken with all the necessary permissions.

4.4 Data preparation

After the survey administration, data were gathered and made ready for analysis. Some steps are taken, and the following subsections describe the steps to ensure data utility. After conducting the experiments, questionnaires are collated, and the data from the questionnaires are organised by the type of respondents and class level. Then, data collection is directly encoded using Microsoft Excel. Student data are encoded first, followed by teacher data, and then from student and teacher interviews. The Excel files of quantitative data are cleaned and were prepared for the data analysis using the Statistical Package for Social Science (SPSS) format using the SPSS software (v.26). The study assigned codes for the demographic and other data. Also, the qualitative interview data are manually transcribed and are entered as text data in Microsoft Word.

4.5 Data analysis techniques

This study involves multiple scales and requires multiple statistical analytical techniques and statistical software packages to cover the wide range of issues to be examined. Consequently, to analyse the data in this study, the data have been organised for student, teacher, and school respondents. Quantitative and qualitative data analyses are done in accordance with the objectives and questions of study and employ the corresponding statistical techniques and procedures.

Student and teacher data are entered in a separate file using the IBM SPSS 26 for the convenience of descriptive analysis. Student and teacher data in the SPSS file format are classified into an appropriate numerical form for use in the statistical software packages employed in this study. For instance, numbers are assigned to items such as gender (namely, 0 for male and 1 for female). The next few chapters provide full details of the codes assigned to the items or scales. Further, data were checked for error before any analysis was conducted (Pallant, 2011). In this study, all data are evaluated using the data screening and cleaning procedure, that is, through the examination of the basic descriptive statistics and the frequency distributions. For example, data are examined if they were within the range of the possible scores. Any error detected through the data screening, and cleaning processes is corrected immediately. This process is repeated until all data are free of mistakes.

4.5.1 Descriptive statistics and inferential statistics

For the descriptive analysis, values for each scale from the questionnaire are described as frequencies for their grouped values. The value or data in categorical scales are described as their frequencies. If the scales are continuous scales, such data are more easily described in terms of their central tendency, variability or spread, and distribution shape.

The measures of central tendency include mean, median, mode, and quartiles. Mean is the arithmetic average of a set of scores, median is the middle value of ranked scores, and mode is the most frequently occurring score. Indices of variability describe the variability of the scores around a measure of central tendency, or more generally, the spread of scores on a scale. The most common measures of variability are variance and standard deviation. Variance is the average sum of squares of the mean deviances, and the standard deviation is the square root of the variance. Graphical displays such as error bars, box plots, and histograms are applied to describe central tendency, variability, and distribution measures. For example, error bars display the means and standard deviations, which are commonly used to display confidence intervals.

As the inferential statistic, the analyses of mean differences are applied to compare the achievement of two or more groups. The t-test is applied to compare the achievement of two groups, whereas the analysis of variance (ANOVA) is applied to compare the achievement of three or more groups. The results of ANOVA reveal the two specific indices: the statistical significance level of F statistics and practical application amount of effect size in terms of the partial eta squared (Partial η^2). The F statistics should be significant at the 0.05 level of p value. The cut-off value of partial eta squared (Partial η^2) is identified based on previous studies by Cohen (1969) and Richardson (2011). According to Cohen (1969) and Richardson (2011), the partial eta squared values of 0.0099, 0.0588, and 0.1379 are viewed as the benchmarks for small, medium, and large effect sizes, respectively. For the practical application of the experiment, medium and large effect sizes are preferred. The error bar or bar chart, or line graph is commonly used to compare the mean difference among the groups.

4.5.2 Confirmatory factor analysis (CFA)

A Confirmatory Factor Analysis (CFA), as the structural level analysis, is carried out to investigate the construct validity of the instruments which were applied in this study. CFA is concerned with how the observed scales represent the underlying latent scales (Byrne, 2010). A CFA statistical analysis can identify the measurement theory for all scales from the instruments. The measurement model from the theory specifies the extent to which "the measured scales logically and systematically represent constructs involved in a theoretical model" (Hair et al, 2010, p. 671). To be simplified, the measurement model examines the relationships between measured, or manifest, or observed scales and the construct or latent or unobserved scales. In other words, CFA is used to examine how well the measured items represent the latent constructs.

Before conducting CFA, it is important to determine whether the measurement model is reflective or formative. There are two measurement theories: a reflective measurement theory and a formative measurement theory. The underlying assumption of the reflective model is that latent constructs cause measured scales, and the arrows go from latent constructs to the measured scales. In the formative model, the latent constructs are caused by the measuring scales, and the arrows go from the measured scales to the constructs (Hair et al. 2010).

MacCallum (1995) proposed three strategies for conducting the CFA analysis: the strictly confirmatory strategy, the model generation strategy, and the model comparison strategy. A model of interest is constructed by using a strictly confirmatory strategy. The model is then examined to see if it fits the data well. A model that fits the data well can be considered a plausible model, but a model that does not fit the data well is not acceptable and no further analysis is conducted. Next, a model is specified and evaluated by the model generation strategy. The fit of the model is improved by using the modification indices. However, these strategies have serious shortcomings. The strictly confirmatory strategy is overly rigid, while the modification made to the model using the model generation strategy may not be meaningful and substantively justifiable (MacCallum, 1995).

An alternative to these strategies is the model comparison strategy. Several alternatives or competing models are proposed in model comparison strategy based on the competing theoretical position or conflicting research findings. After the alternatives of a priori, models are specified to fit the single set of data. One best model is selected which can represent the sample data (Byrne, 2010). After model specification, they are evaluated and compared to see which models fit the data better, based on the model fit indices, and the interpretability and meaningfulness of the parameter estimates (MacCallum, 1995).

As noted by Curtis (2005), for a construct to be compatible with simple measurement, the structure of the construct must reflect a single underlying factor. The simplest model occurs when all observed scales belong to a single factor while other acceptable alternatives include uncorrelated and correlated factors models as well as hierarchical and nested models (Curtis, 2005; Darmawan, 2003; Hair et al., 2010). Further, the model comparison strategy is appropriate when the CFA analysis aims to discover a model that is more consistent with the data. Consequently, in this study, the model comparison strategy was employed to conduct CFA because of the strengths of this strategy and some limitations of the other strategies. There are five types of models for the model comparison strategy: one-factor model, N-correlated factors model, N-orthogonal factors model, hierarchical factor model, and nested factor model.

Firstly, in the single-factor model, the observed scales are loaded on the one-construct or one-factor. Therefore, the model examines to what extent the observed scales load in a single factor. If there are two or more factors in the measurement model, it needs to be considered whether the model should be the N-correlated factors model, the N-orthogonal factors model, the hierarchical factor model, or the nested factor model. Secondly, if factors in the model are assumed to be correlated, such a model will be identified as the N-correlated factors model (N=

the number of the factors). While these factors are not correlated, it is assumed as the N-orthogonal factors (N= the number of the factors).

In the hierarchical factors model, first-order and second-order factors are involved. A nested factors model can examine if the observed scales load on a single-factor or many factors (Darmawan, 2003; Hair et al., 2010). In this study, only the first four models explained earlier (namely, the single factor model, the N-correlated factors model, the N-orthogonal factors model, and the hierarchical model) are constructed.

The IBM SPSS AMOS 26 (Analysis of Moment Structures) software is applied for the CFA. The Maximum Likelihood (ML) estimation method is the default option of the IBM SPSS AMOS 26. The AMOS software can be used to test the construct validity of the measurement model for the scale in the questionnaire. Using the AMOS, a measurement model can be specified by writing a script or drawing the model (Schreiber, 2008). The program uses to analyse a variety of continuous latent scale models but cannot analyse ordinal data (Schreiber, 2008).

In choosing of the appropriate factor model, it is necessary to determine an acceptable value of indices. The criteria of the indices for CFA are described as follows. CFA concerns how well the observed scales represent the underlying latent scale. The strength of the regression paths from the latent scale to the observed scales is of primary interest (Byrne, 2010). Many researchers have suggested different rule-of-thumb in determining factor loading values appropriate for a measurement model to be accepted. This study follows the cut-off values proposed by Tabachnick and Fidell (2013).

Consequently, the scales with loadings of 0.32 and above are considered acceptable for the measurement model to be interpretable. In addition to examining factor loadings, it is also essential to compare the model fit of each model constructed. As mentioned earlier, this study employs the model comparison strategy to identify the best model that fits the data well.

There are also other model fit indices to compare the structure of the model and assess the goodness of fit. According to Hooper et al. (2008), the traditional criterion for model fit was Chi-square ($\chi 2$), and it is called badness of fit (Kline, 2005). Therefore, if the Chi-square model fit is significant, it would indicate that that model does not fit the data. However, there are some limitations for the use of Chi-square for model fit, for example, its sensitivity to sample size (Hooper et al., 2008).

An alternative index to reduce the impact of the sensitivity to sample size is the chi-square ratio to the degree of freedom ($\chi2/df$). Many researchers recommended different values for the ratio ranging from 2.0 to 5.0 (Hooper et al., 2008). This range is used as the acceptable range for ($\chi2/df$) in this study.

Schreiber and co-workers (2006) described that some common fit indices for CFA are Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA). Therefore, in this study, CFI, TLI and RMSEA fit indices are used to assess the model fit of the scales. The CFI is an "incremental fit index that measures the relative improvement in the fit of the researcher's model over that of a baseline model, typically the independence model" (Kline, 2011, p. 208). The main advantage of using CFI as a model fit index is that it is sample-independent and can perform well if the sample size is relatively small in the study (Tabachnick & Fidell, 2013). The range for CFI is between 0.0 to 1.0, and the closer to 1.0, the better the model fit; it is used as the acceptable range for model fit (Hair et al, 2010).

The TLI is another type of incremental fit index that is not normed with the value of the statistics ranging from 0.0 to above 1.0. Like the CFI, the higher the value, the better the model fit, and it is used to be acceptable for the model fit (Hair et al. 2010).

The RMSEA is the fit index used to assess the model fit in this study. Hair and his colleagues suggested that RMSEA is suitable for the comparison strategy of models or competing model strategy. It considers "the error of approximation in the population" (cited in Byrne, 2010, p. 80) and favours parsimony of the model (Hooper et al., 2008). RMSEA values less than 0.05 indicate good fit; the values between 0.08 and 0.10 indicate mediocre fit, and the values higher than 0.10 indicate poor fit (Byrne, 2010). This study uses RMSEA values between 0.05 and 0.10 to assess model fit (**Table 4.1**).

Table 4.1

Fit indices for examination model fit of CFA

Indices	Acceptable cut-off values of model
X^2/df	The smaller the ratio, the better model fit
CFI	Close to or 0.90 for acceptance
TLI	Close to or 0.90 for acceptance
RMSEA	0.05 – 0.08 or below; not more than 0.10 is still acceptable

4.5.3 Construct validity

Construct validity is applied in this study because it indicates 'the extent to which a particular measure relates to other measures consistent with theoretically derived hypotheses concerning the concepts (or constructs) measured (Carminers & Zeller, 1979, p. 23). Cronbach and Meehl (1995) asserted that it is necessary to examine the construct validity if there is no universal agreement or absolute acceptance for defining the constructs to measure. Hair and his colleagues described that the fundamental purpose of CFA is to test the construct validity of the instrument or measurement model (Hair et al., 2010). There are four important elements in the construct validity: (1) face validity, (2) nomological validity, (3) convergent validity, and (4) discriminant validity (Hair et al., 2010). However, in this study, the face validity and convergent validity are applied to validate the instrument.

4.5.3.1 Face validity

Face validity, sometimes called content validity, is "the degree of correspondence between the items selected to constitute a summated scale and its conceptual definition" (Hair et al., 2010, p. 92), which expert judges can assess prior to the theoretical examination. Face validity for the instrument in this study was conducted by two experts from the University of Adelaide, Australia, and two experts from Yangon University of Education, Myanmar. The three times of modification by the four experts help modify the scales of the instruments.

4.5.3.2 Convergent validity

Convergent validity indicates how closely a new scale is related to other scales the same construct in terms of common variance. Three indicators such as factor loadings from the confirmatory factor analysis (CFA), average variance extracted (AVE), and construct reliability (CR) can be used to estimate the value of convergent validity (Hair et al. 2010). The factor loading from CFA will be applied to calculate AVE and CR.

4.5.3.3 Factor loading

Factor loading is one of the important indices for construct validity. Factor loading is the proportion of variance shared among the measured scales, which can be available from Confirmatory Factor Analysis (CFA). The acceptable cut-off value of factor loading value in this study is above 0.32, which is suggested by Tabachnick and Fidell (2013) the acceptable cut-off value for factor loading. The above section has already explained the process and indices of Confirmatory Factor Analysis (CFA).

4.5.3.4 Average variance extracted (AVE)

The extent to which one construct is distinct from others indicates discriminant validity. The discriminant validity can be computed by comparing average variance extracted (AVE) values for two latent constructs with the correlation estimate square between two constructs. The average variance extracted (AVE) is a summary measure of convergence among a set of items representing a latent construct. It is also known as the average percentage of variation explained (*variance extracted*) among the items of a construct. (Hair et al., 2010). AVE can be calculated as the summation of all squared standardized factor loadings divided by the number of items. See the formula below:

$$AVE = \frac{\sum_{i=1}^{n} L_i^2}{n}$$

where,

AVE = Average variance extracted

L_i = Standardized factor loading of an item (i)

n = the number of items

An AVE value of 0.5 or higher suggests as an adequate convergent validity (Hair et al. 2010). The AVE estimates for two factors also should be greater than the square of the correlation between the two factors to provide evidence of discriminant validity (Hair et al., 2010).

4.5.3.5 Construct reliability

Construct reliability is one of the indicators for determining convergent validity. Construct reliability measures the reliability and internal consistency of the measured scales representing a latent construct (Hair et al., 2010). Cronbach's alpha remains a commonly applied estimate, although it may understate reliability. Different reliability coefficients do not produce dramatically different reliability estimates, but a slightly different construct reliability (CR) value is often used with SEM models.

Construct reliability can be computed as the sum of the squared standardized factor loadings for each construct divided by the sum of error variance terms for each construct (e) plus the squared sum of factor loadings. See the formula below:

$$CR = \frac{(\sum_{i=1}^{n} L_i)^2}{(\sum_{i=1}^{n} L_i)^2 + (\sum_{i=1}^{n} e_i)}$$

where,

CR = Construct Reliability

L_i = Standardized factor loading of an item (i)

n = the number of items

e_i = Error variance of an item (i) within a construct

Construct reliability of 0.7 or higher indicates adequate convergence or internal consistency (Hair et al. 2010). The value indicates good reliability meaning that all measured scales consistently belong to a single construct (Hair et al., 2010).

4.5.4 Criteria of reliability: Cronbach's alpha

In terms of reliability, Cronbach's alpha is used to examine the internal consistency of the instrument in this study. The alpha coefficient of reliability provides "a coefficient of inter-item correlations, that is, the correlation of each item with the sum of all the other relevant items" (Cohen et al., 2007, p.148). Alpha values above 0.70 are reliable, and below 0.60 indicates low reliability, which is unacceptable. Therefore, alpha values higher than 0.60 will be used as a cut-off value for the reliability of scales in this study. Cohen et al. (2007, p.506) provide the guidelines for alpha coefficients, and these ranges are stated below (**Table 4.2**).

Table 4.2

Critical values of Cronbach's alpha

Cut-off values or critical values	Identification
> 0.90	Very highly reliable
0.80 – 0.90	Highly reliable
0.70 – 0.79	Reliable
0.60 – 0.69	Marginally/minimally reliable
< 0.60	Unacceptably low reliable

4.5.5 Item verification and calibration: Rasch measurement model

Confirmatory Factor Analysis (CFA) can be used to examine the factor structures of a set of items in which the results can show evidence of the underlying dimensions (Hu & Bentler, 1999). However, it is not enough to conduct a CFA analysis. Keeves and Masters (1999) argued that CFA analysis is just the first step to examining the data, and further data analysis such as person and item fit using Rasch analysis should also be conducted. Consequently, in this study, Rasch analysis is employed to verify the extent to which the structures of the scales confirmed in CFA analysis fit the Rasch model. In addition, it is also used for scoring purposes for use in subsequent analysis, namely, student- and teacher-level structural equation models and hierarchical linear models.

The Rasch analysis is applied for the item-level validation of the questionnaire. The Rasch analysis results have generally been conducted according to the results of CFA. For the single-factor model, the simple Rash measurement model is conducted. The multi-dimensional Rasch measurement model validates the multi-factor model. The findings of Rasch support the theoretical underpinning from which student questionnaire was developed. Moreover, the analysis results prove that most student questionnaire items have acceptable measurement properties. Consequently, it can be concluded that the questionnaire possesses good psychometric properties. As mentioned earlier, some general Rasch models can be found in the family of Rasch models. These include dichotomous model, rating-scale model (RSM), partial credit model (PCM) and multidimensional and multi-facet models.

The RSM can be used for ratings of two or more categories and when the interval between each response category on a particular scale remains the same for all items in an instrument such as the questionnaire (Andrich, 1999). However, a common criticism concerning the RSM is that the units between each response category may be unequal, although Categories are generally ordered (Blais et al. 2011), which has led to the use of the PCM. The PCM can be used for ratings of two or more categories. However, the interval between each response category on a particular scale differs. The RSM model is employed in this study because the purpose of this study is to determine the attitude and perception level of respondents. In addition, the items developed for the questionnaires contain rating scales that have the same number and structure of categories across the items.

The multidimensional model is helpful to examine scales that contain a hierarchical structure. The analysis of a scale containing several sub-scales was conducted in the past by either applying a unidimensional model to each scale separately or ignoring the multidimensionality and, therefore, and treat the scale as unidimensional (Adams and Wu, 2010). Adams et al. (1997) argued that both methods have limitations and consequently a joint, multidimensional calibration is preferred. It is important to note the multidimensionality. In fact, in multidimensional analysis, each subscale is unidimensional (Adams et al. 1997). As noted by Adams and Wu (2010), the multidimensional test enables the examination of several subscales, each measuring related but distinct latent dimensions.

Rasch analysis provides fit statistics to examine how well each item or person fits the 'ideal' Rasch model. As Bond and Fox (2015) indicated, the concept of fit and misfit of Rasch analysis is like a quality control mechanism. Misfit happens when the collected data are not a good match to the 'ideal' or calculated Rasch estimates (Bond & Fox, 2015). Blais et al. (2011) noted that the number of people included in a survey is usually greater than the number of items available in the instrument. Consequently, this enables the identification of data that do not meet the requirement of the Rasch model and to determine if they impede the measurement.

In this study, the person fit, and item fit analyses were conducted to examine if there were any misfitting persons or items. Person fit concerns the relationship between response probabilities for a set of items by a single person. Item fit concerns the relationship between response probabilities for a group of people by a single item (Engelhard, 2013).

The weighted fit mean square (or INFIT MNSQ) can examine the fit indices of a person or an item and the unweighted fit mean square (or OUTFIT MNSQ). The INFIT MNSQ can be examined by having the squared standardised matrix divided by the information function, and this is then nullified by dividing the sum obtained by the sum of the weights (Blais et al. 2011). It is possible to calculate the OUTFIT MNSQ from the matrix of squared standardised residuals (Blais et al. 2011). Both the INFIT MNQ and OUTFIT MNSQ should be used in a complementary manner to examine different problems. However, Bond and Fox (2015) noted that misfitting in INFIT is of great concern than misfitting in OUTFIT. For this study, the INFIT MNSQ is examined and reported.

Different researchers use different threshold values to examine whether a particular item or person is in a good fit. An INFIT MNSQ or OUTFIT MNSQ of 0.75 is a reasonable lower bound, and 1.33 is a reasonable upper bound (Adams and Khoo, 1996). Bond and Fox (2015, p. 273) also proposed different threshold values that can be used for different purposes. The threshold value of an item and person fit are in **Table 4.3** and **Table 4.4,** according to Wright et al. (1994). People with low fit show somewhat less randomness in the data than the expected ones (Wilson, 2005). On the other hand, there are also persons with high fit. Persons with a high fit show that the expected order may be wrong and indicate more problems (Wilson, 2005). Consequently, in this study, individuals with low fit or high fit are removed for the subsequent item analysis.

For items in the questionnaires and achievement test, a range of INFIT MNSQ from 0.60 to 1.40 is used as the cut-off value to determine if an item is fit or misfit. In addition to examining the INFIT and OUTFIT, it is also useful to examine the standard error of the scales. The standard errors indicate the accuracy of each estimate and are related to the confidence interval (Wilson, 2005).

Another useful index to examine the fit of the person or item is the t-statistics. However, Wilson (2005) noted that data with a large sample size tend to obtain significant and prominent t-values. Consequently, an item or a person is considered as a misfit only when it shows misfitting on both the INFIT MNSQ and OUTFIT MNSQ as well as the t-statistics. As the items in the questionnaires and achievement tests involve rating scales and categories, it is also helpful to examine the threshold if it is ordered. As noted by Bond and Fox (2015), thresholds should be increased across the rating scales.

Table 4.3

Ranges of item MNSQ

Types of tests	Range
Multiple-choice test (high stakes)	0.80 – 1.20
Multiple-choice test (run of the mill)	0.70 – 1.30
Rating scale (Likert/survey)	0.60 – 1.40
Partial credit scale (Likert/survey)	0.60 – 1.40
Clinical observation	0.50 – 1.70
Judge (where agreement is encouraged)	0.40 – 1.20

Table 4.4

Ranges of person MNSQ

Interpretation of parameter-level MNSQ	Range
Distorts or degrades the measurement	> 2.00
Unproductive for construction of measurement, but not degrading	1.50 – 2.00
Productive for measurement	0.50 – 1.50
Less productive for measurement, but not degrading; may produce misleadingly good reliabilities and separations	< 0.50

This study employed ACER ConQuest (v.4) (Adams et al. 2015) to conduct the Rasch analysis. It is a computer programme for both unidimensional and multidimensional item response and latent regression models. It provides data analysis based on a comprehensive and flexible range of item response models, allowing the examination of the properties of dichotomous items, polytomous items, and rating scale or Likert-scale items. It also offers measurement that is more comprehensive and research community analysis procedures, based on the methods of multifaceted items response models, multidimensional item response models, latent regression models (Australian Council for Educational Research [ACER], 2016). Then, the ACER ConQuest (v.4) can estimate drawing plausible values.

4.5.6 Scoring

The Weighted Likelihood Estimate (WLE) is used for this study. It is presented in the form of logits. The WLE enables the estimation of the abilities of participants. In the case of this study, for example, the attitude level of students choosing a certain category in the questionnaire, the attitude level of teachers choosing a certain category in the questionnaire, the ability level of students rated by their teachers in a certain category. As noted by Liu and Wilson (2011), theoretically, students who choose a more difficult category should obtain a higher average WLE than those who do not choose category. Similarly, teachers who choose a more difficulty category should also have a higher average WLE, and students who are achieved a more difficult question in mathematics tests should have a higher average WLE.

The ACER ConQuest (v.4) is used to transform raw scores into WLE (Wu et al. 2007). In this study, item analysis could be conducted for effective scales in the questionnaire by the Rasch rating scale model and the Rasch partial credit model. Once the items analyses are performed, the ability estimates of the respondents are anchored to all the respondents to obtain the WLE scores. For achievement tests, the dichotomous items in the achievement test are analysed by the Rasch Dichotomous Measurement Model to estimate WLE scores.

4.5.7 Structural equation modelling

Structural equation modelling (SEM) allows the researcher to test hypothesized the measurement model and the direct and indirect relationships among the latent constructs (i.e., the structural model). According to Ho (2006), SEM enables estimating the multiple and interrelated dependence relationships simultaneously. The model is analysed using the SEM approach to investigate the relationship or model between the independent and dependent scales in each level (for example, student-level or teacher-level).

In this study, multiple arrows indicating multiple direct and indirect effect among the scales depict student- and teacher-level models. The figure below (**Figure 4.1**) shows an example of direct and indirect effect in an SEM model. Path A represents a direct effect of F1 on F3 (F1 → F3). However, path BC represents an indirect effect of F1 on F3, which F2 mediates, is shown by Path B and Path C (F1→ F2 → F3).

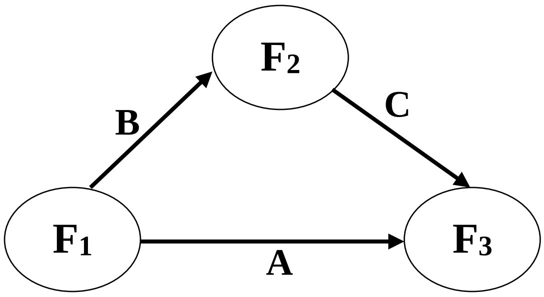

Figure 4.1 Example of direct and indirect effects in SEM model

For this study, model trimming is carried out by examining the critical ratio for significance. Any path coefficient that had a p-value greater than 0.05 or insignificant correlation was removed. In each model, a level of significance of 0.05 is applied. The significance level may be higher due to the small sample size (for example, the teacher-level). The modification indices provided by the MPlus (v.8) are used as a guide to improve the model.

Like CFA, each SEM is assessed to examine whether it fits the data well. The measurement model is also examined on its factor loadings in the model that contains latent scales. To be acceptable, factor loadings of the measurement model should be 0.32 or above (Tabachinick and Fidell, 2013). There are four fit indices that are similar to the fit indices in CFA. The smaller the ratio of chi-square and degree of freedom, the better the model fit. The indices of CFI and TLI, which are close to 0.90, are acceptable (Hu & Bentler, 1999). The index of the RMSEA, which is between 0.05 and 0.08, is acceptable (Hu & Bentler, 1999). The table below provides a summary of the fit indices used to analyse the fit of the SEM model.

Table 4.5

Fit indices for examining the model fit of the structural equation model

Indices	Acceptable cut-off values of model
X^2/df	The smaller the ratio, the better model fit
CFI	Close to or 0.90 for acceptance
TLI	Close to or 0.90 for acceptance
RMSEA	0.05 – 0.08 or blow; not more than 0.10 is still acceptable

SEM is characterised by two basic components: the measurement model and the structural model. The measurement model examines those relationships between the observed and latent unobserved scales while the structural model specifies how particular scales directly or indirectly influence other certain scales in the model (Byrne, 2010; Hair et al., 2010). As this study involves multiple factors that include many independent and dependent scales, the SEM statistical analytical approach was used to first examine the measurement model concerning its construct validity and second, the complex and multiple relationships between the independent and dependent scales (namely, the structural model). SEM is a collection of statistical techniques (Ullman, 2013) that simultaneously estimate multiple and interrelated dependence relationships (Ho, 2006). It is particularly useful when one dependent scale becomes an independent scale in subsequent dependence relationships (Darmawan, 2003, p.82). The AMOS or M-Plus is a covariance-based SEM technique, and it uses a maximum likelihood function to obtain estimators in models. In the following sections, the measurement and structural models are described. In addition, the model specification, model trimming, and model assessment are also illustrated. For identifying models in SEM, this study used MPlus (v 8) because it can handle continuous latent scales, categorical latent scales, or a combination of continuous and categorical latent scales.

4.5.8 Hierarchical linear modelling (HLM)

Hierarchical linear modelling techniques (HLM) become essential because social research dataset is often hierarchical. Also, conventional regression is not a satisfactory analytic approach because it treats the school or student as the unit of analysis. However, that regression method ignores the variation among students within schools or the nested structure within schools (Braun & Clarke, 2006). The hierarchical linear technique can overcome the weakness of aggregated or disaggregated analytic approaches. Analysing multiple-level data using a single-level analysis technique may produce aggregation or disaggregation bias, resulting in a misleading conclusion (Snijders & Bosker, 2012). Both procedures may produce incorrect estimates of the effects among the constructs (Goldstein, 2011; Hox, 2010; Snijders and Bosker, 2012).

The aggregation effect is one of the first statistical problems concerning the level of analysis. When data are aggregated, different data values from many lower-level sub-units are combined into fewer values for fewer higher-level units (Hox, 2010). As a result, much information is lost, which may lead to bias in the estimates of standard errors of means and regression weights related to the dependent scale (Braun & Clarke, 2006). Snijders and Bosker (2012) pointed out that at least four potential errors could occur when data are aggregated. First, a shift of meaning could exist when a scale is aggregated to the higher level and not directly to the lower level. Second, a correlation between higher-level scales cannot be used to assert the lower-level relations as this could introduce ecology fallacy. Third, aggregation could also cause inappropriate significance tests to apply when the original data structure is neglected. Finally, aggregation also prevents the examination of potential cross-level scales in the models.

Another statistics problem with regards to the level of analysis is disaggregation effects. When data are disaggregated, a few data values from a small number of higher-level units are loosened into many more values for more lower-level units (Hox, 2010). Consequently, the assumption of the independence of observation fails to apply as the proper sample size for these data is the number of higher-level units. Disaggregation may lead to severe risks of committing Type I errors (Snijders and Bosker, 2012), that is, rejecting the true null hypothesis. In other words, researchers may come up with many significant results that are false or spurious. When there are variations between levels, incorrect estimates of the effects may exist. This condition may further lead to a more severe risk of providing incorrect or misleading conclusions. Consequently, in the study, the Hierarchical Linear Modelling (HLM) approach is employed to account for the differences in the levels. The dataset of this study includes the information on constructs gathered at student-level and at the teacher-level. Hence, the data files contain information obtained at two different levels. The hierarchical linear modelling (HLM) for this study is conducted by the application of the HLM statistical software package (v.6). The HLM software is due to the pioneering work of Stephen Raudenbush and Anthony Bryk. This software has been one of the leading statistical software packages for hierarchical linear and nonlinear modelling (Garson, 2013). Its default method of parameter estimation is Restricted Maximum Likelihood (REML). The REML can handle high correlations more effectively, and that method is less sensitive and especially useful when the number of higher groups or levels is negligible.

Consequently, REML is commonly used the number of higher groups or levels is small (Garson, 2013). In addition, the HLM (v.6) offers lots of capabilities but more importantly, it provides greater ease in creating multi-level models. HLM (v.6) identifies (1) the estimation for individual effects; (2) cross-level effects or interaction effects; and (3) the partition variance-covariance component across levels of analysis for the appropriate and precise application of the significant tests (Bryk and Raudenbush, 1992).

Three major steps are involved in analysing data using the HLM 6 statistical software package. These include importing the IBM SPSS files into HLM 6 to create a multivariate data matrix (MDM) file, executing analysis based on the MDM file, and evaluating the fitting model based on a residual file. In two-level HLM analysis, two IBM SPSS files are imported into HLM 6 to construct the MDM file. Student-related scales are included in Level-1. In the Level-2 data, there are the teacher-related scales. A common teacher ID links the two IBM SPSS data files because of the two-level HLM analysis. Subsequently, all data lines up in appropriate columns according to the similar sequence of groupings. Once the MDM file is constructed, all subsequent analysis is computed using this MDM file.

The next step of the multilevel modelling analysis specified the model. The scales in these two levels are analysed using the HLM 6 statistical software package. First, a null model or a fully unconditional model is specified to determine whether an HLM analysis would be practically meaningful or needed. This model contains no predictor scales from any level and is also used to obtain the deviance statistics and other coefficients used as a baseline for comparing a final, more complicated model (Garson, 2013).

Then, the critical intra-class correlation coefficient (ICC) is examined to determine whether the HLM analysis to be practically meaningful or needed (Garson, 2013). The ICCs are the proportion of total variance of scale accounted for a higher level (Snijders & Bosker, 2012). The ICC can be calculated as the intercept variances component in the null model divided by the total of variance components (Garson, 2013). In the two-level HLM model, the σ^2 parameter represents student-level variability, and the τ_B parameter represents the teacher-level variability. Niechaus et al. (2014) stated that there is no consensus of how much group-variance is sufficient to conduct the HLM analysis. However, Garson, (2013) suggested that if the ICC is zero or negative, the HLM analysis should not be conducted. The null model allows estimation of the proportions of variation that are within students, among students, and within teachers as follows:

The proportion of variance within students (the ICC at the Level-1 or Student-Level) is $\frac{\sigma^2}{(\sigma^2+\tau_B)}$ and the proportion of variance among students within teachers (the ICC at the Level-2 or Teacher-Level) is $\frac{\tau_B}{(\sigma^2+\tau_B)}$.

In addition, for a reliability value of above 0.05 in a null hierarchical linear model, it is assumed that there is a random effect for the coefficient, and it is appropriate to conduct the HLM analysis (Raudenbush, & Bryk 2002). Therefore, the critical value of the reliability of the model is expected to be close to or more than 0.05 for conducting the HLM analysis. Once the null model is specified, the independent level-1 scales are added to the model. The HLM model for this study is developed using the step-up strategy. As noted by Raudenbush and Bryk (2002), when a relatively small set of Level-1 predictors has been defined, using the step-up strategy, models can be built up from univariate to bivariate, tri-variate, and so on. In addition, the gamma coefficient is examined to determine whether the scales should be included in the model (namely, model trimming). Scales with a t-value of less than 2.00 and a p-value of more than 0.05 were not included in the model.

The model trimming involves the examination of reliability estimates for random coefficients. This is to determine if the effects are random. Low reliability (i.e., below 0.05) indicates that there is too much error in estimating the relationship between the independent and dependent scales (Raudenbush, & Bryk 2002), suggesting the need to treat them as fixed effects. It is important to note a scale from a higher level influencing the effect of an independent scale on the dependent scale from a lower level can only be examined when the effect is random at the lower level.

Alternative models can be compared by examining the deviance and the number of parameters estimated by the models. The final model is compared with the null model to determine the amount of variance explained by the independent scales at each level (Raudenbush & Bryk, 2002). The null and the final models are compared by examining the deviance deduction or reduction with respect to the increased number of estimated parameters (Darmawan & Keeves, 2009). Additionally, the proportion of variance explained at each level represents the percentage of the improved variances of the level. It is calculated by the comparison of the variability of the null model and that of the final model. The formulae are as bellow:

$$\textit{Proportion of variance explained at student-level} = \frac{\sigma^2\ (null) - \sigma^2\ (final)}{\sigma^2\ (null)}$$

$$\textit{Proportion of variance explained at teacher-level} = \frac{\tau_B\ (null) - \tau_B\ (final)}{\tau_B\ (null)}$$

4.5.9 Qualitative data analysis for interview questions

The semi-structured interview is employed after collecting the quantitative data in this study. Its purpose is to explore the perspectives of student participants and teacher participants towards the different test modes and classroom formative assessment. The written protocols or guides that outlined the kind and order of the questions asked, and the manner of the interview is developed for the teacher and student participants.

Likewise, in the construction of survey instruments, this study developed interview questions following some basic guidelines. Each was constructed with references to the topic and the purpose of the study. It is made as brief and as precise as possible. Words that are believed to carry some bias and negative implications to teachers are likewise avoided. All interview responses are analysed by the thematic analysis.

4.6 Summary

This chapter describes how the study was conceived and highlights the steps that were taken to gather, analyse, and interpret the data. The planning and focus of the study were decided based on the researchers' observations and experiences in handling assessment courses, as well as training and reading on the topic area. The embedded mixed-methods design employed the quantitative method as the primary approach, and the qualitative method as the supportive approach is used to collect, examine, and interpret the data. The teacher and student questionnaires comprised adapted/modified scales such as the background questionnaires, technology acceptance scales, attitude towards formative assessment, and attitude towards assessment modes are devised and employed to collect the quantitative data. The achievement tests, which are composed of the Grade-10 mathematics targeted concepts, are employed to collect the outcome scale in the quantitative dataset. The interview questions are also developed and used to gather the qualitative data. The employed instruments are subjected to rigorous validation and reliability using Rasch Model and CFA employing ConQuest (v.4) and AMOS (v.26). These instruments and the interview questions are conducted with the permission of the University of Adelaide's ethics committee and involved the Ministry of Education in Myanmar. The data gathered from the questionnaires and the achievement tests are analysed and interpreted using descriptive and inferential statistics, including SEM and HLM. The hierarchical linear modelling analyses are conducted using the HLM to investigate the direct effect and interaction effect among the scales from students' and the teachers' levels. Further, the interview responses are applied to support the findings of the quantitative approach. The statistical analyses are carried out using SPSS (v.26), MPlus (v.8), and HLM software (v.6.0). The discussion in this chapter provides the fundamental basis for implementing the methods of data analysis for the next few chapters; in particular, instrument validation and verification, demographic and descriptive analysis, structural equation modelling, and hierarchical modelling.

Chapter 5

Item-bank for Test modes

5.1 Introduction

An item-bank is a relatively large collection in which items are stored in a database. Velasco (2017) defined an item-bank as a source of valid, reliable, and defensible test items for any high-quality computer-based testing. Generally, an item-bank can assign items for any form of test modes (for example, paper-and-pencil test (PPT), linear-online test (LOT), and computer-adaptive test (CAT)). There are some benefits to item-bank. They are flexibility, security, and consistency (Umar, 1999). Rudner (1998) suggests the flexibility of an item-bank because it is possible to edit or delete items from an item-bank and if necessary, add new items in the bank. However, according to the comments of Rudner (1998), the researcher or test designer spends some amount of time planning and assembling items in the bank, to calibrate them and then entering them into a database.

In most cases, the item-bank contains good questions or items. Further, meaningful assessment compromised good measures with good psychometric properties. Hence, before the major experiment of this study, the researcher conducted a data collection in two schools to collect the responses of targeted questions and analyse the quality of questions. After examining their quality, the researcher or questioner can assemble questions with good psychometric properties in the item-bank. Hence, an item-bank must contain items with a wide range of item estimates or difficulty.

This study aims to compare the functionality of CAT, LOT, and PPT. Although the conventional test mode (PPT) does not need an item-bank, the item-bank is a key performer for administering CAT and LOT. Mostly, the priority of CAT is the construction of the item-bank. CAT especially selects the items from the bank to determine the abilities of test takers. There are some considerations for assembling items in an item-bank: (1) planning items, (2) assembling items, (3) calibrating items, and (4) entering items into a database for an item-bank. The process of item banking took approximately three months to assemble MCQ items (questions) relating to the targeted mathematics contents (which are f*unction*, *remainder*, and *factor*) for Grade-10 high school students in Myanmar. Two high schools in Yangon participated in the sample. For this study, the Rasch Measurement Model (RMM) does not demand a larger sample size because the model depends not on the sample size. Consequently, this study applied RMM as the method of analysis, and the sample size was sufficient to analyse the items for the item-bank.

This chapter describes the data collection for item planning, assembling and calibration, testing instruments, research methods for assembling item-bank items, psychometric analyses for examining item quality, and the findings.

5.2 Data collection for item planning, assembling and calibration

Out of Myanmar High school students, two high schools took part in the data collection for calibrating the item-bank. There are 250 students from those schools participating in the data collection. The average age of all test takers was 15.5 years. Of the 250 students, 126 students (50%) were from school A and 124 students (50%) from school B. The data collection shows the equal gender distribution of these students. Of the whole sample, 132 (53%) were male students, and 118 (47%) were female students. Sixty-nine students (55%) are males in school A, and 57 students (45%) are female. Also, sixty-three male students (51%) and 61 female students (49%) in school B,

5.3 Testing instruments

This thesis selected 165 items of Grade-10 from the previous matriculation examinations (from 2014 to 2018) that have already been standardised (Myanmar Examination Board, 2018) to collect items of good quality. The study selected items from three sub-topics from the algebra content of the Grade-10 mathematics textbook. It was produced by Ministry of Education and is called 'Grade-10 Mathematics textbook', and its curriculum was changed from 2012 up to now, and all old question sets books, from 2014 to 2018, which were approved by Ministry of Education, Myanmar. According to Blooms' Taxonomy, the items cover three cognitive domains, i.e., knowing, reasoning, and applying. In addition, Blooms Taxonomy can categorise items into different cognitive domains that assess different thinking skills (Granello, 2001).

An item-bank needs to be content valid, and a test should cover the whole construct of the targeted content domain. Again, an item-bank has enough items with a wide range of item statistics and highly precise measurement. The item should function the same way in different subgroups.

After content validation, this work assembled 60 items out of 67 items for the targeted item-bank of the *function* content domain, 42 items out of 50 items for the *remainder* content domain, and 39 items out of 42 items for the *factor* content domain (see **Table 5.1**, **Table 5.2,** and **Table 5.3**). **Table 5.4** represents the number of assembled items according to cognitive domains, and content domains.

Table 5.1

Number of assembled items from function content domain

Objectives of *Function* Content Domain	Number of items from the Matriculation Examination	Number of Selected Items after Content Validation
1. Find the simple functions	9	7
2. Find the unknown value of simple functions in the complex problem	9	8
3. Find the unknown value of the equation of combination operation of functions	9	8
4. Find the unknown function of the combination operation of the functions	9	9
5. Find the inverse functions	9	8
6. Find the unknown number of the equation of inverse function	9	9
7. Find the combination of simple function, combination function and inverse function	13	11
Total	67	60

Table 5.2

Number of assembled items from remainder content domain

Objectives of *Remainder* Content Domain	Number of items from the Matriculation Examination	Number of Selected Items after Content Validity
• Know the remainder of the simple polynomial	7	7
• Find the remainder and substitute the value the polynomial	10	9
• Find the unknown value in the equation of remainder	12	8
• Know the conception of divisible and know the remainder is zero	10	9
• Find the unknown value when the polynomial is divisible by (x+a)	11	9
Total	50	42

Table 5.3

Number of assembled items from factor content domain

Objectives of Factor Content Domain	Number of items from the Matriculation Examination	Number of Selected Items after Content Validity
1. Know the factor concept	10	7
2. Find a factor of the polynomial	8	8
3. Find the unknown value when the polynomial has a factor	11	8
4. Find the unknown value when the polynomial has two factors	9	7
5. Find the unknown value when two polynomials have a common factor	10	9
Total	48	39

Table 5.4

Number of assembled items according to content and cognitive domains

Cognitive Domains	Number of items in *Function* Content	Number of items in *Remainder* Content	Number of items in Factor Content
Knowing	19	13	12
Reasoning	20	14	13
Applying	21	15	14
Total	60	42	39

The sample took seven sets of questions. There were three sets of questions for the *function* content domain, and each set contained 20 MCQ items. The total number of items in the *function* content domain was 60. For the *remainder* content domain, the study contained two sets of questions, and the first set contained 20 items and the second 22 items. There were 42 items from the remainder content domain in total. The thirty-nine items were from the factor content domain, and there were two sets of questions (the first set contains 20 items, and the second contains 19 items). Each item was composed of a stem and five options, including correct and four wrong answers among the options. Test takers were given an hour to respond to each set. In the calibration of items, the common person test equating was applied because the same students took all tests.

This study applied PPT as the mode of test administration for data collection to develop item-bank. The item parameter or item difficulty may depend on test administration mode

(Dillman, 2007). However, previous studies reported data collection in PPT mode for the item-bank (see Burghof, 2001; Chuesathuchon, 2008). Another consideration is the data collection context, where there is insufficient support for administering CBT. Therefore, this study used data collected by the PPT administration to develop an item-bank.

5.4 Research methods of assembling items for item-bank

Rudner (1998) suggests the application of the Rasch Measurement Model (RMM) to examine the item quality for item bank. There is a need for good items in the bank for diagnosis of the abilities of students, if based on the scale-generated items of the model. The assumptions of the RMM are unidimentionality, local independence, and invariance that are the excellent features of measurement (Panayides, Robinson and Tymms, 2010). If the item adequately fits the assumptions of the RMM, it is possible to assume the items as items with good psychometric properties.

Among Rasch models, a simple logistic model is the most appropriate one for analysing dichotomous items because this study mainly focuses on the utility of selected items with a dichotomous score. The formula of the Rasch model is

$$P_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$$

where,

P_{ni} is the probability of a correct response given by person n to item i (Keeves and Alagumalai, 1999). β_n is the person's ability, and δ_i is the item's difficulty.

The equation above can express person ability and item difficulty in the same logit unit. When an ability of a person approaches an item with difficulty, there are more chances of getting the correct response. When an ability of a person matched item difficulty, the chance of getting a correct response is 0.5. In large-scale studies like PISA, the researcher can use the Rasch model to conduct item and test analyses to verify the utility of the test. This study examined item response model fit, item discrimination, item-person map, and separation reliability for the utility of the test (PISA, 2012). This thesis conducted test fairness and equity by differential item functioning (DIF) analysis after examining the quality of items.

5.4.1 Item fit estimates

The performance of persons and the function of items need to fit the theoretical characteristic curve of persons and items in order to confirm the assumption of unidimensionality for the Rasch model (Keeves & Alagumalai, 1999). In the tests of fit, the distribution of the observed probabilities of responses assumed normal. The distribution curve can deviate from the theoretically expected curve within an acceptable range (Keeves & Alagumalai, 1999). However, examinees whose behaviour is unpredictable and poorly constructed items and functions over a narrow range of ability cannot prove Rasch scaling (Keeves & Alagumalai, 1999). Yuan (1999) applied calibration and scored only the items that fit the Rasch scale. The item fit statistics determined how items fit the estimated Rasch model (Boone & Scantlebury, 2005). Item fit statistics indicate the function of individual items (Boone & Scantlebury, 2005). According to Afrassa (1999), INFIT mean square (INFIT MNSQ), or weighted mean square statistics, can examine how consistently item characteristic curve (ICC) of each item are fitted by examinees whose chance of giving the correct answer to this item be close to the 0.5 probability level. Further, an item that all persons correctly respond, or no person responds correctly cannot provide any information for Rasch scaling (Keeves & Alagumalai, 1999). Therefore, it is necessary to investigate whether INFIT MNSQ ranges from 0.70 to 1.30 to identify those items that fit the Rasch model (Adams & Khoo, 1993, cited in Keeves & Alagumalai, 1999; Yuan, 1999). Items with their INFIT MNSQ, which is outside the acceptable range, are deleted or removed from the process of calibration and scoring.

5.4.2 Person-item map

The study transformed raw scores to logits of item difficulty and person ability to equal interval measures in Rasch logit units (Boone & Scantlebury, 2005) and mapped interval measures onto a linear interval scale. This map refers to a *person-item map* or *Wright map* (Boone & Scantlebury, 2005). The distribution of an ability of a person is according to their ability levels, on the left of the map (Boone & Scantlebury, 2005). The most able person is at the top of the map, and the least able person is on the base. Item difficulty distribution is on the right side of the map (Boone & Scantlebury, 2005). Items are plotted as their difficulty levels, with more difficult items at the top of the map and easier items at the base (Boone & Scantlebury, 2005). If an ability level of a person is plotted at the same logit value of an item, the person has a 50/50 chance of responding to the correct answer (Boone & Scantlebury, 2005). However, if the items are plotted above some persons, they are more difficult for these persons, and their chances of responding correctly is less than 50% (Boone & Scantlebury, 2005). Likewise, if the ability of a persons is plotted above some items, the persons are a greater than 50% chance of providing a correct response to these items (Boone & Scantlebury, 2005).

A person-item map helps the process of evaluating and interpreting data (Boone & Scantlebury, 2005). For analysis, it is possible to use a map as a tool for providing information, for building and monitoring the test; for evaluating the effectiveness of instruments; for helping test developers to order and space items, and to remove the redundant items with maintaining the tests' integrity (Boone & Scantlebury, 2005). Item ordering demonstrates which item concepts are easier or more difficult for targeted examinees, as well as the spacing of items, depicts the range of examinee ability for the single measured latent trait (Boone & Scantlebury, 2005). The steps of item ordering and spacing are useful for improving a measurement error (Boone & Scantlebury, 2005).

5.4.3 Separation reliability

Rasch Measurement Model (RMM) provides item and separation reliability indices of a person. These indices provide valuable information related to test utility and reliability. Its range is from 0 to 1 and the item separation reliability depends on the variance of item difficulty (Wright & Stone, 1999; Bond & Fox 2015). Person separation reliability depends on person ability variance (Wright & Stone, 1999; Bond & Fox 2015). The higher the reliability, the better separation of item difficulty or the ability of a person and the more precise the measurement (Wright & Stone, 1999). Acceptable item and person reliability indices should be above 0.8 (Bond & Fox 2015).

5.5 Psychometric analyses for examining item quality

Generally, it is important to conduct the analysis of item quality in order to evaluate the fit of item statistics to RMM. This study conducted three analyses to construct three item-banks according to different concepts: (1) *function* content domain, (2) *remainder* content domain, and (3) *factor* content domain. The researcher of this thesis used ACER ConQuest software and (v.4) to analyse the data (Adams, Wu & Wilson, 2015). Next, items are analysed to examine their psychometric properties and to identify non-fitting items to RMM through the following five steps.

1. The study checked the item separation reliability and person separation reliability as test statistics. The range of item separation reliability and person separation reliability varies from 0 to 1. The higher the reliability, the better separation of time difficulty or person's ability and the more precise the measurement (Wright & Stone, 1999).
2. Whether or not weighted mean square (i.e., INFIT MNSQ) statistics and the standardized statistics (ZSTD) for all items exist within the acceptable range are

checked (Wu & Adams, 2007). If fit statistics of items are outside the acceptable range, the items will violate unidimensionality and represent a threat to measurement.

The evaluation of fit statistics for individual items is based on INFIT MNSQ statistics, and ZSTD for all items that exist within the acceptable range is needed to be checked (Wu and Adams 2007). INFIT MNSQ is a good indicator for item fit statistics (Keeves and Alagumalai 1999). It is desirable that the value of INFIT MNSQ is near one. An item with greater INFIT MNSQ than one is associated with a low discrimination index, displaying data with more variability than expected in the model. If an item has an INFIT MNSQ less than one, it is associated with a high discrimination index, displaying data with less variability than expected in the model. The ZSTD represents how data fit the model (Wright and Linacre 1994). Its expected value is zero. If its value is less than zero, it is too predictable that data fit the model. More than zero represents the lack of predictability. If the INFIT MNSQ is acceptable, it is normal to ignore the ZSTD.

For item fit statistics, INFIT MNSQ range from 0.7 to 1.3 is considered fit while the ZSTD values between -2.00 and +2.00 are considered acceptable, with the 95 % confidence interval level of significance (Keeves and Alagumalai 1999). Wu and Adams (2007) noted in their study if the number of the sample is bigger and bigger, ZSTD may go beyond ±2. In the study, the sample size is 250, and ZSTD was out of the range. However, ZSTD is not considered if the INFIT is within the acceptable range (0.7 to 1.3).

3. Apart from examining INFIT MNSQ, the study examined the point-biserial index for the category (option) of the items. The point-biserial index is a good indicator of item discrimination (Ebel and Frisbie, 1986). The study examined the content and stem of the item and the nature of its options based on their discrimination power. A reliable test on discrimination between students with higher ability and those with lower ability could be a good test. The point-biserial index of a category that indicates the correct option should be positive and higher than 0.20 (Ebel and Frisbie, 1986; Penn, 2009; McGahee and Ball, 2009). An item with the point-biserial index, which is more significant than 0.39, has excellent discrimination index; being between 0.30 and 0.39 represents good discrimination coefficient; being between 0.20 and 0.29 represents mediocre discrimination index; being between 0.00 and 0.20 represent poor discrimination power; and less than (- 0.01) displays worst discrimination index. The distractors with a negative point-biserial index indicate that weaker students are selecting the correct answer. It is reasonable or considered to modify or delete from the

test items with a value lower than 0.20 (Ebel and Frisbie, 1986). After examining the steps above, a good analysis must remove any problematic items.

5.6 Item-bank of *function, remainder* and *factor* content domain

There is a need to store the items with good psychometric properties for each item-bank. Firstly, item separation reliability for item-bank of *function* content domain is 0.992, the *remainder* content domain is 0.943, and *factor* content domain is 0.978. All these values of item separation reliability indicate the excellence of the reliability of the item difficulty estimates. The person separation reliability of *function* content domain is 0.891, the *remainder* is 0.710, and the *factor* is 0.943. All these values of person separation reliability are a good index for the reliability of an ability of a person estimate (Bond and Fox 2015). These sets of items can sufficiently separate that sample size. To sum up, the 60 items for *function*, 42 items for the *remainder*, and 39 items for *factor* are associated with good test statistics.

Secondly, the study examined the fit statistics for all these three sets of MCQ items to fit the RMM (**Table 5.1**). The INFIT MNSQ of items for the *function* range from 0.72 to 1.30 with ZSTD (-4.0 to 10.4); items for the *remainder* from 0.76 to 1.30 with ZSTD (-1.7 to 2.4); and items for the factors from 0.86 to 1.14 with ZSTD (-2.27 to 1.6). Those indicate a good fit to the theoretical item characteristics curve or RMM curve. This result indicates that all items fit the model because students with average ability levels from the sample are responded correctly.

Thirdly, according to their respective point-biserial indices, all these three sets of items have good item discrimination power, not less than 0.200 of point-biserial indices. With good psychometric properties, present study selected 60 items for the item-bank of *function* content domain, 42 for the *remainder* content domain, and 39 for the *factor* content domain. **Figures 5.1, 5.2,** and **5.3** presented the Wright-maps of *function, remainder and factor* content domains and the psychometric properties for each content domain are described in **Tables 5.5, 5.6,** and **5.7**

Figure 5.1 Wright Map of *function* content domain

Table 5.5

Psychometric properties of items from function content domain

No.	Item	Item Difficulty	Error	INFIT MNSQ	CI	ZSTD	Pt Bis for correct answer
1	item1	-1.491	0.112	0.790	(0.82, 1.18)	-5.100	0.580
2	item2	-1.566	0.113	0.790	(0.82, 1.18)	-5.200	0.600
3	item3	-1.415	0.111	1.000	(0.82, 1.18)	-1.200	0.380
4	item4	1.443	0.115	1.240	(0.82, 1.18)	5.000	0.260
5	item5	-0.371	0.105	1.020	(0.82, 1.18)	0.900	0.440
6	item6	-1.586	0.113	0.990	(0.82, 1.18)	2.300	0.360
7	item7	-0.310	0.105	1.010	(0.82, 1.18)	-0.300	0.430
8	item8	0.936	0.110	0.740	(0.82, 1.18)	-4.300	0.670
9	item9	0.845	0.110	0.880	(0.82, 1.18)	-2.700	0.580
10	item10	0.128	0.106	1.100	(0.82, 1.18)	0.900	0.350
11	item11	-3.049	0.134	0.890	(0.82, 1.18)	-6.700	0.350
12	item12	-2.453	0.125	1.010	(0.82, 1.18)	3.300	0.230

No.	Item	Item Difficulty	Error	INFIT MNSQ	CI	ZSTD	Pt Bis for correct answer
13	item13	0.276	0.106	1.020	(0.82, 1.18)	0.400	0.420
14	item14	-1.054	0.108	0.720	(0.82, 1.18)	-5.100	0.690
15	item15	-1.208	0.109	0.800	(0.82, 1.18)	-4.400	0.610
16	item16	-0.061	0.105	0.800	(0.82, 1.18)	-2.100	0.640
17	item17	-1.139	0.109	0.870	(0.82, 1.18)	-2.800	0.540
18	item18	-1.274	0.110	0.880	(0.82, 1.18)	-2.400	0.500
19	item19	-0.038	0.105	0.720	(0.82, 1.18)	-4.000	0.710
20	item20	-0.472	0.106	0.870	(0.82, 1.18)	-2.100	0.590
21	item21	-0.372	0.105	0.720	(0.82, 1.18)	-3.900	0.730
22	item22	-0.550	0.106	0.910	(0.82, 1.18)	-1.600	0.550
23	item23	0.203	0.106	0.910	(0.82, 1.18)	-0.700	0.540
24	item24	0.244	0.106	1.010	(0.82, 1.18)	0.400	0.410
25	item25	-0.174	0.105	0.770	(0.82, 1.18)	-2.800	0.690
26	item26	1.203	0.113	1.300	(0.82, 1.18)	7.100	0.260
27	item27	-1.719	0.115	1.090	(0.82, 1.18)	3.800	0.210
28	item28	0.684	0.109	1.150	(0.82, 1.18)	2.400	0.280
29	item29	-0.772	0.107	1.000	(0.82, 1.18)	-0.400	0.400
30	item30	1.472	0.116	0.940	(0.82, 1.18)	0.100	0.490
31	item31	0.727	0.109	1.240	(0.82, 1.18)	4.000	0.220
32	item32	-0.117	0.105	1.280	(0.82, 1.18)	3.500	0.200
33	item33	0.181	0.106	0.820	(0.82, 1.18)	-2.700	0.610
34	item34	1.073	0.112	1.110	(0.82, 1.18)	2.700	0.300
35	item35	-1.781	0.115	0.810	(0.82, 1.18)	-5.100	0.530
36	item36	-1.276	0.110	1.010	(0.82, 1.18)	-0.300	0.380
37	item37	-0.841	0.107	0.980	(0.82, 1.18)	0.000	0.390
38	item38	-0.299	0.105	0.810	(0.82, 1.18)	-2.800	0.600
39	item39	0.461	0.107	0.820	(0.82, 1.18)	-2.700	0.620
40	item40	0.214	0.106	1.040	(0.82, 1.18)	0.600	0.430
41	item41	0.213	0.106	0.880	(0.82, 1.18)	-1.000	0.560
42	item42	1.931	0.122	1.240	(0.82, 1.18)	6.700	0.200
43	item43	0.847	0.110	1.040	(0.82, 1.18)	1.600	0.390
44	item44	1.690	0.119	1.130	(0.82, 1.18)	8.000	0.200
45	item45	0.709	0.109	1.130	(0.82, 1.18)	2.600	0.280
46	item46	0.409	0.107	1.020	(0.82, 1.18)	0.900	0.420
47	item47	0.366	0.106	0.830	(0.82, 1.18)	-2.300	0.600
48	item48	1.279	0.114	1.300	(0.82, 1.18)	10.400	0.600
49	item49	-2.802	0.130	1.020	(0.82, 1.18)	0.800	0.220
50	item50	0.975	0.111	0.960	(0.82, 1.18)	-1.900	0.490
51	item51	1.739	0.119	1.110	(0.82, 1.18)	5.700	0.300
52	item52	0.612	0.108	1.070	(0.82, 1.18)	2.400	0.360
53	item53	1.194	0.113	1.050	(0.82, 1.18)	4.000	0.350
54	item54	1.580	0.117	1.270	(0.82, 1.18)	7.400	0.240
55	item55	1.836	0.120	1.220	(0.82, 1.18)	10.400	0.220
56	item56	-2.370	0.123	1.170	(0.82, 1.18)	7.500	0.200

No.	Item	Item Difficulty	Error	INFIT MNSQ	CI	ZSTD	Pt Bis for correct answer
57	item57	1.298	0.114	1.010	(0.82, 1.18)	3.400	0.360
58	item58	0.743	0.109	1.010	(0.82, 1.18)	0.400	0.410
59	item59	1.908	0.121	1.010	(0.82, 1.18)	1.100	0.320
60	item60	1.144	0.856	0.970	(0.82, 1.18)	0.300	0.400
	Mean	0.000					
	SD	1.249					
	Minimum	-3.049	0.105				0.200
	Maximum	1.931	0.856				0.710
	Item Separation Reliability	0.992					
	Person Separation Reliability	0.891					

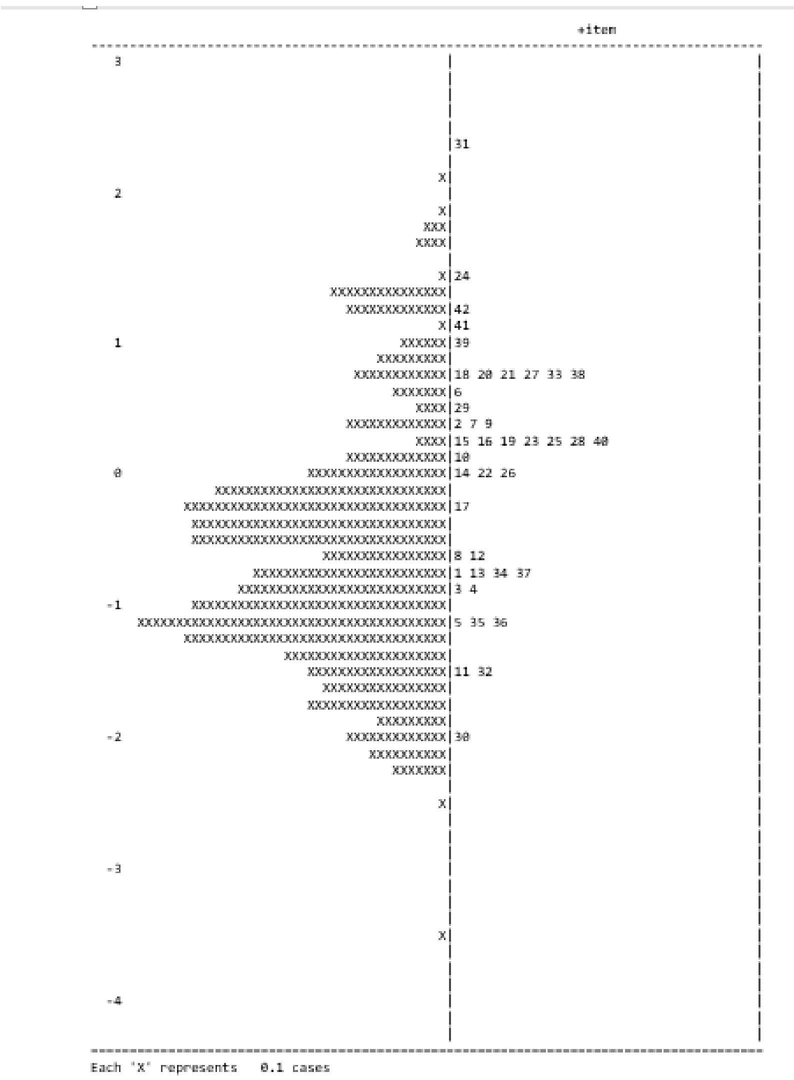

Figure 5.2 Wright Map of *remainder* content domain

Table 5.6

Psychometric properties of items from remainder content domain

No.	Item	Item Difficulty	Error	INFIT MNSQ	CI	ZSTD	Pt Bis for correct answer
1	item1	-0.802	0.200	0.940	(0.83, 1.17)	-0.600	0.420
2	item2	0.332	0.210	1.000	(0.76, 1.24)	0.000	0.320
3	item3	-0.931	0.201	1.120	(0.83, 1.17)	1.300	0.250
4	item4	-0.867	0.200	1.130	(0.83, 1.17)	1.500	0.290
5	item5	-1.126	0.202	0.880	(0.82, 1.18)	-1.300	0.590
6	item6	0.563	0.215	1.070	(0.72, 1.28)	0.500	0.240
7	item7	0.330	0.210	1.040	(0.76, 1.24)	0.300	0.540
8	item8	-0.613	0.200	0.840	(0.83, 1.17)	-1.900	0.730
9	item9	0.330	0.210	0.950	(0.76, 1.24)	-0.400	0.360
10	item10	0.113	0.206	0.940	(0.78, 1.22)	-0.500	0.620
11	item11	-1.466	0.207	0.940	(0.78, 1.22)	-0.500	0.420
12	item12	-0.613	0.200	0.880	(0.83, 1.17)	-1.400	0.570
13	item13	-0.804	0.200	1.070	(0.83, 1.17)	0.800	0.300
14	item14	-0.026	0.204	0.880	(0.80, 1.20)	-1.200	0.490
15	item15	0.256	0.209	0.990	(0.77, 1.23)	0.000	0.460
16	item16	0.184	0.207	0.870	(0.77, 1.23)	-1.200	0.550
17	item17	-0.291	0.202	0.980	(0.82, 1.18)	-0.200	0.420
18	item18	0.646	0.217	1.010	(0.71, 1.29)	0.100	0.310
19	item19	0.258	0.209	1.150	(0.77, 1.23)	1.200	0.200
20	item20	0.732	0.219	1.300	(0.70, 1.30)	2.400	0.420
21	item21	0.733	0.219	0.760	(0.70, 1.30)	-1.700	0.660
22	item22	-0.023	0.204	1.150	(0.80, 1.20)	1.500	0.200
23	item23	0.187	0.208	1.050	(0.77, 1.23)	0.400	0.230
24	item24	1.444	0.239	1.140	(0.56, 1.44)	0.700	0.200
25	item25	0.189	0.208	0.950	(0.77, 1.23)	-0.400	0.500
26	item26	-0.088	0.204	0.900	(0.80, 1.20)	-1.000	0.530
27	item27	0.736	0.219	0.970	(0.70, 1.30)	-0.100	0.550
28	item28	0.190	0.208	1.020	(0.77, 1.23)	0.200	0.220
29	item29	0.490	0.213	0.900	(0.73, 1.27)	-0.700	0.410
30	item30*	-1.996	0.218	0.960	(0.70, 1.30)	-0.200	0.400
31	item31*	2.433	0.272	1.220	(0.24, 1.76)	0.700	0.420
32	item32	-1.531	0.208	1.000	(0.78, 1.22)	0.100	0.360
33	item33	0.652	0.217	1.040	(0.71, 1.29)	0.300	0.430
34	item34	-0.735	0.200	0.800	(0.83, 1.17)	-2.600	0.610
35	item35	-1.121	0.202	0.830	(0.82, 1.18)	-1.900	0.510
36	item36	-1.122	0.202	0.840	(0.82, 1.18)	-1.800	0.510
37	item37	-0.799	0.200	0.890	(0.83, 1.17)	-1.400	0.560
38	item38	0.735	0.219	1.150	(0.70, 1.30)	1.000	0.240
39	item39	0.914	0.224	1.170	(0.67, 1.33)	1.000	0.200

No.	Item	Item Difficulty	Error	INFIT MNSQ	CI	ZSTD	Pt Bis for correct answer
40	item40	0.188	0.208	1.080	(0.77, 1.23)	0.700	0.330
41	item41	1.108	0.229	1.140	(0.63, 1.37)	0.800	0.200
42	item42	1.213	1.353	1.020	(0.61, 1.39)	0.200	0.200
	Mean	0.000					
	SD	0.898					
	Minimum	-1.996	0.218	0.760			0.200
	Maximum	2.433	0.272	1.300			0.730
	Item Separation Reliability	0.943					
	Person Separation Reliability	0.710					

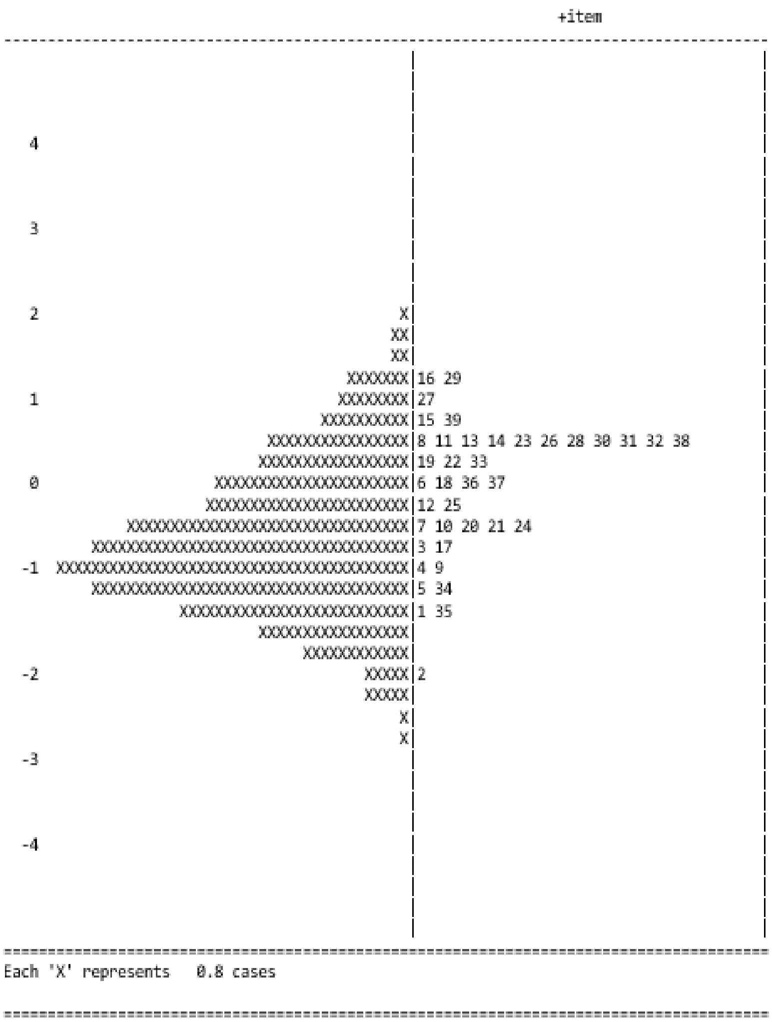

Figure 5.3 Wright Map of *factor* content domain

Table 5.7

Psychometric properties of items from factor content domain

No.	Item	Item Difficulty	Error	INFIT MNSQ	CI	ZSTD	Pt Bis for correct answer
1	item1	-1.480	0.116	0.910	(0.88, 1.12)	-1.500	0.480
2	item2	-2.101	0.123	0.980	(0.81, 1.19)	-0.200	0.410
3	item3	-0.547	0.113	1.040	(0.91, 1.09)	0.900	0.480
4	item4	-0.848	0.113	0.970	(0.91, 1.09)	-0.700	0.480
5	item5	-1.087	0.114	1.000	(0.90, 1.10)	-0.100	0.430
6	item6	-0.014	0.115	0.920	(0.89, 1.11)	-1.500	0.580
7	item7	-0.305	0.113	0.880	(0.91, 1.09)	-2.500	0.580
8	item8	0.662	0.121	1.140	(0.83, 1.17)	1.600	0.480
9	item9	-0.781	0.113	0.970	(0.91, 1.09)	-0.700	0.500
10	item10	-0.522	0.113	1.010	(0.91, 1.09)	0.100	0.490
11	item11	0.498	0.119	0.930	(0.84, 1.16)	-0.900	0.640
12	item12	-0.196	0.114	0.890	(0.90, 1.10)	-2.300	0.580
13	item13	0.496	0.119	1.030	(0.84, 1.16)	0.400	0.560
14	item14	0.494	0.119	0.980	(0.85, 1.15)	-0.200	0.580
15	item15	0.855	0.124	1.110	(0.80, 1.20)	1.100	0.520
16	item16	1.278	0.129	0.910	(0.75, 1.25)	-0.700	0.640
17	item17	-0.684	0.113	0.980	(0.91, 1.09)	-0.500	0.470
18	item18	-0.027	0.115	0.860	(0.89, 1.11)	-2.700	0.620
19	item19	0.303	0.117	0.910	(0.86, 1.14)	-1.400	0.610
20	item20	-0.409	0.113	1.090	(0.91, 1.09)	2.000	0.400
21	item21	-0.345	0.113	0.980	(0.91, 1.09)	-0.500	0.540
22	item22	0.322	0.118	1.020	(0.86, 1.14)	0.300	0.540
23	item23	0.633	0.121	1.000	(0.83, 1.17)	0.000	0.570
24	item24	-0.481	0.113	1.010	(0.91, 1.09)	0.100	0.470
25	item25	-0.263	0.114	0.940	(0.90, 1.10)	-1.300	0.560
26	item26	0.469	0.119	0.970	(0.85, 1.15)	-0.400	0.580
27	item27	1.112	0.127	0.990	(0.77, 1.23)	-0.100	0.590
28	item28	0.684	0.122	1.000	(0.82, 1.18)	0.100	0.570
29	item29	1.216	0.129	0.970	(0.76, 1.24)	-0.200	0.610
30	item30	0.576	0.120	0.980	(0.84, 1.16)	-0.200	0.570
31	item31	0.524	0.120	1.060	(0.84, 1.16)	0.800	0.500
32	item32	0.526	0.120	0.980	(0.84, 1.16)	-0.200	0.580
33	item33	0.322	0.118	1.050	(0.86, 1.14)	0.700	0.510
34	item34	-1.038	0.113	1.130	(0.91, 1.09)	2.700	0.390
35	item35	-1.443	0.116	1.080	(0.88, 1.12)	1.200	0.370
36	item36	0.110	0.116	0.990	(0.88, 1.12)	-0.100	0.520
37	item37	0.065	0.115	0.960	(0.88, 1.12)	-0.700	0.520
38	item38	0.486	0.119	1.080	(0.85, 1.15)	1.000	0.540
39	item39	0.940	0.726	1.030	(0.79, 1.21)	0.400	0.580

No.	Item	Item Difficulty	Error	INFIT MNSQ	CI	ZSTD	Pt Bis for correct answer
	Mean	0.000					
	SD	0.790					
	Minimum	-2.101	0.123	0.860			0.370
	Maximum	1.278	0.129	1.140			0.640
	Item Separation Reliability	0.978					
	Person Separation Reliability	0.943					

5.7 Summary

This chapter provided information about the procedure of constructing an item-bank. It describes the data collection for item-bank, testing instruments, research methods of data analyses, and the processes and findings of psychometric analyses. The researcher selected 165 items of Grade-10 to collect items of good quality. All items cover three content domains, namely *function*, *remainder*, and *factor*, and applied the Rasch Measurement Dichotomous model for the research methods to analyse psychometric properties of these items. This study conducted the analyses using ACER Conquest software, (v.4) (Adams, Wu & Wilson, 2015). According to the processes of psychometric analyses, the researcher of this thesis assembled 60 items for the item-bank of *function* content domain, 42 items for the *remainder* content domain, and 39 items for the *factor* content domain.

Chapter 6

Demographic and Descriptive Information

6.1 Introduction

This chapter considers the demographic background and descriptive analysis of 5 schools, 15 teachers, and 659 students and uses the IBM SPSS Statistics (v.26) to process the demographic information utilizing frequency and percent. In addition, the IBM SPSS (v.26) processed the descriptive data, including error bars, mean and standard deviation, and minimum and maximum values.

The first part of this chapter presents the demographic information and descriptive information of students. The demographic background included the *gender* status of students, their parents*' highest education level*, their *expected education*, and their *ICT familiarity*, followed by students' descriptive analysis of a wide range of scales. These included student-level scales such as *motivation*, *self-efficacy*, *attitude towards mathematics*, *ICT*, *attitude towards formative assessment*, and *attitude towards test modes* (*PPT*, *LOT*, and *CAT*).

The second part of this chapter deals with the demographic and descriptive information of teachers. This part describes the demographic information of teachers, their highest education level and qualification in education, *class size*, *multiple subject teaching*, and *ICT familiarity*. In addition, teacher-level scales such as the *general practice of formative assessment*, *specific practice of formative assessment*, *attitude towards formative assessment*, *attitude towards ICT*, *attitude towards test modes* (*PPT*, *LOT*, and *CAT*) are presented as descriptive information.

This analysis enables a summary of the data to highlight the trends of the scales and ensure better understanding for policymakers, curriculum specialists, national assessors, national examination boards, researchers, principals, and teachers. The demographic backgrounds and descriptive analysis are also important for preparing the subsequent analyses for teacher-level and student-level.

6.2 Number of participants

The respondents of this study consist of 659 Grade-10 students of five high schools in Myanmar. The present work described the distribution of student participants before their demographic information. Of the 659 students in this study, 170 (25.8 %) students are from School 1, 120 (18.2 %) from School 2, 138 (20.9 %) from School 3, 122 (18.5 %) from School

4, and 109 (16.5 %) from School 5 (see **Figure 6.1**). **Figure 6.1** shows the ratios of student participants according to different classrooms and different schools. Consequently, the number of students from five schools and respective classrooms proportionately participate in this study. The researcher selected three classrooms from each participant school and their mathematics teachers involved in this study. In total, there are 15 mathematics teachers. Two students from each school, in total ten students (5 boys, 5 girls), were invited to participate by intensity sampling method of qualitative approach (Gay, et.al., 2012), for interview in the qualitative section. A teacher from each school participated for interview also.

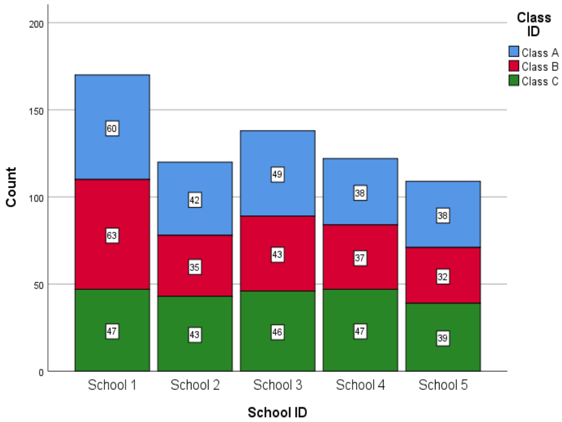

Figure 6.1 Number of students from different classrooms of five schools

6.3 Demographic information from student questionnaire

The following sections describe the number of student participants before presenting their demographic characteristics related to their *gender* status, *highest education level of* their *fathers and mothers*, their *expected education*, and their *ICT familiarity* with frequency and percent.

6.3.1 Gender

Of the participating students in this study, 333 (50.5 %) students are females, and 326 (49.5 %) students are males. The number of male and female students are proportionately participating in this study. **Figure 6.2** shows the number of student participants according to gender, different classrooms of five schools. In this study, the number of students from classrooms of five schools is proportionately participating in terms of gender.

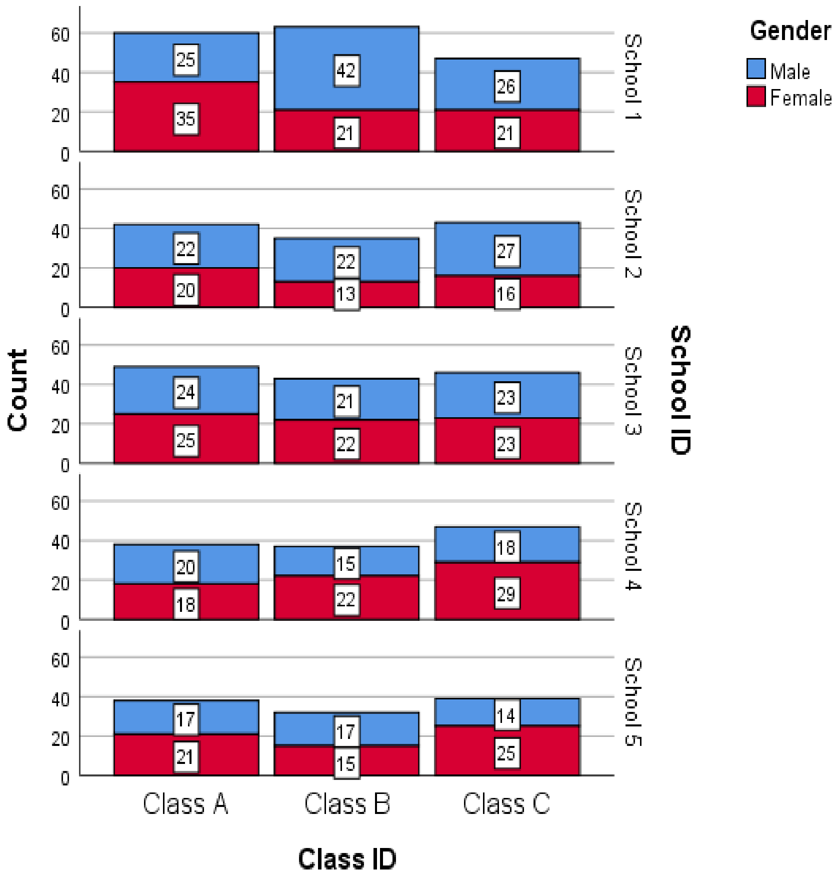

Figure 6.2 Number of male and female students from classrooms of five schools

6.3.2 Parent's highest education level

For the scale of parents' highest education level, there are two sub-scales, namely the highest education level for fathers and mothers. The number of students whose father has finished a *bachelor's degree* is 248 (37.6%), followed by the second largest group whose father has achieved *high school graduate* (n= 145; 22%). These two groups represented 59.6 % of the respondents. The remaining 24.9% consists of students whose father's education level is *diploma* (n= 89; 13.5%) and students whose father has achieved *master's degree* (n= 75; 11.4%). The rest of the participants (n= 80 students; 12.1%) are the group whose fathers have *not achieved high school*. However, the number of students whose fathers have achieved the *doctoral degree* is 22 (3.3%).

Mothers of 272 students (41.3 %) got a *bachelors' degree*. The remaining 34.1% consisted of students whose mothers have achieved *high school graduate* (n= 129; 19.6%), and 89 students (13.5%) consisted of those whose mothers have acquired a *diploma*. This number followed 100 students (15.1%) whose mothers have achieved *master's degree* (n=72; 10.9%) and *doctoral degree* (n= 28; 4.2%). Mothers of 69 students (10.5%) have not finished *high* *school*. **Table 6.1** presents the distribution of respondents according to the highest education levels (n= 659) of their fathers and mothers—the majority of students whose fathers or mothers got at least a *bachelors' degree.*

Table 6.1

Number of students according to their father's and mother's highest education level (n = 659)

	Father's Highest Education Levels		Mother's Highest Education Levels	
	Frequency	Percent	Frequency	Percent
Less than High School	35	5.3	33	5.0
Some High School	45	6.8	36	5.5
High School Graduate	145	22.0	129	19.6
Diploma's Degree	89	13.5	89	13.5
Bachelor's Degree	248	37.6	272	41.3
Master's degree	75	11.4	72	10.9
Doctoral Degree	22	3.3	28	4.2
Total	659	100.0	659	100.0

6.3.3 Expected education level

Most students presumed to acquire a Doctoral Degree (n=288; 43.7%), followed by 276 students (41.8%) that consist of students who want to achieve a bachelor's degree (n=140; 21.2%) and a master's degree (n= 136; 20.6%). The expected education level of 74 students (11.2%) is just the Diploma's Degree. Only 21 students (3.2%) hope to finish High School successfully. **Figure 6.3** show the distribution of students' expected education levels.

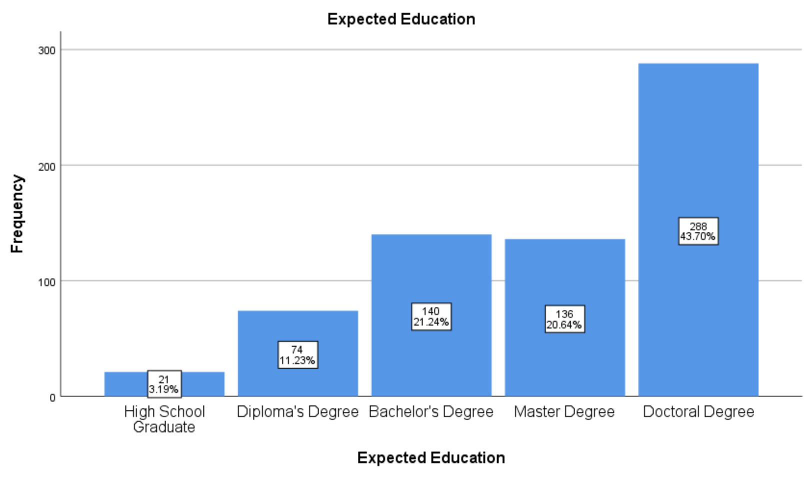

Figure 6.3 Number of students in terms of their expected education level (n = 659)

6.3.4 ICT familiarity

This study investigated the ICT familiarity of students. Questions about ICT access and ICT usage involves in the student background questionnaire Section 3 (see Appendix).

6.3.4.1 ICT access

This aspect of the study investigated whether students have access to ICT both at school and at home. **Table 6.2** presented the demographic information of the institutional and personal ICT accessibility. The study asked students concerning ICT accessibility in their school and home. The results of ICT accessibility in classroom revealed that most students possessed their own computer/laptop/tablets/smartphone (n=604, 91.7%), and personal email address (n=527, 80%). Most students reported that they have Wi-Fi access at home (n=561, 85.1%). However, over 50% described that they did not have access to training for Internet usage, such as email usage (n=534, 81%). According to the scale of ICT access, most students have personal access to the ICT devices, email addresses, and Internet at home, but they do not have Internet and email usage training.

Table 6.2

Frequency and percentage of ICT access (n = 659)

ICT access	I have access		I have no access	
	Frequency	Percent	Frequency	Percent
Computer, Laptop, tablets, smartphone for their own	604	91.7	55	8.3
Personal email address	527	80.0	132	20.0
Wifi access at home	561	85.1	98	14.9
Training for how to use the Internet and email	125	19.0	534	81.0

6.3.4.2 ICT usage

ICT usage is composed of six questions of how to use ICT and two questions of how long to use the Internet and ICT devices. It is interesting to note that over 60% of students frequently apply ICT devices for different purposes. **Table 6.3** can be easily seen that eighty-six percent of students (n=567) frequently use ICT to download learning material, music, film, games, and software. This number followed by 524 students (79.5%) who frequently participate in social media such as Facebook, WhatsApp, Viber. Nearly 70 percent of students (n= 457) often use ICT by communicating using email or social networks such as Facebook, WhatsApp, Viber. Approximately half of the participants frequently engaged ICT by uploading their own-created music, poetry, and videos (n=304, 46.1%) and by playing instructional programs, tutorials, remedial, or mastery learning (n=285, 43.2%). In summary, this result shows that students can use ICT without the introductory ICT training courses.

Table 6.3

Frequency and percentage of ICT usage (n = 659)

ICT usage	I have access		I have no access	
	Frequency	Percent	Frequency	Percent
Frequent Use of ICT by playing on a computer, online games	378	57.4	281	42.6
Frequent Use of ICT by playing instructional programs, tutorial, remedial, or mastery learning	285	43.2	374	56.8
Frequent Use of ICT by communicating by email, or social network such as Facebook, WhatsApp, Viber, etc.	457	69.3	202	30.7
Frequent Use of ICT by participating social network such as Facebook, WhatsApp, Viber, etc.	524	79.5	135	20.5
Frequent Use of ICT by downloading learning material, music, film, games, software	567	86.0	92	14.0
Frequent Use of ICT by uploading your own created music, poetry, videos	304	46.1	355	53.9

The research consisted of the weekly time students spent on ICT-related devices (for example, desktops, laptops, smart phones, and tablets). The average hours spent on ICT devices in a week is ten, and the average hours on Internet surfing in a week is ten (**Table 6.4**). Some students do not use ICT devices and the Internet surfing. However, the maximum hours of using ICT devices and Internet surfing are 35 hours a week. Further, this study regrouped these two items into seven categories. Most students spent about 11 to 15 hours on ICT devices (n= 241; 36.6%), followed by the second largest group who spent about 6 to 10 hours on ICT devices (n= 217; 32.9%). The 6.7% of students do not weekly spend time on ICT devices. The remaining 25 % consisted of students who spent about 16 to 20 hours in a week (n= 55; 8.3%), about 21 to 25 hours in a week (n= 67; 10.2%), about 26 to 30 hours (n=26; 3.9%), and about 31 to 35 hours (n=9; 1.4%).

For Internet surfing, a few students (n=42, 6.4%) do not weekly spend time on Internet surfing. However, the remaining 95 percent of students frequently spend time on surfing the Internet. Most students spent about 1 to 5 hours on the Internet surfing (n= 199; 30.2%). Nearly 20% of students spend about 6 and 10 hours on surfing the Internet (n= 126) and about 16 and 20 hours (n=126), followed by students who spent about 21 to 25 hours on Internet surfing (n= 68; 10.3%), about 16 to 20 hours (n=57, 8.6%), and those spending about 26 to 30 hours (n=26, 3.9%). Only eight students spent more than 31 hours a week on Internet Surfing. **Figures 6.4** and **6.5** illustrate the distribution of student participants according to the weekly time spent on ICT devices and Internet surfing.

Table 6.4

Means and standard deviation of the time spent on ICT Devices and internet surfing (n=659)

Item	N	Minimum	Maximum	Mean	Std. Deviation
Hours spent on ICT devices	659	0	35	10.07	8.480
Hours spent on internet surfing	659	0	35	10.36	8.372

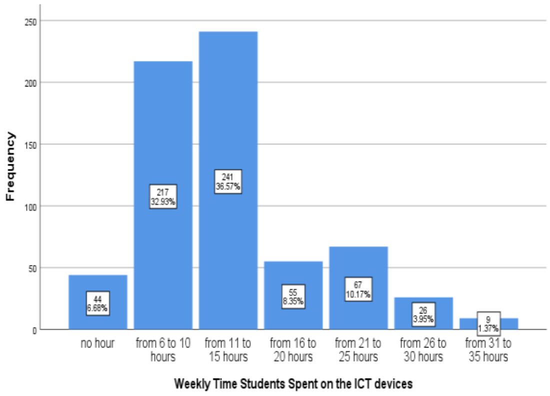

Figure 6.4 Weekly time students Spent on the use of ICT devices (n=659)

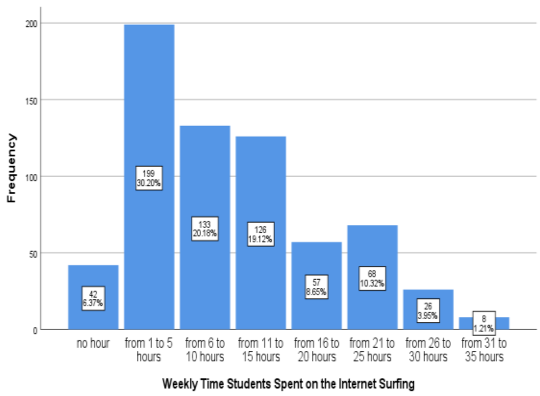

Figure 6.5 Weekly Time Students Spent on the Internet Surfing (n=659)

6.4 Descriptive results of scales from student questionnaire

Descriptive analysis of students is conducted for the scales such as intrinsic motivation, extrinsic motivation, self-efficacy, attitude towards learning mathematics and ICT, attitude towards classroom formative assessment, and attitude towards three type test modes. Hence, it is necessary to discuss the respondents' characteristics in detail. These scales use a four-point Likert-type scale. Item responses are coded as 1, 2, 3, and 4, corresponding to 'strongly disagree', 'disagree', 'agree', and 'strongly agree'. Some items are positively-worded statements, and others are negatively-worded statements. A negatively-worded statement has to be reverse-scored to keep the scale scoring consistency in the data analysis process and are initiated with ending as _R. Non-responded items were counted as 'missing items' and coded

‘9’. **Table 6.5** shows all items of the scales, their nature such as positive statement and negative statement, their items code equivalent to indicate reverse scoring and item texts.

Table 6.5

Items in scales from student questionnaire

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
Motivation			
IntrinM1	Positive	none	I enjoy learning mathematics.
IntrinM2	Negative	IntrinM2_R	I wish I did not have to study mathematics.
IntrinM3	Negative	IntrinM3_R	Mathematics is boring.
IntrinM4	Positive	none	I learn many interesting things in mathematics.
IntrinM5	Positive	none	I like mathematics.
IntrinM6	Positive	none	I like to solve mathematics problems
IntrinM7	Positive	none	I look forward to mathematics class.
IntrinM8	Positive	none	Mathematics is one of my favourite subjects
ExtrinM1	Positive	none	I know what my teacher expects me to do
ExtrinM2	Positive	none	My teacher gives me interesting things to do
ExtrinM3	Positive	none	My teacher has clear answers to my questions
ExtrinM4	Positive	none	My teacher is good at explaining mathematics
ExtrinM5	Positive	none	My teacher does a variety of things to help us learn
ExtrinM6	Positive	none	My teacher tells me how to do better when I make a mistake
ExtrinM7	Positive	none	My teacher links new lessons to what I already know
ExtrinM8	Positive	none	My teacher explains a topic again when we don’t understand
Self-efficacy			
Self-Efficacay_1	Positive	none	I usually do well in mathematics
Self-Efficacay_2	Negative	Self-Efficacay_2R	Mathematics is more difficult for me than for many of my classmates
Self-Efficacay_3	Negative	Self-Efficacay_3R	Mathematics is not one of my strengths
Self-Efficacay_4	Positive	none	I learn things quickly in mathematics
Self-Efficacay_5	Negative	Self-Efficacay_5R	Mathematics makes me nervous

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
Self-Efficacay_6	Positive	none	I am good at working out difficult mathematics problems
Self-Efficacay_7	Positive	none	My teacher tells me I am good at mathematics
Self-Efficacay_8	Negative	Self-Efficacay_8R	Mathematics is harder for me than any other subject
Attitude towards learning mathematics			
Att_LearnMath_1	Positive		I think learning mathematics will help me in my daily life.
Att_LearnMath_2	Positive		I need mathematics to learn other school subjects.
Att_LearnMath_3	Positive		I need to do well in mathematics to get into the college or university of my choice.
Att_LearnMath_4	Positive		I need to do well in mathematics to get the job I want.
Att_LearnMath_5	Positive		It is important to learn about mathematics to get ahead in the world.
Att_LearnMath_6	Positive		Learning mathematics will give me more job opportunities when I am an adult.
Att_LearnMath_7	Positive		My parents think that it is important that I do well in mathematics.
Attitude towards ICT			
Att_ICT_Use_1	Positive	none	ICT devices do not scare me at all.
Att_ICT_Use_2	Negative	Att_ComUse_2R	Working with ICT devices would make me very nervous.
Att_ICT_Use_3	Negative	Att_ComUse_3R	ICT devices make me feel uncomfortable.
Att_ICT_Use_4	Positive	None	I would feel at ease in the ICT class
Att_ICT_Use_5	Positive	none	I would feel comfortable working with a computer.
Att_ICT_Use_6	Negative	Att_ComUse_6R	ICT devices bore me.
Att_ICT_Use_7	Negative	Att_ComUse_7R	ICT devices are difficult to use
Att_ICT_Use_8	Negative	Att_ComUse_8R	Learning about ICT devices is a waste of time.
Att_ICT_Use_9	Positive	none	People that use of ICT devices are seen as being more important than those who don’t.
Att_ICT_Use_10	Positive	none	People who work with ICT devices make really good money.

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
Att_ICT_Use_11	Positive	none	I learn new tasks of use of ICT by trial and error.
Att_ICT_Use_12	Positive	none	When I have a problem with ICT devices, I will usually solve it on my own.
Att_ICT_Use_13	Positive	none	Using the ICT devices has increased my interaction with other students.
Att_ICT_Use_14	Positive	none	I develop short cuts, and more efficient ways to use ICT devices.
Att_ICT_Use_15	Positive	none	I would like to learn more about ICT devices.
Att_ICT_Use_16	Positive	none	If I need ICT skills for my career choice, I will develop them.
Attitude towards formative assessment			
Att_Formative Assessment_1	Positive	none	The use of formative assessment improves my performance.
Att_Formative Assessment_2	Positive	none	Formative assessment makes me to be actively involved in learning process.
Att_Formative Assessment_3	Positive	none	I enjoy my teacher asking questions during lesson.
Att_Formative Assessment_4	Negative	Att_Formative Assessment_4R	Asking me questions when the lesson is going on distracts my attention
Att_Formative Assessment_5	Negative	Att_Formative Assessment_5R	Formative assessment is time consuming.
Att_Formative Assessment_6	Positive	none	Corrective feedback enhances my learning.
Att_Formative Assessment_7	Positive	none	I adopt a deeper approach to learning whenever I am corrected.
Att_Formative Assessment_8	Positive	none	Corrective feedback helps me to know where I am lacking after each feedback
Att_Formative Assessment_9	Positive	none	I like it when my teacher points out my mistakes
Attitude towards PPT			
Att_PPTMode_1	Positive	none	PPT helps me to identify my weak areas
Att_PPTMode_2	Positive	none	PPT helps me to build my confidence.
Att_PPTMode_3	Positive	none	PPT helps me to improve mathematics learning

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
Att_PPTMode_4	Positive	none	In PPT, I was very focused on understanding the questions and tasks
Att_PPTMode_5	Positive	none	I persisted in PPT mode even when it was challenging or difficult.
Att_PPTMode_6	Negative	Att_PPTMode_6R	I was anxious in PPT.
Att_PPTMode_7	Positive	none	The feedback from PPT helps me reach my learning goal.
Att_PTMode_8	Positive	none	The feedback from PPT helps me recognize where I can improve.
Att_PPTMode_9	Positive	none	The feedback from PPT lets me know which types of tasks I should practice
Att_PPTMode_10	Positive	none	The feedback from PPT lets me know whether I should/have to prepare myself better.
Att_PPTMode_11	Positive	none	After receiving the feedback from PPT, I make more effort.
Attitude towards LOT			
Att_LOTMode_1	Positive	none	LOT helps me to identify my weak areas
Att_LOTMode_2	Positive	none	LOT helps me to build my confidence.
Att_LOTMode_3	Positive	none	LOT helps me to improve mathematics learning
Att_LOTMode_4	Positive	none	In LOT, I was very focused on understanding the questions and tasks
Att_LOTMode_5	Positive	none	I persisted in LOT mode even when it was challenging or difficult.
Att_LOTMode_6	Negative	Att_LOTMode_6R	I was anxious in LOT.
Att_LOTMode_7	Positive	none	The feedback from LOT helps me reach my learning goal.
Att_LOTMode_8	Positive	none	The feedback from LOT helps me recognize where I can improve.
Att_LOTMode_9	Positive	none	The feedback from LOT lets me know which types of tasks I should practice
Att_LOTMode_10	Positive	none	The feedback from LOT lets me know whether I should/have to prepare myself better.
Att_LOTMode_11	Positive	none	After receiving the feedback from LOT, I make more effort.

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
Attitude towards CAT			
Att_CATMode_1	Positive	none	CAT helps me to identify my weak areas
Att_CATMode_2	Positive	none	CAT helps me to build my confidence.
Att_CATMode_3	Positive	none	CAT helps me to improve mathematics learning
Att_CATMode_4	Positive	none	In CAT, I was very focused on understanding the questions and tasks
Att_CATMode_5	Positive	none	I persisted in CAT mode even when it was challenging or difficult.
Att_CATMode_6	Negative	Att_CATMode_6R	I was anxious in CAT.
Att_CATMode_7	Positive	none	The feedback from CAT helps me reach my learning goal.
Att_CATMode_8	Positive	none	The feedback from CAT helps me recognize where I can improve.
Att_CATMode_9	Positive	none	The feedback from CAT lets me know which types of tasks I should practice
Att_CATMode_10	Positive	none	The feedback from CAT lets me know whether I should/have to prepare myself better.
Att_CATMode_11	Positive	none	After receiving the feedback from CAT, I make more effort.

6.4.1 Motivation

This aspect of research is an adaptive questionnaire from the TIMSS motivation questionnaire on mathematics learning to study student motivation on mathematics learning. This scale assesses the level of motivation mathematics learning by students and consists of two sub-scales, which are intrinsic motivation and extrinsic motivation. The descriptive information of these sub-scales (**Table 6.6**) indicates that students generally exhibited higher intrinsic motivation and extrinsic motivation to learn mathematics because students tend to agree to the statements use to measure these scales.

Table 6.6

Descriptive information of motivation (n=659*)*

Item	N	Minimum	Maximum	Mean	Std. Deviation
IntrinM1	659	1	4	3.19	.655
IntrinM2_R	659	1	4	3.18	.737
IntrinM3_R	659	1	4	3.10	.670
IntrinM4	659	1	4	3.10	.646
IntrinM5	659	1	4	3.03	.691
IntrinM6	659	1	4	2.98	.724
IntrinM7	659	1	4	2.66	.753
IntrinM8	659	1	4	2.83	.827
ExtrinM1	659	1	4	2.73	.674
ExtrinM2	659	1	4	2.93	.632
ExtrinM3	659	1	4	3.28	.607
ExtrinM4	659	1	4	3.46	.562
ExtrinM5	659	1	4	3.21	.620
ExtrinM6	659	1	4	3.31	.568
ExtrinM7	659	1	4	3.26	.598
ExtrinM8	659	1	4	3.33	.575

6.4.2 Self-efficacy

This study measured general self-efficacy of mathematics students who participated by using TIMSS questionnaire's self-efficacy scale. The descriptive analysis for the scale of self-efficacy reveals that participants are generally confident in learning mathematics because mean values of all the items in this scale are above 2.5 (see **Table 6.7**).

Table 6.7

Descriptive information of self-efficacy (n=659)

Item	N	Minimum	Maximum	Mean	Std. Deviation
SelfEfficacy_1	659	1	4	3.01	.605
SelfEfficacy_2R	659	1	4	2.81	.808
SelfEfficacy_3R	659	1	4	2.64	.844
SelfEfficacy_4	659	1	4	2.76	.724
SelfEfficacy_5R	659	1	4	2.71	.849
SelfEfficacy_6	659	1	4	2.50	.745
SelfEfficacy_7	659	1	4	2.47	.799
SelfEfficacy_8R	659	1	4	2.83	.824

6.4.3 Attitude towards learning mathematics

The attitude of students towards mathematics learning included in the student-level questionnaire and adapted from the TIMSS questionnaire's attitude towards mathematics learning. For this scale, the results of the descriptive analysis exhibit that participant students generally have positive attitude towards learning mathematics because they tend to agree to the statements us to measure this scale (see **Table 6.8**)

Table 6.8

Descriptive information of attitude towards learning mathematics (n=659)

Item	N	Minimum	Maximum	Mean	Std. Deviation
Att_LearnMath_1	659	1	4	3.04	.714
Att_LearnMath_2	659	1	4	3.08	.642
Att_LearnMath_3	659	1	4	3.19	.687
Att_LearnMath_4	659	1	4	2.66	.864
Att_LearnMath_5	659	1	4	2.66	.864
Att_LearnMath_6	659	1	4	2.95	.753
Att_LearnMath_7	659	1	4	3.06	.704

6.4.4 Attitude towards ICT

The scale of attitude towards ICT assesses the level of attitude toward the use of ICT devices and internet surfing were adapted from the computer attitude scale for secondary students (Jones and Clarke, 1994). The mean values of all the items in this scale that are above 2.5 by the descriptive analysis reveal that participant students exhibited a positive attitude towards ICT use (see **Table 6.9**).

Table 6.9

Descriptive information of attitude towards ICT (n=659)

Item	N	Minimum	Maximum	Mean	Std. Deviation
Att_ICT_Use_1	659	1	4	3.17	.766
Att_ICT_Use_2R	659	1	4	3.18	.727
Att_ICT_Use_3R	659	1	4	3.25	.683
Att_ICT_Use_4	659	1	4	3.02	.723
Att_ICT_Use_5	659	1	4	3.12	.648
Att_ICT_Use_6R	659	1	4	3.31	.682
Att_ICT_Use_7R	659	1	4	2.91	.770
Att_ICT_Use_8R	659	1	4	3.27	.696
Att_ICT_Use_9	659	1	4	2.93	.810
Att_ICT_Use_10	659	1	4	2.90	.714
Att_ICT_Use_11	659	1	4	3.18	.684
Att_ICT_Use_12	659	1	4	2.94	.697
Att_ICT_Use_13	659	1	4	2.49	.769
Att_ICT_Use_14	659	1	4	2.77	.747
Att_ICT_Use_15	659	1	4	2.93	.649
Att_ICT_Use_16	659	1	4	3.21	.641

6.4.5 Attitude towards formative assessment

The scale of attitude towards formative assessment scale assesses the level of attitude toward formative assessment was adapted from the scale of students' attitude towards formative assessment and corrective feedback (Fakeye, 2016). **Table 6.10** presents that student has positive attitude towards formative assessment according to the mean values of all the items in this scale, which is above 2.5.

Table 6.10

Descriptive information of attitude towards formative assessment (n=659)

Item	N	Minimum	Maximum	Mean	Std. Deviation
Att_FormativeAssessment_1	659	1	4	3.07	.651
Att_FormativeAssessment_2	659	1	4	3.02	.618
Att_FormativeAssessment_3	659	1	4	2.83	.787
Att_FormativeAssessment_4R	659	1	4	2.83	.829
Att_FormativeAssessment_5R	659	1	4	2.98	.750
Att_FormativeAssessment_6	659	1	4	3.23	.591
Att_FormativeAssessment_7	659	1	4	3.15	.585
Att_FormativeAssessment_8	659	1	4	3.27	.585
Att_FormativeAssessment_9	659	1	4	3.34	.607

6.4.6 Attitudes towards PPT, LOT and CAT

The scales of attitudes to PPT, LOT, and CAT assess the level of attitude toward PPT, LOT, and CAT. These scales are adapted from students' computer-aided assessment survey by Broughton (2017). The descriptive information about these scales is present in **Table 6.11**. According to the attitude scale towards PPT, participants exhibited negative attitude towards PPT, due to the mean value of some items of this scale, which are around two or lower than two (**Table 6.11**). The descriptive result of the scales for attitude towards LOT and CAT reveals that participant students exhibited a positive attitude towards LOT and CAT because of the mean values of all their respective items in the scales above 2.5, see **Table 6.11**.

Table 6.11

Descriptive information of attitudes towards PPT, LOT and CAT

Item	N	Minimum	Maximum	Mean	Std. Deviation
Attitude towards PPT					
Att_PPTMode_1	659	1	4	2.12	.623
Att_PPTMode_2	659	1	4	1.93	.611
Att_PPTMode_3	659	1	4	1.75	.646
Att_PPTMode_4	659	1	4	1.93	.703
Att_PPTMode_5	659	1	4	2.01	.706
Att_PPTMode_6R	659	1	4	1.44	.818
Att_PPTMode_7	659	1	4	2.11	.653
Att_PTMode_8	659	1	4	2.10	.691
Att_PPTMode_9	659	1	4	2.05	.656
Att_PPTMode_10	659	1	4	2.02	.710
Att_PPTMode_11	659	1	4	2.00	.637
Attitude towards LOT					
Att_LOTMode_1	659	1	4	2.53	.697
Att_LOTMode_2	659	1	4	2.60	.700
Att_LOTMode_3	659	1	4	2.66	.676
Att_LOTMode_4	659	1	4	2.64	.697
Att_LOTMode_5	659	1	4	2.64	.715
Att_LOTMode_6R	659	1	4	3.24	.789
Att_LOTMode_7	659	1	4	2.44	.644
Att_LOTMode_8	659	1	4	2.58	.705
Att_LOTMode_9	659	1	4	2.64	.733
Att_LOTMode_10	659	1	4	2.68	.719
Att_LOTMode_11	659	1	4	2.60	.725
Attitude towards CAT					
Att_CATMode_1	659	1	4	2.95	.623
Att_CATMode_2	659	1	4	3.05	.624
Att_CATMode_3	659	1	4	3.30	.532
Att_CATMode_4	659	1	4	3.24	.508
Att_CATMode_5	659	1	4	3.09	.616
Att_CATMode_6R	659	1	4	2.76	.909
Att_CATMode_7	659	1	4	3.11	.500
Att_CATMode_8	659	1	4	2.96	.610
Att_CATMode_9	659	1	4	3.10	.548
Att_CATMode_10	659	1	4	3.18	.554
Att_CATMode_11	659	1	4	3.15	.573

6.5 Demographic information from teacher questionnaire

Teachers' demographic characteristics relate to their highest education levels, qualification in education, class size, multi-subject teaching, and ICT familiarity. The respondents of this study consist of 15 teachers who are currently teaching 659 participant students' mathematics in five different High schools in Myanmar.

6.5.1 Highest education level

Most of the teachers possessed a bachelor's degree of education (BEd) (n=9, 60%). This is followed by the bachelor's degree of science (BSc) (n=4, 27%) and bachelor's degree of arts (BA) (n=2, 13%).

6.5.2 Qualification in education

Sixty percent (n = 9) of teachers obtained a teacher education degree (4-year program), followed by 27% (n = 4) with a teacher education degree (2-year program) and short-term program in teaching education (n=2, 13%).

6.5.3 Class size

The average number of Grade-10 students in 15 classes from five participating schools is 53, with a minimum of 40 students and a maximum of 65 students. Most classes have 41 to 50 Grade-10 students (n=6, 40%), followed by 51 to 60 students (n=5, 33%), 61 to 70 students (n=3, 20%), and 31 to 40 students (n=1, 7%). **Table 6.12** shows the number of Grade-10 students in classes of five participating schools.

Table 6.12

Groups of class sizes

Number of Students	Frequency	Per Cent
CLSZ between 61 and 70	3	20.0
CLSZ between 51 and 60	5	33.3
CLSZ between 41 and 50	6	40.0
CLSZ between 31 and 40	1	6.7
Total	15	100.0
		Number of students
Average		53
Minimum		40
Maximum		65

6.5.4 Multi-subject teaching

Regarding to multi-subject teaching in the current academic year, most teachers reported that they teach only mathematics (n= 8, 53%), followed by teachers teaching three subjects, i.e., mathematics and other two subjects (n=4, 27%). Twenty percent (n=3) of teachers are teaching only two subjects, i.e., mathematics and another subject. **Table 6.13** shows the distribution of number of subjects which participant teachers are currently teaching.

Table 6.13

Number of Subjects that participant teachers are currently teaching

Number of Subjects	Frequency	Per Cent
1 (Maths)	8	53.3
2 (Maths and another subject)	3	20.0
3 (Maths and other two subjects)	4	26.7
Total	15	100.0
	Number of Classes	
Average	2	
Minimum	1	
Maximum	3	

6.5.5 ICT familiarity

This study investigated ICT familiarity among teachers. Questions about ICT Usage and time spent on ICT devices and the internet involve ICT familiarity in the teachers' background questionnaire. This study investigated whether teachers are using ICT for different purposes. According to **Table 6.14,** almost half of the teachers frequently use Microsoft Word Document (n=7, 47%). On the other hand, most teachers do not apply Microsoft Excel (n=10, 67%) and Microsoft PowerPoint (n=9, 60%). Sixty-seven percent (n=10) of teachers do not usually communicate with their colleagues, student, and student parents by email or on social networks such as Facebook, WhatsApp, and Viber. Similarly, Sixty-seven percent (n=10) of teachers do not usually download teaching and learning material, music, films, games, or software from the internet (see **Table 6.14**).

Table 6.14

Frequency and percentage of ICT usage

Item Code	ICT Usage	I have access		I have no access	
		Frequency	Percent	Frequency	Percent
ICT_Use1	Frequent use of Microsoft Word Document.	7	47%	8	53%
ICT_Use2	Frequent use of Microsoft Excel Worksheet.	5	33%	10	67%
ICT_Use3	Frequent use of Microsoft PowerPoint Presentation.	6	40%	9	60%
ICT_Use4	Frequent use of ICT for communicating colleagues, student and students' parent by email or social network (such as Facebook, WhatsApp, Viber, etc.).	5	33%	10	67%
ICT_Use5	Frequent use of ICT for downloading teaching and learning material, music, films, games, or software form the internet.	5	33%	10	67%

The teachers reported the weekly time they spent on using a computer. Teachers spend at least seven hours on ICT, and an average hour spent on ICT devices in a week is twelve. The maximum number of hours spent on computers in a week is 22 hours. Based on the average hours spent on ICT devices, two categories are low user less than 12 hours and high user more than 12 hours. More teachers are low users (n=9, 60%) than those who are high users (n=6, 40%) (**Table 6.15**).

The teachers reported the weekly time they spent accessing the internet. Teachers spend at least four hours on the internet, and the average hour spent on ICT devices in a week is eight. The maximum number of hours spent assessing the internet in a week is 14 hours. There are two categories based on the average hours spent on accessing the internet, the low user, i.e., less than 8 hours, and the high user, i.e., more than 8 hours. **Table 6.15** shows teachers who are low (n=9, 60%) and high (n=6, 40%) ICT users.

Table 6.15

Means and standard deviation of the time spent on ICT devices and internet surfing

Item Code	Items	N	Min	Max	Mean	Std.dev	High User		Low User	
							Frequency	Percent	Frequency	Percent
ICT_Use6	Hours for applying ICT devices	15	7	22	12.00	5.028	6	40%	9	60%
ICT_Use7	Hours for Internet surfing	15	4	14	8.00	3.586	6	40%	9	60%

6.6 Descriptive results of scales from teacher questionnaire

The study conducted descriptive analysis among teachers for general practices formative assessment scales, specific practices formative assessment, attitude towards formative assessment, attitude towards ICT, and attitudes towards PPT, LOT, and CAT. The first two scales use a four-point Likert-type scale, and their item responses are coded 1, 2, 3, and 4 corresponding to 'never', 'sometimes', 'regularly', all of the time'. The rest scales also use four-point Likert-type scale and their item responses are coded 1, 2, 3, and 4 corresponding to 'strongly disagree', 'disagree', 'agree', and 'strongly agree'. Some items are positively-worded statements, and others are negatively-worded statements. Negatively-worded statement has to be reverse-scored to keep the scoring scale consistent in the data analysis process and is initiated with ending as _R. Non-responded items were counted as 'missing items' and coded '9'. **Table 6.16** shows all items of the scales, their items code equivalent to indicate reverse scoring and item texts.

Table 6.16

Items in scales from teacher questionnaire

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
General practices of formative assessment			
GPTech_1	Positive	none	I use classroom discussion as general practices of formative assessment.
GPTech_2	Positive	none	I use classroom observation as general practices of formative assessment.
GPTech_3	Positive	none	I use whole class oral question-and-answer as general practices of formative assessment.
GPTech_4	Positive	none	I use oral question-and-answer with individual student as general practices of formative assessment.
Specific practices of formative assessment			
SPFA_1	Positive	none	I provide your students score on their formative assessments?
SPFA_2	Positive	none	I describe the results/scores of formative assessments?
SPFA_3	Positive	none	I provide the written feedback to students?
SPFA_4	Positive	none	I provide the feedback individual student?

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
SPFA_5	Positive	none	I provide students' weaknesses and strengthens as contents of feedback?
Attitude towards formative assessment			
AttFA_1	Negative	AttFA_1R	I rarely use formative assessment to evaluate my students' achievement.
AttFA_2	Positive	none	Formative assessment is interesting for my class.
AttFA_3	Negative	AttFA_3R	I do not like asking questions while the lesson is going on.
AttFA_4	Negative	AttFA_4R	Formative assessment makes my class boring.
AttFA_5	Positive	none	Formative assessment provides the useful information your students for enhancing students learning progress.
AttFA_6	Positive	none	Formative assessment conceived as feedback loop to close the gap between students' current learning statues and intended learning outcomes.
AttFA_7	Positive	none	Formative assessment provides a valuable learning experience for students.
Attitude towards ICT			
AttICT_1	Positive	none	ICT has the capacity to strongly enhance classroom teaching and learning.
AttICT_2	Positive	none	ICT provides valuable resources and tools to support student learning.
AttICT_3	Positive	none	ICT provides students with efficient presentation and communication tools
AttICT_4	Positive	none	I like challenge of exploring new technology and software.
AttICT_5	Positive	none	I fee apprehensive about using ICT in my classroom teaching.
AttICT_6R	Negative	AttICT_6R	It scares me to think that I could cause the computer to destroy a large amount of information by hitting the wrong key.
AttICT_7R	Negative	AttICT_7R	I hesitate to use ICT tools and equipment for fear of making mistakes that I can't correct.

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
AttICT_8R	Negative	AttICT_8R	Computer and internet technologies are somewhat intimating to me.
AttICT_9	Positive	none	I can effectively use ICT as instructional tool.
AttICT_10	Positive	none	I can effectively manage my classroom when students are using ICT
AttICT_11	Positive	none	I can extend my classroom teaching by using computer and internet.
AttICT_12	Positive	none	I can learn to use ICT for my teaching and learning process.
Attitude towards PPT			
PPTatt_1	Positive	none	PPT Mode examines students' ability to carry out mathematical procedures and methods.
PPTatt_2	Positive	none	PPT Mode examines students' deeper understanding of mathematical concepts.
PPTatt_3	Positive	none	Students immediately receive good quality feedback.
PPTatt_4	Positive	none	PPT Mode provides students with the opportunities to progress their learning.
PPTatt_5	Positive	none	PPT Mode provides students with the motivation to mathematics learning.
Attitude towards LOT			
LOTatt_1	Positive	none	LOT Mode examines students' ability to carry out mathematical procedures and methods.
LOTatt_2	Positive	none	LOT Mode examines students' deeper understanding of mathematical concepts.
LOTatt_3	Positive	none	Students immediately receive good quality feedback.
LOTatt_4	Positive	none	LOT Mode provides students with the opportunities to progress their learning.
LOTatt_5	Positive	none	LOT Mode provides students with the motivation to mathematics learning.
Attitude towards CAT			
CATatt_1	Positive	none	CAT Mode examines students' ability to carry out mathematical procedures and methods.

Item Code	Nature of Statement	Item Code to indicate reverse scoring	Item Text
CATatt_2	Positive	none	CAT Mode examines students' deeper understanding of mathematical concepts.
CATatt_3	Positive	none	Students immediately receive good quality feedback.
CATatt_4	Positive	none	CAT Mode provides students with the opportunities to progress their learning.
CATatt_5	Positive	none	CAT Mode provides students with the motivation to mathematics learning.

6.6.1 General practices of formative assessment

The general practice of formative assessment included in the teacher questionnaire in this study is adapted from a lecturer's computer-aided assessment survey (Broughton, 2017). This scale measures how frequently the teachers apply the general practices of formative assessment. The descriptive information of this scale (**Table 6.17**) indicates that teachers apply these mentioned types of formative assessment regularly as the mean values of all items are close to 2.5.

Table 6.17

Descriptive information of general practices of formative assessment (n=15)

Items	N	Minimum	Maximum	Mean	Std. Deviation
GPTech_1	15	1	4	3.07	1.033
GPTech_2	15	1	4	3.13	.915
GPTech_3	15	1	4	3.13	.915
GPTech_4	15	2	4	2.93	.594

6.6.2 Specific practices of formative assessment

Teachers' specific practices of formative assessment included in the teacher questionnaire in this study is adapted from a national survey examining teachers' formative assessment practices (Fishman, Riconscente, Snider, Tsai, & Plass, 2014). This scale measures how frequently the teachers apply the specific practices of formative assessment. Participant teachers sometimes apply these mentioned types of formative assessment as the mean values of all items are around 2 (**Table 6.18**).

Table 6.18

Descriptive information of specific practices of formative assessment (n=15)

Items	N	Minimum	Maximum	Mean	Std. Deviation
SPFA_1	15	1	4	2.13	.743
SPFA_2	15	1	4	2.07	.458
SPFA_3	15	1	4	1.87	.743
SPFA_4	15	1	4	1.73	.594
SPFA_5	15	1	4	2.33	.724

6.6.3 Attitude towards formative assessment

Attitude towards formative assessment is adapted from the scale of teacher *attitude towards formative assessment* (Karim, 2015). This scale measures the level of attitude towards the application of formative assessment, and **Table 6.19** illustrates that participant teachers generally exhibited a more positive attitude towards ICT as the mean values of all items are around 2.5.

Table 6.19

Descriptive information of attitude towards formative assessment (n=15)

Item	N	Minimum	Maximum	Mean	Std. Deviation
AttFA_1R	15	1	4	2.67	1.234
AttFA_2	15	1	4	2.73	1.033
AttFA_3R	15	1	4	2.80	1.082
AttFA_4R	15	1	4	2.67	.976
AttFA_5	15	1	4	2.73	.884
AttFA_6	15	1	4	2.80	1.082
AttFA_7	15	1	4	2.73	1.033

6.6.4 Attitude towards ICT

Scale of attitude towards ICT is adapted from a computer attitude scale by Jones and Clarke (1994). This scale is for measuring the level of teachers' attitude to the application of ICT devices and internet surfing. According to the result of **Table 6.20**, participant teachers generally exhibited more positive attitude towards ICT because they tend to agree to the statements us to measure this scale.

Table 6.20

Descriptive information of attitude towards ICT (n=15)

Items	N	Minimum	Maximum	Mean	Std. Deviation
AttICT_1	15	1	4	2.87	.915
AttICT_2	15	1	4	2.53	1.060
AttICT_3	15	1	4	2.80	1.014
AttICT_4	15	1	4	3.00	1.000
AttICT_5	15	1	4	2.60	.828
AttICT_6R	15	1	4	2.87	1.060
AttICT_7R	15	1	4	2.67	1.047
AttICT_8R	15	1	4	2.73	.961
AttICT_9	15	1	4	2.67	.976
AttICT_10	15	1	4	2.60	1.183
AttICT_11	15	1	4	2.80	1.082
AttICT_12	15	1	4	2.87	1.125

6.6.5 Attitudes towards PPT, LOT, and CAT

The researcher of this thesis measured the scale of attitude towards test mode with three different scales: attitude towards PPT, attitude towards LOT, and attitude towards CAT. These scales are adapted from the survey of lecturers' computer-aided assessment by Broughton, (2017). They measure the level of teachers' attitude towards the application of PPT, LOT, and CAT as formative assessment. According to **Table 6.21**, participant teachers have more positive attitude towards PPT, LOT and CAT because they tend to agree to the statements us to measure these scales.

Table 6.21

Descriptive information of attitudes towards PPT, LOT and CAT

Items	N	Minimum	Maximum	Mean	Std. Deviation
Attitude towards PPT					
PPTatt_1	15	1	4	2.73	1.033
PPTatt_2	15	1	4	2.60	.910
PPTatt_3	15	1	4	2.80	.941
PPTatt_4	15	1	4	2.73	1.033
PPTatt_5	15	1	4	2.73	1.033
Attitude towards LOT					
LOTatt_1	15	1	4	3.07	.884
LOTatt_2	15	1	4	3.13	.915
LOTatt_3	15	1	4	3.00	1.000
LOTatt_4	15	1	4	3.13	1.060
LOTatt_5	15	1	4	3.07	.884
Attitude towards CAT					
CATatt_1	15	1	4	2.73	.884
CATatt_2	15	2	4	3.27	.884
CATatt_3	15	1	4	2.87	.834
CATatt_4	15	1	4	2.67	.816
CATatt_5	15	1	4	3.07	.961

6.7 Summary

The results presented in this chapter show the demographic information of students and teachers and the descriptive analyses of the scales employed. The number of male and female students was almost equal in each classroom and school, and most parents of students finished either high school graduate or bachelors' degrees. Half of the participant students expected their education to reach the Doctoral Degree level. More than 80 percent of students have already accessed the ICT. Most of the students use ICT devices in the 5 to 15 hours range and do surf the internet from 1 to 15 hours a week.

The descriptive analyses of scales from student questionnaire are as follows: participant students generally exhibited higher intrinsic motivation and extrinsic motivation to learn mathematics because mean values of all of the items on this scale are above 2. They are generally confident in learning mathematics. They have a positive attitude towards learning mathematics, ICT, formative assessment, LOT, and CAT due to all of the items on this scale above the average. On the other hand, they have negative attitude towards PPT due to mean values of some items on this scale that are lower than 2.

As the demographic information about teachers, their highest education level, 9 out of 15 teachers finished a bachelor's degree of education, a 4-year program. Their class sizes are above 30 students, which means all participant teachers handle the large class size. Seven out of 15 participant teachers teach one or two other subjects apart from mathematics. Less than 50 percent of teachers have highly frequent use of ICT devices. Sixty percent of teachers are lower users of ICT devices and internet surfing. Concerning the result of descriptive analyses of participant teachers, they apply general practices and specific practices of formative assessment more frequently in the average scale. They have positive attitudes towards formative assessment, ICT, PPT, LOT, and CAT.

The demographic and descriptive analyses provide a basis for the subsequent analyses such as student- and teacher-level SEM models and hierarchical linear models due to the three test modes.

Chapter 7

Instrument Validation: Confirmatory Factor Analysis

7.1 Introduction

This study examines student-level and teacher-level factors, their interrelationships, and their impact on mathematics achievement improvement of students due to the different test modes. As discussed earlier in Chapter 5, this study derived student-level and teacher-level constructs from abstract concepts that are clearly defined and well operationalised. However, we know little concerning the items constructed if the items are good indicators of the latent scales or if the items constructed fit the structure of the latent scales for the context of five Myanmar High Schools. In addition, establishing the reliability and validity of the constructs used in this study is fundamental in providing accurate estimates for the subsequent analysis of student- and teacher-level models and hierarchical linear models. Consequently, it is vital to examine the constructs employed in the instruments.

This chapter describes the validation of instruments used for this study through Confirmatory Factor Analysis (CFA). This instrument includes student-level instrument: Student questionnaire and Student Achievement Tests, and Teacher-level instrument: Teacher questionnaire. The scales from the student questionnaire involved were *motivation*, *self-efficacy*, *attitude towards mathematics*, *attitude towards formative assessment*, *attitudes towards paper-and-pencil test (PPT)*, *linear-online test (LOT)*, *computer-adaptive test (CAT)*, and *ICT familiarity*, and *attitude towards ICT*. All scales used in the student questionnaire are a four-point Likert scale apart from the ICT Familiarity scale. The items from that scale were dichotomous items with Yes or No answers. The four-point Likert responses are 1, 2, 3, and 4, corresponding to 'strongly disagree, disagree, agree and strongly agree', while the dichotomous item responses were coded as 0 and 1. The items from the Student Achievement Tests are MCQ items. The items are dichotomous scored. The correct answer is coded as 1, while the incorrect answer is coded as 0. Moreover, the study presented a fit comparison of models and the structures.

The study validated the scales employed by examining their construct validity for these instruments. Firstly, Confirmatory Factor Analysis (CFA), followed by convergent validity, another part of construct validity, examines the construct validity. For the analyses, the researcher employed CFA statistics to investigate the construct validity of the instrument and helped to identify the measurement model. The study examined the relationships between measured or observed scales and the constructs or unobserved scales in the measurement model. CFA results test the measurement theory by comparing the theoretical measurement model against reality for this current sample. As indicated in **Chapter 4**, this study employed a model comparison approach to identify the best structure for the scales. In addition, the study examined four different models and included the one-factor model, N-orthogonal factors model, N-correlated factors model, and the hierarchical model. Subsequently, this chapter discussed the fit comparison of the alternative models and their model fit indices as well as the final structure of the scales and presented in the following sections. As mentioned in **Chapter 4**, this chapter examined the acceptable range of incremental and absolute fit indices, including the χ^2 statistics, comparative fit index (CFI), Tucker-Lewis Index (TLI), and root mean square error of approximation (RMSEA) to evaluate model fit.

The result from CFA enables the researcher to confirm the structure of the observed scales and the relationship between the observed scales and their underlying latent traits. These confirmed structures for subsequent Rasch analysis as described in the next chapter. Moreover, the measurement model by the CFA causes the structural theory of the Structural Equation Modelling (SEM) model to be fully specified.

There are three components in convergent validity: factor loadings, average variance extracted (AVE), and construct reliability. In addition, Cronbach's alpha is examined for the inter consistency of each instrumental scale. See **Chapter 4** for a detailed description of all Cronbach's alpha criteria for scales of the instruments.

This study employed data analysis software, IBM SPSS AMOS (Analysis of Moment Structures) (v.26) and MPlus (v.8) to test the construct validity of the instrument. Regarding the data preparation, the researcher of this study entered data from the questionnaire in a file using the IBM SPSS (v.26) before conducting the analysis of CFA. For the CFA analysis, the researcher employed IBM SPSS AMOS (v.26) statistical software package to produce a graphical image of each scale and MPlus (v.8) for generating the fit indices values of the scales and the factor loading and the error variance. The study applied factor loading and the error variance to calculate average variance extracted (AVE) and construct reliability according to their respective formulae (see in Chapter 5). The study employed IBM SPSS (v.26) to conduct Cronbach's Alpha.

7. 2 CFA for scales from student questionnaire

7.2.1 Motivation

The scale of motivation from student questionnaire consists of two sub-scales: intrinsic motivation and extrinsic motivation. Each sub-scale consists of eight items. An examination of the model fit indices of the first four models, as shown in **Table 7.1,** indicated that the three-correlated factors model (Model 3) and hierarchical model (Model 4) exhibited better model fit when compared to the other two models ((Model 1) and (Model 2)). Although the two-correlated factors and hierarchical models provided similar results in their model fit indices, the two-correlated factor model was preferred and used for subsequent analysis. The use of the correlated model for this scale enables this study to compare the Rasch Analysis more. The model has a higher RMSEA value (0.103), a higher CFI value (0.950), and a higher TLI value (0.941). In addition, the model has a significant change in chi-square ($\boldsymbol{\chi^2}$ = 821.112, $p < 0.05$) given the change in degrees of freedom. The study selected Model 3 as the best fit for the data (**Figure 7.1**).

Table 7.1

Fit indices for CFA models in motivation in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	2463.834	104		0.835	0.809	0.186
2. Orthogonal Model	2344.439	104	22.543	0.843	0.819	0.181
3. Correlated Factor Model	821.112	103	7.972	0.950	0.941	0.103
4. Hierarchical Model	821.112	103	7.972	0.950	0.941	0.103

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.2 displays standardized factor loadings for scales. The results of this analysis confirm that factor-loading values of all items in this construct are above the cut-off value (0.30), referring to these items measured the targeted latent construct. **Figure 7.1** and **Table 7.2** show the final structure and factor loadings of the model. Further examination of the correlation factor loading between intrinsic and extrinsic motivation indicated a medium correlation value (r = 0.429).

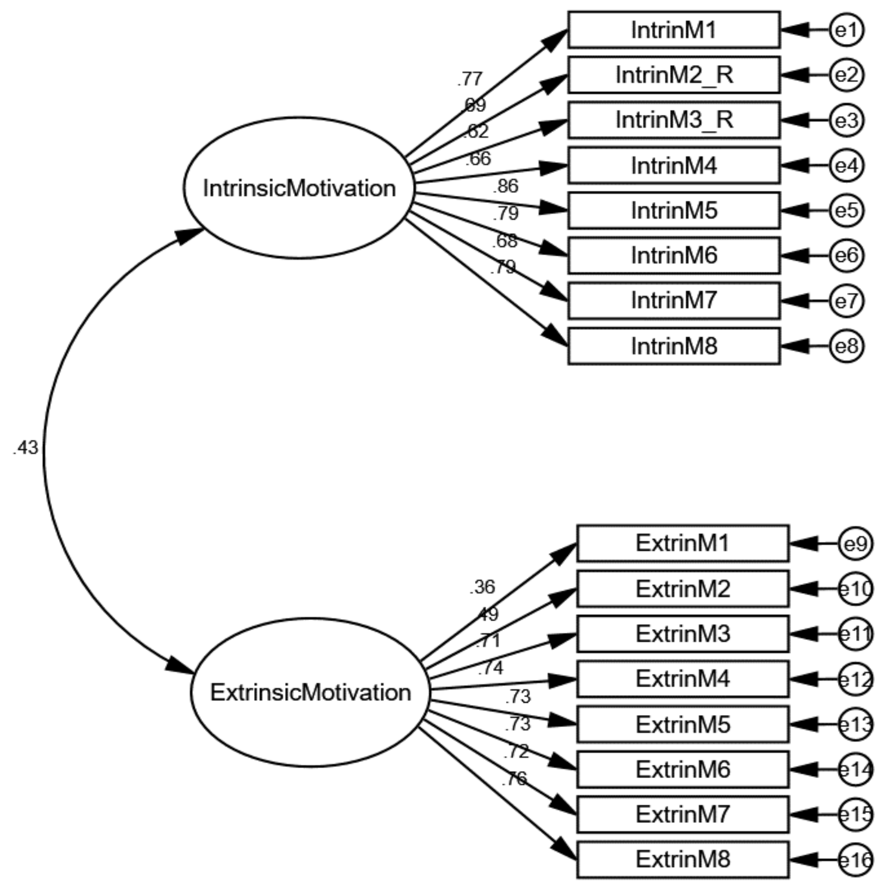

Figure 7.1 Correlated factor model of motivation

Table 7.2

Factor loadings of the correlated factor model of motivation

Corelated Factors' Loading	Correlated Factors	Items	Factor Loadings
0.429***	Intrinsic Motivation	IntrinM1	.769***
		IntrinM2_R	.695***
		IntrinM3_R	.622***
		IntrinM4	.664***
		IntrinM5	.862***
		IntrinM6	.788***
		IntrinM7	.676***
		IntrinM8	.791***
	Extrinsic Motivation	ExtrinM1	.364***
		ExtrinM2	.489***
		ExtrinM3	.707***
		ExtrinM4	.739***
		ExtrinM5	.734***
		ExtrinM6	.733***
		ExtrinM7	.718***
		ExtrinM8	.762***

*** $p < 0.001$

7.2.2 Self-efficacy

The scale of self-efficacy of mathematics students who participated in this study consists of eight items. The study examined only a one-factor model or single factor model or Model 1 for Self-Efficacy, and the study assumed scale to have a single-factor structure.

The model fit indices of Model 1 are closer to the acceptable fit indices than other alternative models (**Table 7.3**). Then, the one error covariance between item 6 and 7 according to the modification indices modified Model 1. As a result, the entire model fit indices of the last model (Model 2) or the single factor with one error covariance reach the acceptable range.

Table 7.3

Fit indices for CFA models in self-efficacy in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	553.478	20	27.674	0.902	0.863	0.201
2. Single Factor Model with error correlated	185.947	19	9.787	0.969	0.955	0.115

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

According to the fit indices (**Table 7.3**), Model 2 significantly improved over Model 1. Although that model had a higher RMSEA value (0.115), there are a higher CFI value (0.969), and a higher TLI value (0.955) and a significant change in chi-square (χ^2 = 185.947, $p < 0.05$) given the change in degrees of freedom when compared to Model 1 (χ^2 = 553.478, $p < 0.05$). Model 2 shows best fit for the data analysed (**Figure 7.2**). **Table 7.4** shows standardized factor loadings for all models. The results confirm that factor-loading values of all items in this construct are above the cut-off value (0.30), referring to all of the items measured in the latent construct. **Figure 7.2** and **Table 7.4** show the final structure and factor loadings of the model.

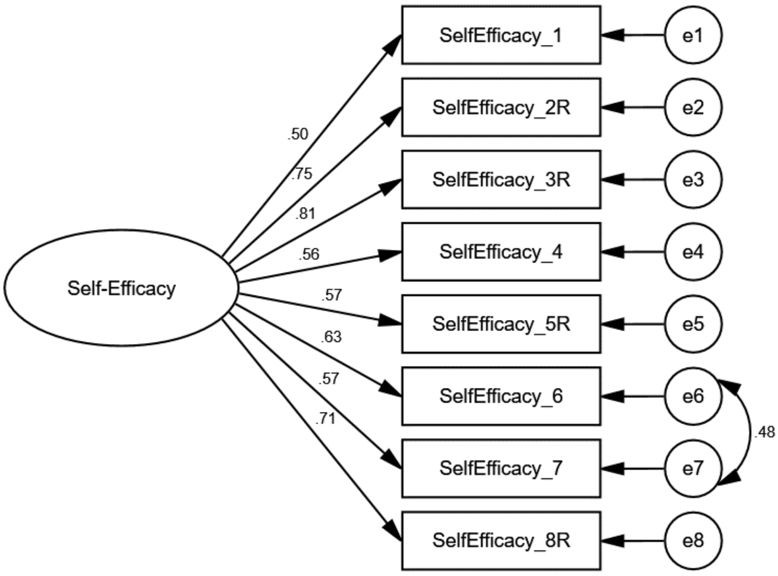

Figure 7.2 Single factor model of *self-efficacy towards mathematics*

Table 7.4

Factor loadings of the single factor model of self-efficacy towards mathematics

First-order Factor	Items	Loading (Standardised Regression Weights)
Self-Efficacy	Self-Efficacy_1	.501***
	Self-Efficacy_2R	.730***
	Self-Efficacy_3R	.800***
	Self-Efficacy_4	.574***
	Self-Efficacy_5R	.558***
	Self-Efficacy_6	.692***
	Self-Efficacy_7	.641***
	Self-Efficacy_8R	.686***

*** $p < 0.001$

7.2.3 Attitude towards mathematics learning

There are seven items in the scale of attitude towards mathematics learning included in student questionnaire. This study examined only one-factor model or single factor model or Model 1 for attitude towards Mathematics Learning. This is because the study assumed the scales to have a single-factor structure and the model fit indices of Model 1 are closer to the acceptable fit indices than other alternative models (**Table 7.5**). According to the modification indices, this study used the one error covariance between the item 4 and 8 to modify Model 1. As a result, Model 2 had a lower RMSEA value (0.099), a higher CFI value (0.973), and a higher TLI value (0.956) and a significant change in chi-square (χ^2 = 97.381, p < 0.05) given the change in degrees of freedom when compared to Model 1 (χ^2 = 127.27, p < 0.05).

Therefore, the modification indices indicate another error covariance between the item 1 and 2, and Model 3 showed a significant improvement over other models according to the fit indices (χ^2 = 74.237, p < 0.05, CFI = 0.980; TLI= 0.968; RMSEA = 0.089). Hence, the last model (Model 3) or the single factor with one error covariance reaches the acceptable range.

Table 7.5

Fit indices for CFA models in attitude towards mathematics learning in overall sample

Model	CMIN (χ^2)	df	CMIN (χ^2)/df	CFI	TLI	RMSEA
1. Single Factor Model	124.27	14	8.876	0.964	0.946	0.109
2. Single Factor Model with correlated error 4 and 8	97.381	13	7.491	0.973	0.956	0.099
3. Single Factor Model with correlated error 4 and 8, and 1 and 2	74.237	12	6.186	0.980	0.965	0.089

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

The results of this analysis confirm the factor loading values of all items in this construct to be above the cut-off value (0.30), referring that these seven items measured the targeted latent construct. **Figure 7.3** and **Table 7.6** show the final structure and factor loadings of the model.

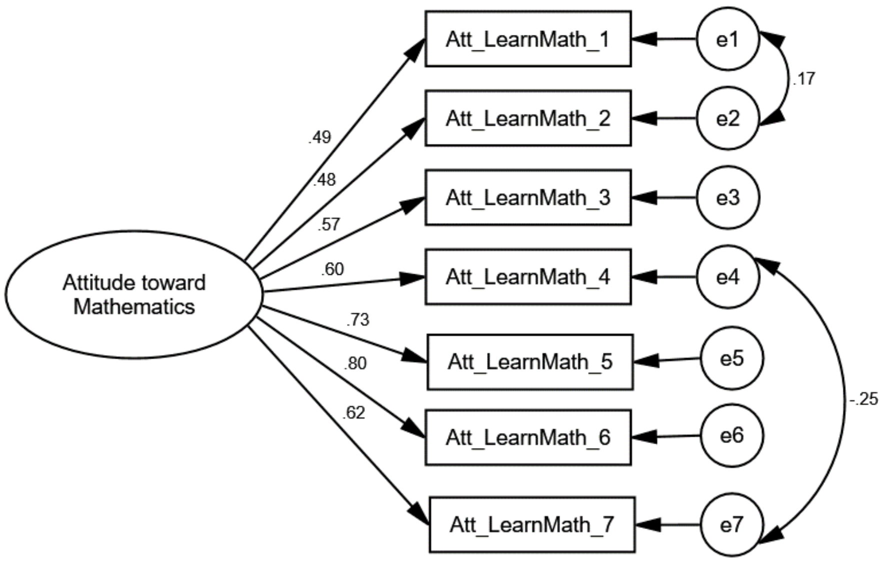

Figure 7.3 Single factor model of attitude towards Mathematics

Table 7.6

Factor loadings of the single factor model of attitude towards mathematics

First-order Factor	Items	Loading (Standardised Regression Weights)
Attitude towards Mathematics	Att_LearnMath_1	.489***
	Att_LearnMath_2	.481***
	Att_LearnMath_3	.569***
	Att_LearnMath_4	.600***
	Att_LearnMath_5	.726***
	Att_LearnMath_6	.800***
	Att_LearnMath_7	.624***

*** $p < 0.001$

7.2.4 Attitude towards formative assessment

Attitude towards formative assessment scale assesses the level of attitude towards formative assessment by students and consists of nine items. An examination of the model fit indices of the first four models indicated that the two-correlated factors model (Model 3) and hierarchical model (Model 4) exhibited better model fit when compared to Model 1 and Model 2 (**Table 7.7**). Although the two-correlated factors and hierarchical models provided similar results in their model fit indices, the two-correlated factors model was preferred and subsequently used for the analysis. In addition, the principles of parsimony in the CFA and SEM practices (Ho, 2008; Kline, 2011) supported present analysis. The use of the correlated model for this scale enables more comparison with the multidimensional Rasch Analysis.

Further, this study modified Model 3 four times by four error-covariance between items 4 and 5; items 3 and 4; items 1 and 3; and items 6 and 7 according to the modification indices (see **Table 7.7**). As a result, all the models fit indices of the last model (Model 8) or the two-correlated factors model with four error-covariance reaches the acceptable range. According to the fit indices, model 8 improves significantly over other models (**Table 7.8**). Model 8 has a lower RMSEA value (0.081), a higher CFI value (0.987), and a higher TLI value (0.979), and a significant change in chi-square (χ^2 = 116.732, $p < 0.05$) given the change in degrees of freedom (**Table 7.7**). The study assumed Model 8 as the best-fit model for the dataset from these results.

Table 7.7

Fit indices for CFA models in attitude towards formative assessment in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	674.353	27	24.976	0.912	0.882	0.191
2. Orthogonal Model	1918.265	27	71.047	0.742	0.656	0.326
3. Correlated Factor Model	404.834	26	15.571	0.948	0.929	0.149
4. Hierarchical Model	404.834	26	15.571	0.948	0.929	0.149
5. Correlated Factor Model with error correlated of item 4 and 5	311.529	25	12.461	0.961	0.944	0.132
6. Correlated Factor Model with error correlated of item 4 and 5, and item 3 and 4	199.810	24	8.325	0.976	0.964	0.105
7. Correlated Factor Model with error correlated of item 4 and 5, and item 3 and 4, and item 1 and item 3	171.547	23	7.459	0.980	0.968	0.099
8. Correlated Factor Model with error correlated of item 4 and 5, and item 3 and 4, and item 1 and item 3, item 6, and 7	116.732	22	5.306	0.987	0.979	0.081

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.8 displays standardized factor loadings for all models. The results of this analysis confirm that factor-loading values of all the items in this construct are above the cut-off value (0.30), referring that all the items measured the latent construct. **Figure 7.4** and **Table 7.8** show the final structure and factor loadings of the model. Further examination of the correlation of the two-correlated factors model indicated a strong correlation (r = 0.681) between the two factors.

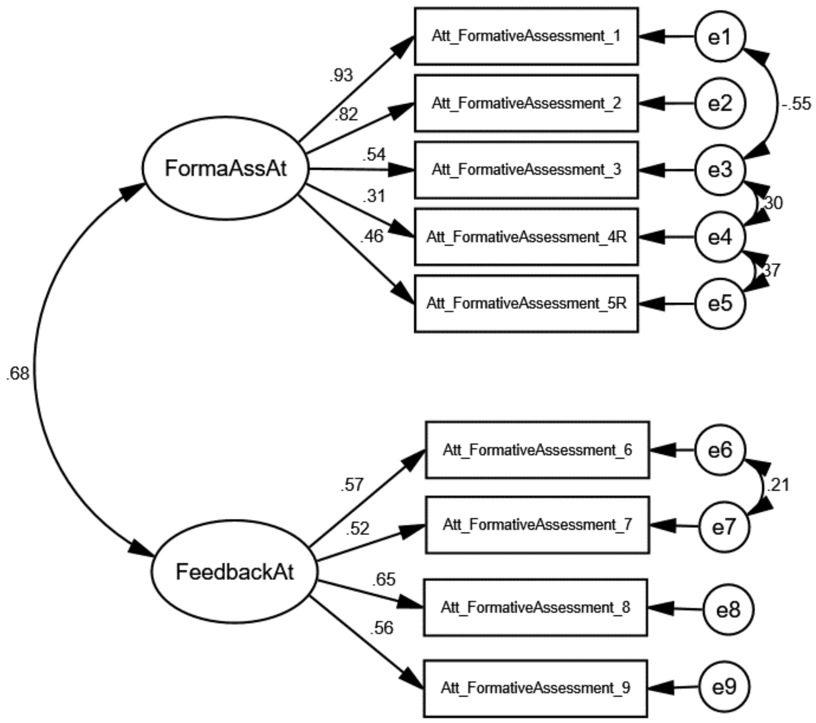

Figure 7.4 Correlated factor model of attitude towards formative assessment

Table 7.8

Factor loadings of correlated factor model of attitude towards formative assessment

Corelated Factors' Loading	Correlated Factors	Items	Factor Loadings
0.681	Attitude to Formative test	Att_FormativeAssessment_1	.927***
		Att_FormativeAssessment_2	.816***
		Att_FormativeAssessment_3	.544***
		Att_FormativeAssessment_4R	.311***
		Att_FormativeAssessment_5R	.459***
	Attitude to Formative Feedback	Att_FormativeAssessment_6	.567***
		Att_FormativeAssessment_7	.522***
		Att_FormativeAssessment_8	.650***
		Att_FormativeAssessment_9	.559***

*** $p < 0.001$

7.2.5 Attitude towards PPT

Attitude towards PPT scale assesses the level of attitude towards PPT from students and consists of eleven items. This scale consists of two sub-scales that are PPT and its feedback. The first sub-scale has six items, and another sub-scale has five items. Table 7.9 examines the model fit indices of the first four models. These results indicated two-correlated factors model

(Model 3) and hierarchical model (Model 4) exhibited better model fit when compared to the other two models—Model 1 and Model 4. Although the two-correlated factors and hierarchical models provided similar results in their model fit indices, the two-correlated factors model was preferred and subsequently used for analysis. In addition, the principles of parsimony in the CFA and SEM practices (Ho, 2008; Kline, 2011) supported this analysis. The use of the correlated model for this scale enables more comparison with the multidimensional Rasch Analysis.

According to the modification indices, the study modified Model 3 three times by three error-covariance between items 2 and 3, items 7 and 8, and items 9 and 11. As a result, all fit indices of the last model (Model 7) or the two-correlated factors model with three error-covariance reaches the acceptable range. According to the fit indices (**Table 7.9**), Model 7 improves significantly over other six models and has a higher RMSEA value (0.107), a higher CFI value (0.990), a higher TLI value (0.986), and a significant change in chi-square (χ^2 = 342.316, $p < 0.05$) given the change in degrees of freedom. Therefore, the study selected Model 7 as the best fit for the data (**Figure 7.5**) from these results.

Table 7.9

Fit indices for CFA models in attitude towards PPT in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	1402.569	44	31.877	0.955	0.944	0.216
2. Orthogonal Model	16779.66	44	381.356	0.445	0.306	0.76
3. Correlated Factor Model	1306.169	43	30.376	0.958	0.946	0.211
4. Hierarchical Model	1306.169	43	30.376	0.958	0.946	0.211
5. Correlated Factor Model with error correlated of item 2 and 3	563.318	42	13.412	0.983	0.977	0.137
6. Correlated Factor Model with error correlated of item 2 and 3, and item 7 and 8	385.866	41	9.411	0.989	0.985	0.113
7. Correlated Factor Model with error correlated of item 2 and 3, and item 7 and 8, and item 9 and item 11	342.316	40	8.558	0.990	0.986	0.107

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.10 displays standardized factor loadings for all models. This analysis confirms that factor loading values of all items in this construct are above the cut-off value (0.30), referring that all the items measured the latent construct. **Figure 7.5** and **Table 7.10** show the final structure and factor loadings of the model. However, further examination of the correlation of the two-correlated factors model indicated high correlation (r = 0.958) between two factors.

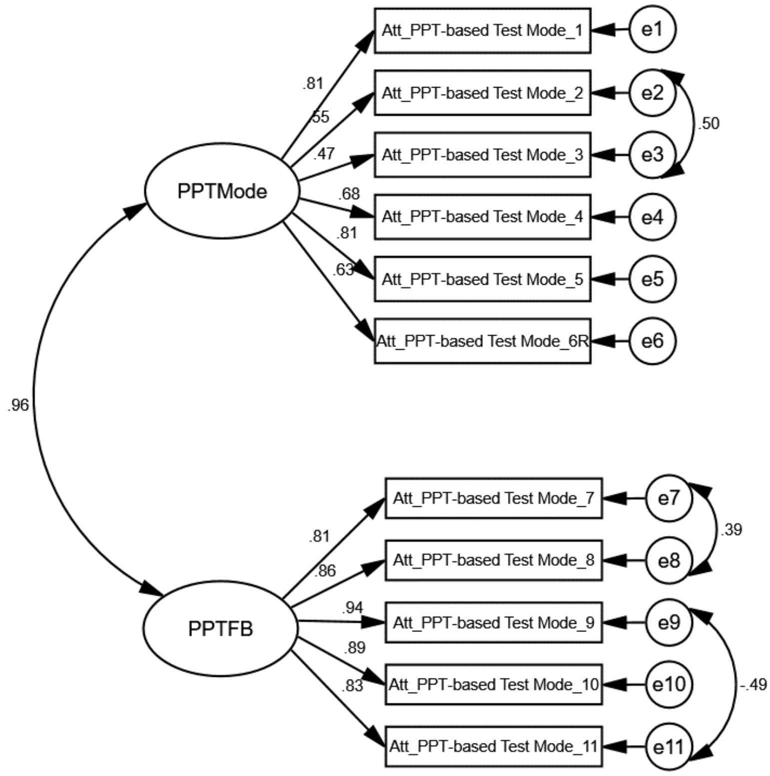

Figure 7.5 Correlated factor model of attitude towards PPT

Table 7.10

Factor loadings of the correlated factor model of attitude towards PPT

Corelated Factors' Loading	Correlated Factors	Items	Factor Loadings
0.958***	Attitude to PPT	Att_PPTTestMode_1	.814***
		Att_PPTTestMode_2	.554***
		Att_PPTTestMode_3	.474***
		Att_PPTTestMode_4	.683***
		Att_PPTTestMode_5	.813***
		Att_PPTTestMode_6R	.634***
	Attitude to Feedback from PPT	Att_PPTTestMode_7	.807***
		Att_PPTTestMode_8	.865***
		Att_PPTTestMode_9	.938***
		Att_PPTTestMode_10	.889***
		Att_PPTTestMode_11	.830***

*** $p < 0.001$

7.2.6 Attitude towards LOT

The scale of attitude towards LOT assesses the level of attitude towards LOT from students consists of eleven items. This scale consists of two sub-scales, which are CAT and its feedback. The first sub-scale has six items, and another sub-scale has five items. An examination of the model fit indices of the first four models as shown in **Table 7.11** indicated that the two-correlated factors model (Model 3) and hierarchical model (Model 4) exhibited better model fit when compared to the other two models ((Model 1) and (Model 4)). Although the two-correlated factors and hierarchical models provided similar results in their model fit indices, the two-correlated factors model was preferred and used for subsequent analysis. In addition, the principles of parsimony in the CFA and SEM practices (Ho, 2008; Kline, 2011) supported the result of this analysis. The use of the correlated model for this scale enables more comparison with the multidimensional Rasch Analysis.

The study modified Model 3 three times by four error-covariance between items 3 and 4, items 2 and 3, items 9 and 11, and items 7 and 9, according to the modification indices. As a result, the entire model fit indices of the last model (Model 8) or the two-correlated factors model with three error-covariance reaches the acceptable range. According to the fit indices (**Table 7.11**), Model 8 improve significantly over other 7 models and model 8 has a higher RMSEA value (0.171), higher CFI value (0.986), higher TLI value (0.980) and a significant change in chi-square (χ^2 = 794.254, $p < 0.05$) given the change in degrees of freedom. From these results, this study selected Model 8 as the best fit for the data (**Figure 7.6**).

Table 7.11

Fit indices for CFA models in attitude towards LOT in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	1392.116	44	31.639	0.975	0.969	0.216
2. Orthogonal Model	3037.754	44	69.040	0.434	0.292	1.023
3. Correlated Factor Model	1344.617	43	31.270	0.976	0.969	0.214
4. Hierarchical Model	1344.617	43	31.270	0.976	0.969	0.214
5. Correlated Factor Model with error correlated of item 3 and 4	951.101	42	22.645	0.983	0.978	0.181
6. Correlated Factor Model with error correlated of item 3 and 4, and item 2 and 3	860.897	41	20.998	0.985	0.979	0.174
7. Correlated Factor Model with error correlated of item 3 and 4, and item 2 and 3, and item 9 and item 11	846.284	40	21.157	0.985	0.979	0.175
8. Correlated Factor Model with error correlated of item 3 and 4, and item 2 and 3, item 9 and item 11, and item 7 and 9	794.254	39	20.366	0.986	0.980	0.171

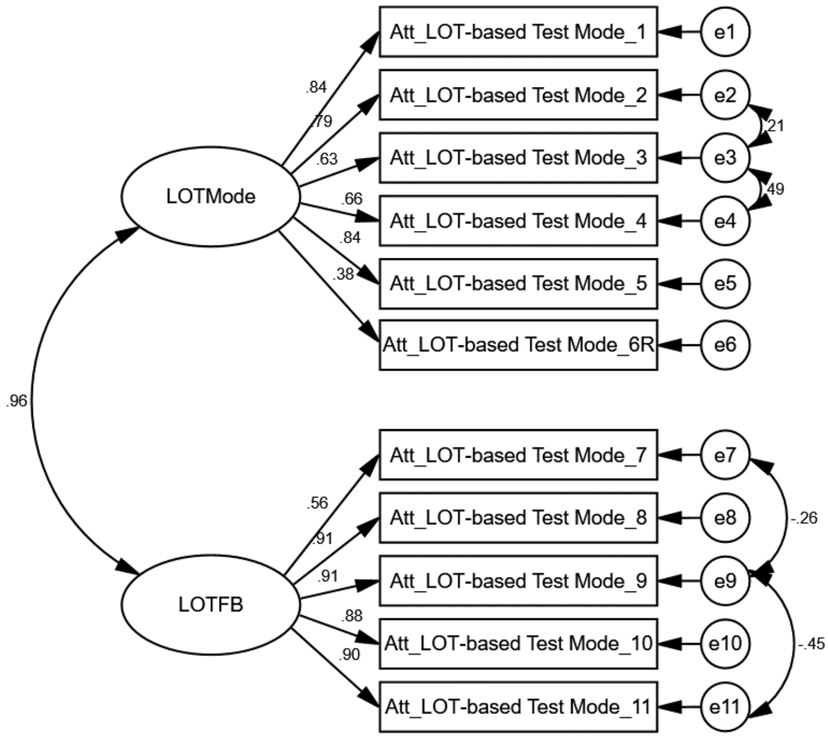

Figure 7.6 Correlated factor model of attitude towards LOT

Table 7.12

Factor loadings of correlated factor model of attitude towards LOT

Corelated Factors' Loading	Correlated Factors	Items	Factor Loadings
0.961***	Attitude to LOT	Att_LOTTestMode_1	.842***
		Att_LOTTestMode_2	.789***
		Att_LOTTestMode_3	.625***
		Att_LOTTestMode_4	.662***
		Att_LOTTestMode_5	.840***
		Att_LOTTestMode_6R	.559***
	Attitude to Feedback from LOT	Att_LOTTestMode_7	.909***
		Att_LOTTestMode_8	.908***
		Att_LOTTestMode_9	.883***
		Att_LOTTestMode_10	.900***
		Att_LOTTestMode_11	.379***

*** $p < 0.001$

7.2.7 Attitude towards CAT

Attitude towards CAT scale assesses the level of attitude towards such test mode by students and consists of eleven items. This scale consists of two sub-scales that are CAT and its feedback. The first sub-scale has six items, and another sub-scale has five items. An examined model fit indices of the first four models indicated that the two-correlated factors model (Model 3) and hierarchical model (Model 4) exhibited better model fit when compared to the other two models—Model 1 and Model 4 (**Table 7.13**). Although the two-correlated factors and hierarchical models provided similar results in their model fit indices, the two-correlated factors model was preferred and used for subsequent analysis. In addition, the principles of parsimony in the CFA and SEM practices (Ho, 2008; Kline, 2011) supported the results of this analysis. The use of the correlated model for this scale enables this study to compare more with the Rasch Analysis.

According to the modification indices, model 3 was modified by the error-covariance between items 9 and 10. As a result, the entire model fit indices of the last model (Model 5) or the two-correlated factors model with the error covariance reach the acceptable range. According to the fit indices (**Table 7.13**), Model 5 improved significantly over other models and has a lower RMSEA value (0.055), a higher CFI value (0.989), and a higher TLI value (0.985), and a significant change in chi-square (χ^2 = 125.153, p < 0.05) given the change in degrees of freedom. Therefore, the best fit for the data in these results is Model 5 (**Figure 7.7**).

Table 7.13

Fit indices for CFA models in attitude towards CAT in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	206.315	44	4.689	0.978	0.972	0.075
2. Orthogonal Model	4462.044	44	101.410	0.399	0.248	0.390
3. Correlated Factor Model	174.008	43	4.047	0.982	0.977	0.068
4. Hierarchical Model	174.008	43	4.047	0.982	0.977	0.068
5. Correlated Factor Model with error correlated of item 9 and 10	125.153	42	2.980	0.989	0.985	0.055

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.14 displays standardized factor loadings for all models. This analysis confirms that factor loading values of all items in this construct are above the cut-off value (0.30), referring that the entire items measured the latent construct. **Figure 7.7** and **Table 7.14** show the final structure and factor loadings of the model. Further examination of the correlation of the two-correlated factors model indicated a high correlation (r = 0.936) between the two factors.

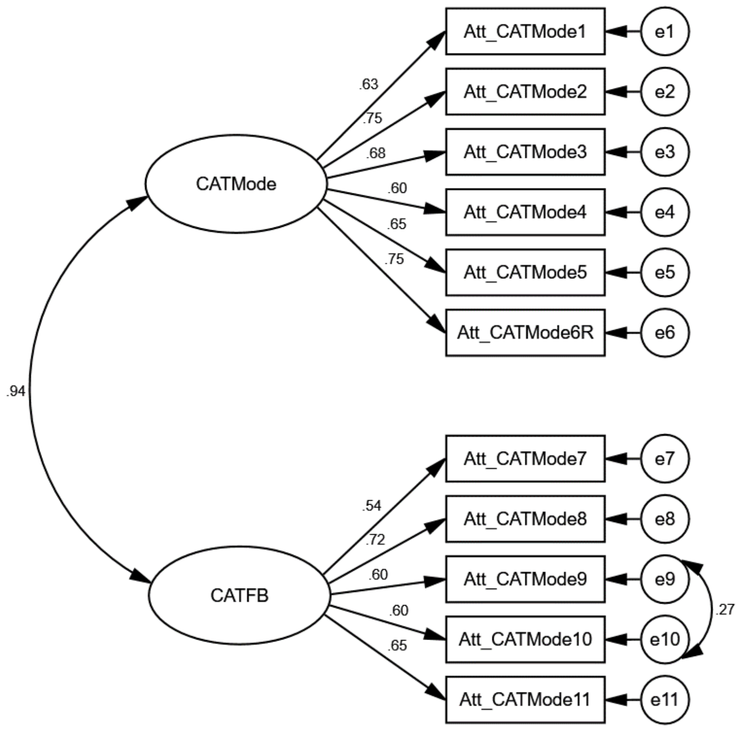

Figure 7.7 Correlated factor model of attitude towards CAT

Table 7.14

Factor loadings of the correlated factor model of attitude towards CAT

Corelated Factors' Loading	Correlated Factors	Items	Factor Loadings
0.936***	Attitude to CAT	Att_CATTestMode_1	.629***
		Att_CATTestMode_2	.748***
		Att_CATTestMode_3	.680***
		Att_CATTestMode_4	.600***
		Att_CATTestMode_5	.653***
		Att_CATTestMode_6R	.748***
	Attitude to Feedback from CAT	Att_CATTestMode_7	.543***
		Att_CATTestMode_8	.724***
		Att_CATTestMode_9	.602***
		Att_CATTestMode_10	.597***
		Att_CATTestMode_11	.651***

*** $p < 0.001$

7.2.8 ICT Familiarity

This study adapted the ICT Familiarity of students in the student questionnaire from the TIMSS questionnaire's ICT Familiarity. This scale assesses the level of ICT Familiarity by students and consists of two sub-constructs: ICT Access and ICT Usage. The first four items measured ICT Access, and the rest eight items measured ICT Usage.

An examination of the model fit indices of the first four models, as shown in **Table 7.15**, indicated that the two-correlated factors model (Model 3) and hierarchical model (Model 4) exhibited better model fit when compared to the other two models ((Model 1) and (Model 2)). Although the two-correlated factors and hierarchical models provided similar results in their model fit indices, the two-correlated factor model was preferred and used for subsequent analysis. In addition, the principles of parsimony in the CFA and SEM practices (Ho, 2008; Kline, 2011) supported this analysis. The use of the correlated model for this scale enables this study to compare more with the Rasch Analysis.

According to the modification indices, this study modified model 3 four times by an error-covariance between the items 11 and 12. As a result, all the model fit indices of the last model (Model 5) or the two-correlated factors model with an error covariance reach the acceptable range. According to the fit indices (**Table 7.15**), Model 5 improved significantly over the other four models and has a lower RMSEA value (0.054), a higher CFI value (0.995), a higher TLI value (0.993), and a significant change in chi-square (χ^2 = 153.393, $p < 0.05$)

given the change in degrees of freedom. This analysis defined Model 5 as the best fit for the data (**Figure 7.8**).

Table 7.15

Fit indices for CFA models in ICT familiarity in overall sample

	Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1.	Single Factor Model	1090.154	54	20.188	0.946	0.934	0.171
2.	Orthogonal Model	1533.368	54	28.396	0.923	0.906	0.204
3.	Correlated Factor Model	1007.450	53	19.009	0.950	0.938	0.165
4.	Hierarchical Model						
5.	Correlated Factor	1007.450	53	19.009	0.950	0.938	0.165
	Model	153.393	52	2.950	0.995	0.993	0.054
6.	with error variance						

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.16 displays standardized factor loadings for all models. This analysis confirms factor loading values of all items in this construct to be above the cut-off value (0.30), referring that all the items measured the latent construct. **Figure 7.8** and **Table 7.16** show the final structure and factor loadings. Further examination of the correlation factor loading between ICT access and ICT usage indicated a strong correlation value (r = 0.686).

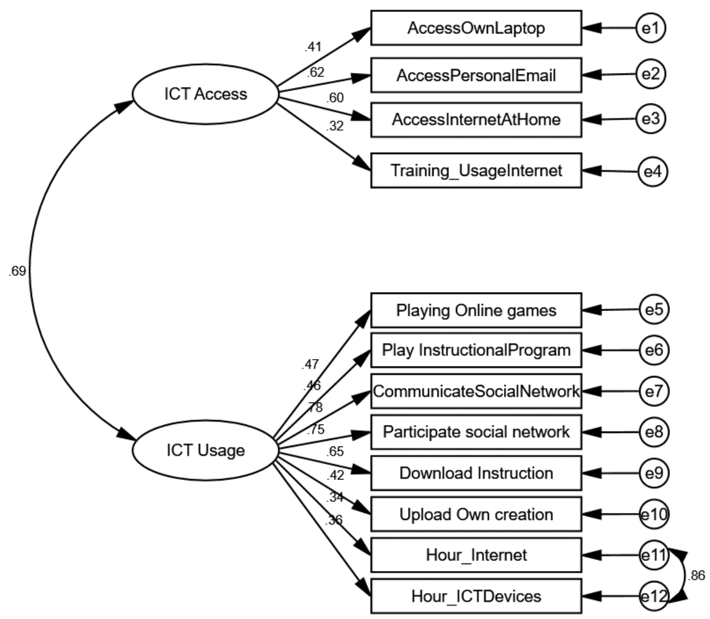

Figure 7.8 Correlated factor model of ICT Familiarity

Table 7.16

Factor loadings of correlated factor model of ICT Familiarity

Corelated Factors' Loading	Correlated Factors	Items	Factor Loadings
0.686***	ICT Access	Access Own Laptop	.408***
		Access Personal Email	.618***
		Access Internet at Home	.602***
		Training Usage Internet	.321***
	ICT Usage	Playing Online games	.475***
		Play Instructional Program	.457***
		Communication Social Network	.777***
		Participate Social Network	.753***
		Download Instruction	.650***
		Upload Own Creation	.419***
		Hour_internet	.338***
		Hour_ICTDevices	.361***

*** $p < 0.001$

7.2.9 Attitude towards ICT

Attitude towards the use of ICT devices assesses the level of attitude towards the use of a computer by students and consists of sixteen items. There are three sub-scale in this scale: Affective, Behavioural, and Cognition scales of a ICT attitude scale. Six items measured an affective scale of computer attitude scale, from Att_ComUse_1 to Att_ComUse_6. Six items measured a cognition scale of computer attitude scale, from Att_ComUse_7 to Att_ComUse_10 and six items measured the behaviour scale of computer attitude scale, from Att_ComUse_1 to Att_ComUse_16.

An examination of the model fit indices of the first four models, as shown in **Table 7.17**, indicated that the three-correlated factors model (Model 3) and hierarchical model (Model 4) exhibited better model fit than the other two models—Model 1 and Model 2. Although the three-correlated factors and hierarchical models provided similar results in their model fit indices, the hierarchical models provided similar results in their model fit indices, and the hierarchical model was preferred and used for subsequent analysis. In addition, the principles of parsimony in the CFA and SEM practices (Ho, 2008; Kline, 2011) supported the analysis of these results. The use of the correlated model for this scale enables comparison with the Rasch Analysis.

According to the modification indices, this study modified model 3 by the nine error-covariance between items 1 and 6, items 2 and 3, items 4 and 5, items 5 and 6, items 7 and 9, items 9 and 10, items 12 and 13, items 13 and 14, and items 15 and 16. As a result, the entire model fit indices of the last model (Model 5) or the three-correlated factors model with the nine error-covariance reach the acceptable range. According to the fit indices (**Table 7.17**), Model 5 improved significantly over the other four models and has a higher RMSEA value (0.100), a higher CFI value (0.936), a higher TLI value (0.916), and a significant change in chi-square (χ^2 = 702.927, $p < 0.05$) given the change a degree of freedom. These results defined Model 5 as the best fit for the data (**Figure 7.9**).

Table 7.17

Fit indices for CFA models in attitude towards ICT in overall sample

Model	CMIN (χ^2)	df	CMIN (χ^2)/df	CFI	TLI	RMSEA
1. Single Factor Model	2104.199	104	20.233	0.790	0.757	0.171
2. Orthogonal Model	4225.634	104	40.631	0.567	0.500	0.245
3. Correlated Factor Model	1132.548	101	11.213	0.892	0.871	0.124
4. Hierarchical Model	1604.156	101	15.727	0.842	0.814	0.149
5. Correlated Factor Model with error covariance	702.927	92	7.641	0.936	0.916	0.100

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.18 displays standardized factor loadings for all models. This analysis confirms that the factor loading values of all items in this construct to be above the cut-off value (0.30), referring that all the items measured the latent construct. **Figure 7.9** and **Table 7.18** show the final structure and factor loadings of the model. Further examination of the correlations among the Affective, Cognitive and Behavioural factors found that there were strong correlations between Affective and Cognitive factors (r = 0.705), and between Affective and Behavioural factors (r= 0.583) **(**see **Table 7.19**). However, there is only a medium correlation between Cognitive and Behavioural factors (r=0.370) (see **Table 7.19**). It is also found that the correlation of Cognitive factor towards Affective factor is higher than its correlation with the Behavioural factor. This may be because participant students responded to the observed variables (i.e., items) from Cognitive and Affective factors are in a similar way.

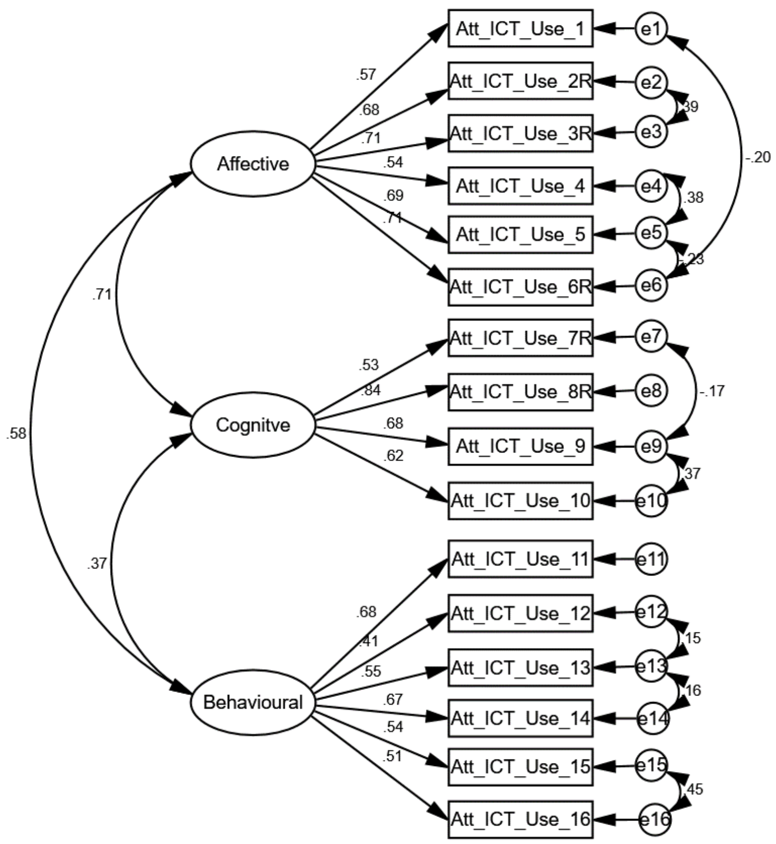

Figure 7.9 Correlated factor model of attitude towards ICT

Table 7.18

Factor loadings of correlated factor model of attitude towards ICT

Correlated Factors	Items	Factor Loadings
Affective	Att_ICT1	0.568***
	Att_ICT2	0.676***
	Att_ICT3	0.709***
	Att_ICT4	0.539***
	Att_ICT5	0.687***
	Att_ICT6	0.707***
Cognitive	Att_ICT7	0.527***
	Att_ICT8	0.840***
	Att_ICT9	0.677***
	Att_ICT10	0.625***
Behavioural	Att_ICT11	0.684***
	Att_ICT12	0.413***
	Att_ICT13	0.553***
	Att_ICT14	0.668***
	Att_ICT15	0.539***
	Att_ICT16	0.515***

*** $p < 0.001$

Table 7.19

Correlated Factor Loadings of the factor model of attitude towards ICT

	Correlated Factor Loading
Factor Loading Between Affective and Cognitive Factors	0.705***
Factor Loading Between Cognitive and Behavioural Factors	0.370***
Factor Loading Between Affective and Behavioural Factors	0.583***

*** $p < 0.001$

7.2.10 Reliability and validity of scales of student questionnaire

As shown in **Table 7.20**, Cronbach's alpha values for all scales are above 0.7. This value indicates that all scales are reliable and will produce similar results if utilized for similar populations over time. Moreover, average variance extracted (AVE) values for all scales are described in the fourth column of **Table 7.20**. The AVE values for some scales are higher than 0.50, indicating high convergent validity. However, the rest scales have an AVE of less than 0.5 and greater than 0.3, indicating adequate convergent validity. Finally, the composite or construct reliability results are presented in the fifth column of **Table 7.20.** As mentioned in Chapter 4, the construct reliability value higher than 0.70 indicate that all measured scales or items consistently belong to a single construct (Hair et. al., 2010).

Table 7.20

Reliability and convergent validity of the scales of student questionnaire

Scales from student questionnaire	No of Items	Cronbach's Alpha	Average Variance Extracted (AVE)	Construct Reliability
Motivation	16	0.898	0.496	0.972
Self-Efficacy	8	0.853	0.428	0.907
Attitude towards Mathematics	7	0.805	0.387	0.893
Attitude towards Formative Assessment	9	0.803	0.384	0.916
Attitude towards PPT	11	0.933	0.589	0.971
Attitude towards LOT	11	0.929	0.560	0.965
Attitude towards CAT	11	0.886	0.430	0.941
ICT Familiarity	12	0.723	0.389	0.877
Attitude towards ICT	16	0.861	0.395	0.952

7.3 CFA for scales from teacher questionnaire

7.3.1 General practices of formative assessment

General practices of formative assessment are included in teacher questionnaire. This scale assesses the general techniques of formative assessment applied by the teachers and consists of four items. **Table 7.21** shows an examination of fit indices of a single model. This model has a low RMSEA value (0.038), high CFI value (0.995), high TLI value (0.977) and a significant change in chi-square (χ^2 = 0.756, p < 0.05) given the change in degrees of freedom.

Table 7.21

Fit indices for CFA model in general practices of formative assessment in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	0.756	2	0.378	0.995	0.977	0.038

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.22 shows standardized factor loadings for scales. The results confirm that the factor loading values of all items in this construct are above the cut-off value (0.70), referring that all the items represented the latent construct they intended to measure. **Figure 7.10** and **Table 7.22** show the final structure and factor loadings of the model.

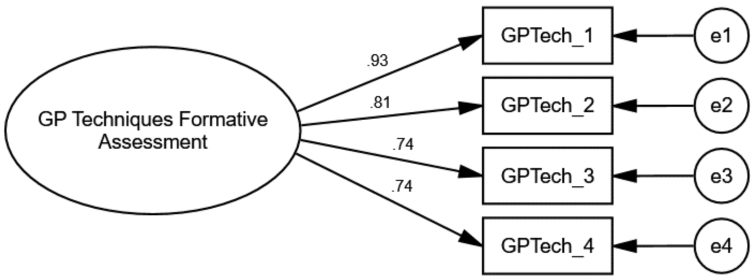

Figure 7.10 Single factor model of general practices of formative assessment

Table 7.22

Factor loadings of single factor model of general practices of formative assessment

First-order Factor	Items	Loading (Standardised Regression Weights)
General practices of formative assessment	GPTech_1	.926***
	GPTech_2	.807***
	GPTech_3	.744***
	GPTech_4	.740***

*** p < 0.001

7.3.2 Specific practices of formative assessment

In this study, the specific formative assessment practices of teachers are included in teacher questionnaire. This scale assesses teachers' specific practices for formative assessment and consists of five items. **Table 7.23** shows an examination of the fit indices of a single model. This model has a low RMSEA value (0.081), high CFI value (0.984), high TLI value (0.968), and a significant change in chi-square (χ^2 = 5.460, p < 0.05) given the change in degrees of freedom.

Table 7.23

Fit indices for CFA model in specific practices of formative assessment in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	5.460	5	1.092	0.984	0.968	0.081

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.24 shows standardized factor loadings for scales. The results confirm that the factor loading values of all items in this construct are above the cut-off value (0.50), referring that all the items represented the latent construct they intended to measure. **Figure 7.11** and **Table 7.24** show the final structure and factor loadings of the model.

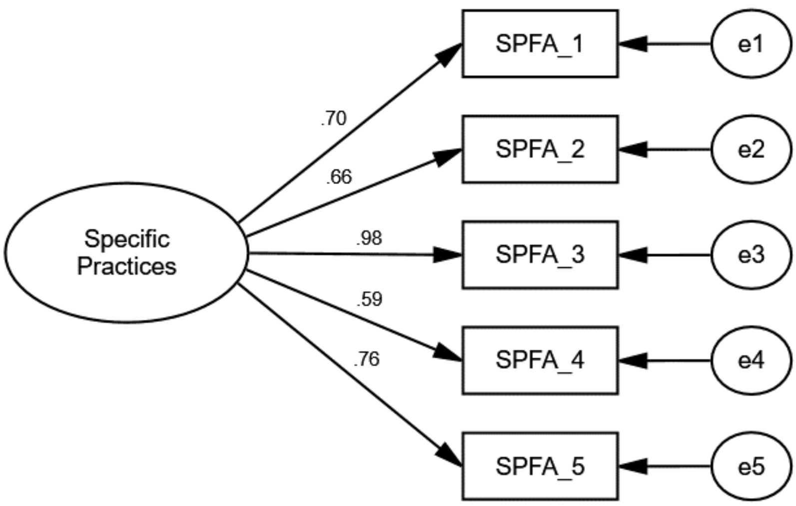

Figure 7.11 Single factor model of specific practices of formative assessment

Table 7.24

Factor loadings of single factor model of specific practices of formative assessment

First-order Factor	Items	Loading (Standardised Regression Weights)
Specific practices formative assessment	SPFA_1	.700***
	SPFA_2	.664***
	SPFA_3	.981***
	SPFA_4	.587***
	SPFA_5	.761***

*** $p < 0.001$

7.3.3 Attitude towards formative assessment

This study included the attitude towards formative assessment in the teacher questionnaire. This scale assesses the level of attitude towards formative assessment and consists of seven items. **Table 7.25** shows an examination of fit indices of the single model. This model has a low RMSEA value (0.224), high CFI value (0.905), high TLI value (0.858), and a significant change in chi-square (χ^2 = 23.818, $p < 0.05$) given the change in degrees of freedom.

Table 7.25

Fit indices for CFA models in attitude towards formative assessment in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	23.818	14	1.701	0.905	0.858	0.224

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.26 shows standardized factor loadings for scales. The results confirm that the factor loading values of all items in this construct are above the cut-off value (0.80), referring that all the items represented the latent construct they intended to measure. **Figure 7.12** and **Table 7.26** show that final structure and factor loadings of the model.

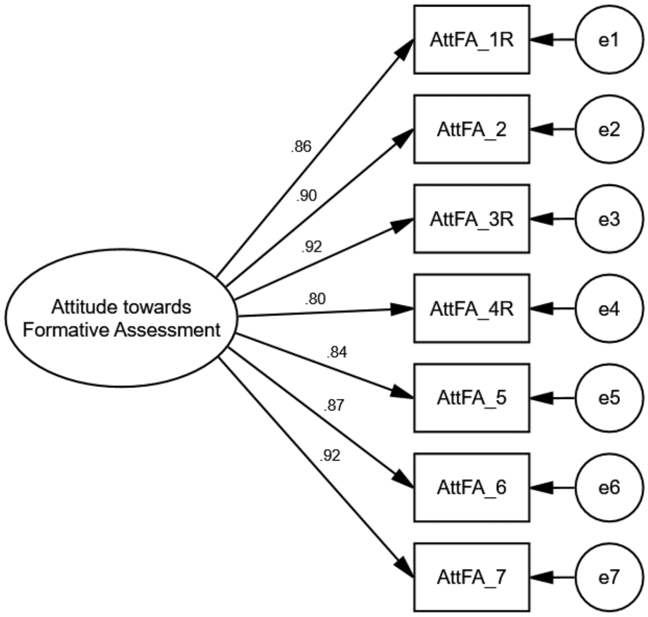

Figure 7.12 Single factor model of attitude towards formative assessment

Table 7.26

Factor loadings of single factor model of attitude towards formative assessment

First-order Factor	Items	Loading (Standardised Regression Weights)
Attitude towards formative assessment	AttFA_1R	.859***
	AttFA_2	.897***
	AttFA_3R	.922***
	AttFA_4R	.804***
	AttFA_5	.841***
	AttFA_6	.872***
	AttFA_7	.924***

*** $p < 0.001$

7.3.4 ICT familiarity

The study adapted teachers' ICT familiarity from the TIMSS questionnaire ICT familiarity. This scale assesses the level of ICT familiarity by teachers. There are seven items on this scale. This study examined only one-factor model or single factor model, or Model 1 for the ICT familiarity. This is because present study assumed these scales to have a single-factor structure, and the model fit indices of Model 1 are not close to the acceptable fit indices (**Table 7.27**). Therefore, according to the modification indices, one error covariance modified Model 1 between items 3 and 4. As a result, the entire model fit indices of model 2 or the single factor with one error covariance reach the acceptable range. According to the fit indices (**Table 7.27**), Model 2 improved significantly over Model 1. Model 2 had a lower RMSEA value

(0.139), a higher CFI value (0.962), a higher TLI value (0.938) and a significant change in chi-square (χ^2 = 16.510, p < 0.05) given the change in degrees of freedom when compared to Model 1 (χ^2 = 44.990, p < 0.05). According to **Figure 7.13,** the study selected Model 2 as the best fit for the data.

Table 7.27

Fit indices for CFA models in ICT familiarity in overall sample

Model	CMIN (χ^2)	df	CMIN (χ^2)/df	CFI	TLI	RMSEA
1. Single Factor Model	44.990	14	3.214	0.660	0.490	0.398
2. Single Factor Model with Items 3 and 4 error correlated	16.510	13	1.270	0.962	0.938	0.139

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.28 shows standardized factor loadings for scales. The results confirm that the factor loading values of all items in this construct are above the cut-off value (0.40), referring that all the items represented the latent construct they intended to measure. **Figure 7.18** and **Table 7.28** show the final structure and factor loadings of the model.

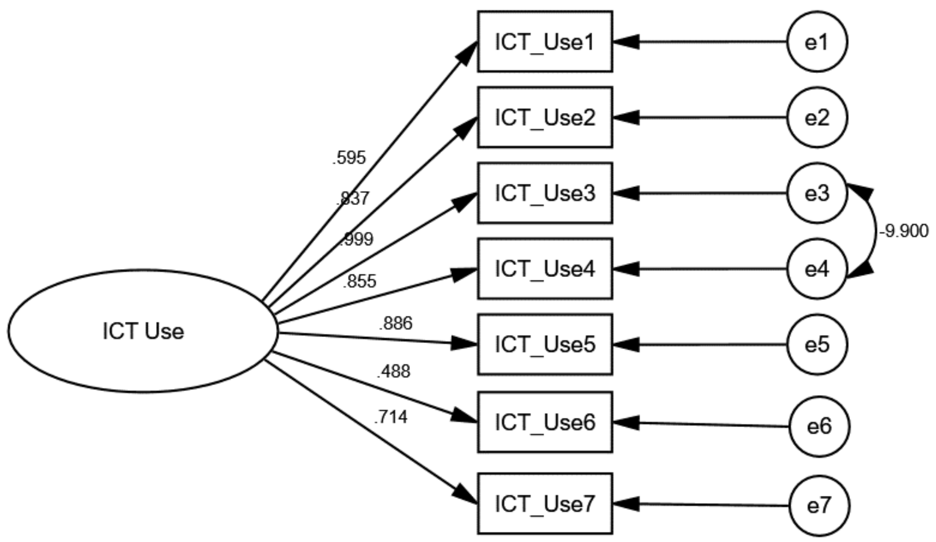

Figure 7.13 Single factor model of ICT familiarity

Table 7.28

Factor Loadings of single factor model of ICT familiarity

First-order Factor	Items	Loading (Standardised Regression Weights)
ICT Familiarity	ICT_Use1	.595***
	ICT_Use2	.837***
	ICT_Use3	.999***
	ICT_Use4	.855***
	ICT_Use5	.886***
	ICT_Use6	.488***
	ICT_Use7	.714***

*** $p < 0.001$

7.3.5 Attitude towards ICT

Attitude towards the use of ICT assesses the level of attitude towards using ICT by teachers and consists of twelve items. There are three sub-scale in this scale: namely, cognition, affective, and behavioural scales of attitude towards ICT. Cognition factor of this scale is measured by three items, AttICT_1, AttICT_2, and AttICT_3; Affective factor of this scale is measured by five items, AttICT_4, AttICT_5, AttICT_6R, AttICT_7R and AttICT_8R; and Behavioural factor of this scale is measured by four items, AttICT_9, AttICT_10, AttICT_11, and AttICT_12.

An examination of the model fit indices of the four models as shown in **Table 7.28** indicated that the three-correlated factors model (Model 4) exhibited better model fit when compared to the three models, i.e., Model 1, Model 2, and Model 3 (**Table 7.29**). Although the two-correlated factors and hierarchical models provided similar results in their model fit indices, the hierarchical factor model was preferred and used for subsequent analysis. In addition, the principles of parsimony in the CFA and SEM practices (Ho, 2008; Kline, 2011) supported the findings of this study. The model has a lower RMSEA value (0.234), a higher CFI value (0.812), and a slightly lower TLI value (0.756), and a significant change in chi-square (χ^2 = 89.929, $p < 0.05$) given the change in degrees of freedom (**Table 7.29**).

Table 7.29

Fit indices for CFA models in attitude towards ICT in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	94.946	54	1.758	0.802	0.758	0.233
2. Orthogonal Model	146.467	54	2.712	0.552	0.453	0.350
3. Correlated Factor Model	91.456	51	1.793	0.807	0.757	0.238
4. Hierarchical Model	89.929	51	1.763	0.812	0.756	0.234

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.30 shows standardized factor loadings for all models. The results of this analysis confirm that the factor loading values of all items in this construct are above the cut-off value (0.80), referring that all the items represented the latent construct they intended to measure. **Figure 7.14** and **Table 7.30** show the final structure and factor loadings of the model.

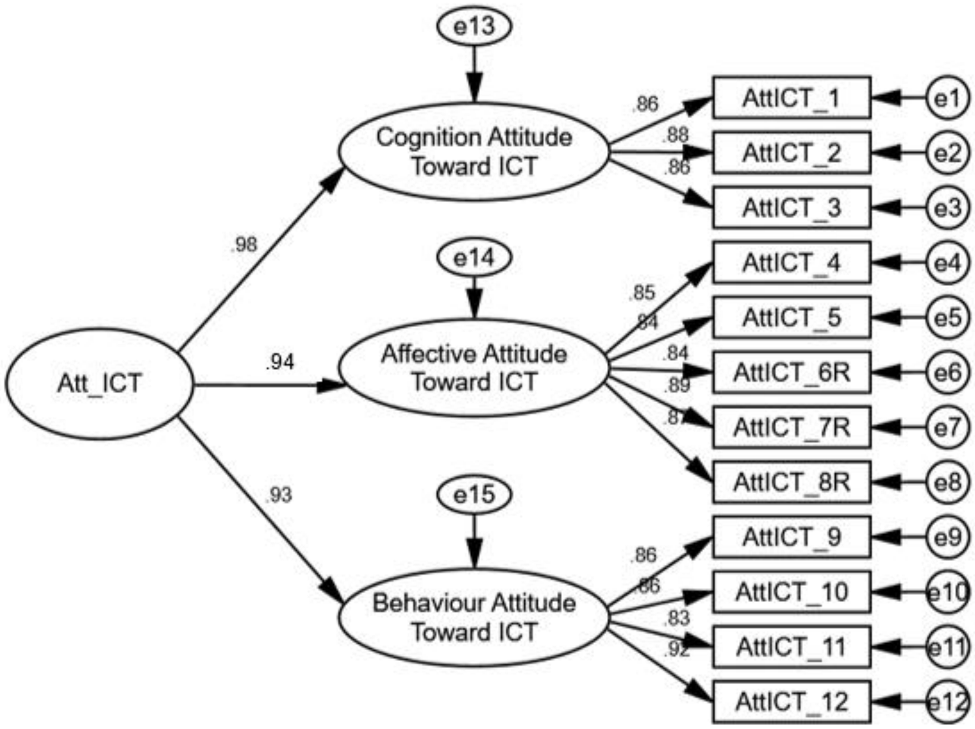

Figure 7.14 Hierarchical factor model of attitude towards ICT

Table 7.30

Factor loadings of hierarchical factor model of attitude towards ICT

Second-order Factor	*First-order Factor*	*Loading*	*Items*	*Loading*
Attitude towards ICT	Cognition Factor	0.978***	AttICT_1	.863***
			AttICT_2	.877***
			AttICT_3	.858***
	Affective Factor	0.938***	AttICT_4	.852***
			AttICT_5	.840***
			AttICT_6R	.840***
			AttICT_7R	.891***
			AttICT_8R	.871***
	Behavioural Factor	0.928***	AttICT_9	.858***
			AttICT_10	.865***
			AttICT_11	.832***
			AttICT_12	.917***

*** $p < 0.001$

7.3.6 Attitude towards PPT

This study included teachers' attitudes towards PPT in the teacher questionnaire. This scale assesses the level of attitude towards PPT and consists of five items. An examination of the model fit indices of the single model, as shown in **Table 7.31**, indicated that the single factor model exhibited the best model. This model has a low RMSEA value (0.070) and has a high CFI value (0.994), and a high TLI value (0.987), and a significant change in chi-square (χ^2 = 5.347, p < 0.05) with the degrees of freedom.

Table 7.31

Fit indices for CFA model in attitude towards PPT in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	5.347	5	1.069	0.994	0.987	0.070

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.32 shows standardized factor loadings for scales. The results of this analysis confirm that the factor loading values of all items in this construct are above the cut-off value (0.70), referring that all the items represented the latent construct they intended to measure. **Figure 7.15** and **Table 7.32** show the final structure and factor loadings of the model.

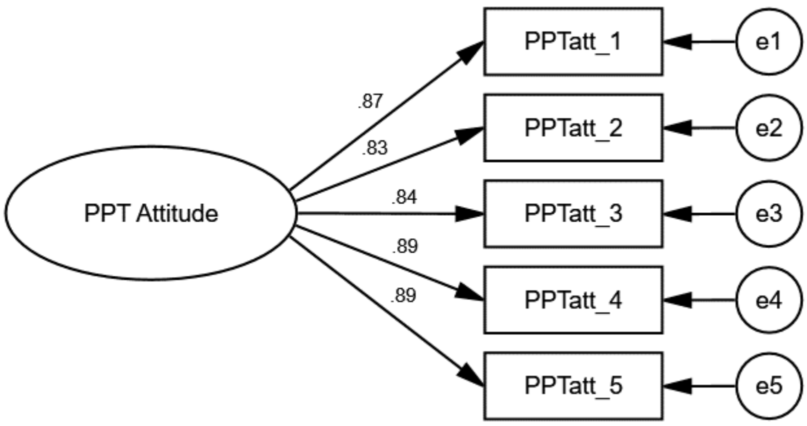

Figure 7.15 Single factor model of attitude towards PPT

Table 7.32

Factor loadings of single factor model of attitude towards PPT

First-order Factor	Items	Loading (Standardised Regression Weights)
Attitude towards PPT	PPTatt_1	.874***
	PPTatt_2	.825***
	PPTatt_3	.841***
	PPTatt_4	.887***
	PPTatt_5	.890***

*** $p < 0.001$

7.3.7 Attitude towards LOT

Teachers' attitude towards LOT assesses the level of attitude towards LOT by teachers and consists of five items. Hence, an examination of the model fit indices of the single model, as shown in **Table 7.33**, indicated that the single factor model exhibited the best model. This model has a low RMSEA value (0.014), a high CFI value (0.987), and a high TLI value (0.937), and a significant change in chi-square (χ^2 = 3.458, $p < 0.05$) given the change in degrees of freedom.

Table 7.33

Fit indices for CFA model in attitude towards LOT in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	3.458	5	0.692	0.987	0.937	0.014

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.34 displays standardized factor loadings for scales. The results of this analysis confirm that the factor loading values of all items in this construct are above the cut-off value (0.70), referring that all the items represented the latent construct they intended to measure. **Figure 7.16** and **Table 7.34** show the final structure and factor loadings of the model.

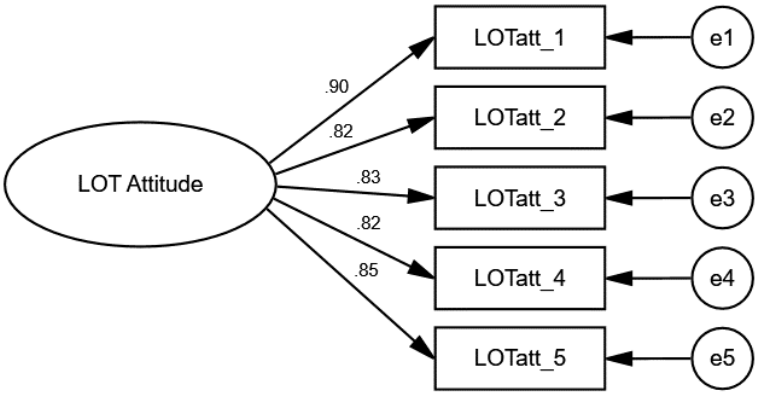

Figure 7.16 Single factor model of attitude towards LOT

Table 7.34

Factor loadings of single factor model of attitude towards LOT

First-order Factor	Items	Loading (Standardised Regression Weights)
Attitude towards LOT	LOTatt_1	.896***
	LOTatt_2	.821***
	LOTatt_3	.833***
	LOTatt_4	.816***
	LOTatt_5	.852***

*** $p < 0.001$

7.3.8 Attitude towards CAT

Attitude towards CAT measures teachers' attitude towards CAT and consists of five items. An examination of the model fit indices of the single model as shown in **Table 7.35** indicated that the single factor model exhibited the best model. This model has a low RMSEA value (0.064), and there is a high CFI value (0.999), and a high TLI value (0.988), and a significant change in chi-square (χ^2 = 2.018, p < 0.05) given the change in degrees of freedom.

Table 7.35

Fit indices for CFA model in attitude towards CAT in overall sample

Model	CMIN (χ^2)	*df*	CMIN (χ^2)/*df*	CFI	TLI	RMSEA
1. Single Factor Model	2.018	5	0.404	0.999	0.988	0.064

Note: CFI = Comparative fit index; TLI = Tucker-Lewis's index; RMSEA = Root Mean Squared Error of Approximation

Table 7.36 shows standardized factor loadings for scales. The results confirm that the factor loading values of all items in this construct are above the cut-off value (0.60), referring that all the items represented the latent construct they intended to measure. **Figure 7.17** and **Table 7.36** show the final structure and factor loadings of the model.

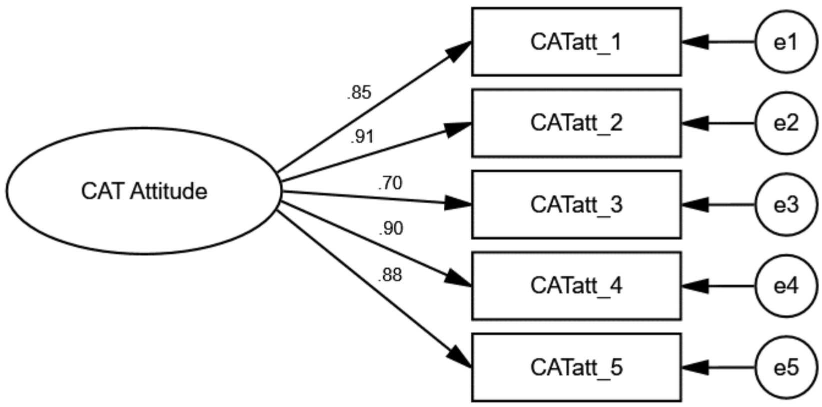

Figure 7.17 Single factor model of attitude towards CAT

Table 7.36

Factor loadings of single factor model of attitude towards CAT

First-order Factor	Items	Loading (Standardised Regression Weights)
Attitude towards CAT	CATatt_1	.847***
	CATatt_2	.912***
	CATatt_3	.699***
	CATatt_4	.895***
	CATatt_5	.877***

*** $p < 0.001$

7.3.9 Reliability and validity of scales of teacher questionnaire

As shown in **Table 7.37**, Cronbach's alpha values for all scales are above 0.70, indicating that all the scales are reliable and similar results will acquire if these scales are used for a similar population over time (Cronbach's alpha = 0.742, which is above the acceptable range).

Moreover, **Table 7.37** described AVE values for all scales in the fourth column. The AVE values for some sub-scales are higher than 0.50 indicates that these scales have high convergent validity. However, the rest sub-scales have an AVE value less than 0.5 and greater than 0.3, indicating adequate convergent validity. The AVE estimates range from 44 percent for the sub-scale attitude towards ICT to 76.6 percent for the sub-scale of Attitude towards Formative Assessment.

The composite or construct reliability results are presented in the fifth column in **Table 7.37**. As mentioned in Chapter 4, the construct reliability value of higher than 0.90 indicates that all measured scales or items consistently belong to a single construct (Hair et al. 2010).

Table 7.37

Reliability and convergent validity of the scales of teacher questionnaire

Scales from Students' Questionnaire	No of items	Reliability	Convergent Validity	
		Cronbach's Alpha	Average Variance Extracted (AVE)	Construct Reliability
General practices of formative assessment	4	0.871	0.653	0.964
Specific practices of formative assessment	5	0.858	0.563	0.944
Attitude towards formative assessment	7	0.956	0.766	0.958
ICT familiarity	7	0.918	0.617	0.978
Attitude towards ICT	12	0.970	0.440	0.973
Cognition factor of attitude towards ICT	3	0.898	0.750	0.906
Affective factor of attitude towards ICT	5	0.932	0.738	0.940
Behavioural factor of attitude towards ICT	4	0.925	0.754	0.917
Attitude towards PPT	5	0.936	0.746	0.942
Attitude towards LOT	5	0.922	0.713	0.935
Attitude towards CAT	5	0.926	0.722	0.948

7.4 Summary

This chapter discusses the validation of instruments through the CFA approach for scales employed in student and teacher questionnaires. These included student-level scales such as motivation, self-efficacy, and attitude towards learning mathematics, ICT familiarity, and attitudes towards ICT, formative assessment, PPT, LOT, and CAT. Scales such as self-efficacy and attitude towards learning mathematics were only examined on its one-factor model as they were assumed to have a single-factor structure. This study examined other scales by considering their factor loadings and model fit indices. For motivation, ICT familiarity, attitudes towards ICT, formative assessment, PPT, LOT, and CAT scales, the results indicated correlated factor models to be the best fitting model compared to other models for subsequent analysis in the next chapter. All scales of student-level have either high convergent validity or adequate one. Concerning reliability, all the scales are reliable with the value of Cronbach's alpha greater than 0.70 and construct reliability value of higher than 0.70.

Scales of a teacher-level questionnaire, namely general practices of formative assessment, specific formative assessment practices, ICT familiarity, and attitudes towards formative assessment, PPT, LOT, and CAT, were only examined on its one-factor model as they were assumed to have a single-factor structure. By considering their factor loadings and model fit indices, hierarchical factor models were the best fitting model for attitude towards ICT compared to other models. All scales have either high convergent validity or adequate one. The Cronbach's alpha values for all scales of teacher-level are above 0.70, and their construct reliability values are higher than 0.90; all scales indicate good reliable scales.

For all scales from student- and teacher-level, the results of CFA indicated the best fitting model for subsequent analysis of the next chapter in the next chapter, the Rasch analysis that provides the result of the instrument calibration and verification.

Chapter 8

Instrument Calibration and Verification: Rasch Analysis

8.1 Introduction

At the item level, it is fundamental to establish the reliability and validity of all scales employed in the study to provide an accurate estimate for the subsequent analysis. Therefore, the study conducted the Rasch analysis of each scale to verify whether the data already been validated in, CFA fit the Rasch Model. In this chapter, the Rasch analysis was conducted to verify the extent of the structures of each scale, which have already been confirmed in the CFA analysis in the previous chapter, fit the Rasch model, that is, the 'ideal measurement model'.

The study conducted Rasch analysis, followed by examining the item weighted fit mean square to see how well an item in each scale fit the 'ideal Rasch measurement model. Item fit can be determined by examining the in fit statistics, called the weighted mean square (INFIT MNSQ). The INFIT MNSQ is a basis for model fitting items or non-fitting items. This study chose the 0.60 to 1.40 range for the survey questionnaire and was justified in Chapter 5 (Bond & Fox, 2015). Item with INFIT MNSQ fall outside the acceptable range are considered not fitting the model and consequently should be deleted. If items have the INFIT MNSQ outside the acceptable range and show item delta in order, these items should be carefully examined as to whether they appear to measure what is needed for this study or not. It is necessary to take caution and thoroughly examine the list of items before removing any misfitting items as they may contain valuable information to substantiate the study. The following section presents the results of the item fit analysis.

Item separation reliability and weighted likelihood estimate (WLE) Person separation reliability provided from Rasch analysis identify the internal consistency of each scale and provide better measure for test reliability. For student and teacher questionnaire scales, the study conducted Rasch analysis to examine item fit after removing the misfitting items according to the acceptable range of the Weighted Mean Square.

The Weighted Likelihood Estimate (WLE) scores for each scale were obtained, after Rasch item analysis, for subsequent analysis, which are student and teacher-level models as well as the hierarchical linear model. In this study, Rasch analysis based on the Rating Scale Model (RSM) was undertaken using the ConQuest 4 statistical software package.

8.2 Rasch analysis for the scales from student questionnaire

8.2.1 Motivation

The motivation scale consists of two sub-scales, 'intrinsic motivation' and 'extrinsic motivation'. The first eight items belong to 'intrinsic motivation', and the last eight belong to 'extrinsic motivation'. Based on the hierarchical model of the scale from the CFA analysis, a multidimensional rating scale analysis was carried out to examine if the items fit the model. The correlation between the two sub-scales was 0.845, indicating the strength of the relationship between the two sub-scales in **Table 8.1**. The infit mean square (INFIT MNSQ) of each sub-scale item was within the acceptable range of 0.60 to 1.40, presented in **Table 8.2**. Item deltas of all items showed that their options on each scale were in order. The study used the structure of this scale with its items for subsequent analysis.

Table 8.1

Correlation matrix for two subscales of motivation

	Intrinsic motivation	Extrinsic motivation
Intrinsic motivation	1	
Extrinsic motivation	0.845	1

Table 8.2

Parameter estimates of motivation from multidimensional rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
IntrinM1	-0.775	0.057	0.87	(0.88, 1.12)	-2.3	-3.68	-1.52	2.87
IntrinM2_R	-0.739	0.057	1.25	(0.88, 1.12)	4.0	-3.65	-1.48	2.91
IntrinM3_R	-0.393	0.056	1.17	(0.88, 1.12)	2.8	-3.3	-1.14	3.26
IntrinM4	-0.351	0.056	1.03	(0.88, 1.12)	0.6	-3.26	-1.1	3.3
IntrinM5	-0.069	0.056	0.70	(0.88, 1.12)	-5.6	-2.97	-0.81	3.58
IntrinM6	0.165	0.056	0.89	(0.88, 1.12)	-1.9	-2.74	-0.58	3.82
IntrinM7	1.403	0.054	1.05	(0.89, 1.11)	0.8	-1.5	0.66	5.05
IntrinM8	0.759*	0.148	1.03	(0.88, 1.12)	0.5	-2.15	0.01	4.41
ExtrinM1	2.053	0.057	1.28	(0.89, 1.11)	4.5	-0.81	1.08	5.88
ExtrinM2	1.263	0.059	1.12	(0.88, 1.12)	1.9	-1.6	0.29	5.09
ExtrinM3	-0.385	0.063	1.00	(0.88, 1.12)	0.5	-3.25	-1.36	3.45
ExtrinM4	-1.374	0.064	0.88	(0.89, 1.11)	-2.4	-4.23	-2.34	2.46
ExtrinM5	-0.051	0.063	0.94	(0.88, 1.12)	-1.2	-2.91	-1.02	3.78
ExtrinM6	-0.567	0.063	0.92	(0.88, 1.12)	-1.3	-3.43	-1.54	3.26
ExtrinM7	-0.291	0.063	1.01	(0.88, 1.12)	0.2	-3.15	-1.26	3.54
ExtrinM8	-0.647*	0.163	0.87	(0.88, 1.12)	-2.3	-3.51	-1.62	3.18

Note *Constraint

8.2.2 Self-efficacy

The self-efficacy scale consists of eight items. Based on the one-factor model from CFA analysis, the study used the rating scale modelling to conduct the item analysis for this scale. As shown in **Table 8.3**, all items were within the acceptable INFIT MNSQ range of 0.6 and 1.4, and item deltas were in order. This final structure of this scale with its eight items was used for subsequent analysis.

Table 8.3

Parameter estimates of self-efficacy from rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
Self-Efficacy_1	-0.87	0.048	0.86	(0.89, 1.11)	-2.6	-3.43	1.02	1.84
Self-Efficacy_2R	-0.27	0.047	0.98	(0.89, 1.11)	-0.4	-2.83	0.42	2.44
Self-Efficacy_3R	0.234	0.047	0.92	(0.90, 1.10)	-1.6	-2.33	0.09	2.94
Self-Efficacy_4	-0.112	0.047	0.99	(0.90, 1.10)	-0.2	-2.67	0.26	2.6
Self-Efficacy_5R	0.035	0.047	1.3	(0.90, 1.10)	5.2	-2.53	0.11	2.74
Self-Efficacy_6	0.619	0.047	0.78	(0.90, 1.10)	-4.4	-1.94	0.47	3.33
Self-Efficacy_7	0.69	0.047	0.99	(0.90, 1.10)	-0.3	-1.87	0.54	3.4
Self-Efficacy_8R	-0.325*	0.124	1.12	(0.89, 1.11)	2.2	-2.89	0.47	2.38

Note *Constraint

8.2.3 Attitude towards learning mathematics

The scale of the attitude towards learning mathematics consists of seven items. Based on the one-factor model from the CFA analysis, present work conducted an item analysis for this scale by the rating scale modelling. As shown in **Table 8.4**, all items were within the acceptable INFIT MNSQ range of 0.6 and 1.4, and item deltas were in order. This final structure of this scale with its seven items was used for subsequent analysis.

Table 8.4

Parameter estimates of attitude towards learning mathematics from rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
Att_LearnMath_1	-0.038	0.05	1.13	(0.89, 1.11)	2.1	-2.1	-0.64	2.62
Att_LearnMath_2	-0.144	0.05	0.93	(0.89, 1.11)	-1.3	-2.2	-0.74	2.52
Att_LearnMath_3	-0.507	0.051	1.03	(0.89, 1.11)	0.5	-2.57	-1.11	2.15
Att_LearnMath_4	1.074	0.047	1.26	(0.89, 1.11)	4.4	-0.99	0.47	3.74
Att_LearnMath_5	0.258	0.049	0.92	(0.89, 1.11)	-1.4	-1.8	-0.34	2.92
Att_LearnMath_6	-0.096	0.05	0.74	(0.89, 1.11)	-4.9	-2.16	-0.7	2.57
Att_LearnMath_7	-0.548*	0.121	1.00	(0.89, 1.11)	0.1	-2.61	-1.15	2.11

Note *Constraint

8.2.4 Attitude towards formative assessment

The attitude towards formative assessment scale consists of two sub-scales: 'the attitude towards formative test' and 'the attitude towards formative feedback'. The first five items belong to 'the attitude towards formative test', and the last four items belong to 'the attitude towards formative feedback'. Based on the hierarchical model of the scale from the CFA analysis, this study carried out a multidimensional rating scale analysis to examine if the items fit the model. The correlation between the two sub-scales was 0.803, indicating the strength of the relationship between the two sub-scales in **Table 8.5**. The infit mean square (INFIT MNSQ) of each sub-scale item was within the acceptable range of 0.60 to 1.40, presented in **Table 8.6**. Item deltas of all items showed that their options on each scale were in order. The present chapter used this scale structure with its items for subsequent analysis.

Table 8.5

Correlation matrix for two subscales of attitude towards formative assessment

	Attitude towards questions from formative assessment	Attitude towards formative feedback
Attitude towards questions from formative assessment	1	
Attitude towards formative feedback	0.803	1

Table 8.6
Parameter estimates of attitude towards formative assessment from multidimensional rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
Att_FormativeAssessment_1	-0.412	0.051	0.81	(0.87, 1.13)	-3.1	-1.46	-1.09	2.77
Att_FormativeAssessment_2	-0.242	0.051	0.75	(0.87, 1.13)	-4.1	-1.3	-0.93	2.93
Att_FormativeAssessment_3	0.379	0.049	1.18	(0.88, 1.12)	2.9	-0.73	-0.35	3.5
Att_FormativeAssessment_4R	0.383	0.049	1.33	(0.88, 1.12)	5	-0.72	-0.35	3.51
Att_FormativeAssessment_5R	-0.109*	0.099	1.10	(0.87, 1.13)	1.5	-1.18	-0.81	3.05
Att_FormativeAssessment_6	0.07	0.056	0.90	(0.88, 1.12)	-1.6	-2.04	-1.67	2.19
Att_FormativeAssessment_7	0.392	0.055	0.94	(0.87, 1.13)	-0.9	-1.74	-1.36	2.49
Att_FormativeAssessment_8	-0.072	0.056	0.92	(0.88, 1.12)	-1.3	-2.17	-1.8	2.06
Att_FormativeAssessment_9	-0.390*	0.097	1.06	(0.88, 1.12)	1.1	-2.47	-2.1	1.76

Note *Constraint

8.2.5 ICT Familiarity

The sub-scales of ICT access and ICT Usage were involved in the scale of ICT Familiarity. ICT Access consisted of four items (from Item#1 to Item#4). ICT Usage consisted of eight items (from Item#5 to Item#12). Based on the hierarchical model of the scale from the CFA analysis, this study carried out a multidimensional rating scale analysis to examine if the items fitted the model (**Table 8.7**). An item analysis examined whether all items fit the model. As shown in **Table 8.8**, all items were within the acceptable range of 0.60 to 1.40 for INFIT MNSQ. Item deltas showed that response choices on the scale were in order. The present study used this scale structure with its items for subsequent analysis.

Table 8.7

Correlation matrix for two subscales of ICT Familiarity

	ICT Access	ICT Usage
ICT Access	1	
ICT Usage	0.707	1

Table 8.8
Parameter estimates of ICT familiarity from multidimensional rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta
Access Own Laptop	-2.020	0.096	1.05	(0.78, 1.22)	0.5	-1.80
Access Personal Email	-0.589	0.083	0.95	(0.87, 1.13)	-0.7	-0.43
Access Internet at Home	-1.115	0.088	0.98	(0.85, 1.15)	-0.3	-0.93
Training Usage Internet	3.724*	0.154	1.12	(0.87, 1.13)	1.8	3.59
Playing Online games	1.217	0.085	1.13	(0.91, 1.09)	2.7	1.14
Play Instructional Program	2.119	0.085	1.10	(0.91, 1.09)	2.1	1.96
Communication Social Network	0.426	0.089	0.86	(0.89, 1.11)	-2.6	0.38
Participate Social Network	-0.373	0.095	0.85	(0.87, 1.13)	-2.3	-0.39
Download Instruction	-1.029	0.104	0.81	(0.84, 1.16)	-2.5	-1.03
Upload Own Creation	1.942	0.085	1.10	(0.91, 1.09)	2.2	1.79
Hour_internet	-2.181	0.122	0.84	(0.77, 1.23)	-1.4	-2.16
Hour_ICT Devices	-2.120*	0.254	0.87	(0.77, 1.23)	-1.1	-2.10

Note *Constraint

8.2.6 Attitude towards ICT

The scale attitude towards ICT consists of three sub-scales. The first subscale ‘Affective Factor’ was composed of the first six items; the seventh item to the tenth item belong to the second sub-scale ‘Cognitive Factor’; the last six items related to the third sub-scale ‘Behavioural Factor’. Based on the hierarchical model of the scale from the CFA analysis, this study used a multidimensional rating scale analysis to examine if the items fitted the model

(**Table 8.9**). The researcher used an item analysis to examine if all items fitted the model. As shown in **Table 8.10**, all items were within the acceptable range of 0.60 to 1.40 for INFIT MNSQ. Item deltas showed that response choices on the scale were in order. The present study used this scale structure with its items for subsequent analysis.

Table 8.9

Correlation matrix for three subscales of attitude towards ICT

	Affective Factor	Cognitive Factor	Behavioural Factor
Affective	1		
Cognition	0.905	1	
Behaviour	0.687	0.791	1

Table 8.10

Parameter estimates of attitude towards ICT from multidimensional rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
Att_ICT_Use_1	0.023	0.051	1.23	(0.88, 1.12)	3.6	-2.15	-0.70	2.06
Att_ICT_Use_2R	-0.002	0.051	0.91	(0.88, 1.12)	-1.5	-2.17	-0.72	2.04
Att_ICT_Use_3R	-0.267	0.052	0.86	(0.88, 1.12)	-2.4	-2.41	-0.96	1.80
Att_ICT_Use_4	0.512	0.05	0.96	(0.89, 1.11)	-0.7	-1.71	-0.27	2.49
Att_ICT_Use_5	0.193	0.051	0.75	(0.88, 1.12)	-4.7	-2.00	-0.55	2.21
Att_ICT_Use_6R	-0.458*	0.115	1.00	(0.88, 1.12)	0.5	-2.57	-1.13	1.63
Att_ICT_Use_7R	0.505	0.048	1.17	(0.89, 1.11)	2.8	-1.44	0.01	2.77
Att_ICT_Use_8R	-0.626	0.051	1.18	(0.89, 1.11)	3.0	-2.46	-1.02	1.74
Att_ICT_Use_9	0.445	0.048	1.38	(0.89, 1.11)	6.0	-1.49	-0.05	2.71
Att_ICT_Use_10	-0.323*	0.085	1.26	(0.88, 1.12)	4.1	-2.19	-0.74	2.02
Att_ICT_Use_11	0.077	0.049	0.88	(0.89, 1.11)	-2.2	-1.52	-0.08	2.69
Att_ICT_Use_12	1.256	0.047	1.10	(0.90, 1.10)	1.9	-0.45	1.00	3.76
Att_ICT_Use_13	0.558	0.048	1.00	(0.89, 1.11)	0.8	-1.08	0.36	3.12
Att_ICT_Use_14	0.102	0.049	0.72	(0.89, 1.11)	-5.4	-1.5	-0.05	2.71
Att_ICT_Use_15	-0.742	0.051	0.96	(0.89, 1.11)	-0.8	-2.27	-0.82	1.94
Att_ICT_Use_16	-1.251*	0.109	1.00	(0.89, 1.11)	0.9	-2.74	-1.29	1.47

Note *Constraint

8.2.7 Attitude towards test modes (PPT, LOT, and CAT)

The scale of the attitude towards test modes (PPT, LOT and CAT) has two sub-scales: attitude towards PPT/LOT/CAT format and attitude towards PPT/LOT/CAT feedback respectively. The first six items of each scale (from Item#1 to Item#6R) belong to 'attitude towards PPT/LOT /CAT format', and another five items of each scale (from Item#7 and Item#12) belong to 'attitude towards PPT/LOT/CAT feedback'. Based on their hierarchical models of these scales from their CFA analyses, this study carried out three multidimensional rating scale analyses to examine if the items fitted the models (**Tables 8.11**). This chapter designed an item analysis to examine if all items fit the model. As shown in **Tables 8.12**, all items of scales were within the acceptable range of 0.60 to 1.40 for INFIT MNSQ. Item deltas showed that response choice on the scale were in order. This chapter used these scale structure with its items for subsequent analysis.

Table 8.11

Correlation matrix for two subscales of attitude towards PPT, LOT and CAT

	Attitude towards PPT format
Attitude towards PPT feedback	0.943
	Attitude towards LOT format
Attitude towards LOT feedback	0.983
	Attitude towards CAT format
Attitude towards CAT feedback	0.901

Table 8.12
Parameter estimates of attitude towards PPT, LOT, CAT from multidimensional rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
Attitude towards PPT								
Att_PPTTestMode_1	-1.279	0.064	0.75	(0.87, 1.13)	-4.1	-4.7	0.13	2.23
Att_PPTTestMode_2	-0.314	0.064	1.13	(0.88, 1.12)	2.1	-3.81	1.02	3.12
Att_PPTTestMode_3	0.527	0.063	1.29	(0.90, 1.10)	5.2	-3.02	1.80	3.90
Att_PPTTestMode_4	-0.337	0.064	1.37	(0.88, 1.12)	5.6	-3.83	1.00	3.10
Att_PPTTestMode_5	-0.735	0.064	0.94	(0.88, 1.12)	-1.0	-4.19	0.63	2.74
Att_PPTTestMode_6R	2.138*	0.143	1.11	(0.90, 1.10)	2.9	-1.5	3.33	5.43
Att_PPTTestMode_7	-0.264	0.062	0.66	(0.87, 1.13)	-6.0	-4.65	0.18	2.28
Att_PPTTestMode_8	-0.225	0.062	0.68	(0.87, 1.13)	-5.6	-4.62	0.21	2.31
Att_PPTTestMode_9	0.022	0.062	0.65	(0.87, 1.13)	-4.7	-4.39	0.43	2.54
Att_PPTTestMode_10	0.168	0.062	0.76	(0.88, 1.12)	-4.1	-4.26	0.57	2.67
Att_PPTTestMode_11	0.298*	0.124	0.78	(0.88, 1.12)	-3.9	-4.14	0.69	2.79
Attitude towards LOT								
Att_LOTTestMode_1	-0.291	0.062	0.81	(0.88, 1.12)	-3.1	-4.59	0.08	4.41
Att_LOTTestMode_2	-0.606	0.062	0.83	(0.88, 1.12)	-2.8	-4.9	-0.23	4.11
Att_LOTTestMode_3	-0.875	0.061	1.15	(0.88, 1.12)	2.4	-5.17	-0.5	3.84
Att_LOTTestMode_4	-0.795	0.062	1.14	(0.88, 1.12)	2.3	-5.09	-0.42	3.92
Att_LOTTestMode_5	-0.782	0.062	0.84	(0.88, 1.12)	-2.7	-5.08	-0.41	3.92
Att_LOTTestMode_6R	3.349*	0.138	1.02	(0.89, 1.11)	1.4	-0.97	3.7	8.04
Att_LOTTestMode_7	0.699	0.06	1.33	(0.88, 1.12)	2.5	-4.17	0.5	4.83
Att_LOTTestMode_8	0.055	0.06	0.87	(0.88, 1.12)	-3.3	-4.82	-0.15	4.19
Att_LOTTestMode_9	-0.259	0.06	0.70	(0.88, 1.12)	-5.7	-5.13	-0.46	3.88
Att_LOTTestMode_10	-0.418	0.06	0.71	(0.89, 1.11)	-5.5	-5.28	-0.61	3.73
Att_LOTTestMode_11	-0.077*	0.121	0.79	(0.88, 1.12)	-3.8	-4.94	-0.27	4.07
Attitude towards CAT								
Att_CATTestMode_1	0.499	0.051	1.13	(0.89, 1.11)	2.2	-1.35	-0.53	3.20
Att_CATTestMode_2	-0.098	0.051	0.91	(0.89, 1.11)	-1.7	-1.91	-1.1	2.64
Att_CATTestMode_3	-0.584	0.052	1.15	(0.90, 1.10)	2.8	-2.37	-1.56	2.17
Att_CATTestMode_4	-0.092	0.051	1.00	(0.89, 1.11)	0.0	-1.9	-1.09	2.64
Att_CATTestMode_5	0.200	0.051	0.98	(0.89, 1.11)	-0.3	-1.63	-0.82	2.92
Att_CATTestMode_6R	0.076*	0.115	1.24	(0.89, 1.11)	4.1	-1.75	-0.93	2.80
Att_CATTestMode_7	0.150	0.051	0.96	(0.89, 1.11)	-0.7	-1.61	-0.8	2.94
Att_CATTestMode_8	0.211	0.051	0.83	(0.89, 1.11)	-3.1	-1.55	-0.74	3.00
Att_CATTestMode_9	-0.097	0.051	1.04	(0.89, 1.11)	0.7	-1.84	-1.03	2.71
Att_CATTestMode_10	-0.263	0.052	1.04	(0.89, 1.11)	0.8	-1.99	-1.18	2.55
Att_CATTestMode_11	-0.001*	0.103	0.93	(0.89, 1.11)	-1.3	-1.75	-0.94	2.80

Note *Constraint

8.2.8 Item and person separation reliability

This study measured the internal consistency of the scale by the item separation reliability and the WLE person separation reliability (**Table 8.13**). Although there is no cut-off value for good reliability, it is generally accepted that higher values of items and person separation reliability indicate higher reliability. Two scales that are ICT access and ICT usage had relatively lower reliability. The WLE person separation reliability of all scales is lower than their item separation reliability of the scale, respectively.

Table 8.13

Item and person separation reliabilities for all scales from student questionnaire

Scale	Item Separation Reliability	WLE Person Separation Reliability
Single Dimension		
Self-Efficacy	0.992	0.848
Attitude towards mathematics learning	0.992	0.784
Multi Dimensions		
Motivation	0.995	
Intrinsic Motivation		0.873
Extrinsic Motivation		0.829
Attitude towards formative assessment	0.974	
Attitude towards question in formative assessment		0.797
Attitude towards formative feedback		0.724
Attitude towards PPT	0.985	
Attitude towards questions in PPT		0.884
Attitude towards feedback in PPT		0.910
Attitude towards LOT	0.985	
Attitude towards questions in LOT		0.934
Attitude towards feedback in LOT		0.906
Attitude towards CAT	0.973	
Attitude towards questions in CAT		0.870
Attitude towards feedback in CAT		0.803
ICT familiarity	0.996	
ICT Access		0.626
ICT Usage		0.718
Attitude towards ICT	0.991	
Affective		0.845
Cognition		0.846
Behaviour		0.790

8.3 Rasch Analysis for the scales from teacher questionnaire

8.3.1 Scales related to formative assessment

There are three scales related to formative assessment which are *General practices of formative assessment, Specific practices of formative assessment and Attitude towards formative assessment.* This study conducted an item analysis for these scales by Rasch's rating scale modelling based on their one-factor models from their CFA analyses. As shown in **Table 8.14**, all items of three scales were within the acceptable INFIT MNSQ range of 0.6 and 1.4

and item deltas were in order. This study used their final scale structures for subsequent analysis.

Table 8.14

Parameter estimates of scales related to formative assessment from rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
General practices of formative assessment								
GPTech_1	0.003	0.329	0.77	(0.37, 1.63)	-0.7	-2.12	-0.62	2.75
GPTech_2	-0.218	0.33	1.12	(0.38, 1.62)	0.5	-2.34	-0.84	2.53
GPTech_3	-0.218	0.33	0.72	(0.38, 1.62)	-0.9	-2.34	-0.84	2.53
GPTech_4	0.434*	0.571	0.62	(0.33, 1.67)	-1.2	-1.69	-0.19	3.18
Specific practices of formative assessment								
SPFA_1	-0.59	0.427	0.82	(0.36, 1.64)	-0.5	-3.13	-0.63	2.45
SPFA_2	-0.219	0.427	0.96	(0.27, 1.73)	0.0	-2.76	-0.26	2.82
SPFA_3	0.88	0.432	1.16	(0.05, 1.95)	0.5	-1.66	0.84	3.92
SPFA_4	1.629	0.437	1.13	(0.02, 1.98)	0.4	-0.91	1.59	4.67
SPFA_5	-1.700*	0.861	0.84	(0.50, 1.50)	-0.6	-4.24	-1.74	1.34
Attitude towards formative assessment								
AttFA_1R	-0.252	0.404	0.93	(0.32, 1.68)	-0.1	-3.37	-0.26	5.3
AttFA_2	-0.633	0.395	1.27	(0.36, 1.64)	0.9	-4.6	-1.5	4.06
AttFA_3R	1.224	0.412	0.78	(0.34, 1.66)	-0.6	-3.96	-0.86	4.71
AttFA_4R	-0.386	0.399	0.98	(0.36, 1.64)	0.1	-3.66	-0.56	5.01
AttFA_5	-0.424	0.396	1.14	(0.38, 1.62)	0.5	-3.97	-0.87	4.7
AttFA_6	0.968	0.409	1.15	(0.37, 1.63)	0.6	-4.27	-1.16	4.4
AttFA_7	-0.498*	0.986	0.74	(0.36, 1.64)	-0.8	-3.64	-0.54	5.03

Note *Constraint

8.3.2 ICT Familiarity

The study conducted an item analysis for this scale by modelling the rating scale based on a one-factor model from the CFA analysis. As shown in **Table 8.15**, all items were within the acceptable INFIT MNSQ range of 0.6 and 1.4 and item deltas were in order. This study used a final scale structure with its five items for subsequent analysis.

Table 8.15

Parameter estimates of ICT familiarity from rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta
ICT_Use1	-0.972	0.628	0.9	(0.00, 2.30)	0.1	-0.97
ICT_Use2	0.543	0.633	0.74	(0.24, 1.76)	-0.6	0.54
ICT_Use3	-0.215	0.636	0.69	(0.00, 2.06)	-0.5	-0.21
ICT_Use4	0.548	0.633	1.23	(0.25, 1.75)	0.7	0.55
ICT_Use5	0.542	0.633	0.76	(0.24, 1.76)	-0.5	0.54
ICT_Use6	-0.223	0.636	2.5	(0.00, 2.06)	2.2	-0.22
ICT_Use7	-0.223*	1.551	1.17	(0.00, 2.06)	0.5	-0.22

Note *Constraint

8.3.3 Attitude towards ICT

Based on the hierarchical model of the scale from the CFA analysis, a multidimensional rating scale analysis examined if the items fitted the model (**Table 8.16**). The study used an item analysis to examine if all items fit the model. As shown in **Table 8.17**, all items were within the acceptable range of 0.60 to 1.40 for INFIT MNSQ. Item deltas showed that response choices on the scale were in order. The present study used this scale structure with its items for subsequent analysis.

Table 8.16

Correlation matrix showing correlation coefficients for three subscales of attitude towards ICT

	Attitude towards ICT sub-scale 1	Attitude towards ICT sub-scale 2	Attitude towards ICT sub-scale 3
Cognition	1		
Affective	0.897	1	
Behaviour	0.864	0.881	1

Table 8.17

Parameter estimates of attitude towards ICT from multidimensional rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
AttICT_1	-0.517	0.351	1.26	(0.34, 1.66)	0.8	-4.96	-0.38	3.7
AttICT_2	0.712	0.372	1.00	(0.04, 1.96)	0.2	-3.52	1.06	5.13
AttICT_3	-0.195*	0.512	0.65	(0.35, 1.65)	-1.1	-4.68	-0.11	3.97
AttICT_4	-0.698	0.357	0.63	(0.25, 1.75)	-1	-5.45	-0.88	3.2
AttICT_5	0.564	0.378	0.92	(0.15, 1.85)	0	-3.75	0.83	4.91
AttICT_6	-0.24	0.347	1.17	(0.37, 1.63)	0.6	-4.85	-0.27	3.81
AttICT_7	0.311	0.355	0.83	(0.32, 1.68)	-0.4	-4	0.57	4.65
AttICT_8	0.064*	0.719	0.73	(0.35, 1.65)	-0.8	-4.29	0.29	4.36
AttICT_9	0.084	0.356	0.81	(0.38, 1.62)	-0.5	-4.03	0.55	4.63
AttICT_10	0.498	0.356	1.06	(0.27, 1.73)	1.1	-3.78	0.8	4.87
AttICT_11	-0.221	0.351	1.12	(0.34, 1.66)	0.5	-4.68	-0.1	3.98
AttICT_12	-0.360*	0.614	1.16	(0.26, 1.74)	0.5	-4.95	-0.37	3.71

Note *Constraint

8.3.4 Attitude towards test modes (PPT, LOT)

There are three scales related to attitude towards test mode (PPT, LOT and CAT). Rasch's rating scale modelling was used to estimates of items of the scales based on their one-factor models resulting from their CFA analyses. **Table 8.18** shows all items of the scales were within the acceptable INFIT MNSQ range of 0.6 and 1.4, and item deltas were in order. This study used their final scale structures with its five items for subsequent analysis.

Table 8.18

Parameter estimates of attitude towards test modes (PPT, LOT and CAT) from rating scale modelling

Item	Estimate	Error	MNSQ	CI	t	Item Delta		
Attitude towards PPT								
PPTatt_1	-0.059	0.375	1.31	(0.29, 1.71)	0.9	-4.65	0.23	4.25
PPTatt_2	0.517	0.381	1.02	(0.22, 1.78)	0.2	-4.07	0.8	4.82
PPTatt_3	-0.340	0.375	0.95	(0.33, 1.67)	0	-4.93	-0.05	3.97
PPTatt_4	-0.059	0.375	1.18	(0.29, 1.71)	0.6	-4.65	0.23	4.25
PPTatt_5	-0.059*	0.753	1.07	(0.29, 1.71)	0.3	-4.65	0.23	4.25
Attitude towards LOT								
LOTatt_1	0.049	0.351	0.63	(0.34, 1.66)	-1.1	-2.12	-0.71	2.98
LOTatt_2	-0.197	0.351	0.88	(0.35, 1.65)	-0.3	-2.37	-0.95	2.73
LOTatt_3	0.295	0.351	1.04	(0.32, 1.68)	0.2	-1.88	-0.46	3.22
LOTatt_4	-0.197	0.351	1.08	(0.35, 1.65)	0.4	-2.37	-0.95	2.73
LOTatt_5	0.049*	0.702	0.95	(0.34, 1.66)	0	-2.12	-0.71	2.98
Attitude towards CAT								
CATatt_1	0.870	0.441	1.44	(0.31, 1.69)	1.2	-4.01	0.28	6.34
CATatt_2	-1.719	0.235	0.62	(0.36, 1.64)	-1.7	-6.6	-2.31	3.75
CATatt_3	0.303	0.350	1.03	(0.27, 1.73)	1.4	-4.58	-0.29	5.78
CATatt_4	1.204	0.382	0.64	(0.31, 1.69)	-1.5	-3.68	0.61	6.68
CATatt_5	-0.658*	0.602	1.11	(0.25, 1.75)	0.4	-5.54	-1.25	4.82

Note *Constraint

8.3.5 Item and person separation reliability

This study measured the internal consistency of scale by the item separation reliability and the WLE person separation reliability (**Table 8.19**). Two scales, ICT access and ICT usage, had relatively lower reliability. The WLE person separation reliability of all the scales is lower than their item separation reliability of the scale, respectively.

Table 8.19

Item and person separation reliabilities for all scales from teacher questionnaire

Scale	Item Separation Reliability	WLE Person Separation Reliability
Single Dimension		
General Practices of Formative assessment	0.860	0.780
Specific Practices of Formative Assessment	0.821	0.842
Attitude towards Formative Assessment	0.975	0.894
ICT Familiarity	0.693	0.697
Attitude towards PPT	0.946	0.856
Attitude towards LOT	0.995	0.860
Attitude towards CAT	0.976	0.840
Multi Dimensions		
Cognition	0.998	0.896
Affective		0.872
Behaviour		0.834

8.4 Students' achievement tests: Function, Remainder and Factor

There are pre-test and post-test for each content area: function, remainder and factor. Each test contains 20 items. The pre-test and post-test are identical for each content. All items for these tests are selected from the respective item-bank (see **Chapter 5**) because they have already been calibrated by the Rasch dichotomous model.

8.5 Scoring

The focus of this section includes the scoring process for scales employed in student questionnaires, teacher questionnaires, and student achievement tests. In the previous sections, the researcher of this thesis conducted the calibration of items for each scale after removing the misfit items. Conquest software produce WLE score. For all scales, the analysis obtained weighted likelihood estimate (WLE) scores through item calibration in Rasch analyses. Refer to section 4.5.6 in Methodology chapter.

8.6 Summary

This chapter discusses the verification, calibration, and scoring of the scales employed in student and teacher questionnaires. The questionnaires on Rasch analysis focuses on the Rasch rating scale and multi-dimensional rating scale models. Items from the student achievement test focus on Rasch dichotomous model. This study conducted an item fit analysis to examine the INFIT MNSQ for each item. The results indicated that all items were within the threshold value of 0.4 and 1.6. No misfit items in all constructs. The scoring of scales employed in student and teacher questionnaires and students' achievement tests involved anchoring the item difficulty estimates obtained from the calibrated item to obtain the WLE scores for subsequent analysis. Before presenting the results of student- and teacher-level models and the hierarchical linear model, it is necessary to examine the demographic characteristics of the respondents involved in this study. The following chapter presents the difference in achievement due to the test mode effect.

Chapter 9

Difference in Achievement Improvement due to Effect of Test modes

9.1 Introduction

In Myanmar, teachers apply the Paper-and-pencil test (PPT) as the common assessment technique to test and enhance students' achievement. However, this technique leads to a more significant workload for teachers in administering the test, scoring the answer sheets, providing scores and feedback. Therefore, this technique is ineffective in the learning progress of the students due to a loss of time. The overall aim of this project is to explore the solution to the abovementioned problem and solve it by integrating two technological-integrated approaches: linear-online test (LOT) and computer-adaptive test (CAT). These tests can save time in the test administration and scoring and provide scores and feedback immediately to each student. The researcher of this thesis applied PPT, LOT, and CAT as the classroom formative assessment techniques in the five High Schools in Myanmar. The experiment period was from July to August 2019.

This chapter will identify the framework for transitioning the common assessment technique, PPT, to two technological-integrated ones: LOT and CAT for the classroom formative assessment in high schools. It will mainly contribute to understanding the effectiveness in the practical application of these two technological-integrated techniques for the basic education system in Myanmar. Its methodological contribution will be the data analytical methods for further studies and provide education policymakers and administrators a better comprehensive suggestion of these new test modes for students learning progress in Myanmar.

However, this study applied a counter-balance experimental research design for examining the effect of three test modes (PPT, LOT, and CATs). **Table 9.1** describes a summary of the experimental research design. There are three experimental groups in the study, which were exposed to all three test modes as a classroom formative assessment in learning three mathematics concepts in a different order (see **Table 9.1**). There are 227 participants in Group-1; 210 participants in Group-2; and 222 participants in Group-3.

The three test modes are PPT, LOT, and CAT. Student participants learned three concepts (*Function*, *Remainder*, and *Factor*). In Group-1, the participants faced PPT while learning about Function, LOT while learning about Remainder, and CAT while learning about Factor.

The Group-2 was in the order of LOT, CAT, and PPT. Finally, the Group-3 was in the order of CAT, PPT, and LOT.

Table 9.1 shows three pairs of pre-tests and post-tests. This study used a pre-test to identify the mathematics ability of participants before exposing them to a test mode. Then, the researcher conducted a post-test after exposing each test mode. Finally, to measure achievement improvement, that is, the change in the score of their mathematics ability, the researcher examined the difference between post- and pre-test scores by using WLE.

Before exploring the score difference, the study equated a pair score from the pre-test and post-test on the same scale through a series of Rasch analyses to produce the WLE logit score. This step was undertaken to make it possible to measure the improved achievement over time and to examine possible contributing factors. The results of the analyses confirmed that all those items conformed to the standard item characteristic curve with mean square (MNSQ) values within the acceptable range. The analyses provided the weighted likelihood estimate (WLE) logit scores for each test. The study named the differences of the WLE logit scores from a pair of pre-test and post-test as the gained WLE logit scores or WLE logit score improvement.

Table 9.1

Counter-balanced experimental research due to the three test modes (from Chapter 3)

Experimental Group	Pre-test	Experiment-1	Post-test	Pre-test	Experiment -2	Post-test	Pre-test	Experiment -3	Post-test
Experimental Group-1 (Test-mode-order-1)	Pre-1	F/PPT-1 with Score report and Feedback for Function	Post-1	Pre-2	F/LOT-2 with Score report and Feedback for Remainder	Post-2	Pre-3	F/CAT-3 with Score report and Feedback for Factor	Post-3
Experimental Group-2 (Test-mode-order-2)	Pre-1	F/LOT-1 with Score report and Feedback for Remainder	Post-1	Pre-2	F/CAT-2 with Score report and Feedback for Factor	Post-2	Pre-3	F/PPT-3 with Score report and Feedback for Function	Post-3
Experimental Group-3 (Test-mode-order-3)	Pre-1	F/CAT-1 with Score report and Feedback for Factor	Post-1	Pre-2	F/PPT-2 with Score report and Feedback for Function	Post-2	Pre-3	F/LOT-3 with Score report and Feedback for Remainder	Post-3

Note**:** F/ refers to a formative assessment during the experiment (For example, F/PPT-1 means formative assessment in PPT mode in the first experiment and provide the score report and feedback after assessment); Pre refers to Pre-test; Post refers to Post-test

9.2 Research question

The main research question is *which test mode (PPT, LOT, or CAT) improves students' achievement more when the test mode is applied as classroom formative assessments in Myanmar high schools.* On the other hand, before identifying the main research question, this study should identify the statistical and practical significance of independent factors: test modes, mathematics-concepts, and groups on the achievement improvement. Moreover, it needs to identify whether the specific effect size reaches practical application and examine which of their sub-groups among the individual independent factors affects achievement improvement. This is for examining whether the main effects and interaction effects of mathematical-concepts and groups impacts achievement improvement more than test modes does.

9.3 Analysis of variance (ANOVA)

The analysis of variance (ANOVA) model provides the user with a statistically based technique capable of producing meaningful models on the importance of the factor studied in the experiment. Consequently, ANOVA allows possible interactions among factors within the data set that may have a bearing on the effect of test mode to study. The use of modelling in this context has proved to be highly relevant due mainly to the opportunity of the finding the most suitable test mode for use in the classroom formative assessment interventions.

According to the counter-balance research design, there are three independent factors: test modes, mathematics-concepts, and groups (or other terms: test mode-orders) on the achievement improvement. Consequently, this study applies the three-way between-groups analysis of variance (three-way ANOVA). The three-way ANOVA can reveal the three-factor interaction effect and the main effect of each factor: the test modes, the mathematics-concepts, and the groups and their specific effect sizes. It is necessary to examine whether mathematics-concepts or groups have the main effects on achievement improvement. A statistical interaction occurs when the effect of one factor on the dependent variable changes depending on the level of another factor. The result cannot provide the information of the two-factor interaction effect on the achievement improvement and their interaction effect size because the counterbalance research design is not the full design. Thus, another three two-factor interaction analyses among the test mode, the mathematics-concepts and the groups are examined. The analyses provide the two-factor interaction effect of the test modes and the mathematics-concepts; that of the mathematics-concepts and the groups, and that of the test modes and the groups and their effect sizes. The analysis also assesses the three-way interaction effects of the test modes, the mathematics-concepts, and the groups to interrupt the achievement improvement and the significant interaction effect size. Consequently, the result from the analysis help to examine whether there are interaction effects apart from the main effect of test modes on the achievement improvement, and the amount and significance of their effect sizes can reveal the practical application.

The results of ANOVA analysis reveal the two specific indices: the statistical significance level of *F statistics* and practical application amount of effect size in terms of the *partial eta squared (Partial* η^2*).*

The cut-off value of partial eta squared (Partial η^2) is identified based on previous studies by Cohen (1969) and Richardson (2011). According to Cohen (1969) and Richardson (2011), the partial eta squared values of 0.0099, 0.0588, and 0.1379 are the benchmarks for small, medium, and large effect sizes, respectively. For the practical application of the experiment, medium and large effect sizes are preferred.

Moreover, three-way ANOVA analysis can provide the three-way interaction effect of three independent factors and their two-way interaction effects. Consequently, it is necessary to identify their statistical significance and effect size, which are larger than the main effect of test modes.

9.4 Results of analyses of variance

This three-way ANOVA provides a three-way interaction effect among independent factors: test modes, mathematical concepts, groups on the achievement improvement. Also, the analysis provides information about the main effects of independent factors: test modes, mathematical concepts, groups on the achievement improvement. However, this three-way ANOVA is not a full factorial design, and it cannot provide information about the three two-way interaction effects among the factors on the achievement improvement. Therefore, the study conducted two-way ANOVAs for the two-way interaction effects among main factors. In addition, **Table 9.2** provides all the results from three-way ANOVA and three two-way ANOVA. The above three-way ANOVA provides the information of the main effects of test mode on their achievement improvement in measuring its effect size. The information is applicable to examine whether the main effect of the test mode, and which test mode is more effective for achievement improvement. In addition, the result from the analysis can examine whether its effect is statistically significant and whether its effect size is large enough in the practical application.

Table 9.2

Results of three-way ANOVA and two-way ANOVAs in the relationship among the test modes, the mathematics-concepts, and the groups

Source	Type III Sum of Squares	*df*	Mean Square	*F*	*Sig.*	Partial η^2
Corrected Model	1349.992[a]	8	168.749	102.590	0.000	0.294
Intercept	3993.265	1	3993.265	2427.694	0.000	0.552
Test modes	1188.776	2	594.388	345.263	0.000	0.259
Mathematics-Concepts	66.022	2	33.011	14.413	0.000	0.014
Groups	65.211	2	32.605	14.234	0.000	0.014
Test modes x Mathematics-concepts	85.424	4	21.356	12.983	0.000	0.026
Test modes x Group	96.005	4	24.001	14.591	0.000	0.029
Mathematics-concepts x Groups	1218.759	4	304.690	185.235	0.000	0.274
Groups x Mathematics-Concepts x Test modes	20.213	2	10.106	6.144	0.002	0.006
Error	3237.124	1968	1.645			
Total	8614.247	1977				
Corrected Total	4587.116	1976				

[a]R^2 Squared = 0.294 (adjusted R^2 = 0.291)

9.4.1 Three-way ANOVA interaction effect among test modes, mathematics-concepts, and groups

Overall, R^2 for the corrected model in the three-way ANOVA analysis was 0.294, indicating that almost 29.4% of the variation was explained by the three-way ANOVA among test mode, mathematics-concepts, and Group model.

The three-way interaction effect among the independent factors, was statistically significant at the 5% level, F (2, 1968) = 6.144, p = 0.002 (see **Table 9.2**). It means the statistical significance of a three-way interaction effect among the independent factors.

The partial eta squared (Partial η^2) is 0.006 (**Table 9.2**). It indicated that the three-way interaction effect explained only 0.6 % of the variance of WLE scores among the factors. That interaction effect was still small effect size because it is much lower than the cut-off value (η^2 = 0.0099, which represents as the small partial eta squared) according to the benchmarks of

Cohen (1969) and Richardson (2011). Consequently, that minimal interaction effect due to the three-way interaction effect among the factors could not reach for the practical application. In addition, it is important to examine an interaction effect on the achievement improvement. **Figure 9.1** shows the mean comparison of the achievement improvement due to the three-way interaction effect of the test -modes, the mathematics-concepts, and the groups (test mode-orders). Group-1 gained the lowest achievement in function concept due to the PPT mode and the highest achievement in remainder concept due to LOT mode. Group-2 gained the lowest achievement in factor concept due to the PPT mode and the higher achievement in function due to the LOT mode and remainder in CAT mode. Group I gained the lowest achievement in remainder in PPT mode and the highest achievement in function concept in CAT mode. To sum up, different groups of students achieve higher in different concepts and LOT and CAT make students achieve more than PPT does.

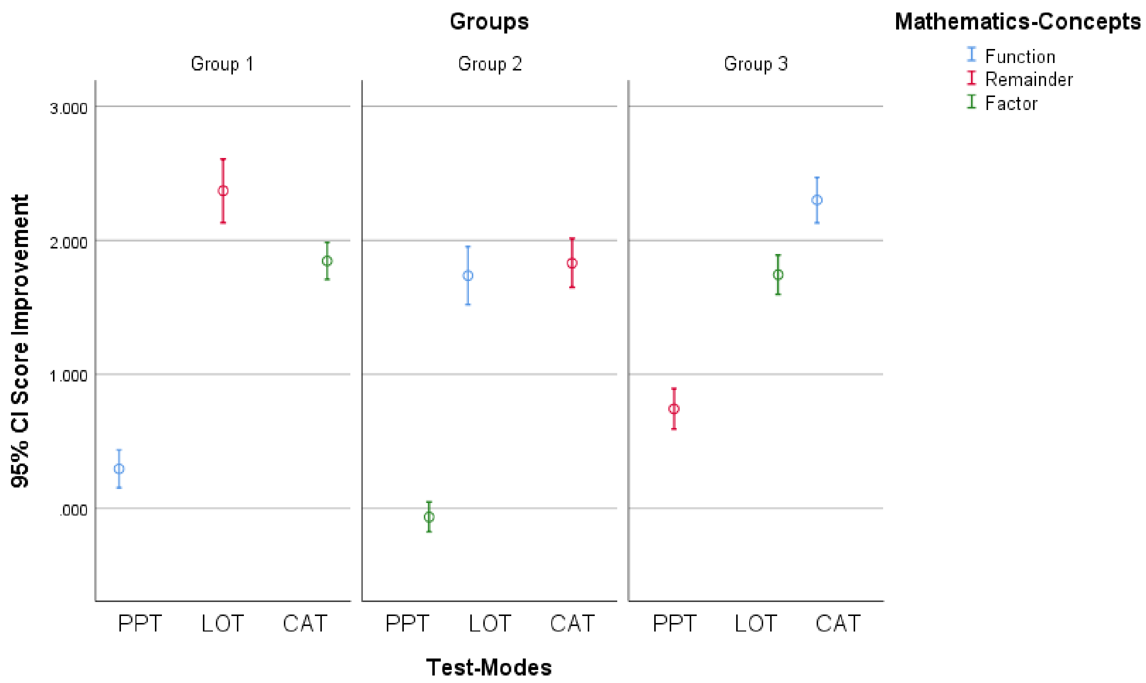

***Figure* 9.1** Error Plots of the mean comparison of the achievement improvement due to the three-way interaction effect of test -modes, mathematics-concepts, and groups

9.4.2 Two-way ANOVA interaction effect between test modes and mathematics-concepts

The provided information about the two-way interaction effect between the test modes and mathematics-concepts was from the three-way ANOVA analysis. To examine whether there is a two-way interaction effect, it needs to identify the significant level of that interaction effect on their achievement improvement and measure its effect size.

In the exploration of the impact of the two-way interaction effect between test modes and the mathematics-concepts on the achievement improvement, there was a statistically significant interaction effect due to the test modes and the mathematics-concepts, F (4, 1968) = 12.983, p = 0.000, reaching statistical significance at the 5% level (**Table 9.2**). It means the statistical significance of the two-way interaction effect between test modes and the mathematics-concepts on the achievement improvement.

Figure 9.2 shows a graph of an average comparison of achievement improvement due to the test modes for the three mathematics-concepts groups. It reveals the achievement in all three concepts; PPT mode is much lower than LOT and CAT modes. The results provide evidence that LOT and CAT modes can help students to achieve more across the three concepts. However, it cannot reveal which one is more effective between LOT and CAT modes. Across the three test modes, achievement in the Factor concept is still lower as compared to those in Remainder and Function concepts. Therefore, it may disclose that Factor concept is more difficult to achieve than the Remainder and Function concepts are.

Due to PPT, the highest gain score is for a reminder, followed by function and factor. The gain score for the remainder is still the highest in LOT, but the gain scores for function and factor are similar. In CAT, the highest gain score is for function, while the gain scores for factor and reminder are similar.

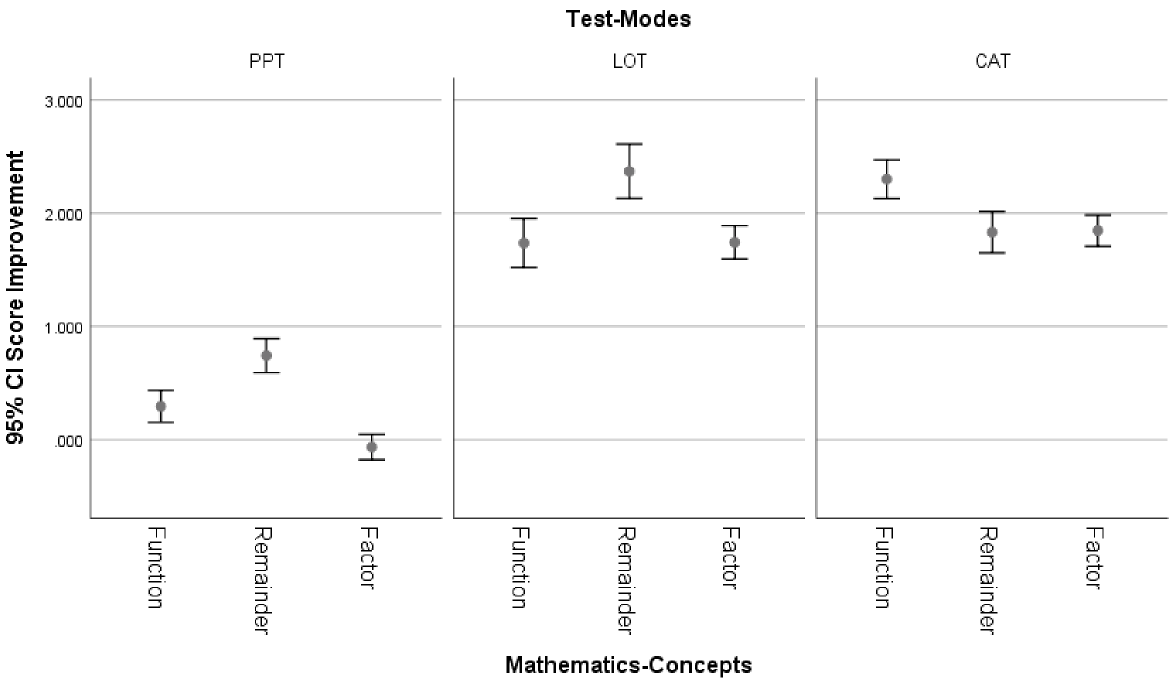

Figure 9.2 Error-plots of the mean comparison of the WLE score improvement due to the interaction effect between the test modes and the mathematics concepts

However, the interaction effect size was described as the partial eta squared (η^2 = 0.026) in **Table 9.2**. The partial eta squared referred that the interaction effect of the test modes and the mathematics-concepts explained only 2.6% of the variance of the WLE score. Although the effect size is statically significant, the interaction effect was considered to be small effect size because it is slightly greater than the cut-off value (η^2 = 0.0099, which represents as the small partial eta squared) (Cohen, 1969; Richardson, 2011). Consequently, there is an interaction effect between test modes and mathematics concepts apart from the main effect of test modes, but its effect size could not reach the practical application.

9.4.3 Two-way ANOVA interaction effect between test modes and groups

Secondly, the two-way ANOVA analysis can also provide information about the two-way interaction effect due to the test modes and groups on the achievement improvement. That information helps to examine whether there is an interaction effect between test modes and groups, apart from the main effect of the test modes on achievement improvement. Consequently, it needs to identify the significant level of that interaction effect on their achievement improvement and measure its effect size.

Table 9.2 shows the results of the two-way interaction effect and was statistically significant, F (4, 1968) = 14.591, p = 0.000. Consequently, the interaction effect reveals the statistical significance at the 5% level. Three error plots mentioned comparing the mean differences in the achievement improvement by the interaction effect between the test modes and the groups, as shown in **Figure 9.3**. In the Group-1 plot, the achievement in LOT and CAT is much higher than that in PPT. The achievement in LOT is slightly higher than that in CAT. Of Group-2, the achievement in LOT and CAT is much higher than that in PPT. The achievement in CAT is somewhat higher than that in LOT. The achievement in CAT mode in Group-3 is the highest, followed by LOT mode. PPT mode is the lowest.

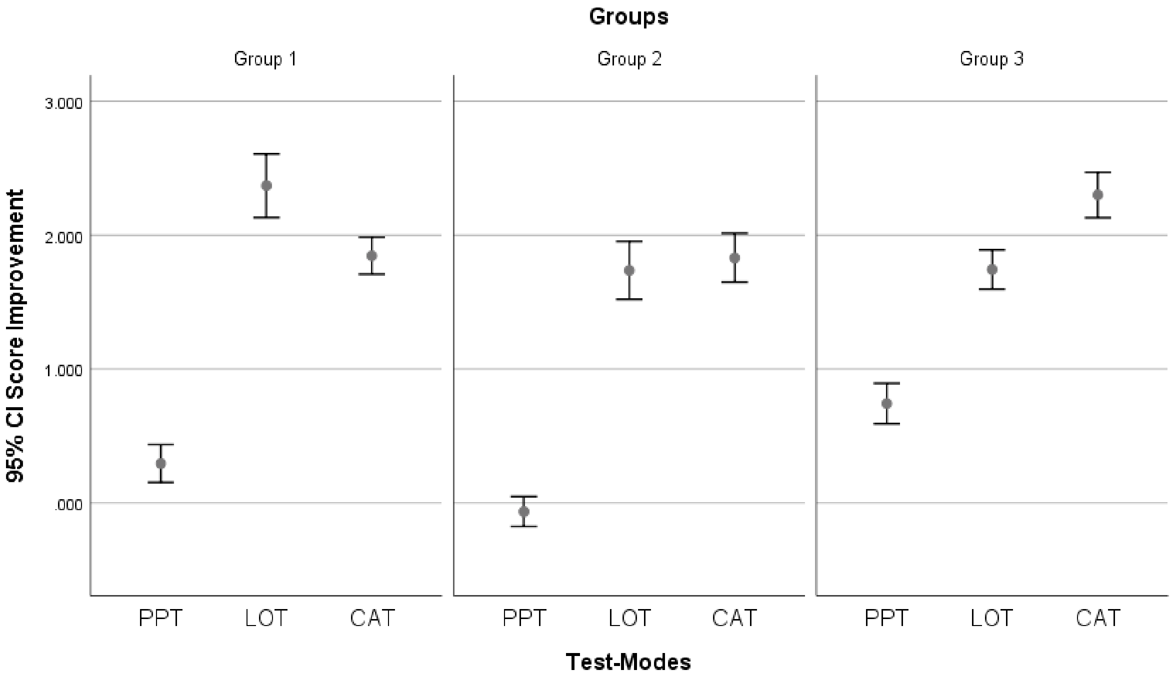

Figure 9.3 Error-plots of the mean comparison of the WLE score improvement due to the interaction effect between the test modes and the groups

However, the amount of effect size can reveal the practical application. For this study, **Table 9.2** exhibits the partial eta squared (η^2) of the two-way interaction effect between the test modes and groups to be 0.029, and the interaction of test modes and the groups explained a 2.9 % of the variance of the WLE score. However, that interaction effect was still small effect size because it is slightly greater than the cut-off value of the small effect (η^2 = 0.0099) and it is much lower than that of the medium effect (η^2 = 0.0588) (Cohen, 1969; Richardson, 2011). Consequently, there is an interaction effect between test modes and groups apart from the main effect due to the test modes, but its effect size could not affect the practical application.

9.4.4 Two-way ANOVA interaction effect between mathematics-concepts and groups

Thirdly, the two-way ANOVA analysis analyses the two-way interaction effect due to the mathematics-concepts and groups on the achievement improvement. That information helps to examine whether that interaction effect could disrupt the main effect of the test modes on the achievement improvement. Consequently, it needs to identify the significant level of that interaction effect on their achievement improvement and measure its effect size.

Table 9.2 shows the effect of the interaction of the mathematics-concepts and the groups on the achievement improvement. The interaction effect was statistically significant, F (4, 1968) = 185.235, p = 0.000 (see **Table 9.2**). Consequently, the interaction effect statistically affects the achievement improvement at the 5% level. The two-way interaction effect depicted the line graph of the mean differences in the WLE score improvement due to the groups for the three mathematics-concepts groups (**Figure 9.4**). In Function Concept, the Group-3 achieved more than other two group. In Remainder Concept, the Group-1 is the highest achievers and in Factor concept, Group-2 is the lowest achiever. Different groups of students achieve the most in different concepts.

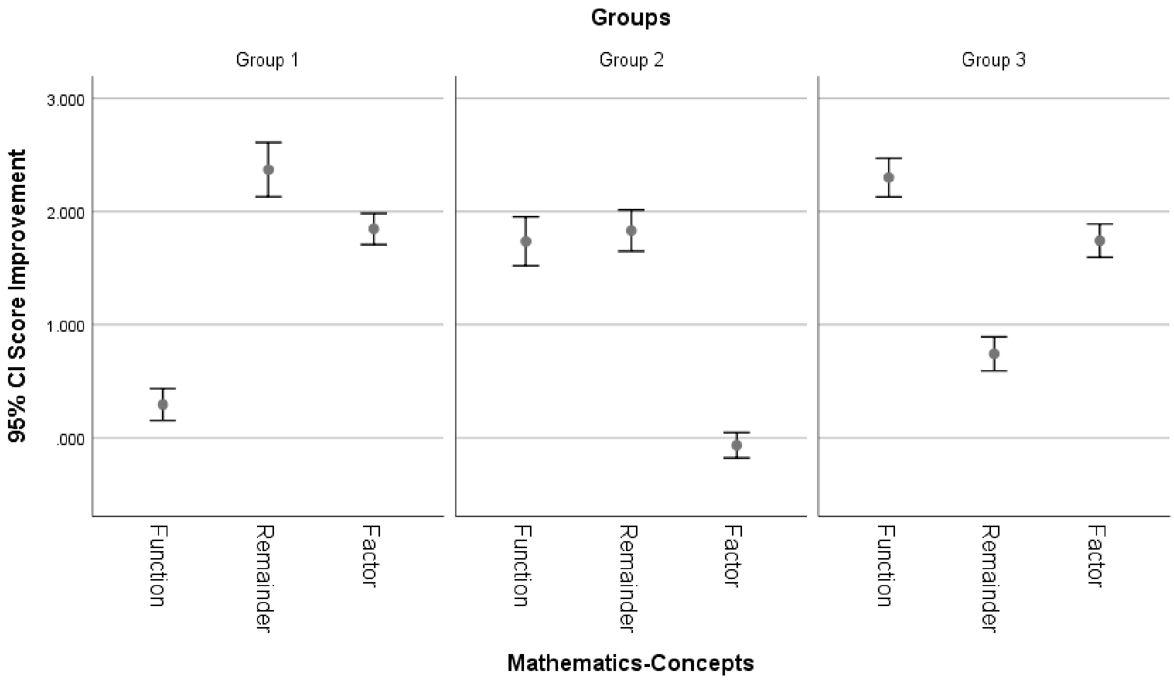

Figure 9.4 Error-plots of the mean comparison of the WLE score improvement due to the interaction effect between the mathematics concepts and the groups

However, the amount of the interaction effect size can reveal the practical application. The partial eta squared is 0.274, indicating that 27.4 % of the variability in the WLE score was attributed to the interaction of the groups and the mathematics-concepts (**Table 9.2**). That interaction effect was significant enough because it is greater than the cut-off value of the significant effect (η^2 = 0.1379) according to the critical values from Cohen, (1969) and Richardson (2011). Comparing with above two-interaction effect, only the interaction effect size could reach the practical application. Consequently, it could affect the achievement improvement because each group has its strength and weakness. For instance, Group-1 has the highest gain score in a remainder; Group-3 has the highest gain score in function. However, Group-2 has the lowest gain score in the factor.

9.4.5 Main effect of mathematics-concepts

The study examined the main effect of mathematics-concepts from the three-way ANOVA analysis and identified the significance level of its main effect and effect size. The information can help to examine whether the main effect affects or influences the mathematics-concepts. The F statistics in the main effect due to the mathematics-concepts, F (2, 1977) = 14.413, p = 0.000, reach statistical significance (**Table 9.2**). Consequently, that main effect on the achievement improvement is statistical significance at the 5% level. After the significant overall F statistics, the study conducted the pairwise differences among the means. As the result of the *post-hoc* comparisons, the mean increased WLE score in the Remainder concept (M = 1.650, SD = 1.618) was significantly different from the one in the Function concept (M = 1.430, SD = 1.576) and that of the Factor concept (M = 1.202, SD = 1.330) (see **Table 9.3**). The mean WLE score improvement in the Factor concept was significantly 0.227 logits lower than those in the Function concept, and it was also significantly 0.448 lower than those in the Remainder concept (**Table 9.4**). Post Hoc comparison results in less difference in the gained WLE logit score in learning the three mathematics- concepts referred to the equivalent ability of three experimental groups.

Table 9.3

Descriptive statistics by score improvement in mathematics-concepts

Score improvement in Mathematics-concepts	Mean	SD	N
Function	1.430	1.576	659
Remainder	1.650	1.618	659
Factor	1.202	1.330	659

Table 9.4

Mean difference between score improvement in mathematics-concepts

Score improvement in Mathematics concepts	Function	Remainder
Remainder	0.220 (0.088) ***	
Factor	-0.227 (0.080) ***	-0.448 (0.082) ***

*** $p < 0.001$

However, its main effect size (Partial η^2) is 0.014 (**Table 9.3**) and indicates that the effect explained only 1.4 % of the variance of WLE scores due to the mathematics-concepts. Further, that interaction effect was still small effect size because it is slighter greater than the cut-off value of the small effect (η^2 = 0.0099) and it is much lower than of the medium effect

(η^2 = 0.0588) (Cohen, 1969; Richardson, 2011). As a result, its main effect size could not reach the effect size of practical application. This will lead to check whether the effect sizes of groups are large enough in the next section.

9.4.6 Main effect of groups

Examining the main effect due to groups from the three-way ANOVA analysis helps to examine whether its effect size could not disrupt the main effect due to the test modes. The main effect for groups, F (2, 1977) = 14.234, p = 0.000, provided as the statistical significance at the 5% level (**Table 9.5**).

After the overall F test statistics were significant, it was necessary to conduct the pairwise differences among the means. As the result of the *post-hoc* comparisons, the mean increased WLE score due to the Group-1 (M = 1.503, SD = 1.625) was significantly 0.336 logits greater than the achievement improvement due to the Groups-2 (M = 1.168, SD = 1.557) (see **Table 9.5** and **Table 9.6**). However, the mean scores due to Group-2 was significantly 0.428 logits lower from the achievement improvement due to Group-3 (M = 1.595, SD = 1.344).

Table 9.5

Descriptive statistics by score improvement in groups

Score improvement in Groups	Mean	SD	N
Group 1	1.503	1.625	681
Group 2	1.168	1.557	630
Group 3	1.595	1.344	666

Table 9.6

Mean difference between score improvement in Groups

Score improvement in Groups	Group 1	Group 2
Group 2	-0.336(0.071) ***	
Group 3	0.092(0.070)	0.428 (0.071) ***

*** p < 0.001

However, the effect size was small (Partial η^2 = 0.014) and indicates that only 1.4 % of the variance of WLE scores was explained by the effect of the groups (**Table 9.2**). Further, that interaction effect was still small effect size because it is slighter greater than the cut-off value of the small effect (η^2 = 0.0099), and it is much lower than the medium effect (η^2 = 0.0588) (Cohen, 1969; Richardson, 2011). Consequently, the main effect size due to the groups could not reach the practical application although the groups significantly affect the achievement improvement.

9.4.7 Main effect of test modes

Finally, the study examined the main effect due to the test modes. The partial eta squared (0.259) indicated its main effect size to be large enough (**Table 9.2**) and indicated that the impact of the test modes explained about 26 % of the variance of gained WLE scores. That main effect was a much large effect size because it is greater than the cut-off value of the large effect (η^2 = 0.1379) (Cohen, 1969; Richardson, 2011). The significant effect size of test mode reached the practical application.

In the main effect of the test modes on the achievement improvement, there was a statistical significance at the 5% level, as the result of F statistics, F (2, 1977) = 345.263, p = 0.000 (**Table 9.2**). Then, the study conducted a follow-up test to evaluate pairwise differences among the means of three test modes and applied a Dunnett C test due to the significance of overall F test statistics, among the methods of post-hoc comparisons. The results of Dunnett C indicated that the mean WLE score due to PPT mode (M = 0.331, SD = 1.077) was significantly different from that of LOT mode (M = 1.957, SD = 1.564) and that of CAT (M = 1.994, SD = 1.249) (see **Table 9.7**).

The mean WLE score improvement due to PPT was significantly 1.626 logit lower than that of LOT, and it was significantly 1.663 logit lower than that of CAT (see **Table 9.8**). There was a 0.038 logit of WLE score between the mean score due to LOT and CATs, but there was no significant difference in mean WLE score due to LOT and CATs as conveyed in **Table 9.8**.

Table 9.7

Descriptive statistics by score improvement due to test modes

Improved WLE score due to Test mode	Mean	SD	N
PPT	0.331	1.077	659
LOT	1.957	1.564	659
CAT	1.994	1.249	659

Table 9.8

Mean difference between score improvement due to test modes

Score improvement due to test mode	PPT	LOT
LOT	1.626 (0.074) ***	
CAT	1.663 (0.064) ***	0.038 (0.078)

*** $p < 0.000$

In summary, the achievement improvement due to CAT is much more significant than that of PPT, but not due to LOT. Therefore, CAT is more effective than PPT, but not LOT for classroom formative assessments in Myanmar high schools. On the other hand, the achievement improvement due to LOT and CATs was significantly higher or greater than that of PPT.

9.5 Summary

In this chapter, from the three-way ANOVA analysis results, the effect size by the three-way interaction effect among the independent factors, was the smallest, (η^2 = 0.006) throughout the analyses. Again, the two-way interaction effect size between the groups and the mathematics-concepts is the largest effect, (η^2 = 0.274), followed by the two-way interaction effect size of other two pairs, between test modes and groups (η^2 = 0.029), and between, test modes and mathematics concepts (η^2 = 0.026). The main effect of test modes is the largest one, (η^2 = 0.259), followed by mathematics concepts (η^2 = 0.014) and groups (η^2 = 0.014). Among the test mode categories, achievement improvement due to CAT could make the most (mean score = 1.994 logit score); the achievement improvement due to LOT was slightly lower (0.038 logit score) than that of CAT; and achievement improvement due to PPT had the smallest (mean score = 0.331 logit score). It can be said that the average score improvement for a student on a test marked out of 100 is +30% for CAT, +27% for LOT, and +8% for PPT, and this across each of the content areas. Students who received immediate and specific scores and feedback from LOT and CAT as their formative assessment improved their mathematics achievement significantly higher than those who receive the delayed score and feedback from PPT. CAT is the most effective for formative assessments in Myanmar high schools. The improved mathematics achievement for LOT is slightly lower than that of CAT. It is also found that groups of students have different ability to achieve in different concepts through the interaction effect size of group and mathematics concepts.

Chapter 10

Impacts on Achievement Improvement

10.1 Introduction

This chapter is concerned with the research question: "*What are the factors that significantly influence student mathematics achievement progress across the application of formative assessment in the different test modes*?" The results of the student-level structural equation models (SEMs) examined the relationships among the scales to answer this research question. There are three outcome measures in this study, which include the *score improvement due to (1) PPT, (2) LOT, (3) CAT*. For each of the outcome measures, this chapter examines and reports on the separated SEMs.

The hypothesised models at the student-level were analysed using M-plus (v.8). This investigation took an exploratory approach and included all possible paths. The researcher trimmed the models by removing insignificant paths based on the initial results. **Chapter 4** described these procedures for the single-level models. For each model, this chapter presents the scales included in the model first, followed by the hypothesised student-level model. First, the study presents the results of its structural models followed by the fit indices for the model. The original item responses or raw scores are only applied to demographic scales such as *gender, expected education, parents' education.* The study used the Weighted Likelihood Estimates (WLE) score for the remaining scales. The WLE scores have already estimated during the validation of the scales using the Rasch Model. The study adopted this approach to simplify the SEM model. **Chapter 7** confirmed some scales, such as *attitude towards mathematics learning* and *self-efficacy*, to measure only one construct. Others measure two correlated constructs, such as *motivation, attitude towards formative assessment, ICT familiarity, and attitude towards test modes (PPT, LOT, and CAT). Attitude towards ICT is in* hierarchical constructs. A single set of WLE scores represents the scales with one-factor structure and are consequently treated as observed scales in the proposed SEM models. The scales with a hierarchical structure are included in the model as latent scales, with a set of indicators. The WLE scores for the corresponding sub-scales represent the scales.

10.2 Scales in student-level models

The models for student–level factors aim to investigate the impact of the hypothesised scales on the size of students' improvement in their mathematics achievement. It is hypothesised that this score improvement is due to the application of a test mode (in the experiment, there are three test modes applied: PPT, LOT, and CAT) as the classroom formative assessment in this study. There are 6 latent (or unobserved) scales and 18 manifest (or observed) scales included in each of the four SEM models at the student-level.

The four demographic scales included in this study are *Students' gender (gen), Fathers' Education Level (faedu), Mothers' Education Level (moedu), expected education level (expedu)*. F*athers' Education Level and Mothers' Education Level* reflect a latent scale labelled *parents' education level (predu)*. In addition, scales reflected as a one-factor structure such as *attitude towards mathematics learning (attm) and self-efficacy (selfeffi)* are treated as manifest scales using their WLE scores.

Scales with a hierarchical structure involve *motivation, attitude towards formative assessment (attfa), ICT familiarity (ictfam), attitude towards ICT (attict), and Attitude towards Test modes (attm)*. Such scales are composed of at least two sub-scales, their first-order latent scales. A latent scale, *motivation (Moti)*, is reflected by two manifest scales: *intrinsic motivation (intrim)* and *extrinsic motivation (extrim)* using their estimated WLE score. *Attitude towards Formative tests (attfass) and Attitude towards Formative Feedbacks (attfafb)* reflect the *attitude towards formative assessment (attfa)*, that is, second latent scale.

There are two latent scales related to ICT, the *ICT familiarity (ictfam)* and the *attitude towards ICT (attict)*. Two manifest scales, *Accessibility to ICT Devices (ictacc)* and *Usage of ICT Devices (ictuse)* reflect the scale of *ictfam*. Their WLE scores estimated using the Rasch modelling procedure for these two sub-scales. Three manifest scales also reflect another latent scale, the attict: *Affective Domain in attitude towards ICT (attaff), Cognitive Domain in attitude towards ICT (attcog)*, and *Behavioural Domain in attitude towards ICT (attbeh)*.

The *attitude towards test mode (attmode)* is one of the essential latent scales in this study. The *attmode* is the formation of the *Attitude towards Test Technique (atttest)* and *Attitude towards Feedback from the Test Technique (attfb)*. For PPT, *attppt* is for the *attitude towards PPT*; *attlot* is for the *attitude towards LOT*, and *attcat* is for the *attitude towards CAT*.

There are three main targeted outcome scales. The scales are achievement improvement due to receiving feedback through PPT, LOT, and CAT (*ppt_ws, lot_ws,* and *cat_ws*). **Table 10.1** shows the list of these latent and manifest scales.

Table 10.1

Latent and manifest scales in student-level model

Latent (Unobserved Scales)	Description	Manifest (Observed) Scales	Description	Code
		gen	Gender	0 = Male 1 = Female
		expedu	Expected Education Level	0 = High School Graduate 1 = Diploma's Degree 2 = Bachelor's Degree 3 = Master's degree 4 = Doctoral Degree
predu	Parents' Education Level	faedu	Father Education Level	0= Less than High School 1 = Some High School 2 = High School Graduate 3 = Diploma's Degree 4 = Bachelor's Degree 5 = Master's degree 6 = Doctoral Degree
		moedu	Mother Education Level	
		attm	Attitude towards mathematics learning	WLE scores
		selfeffi	Self-Efficacy	WLE scores
moti	Motivation	intrm	Intrinsic Motivation	WLE scores
		extrm	Extrinsic Motivation	
attfa	Attitude towards Formative Assessment	attfaass	Attitude towards Formative tests	WLE scores
		attfafb	Attitude towards Formative Feedback	
ictfam	ICT Familiarity	ictacc	Accessibility to ICT	WLE scores
		ictuse	Usage of ICT devices	
attict	Attitude towards ICT	attaff	Affective Domain in Attitude towards ICT	WLE scores
		attcog	Cognitive Domain in	

Latent (Unobserved Scales)	Description	Manifest (Observed) Scales	Description	Code
			Attitude towards ICT	
		attbeh	Behavioural Domain in Attitude towards ICT	
attppt	Attitude Towards PPT	atttest	Attitude Towards PPT Technique	WLE scores
		attfb	Attitude Towards Feedback from PPT	
		g_ppt	Achievement improvement due to PPT	WLE scores
attlot	Attitude Towards LOT	atttest	Attitude Towards LOT Technique	WLE scores
		attfb	Attitude Towards Feedback from LOT	
		g_lot	Achievement improvement due to LOT	WLE scores
attcat	Attitude Towards CAT	atttest	Attitude Towards CAT Technique	WLE scores
		attfb	Attitude Towards Feedback from CAT	
		g_cat	Achievement improvement due to CAT	WLE scores

10.3 Hypothesised model

This study derived the general hypothesised model from the theoretical framework presented in **Chapter 2**. The hypothesised model has three stages: the presage, process, and product stages. This chapter hypothesised the scales at the presage stage as exogenous scales because other scales do not influence these scales in the model. The other scales in the process and product stages are assumed (to be) endogenous scales because they may interact in the process and product stages. In addition, the scales from the presage stage may influence them. The graphical description of the hypothesised model (**Figure 10.1**)

In the theoretical framework of the main study, the exogenous scales in the presage are students' demographic characteristics scales. The demographic characteristics scales are *gender (gen), Fathers' Education Level (faedu), Mothers' Education Level (moedu)* and *expected education level (expedu)*. The scales from the presage stage affect the scales in the process stage. In the process stage, these scales involve their psychological factors towards mathematics learning, their psychological factors towards formative assessment, ICT-related scales, and the specific factors towards test modes (PPT, LOT, and CAT). Their psychological factors towards mathematics learning involve *attitude towards mathematics learning (attm),* and *self-efficacy (selfeff),* and *motivation (moti)*. The psychological factors towards formative assessment are the *attitude towards formative assessment (attfa)*. *ICT familiarity (ICTFAM)* and the *attitude towards ICT (attict)* are the two scales related to ICT. Finally, the psychological factors related to the test modes (PPT, LOT, and CAT) are the attitude towards each test mode (generally, *attmod*; specifically, *attppt*, *attlot*, and *attcat*), respectively.

The product stage has three outcome scales (*g_ppt, g_lot, g_cat*). They are for the mathematics score improvements due to applying three different test modes (PPT, LOT, and CAT). This study viewed the outcome scales as endogenous scales since scales influence them from the presage and process stages in the models and compares effects on different test modes. Consequently, the following sections will explore three SEM models. Further, the researcher used the student-level model to examine the causal relationships among the student-level factors discussed earlier. Finally, the structural model examines the strength of the relationships among the latent scales and manifest scales included in this study. The following sections show the results of the structural models.

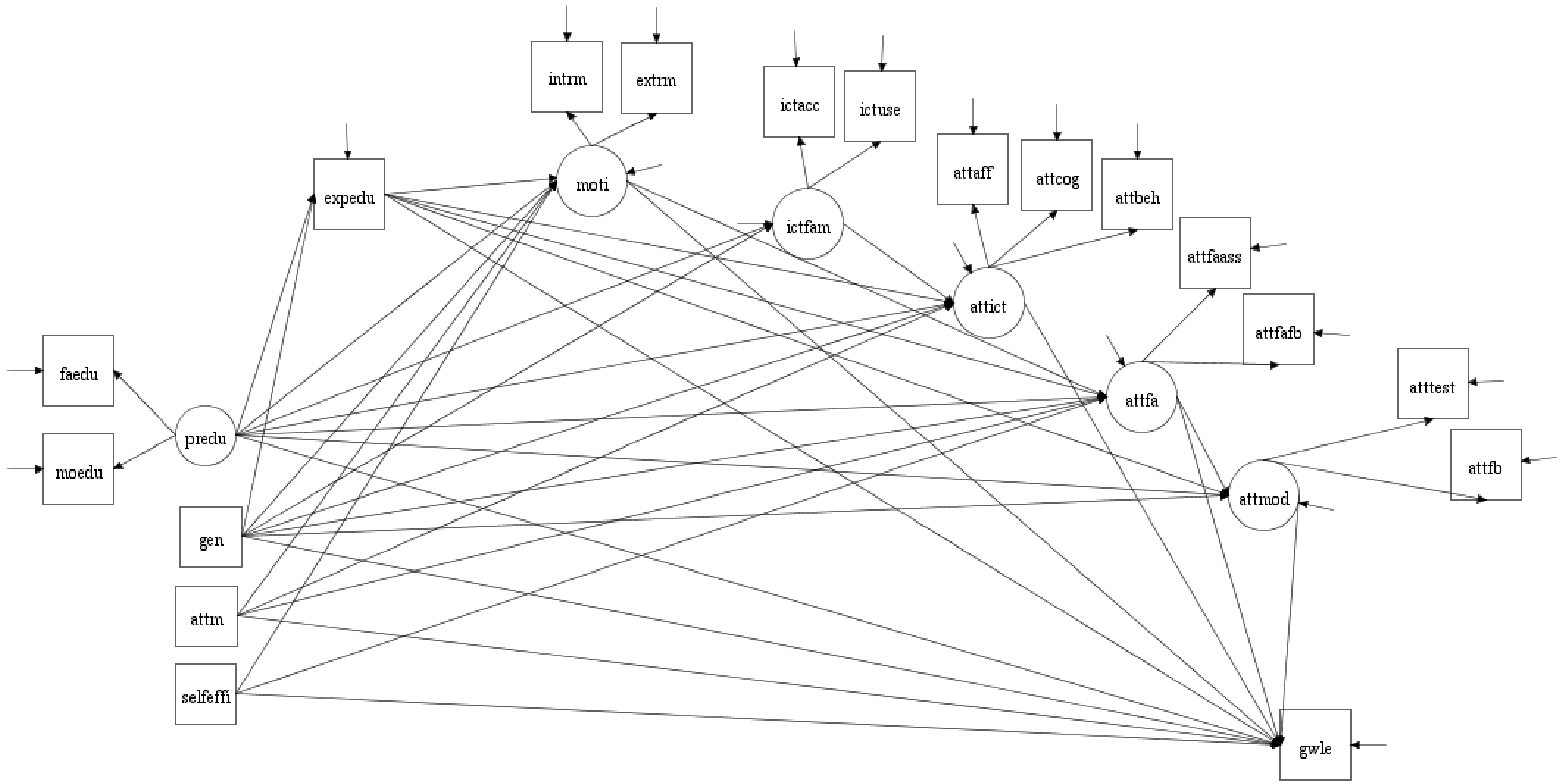

Figure 10.1 Hypothesised model at student-level

10.4 Results of Student_PPT model

This structural equation model at the student-level (Student_PPT model) is examined to explore the factor influence on the achievement improvement due to the application of PPT. The results in the structural equation model and the major outcomes and their predictors and the model are in **Table 10.2** based on fit indices. **Figure 10.2** shows the graphical diagram of this structural equation model.

Table 10.2

Results of Student_PPT model

Outcome(s)	Predictor(s)	β	se	C.R	p
expedu	gen	0.138	0.035	3.922	0.000
	predu	0.445	0.037	12.149	0.000
moti	predu	-0.137	0.038	-3.573	0.000
	attm	0.469	0.034	13.880	0.000
	selfeffi	0.588	0.030	19.332	0.000
ictfam	gen	-0.200	0.037	-5.343	0.000
	predu	0.192	0.042	4.569	0.000
attict	gen	-0.243	0.040	-6.072	0.000
	predu	0.176	0.052	3.373	0.001
	attm	0.148	0.041	3.644	0.000
	ictfam	0.227	0.043	5.330	0.000
attfa	expedu	0.132	0.053	2.510	0.012
	selfeffi	-0.368	0.139	-2.648	0.008
	moti	0.830	0.212	3.914	0.000
g_ppt	gen	0.090	0.039	2.270	0.023
	predu	-0.178	0.055	-3.257	0.001
	attppt	0.359	0.034	10.423	0.000

Note: b = unstandardised estimate; β = standardised estimate

se = standard error; C.R = Critical Ratio (C.R) of standardised estimate and standard error

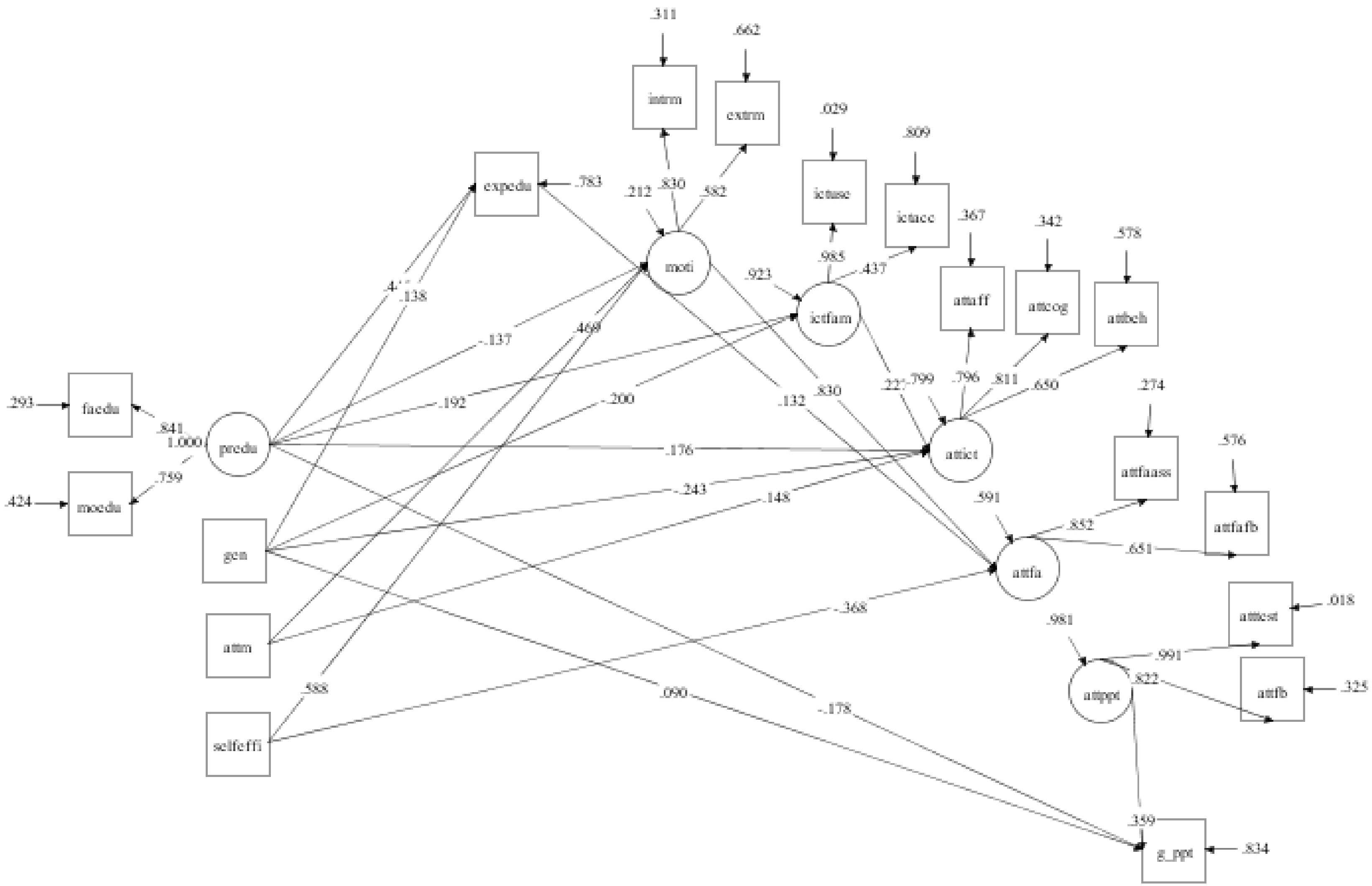

Figure 10.2 Student_PPT Model

10.4.1 Expected education level (expedu)

The scales that directly affect students' *expected education level* are their *gender* status (β = 0.138) and their *parents' education level* (β = 0.455). The positive path coefficient of their *parents' education level* indicates that a student whose parents have a higher education level expects to achieve a higher education level. The direct positive path coefficient of the *gender* on their *expected education level* indicates that girls expect to accomplish a higher education level than boys do.

10.4.2 Motivation (moti)

There are three scales, namely *attitude towards mathematics learning* (β = 0.469), *self-efficacy* (β = 0.588), and *parents' education level* (β = -0.138), having a direct effect on *motivation*. The positive path coefficients indicate that *attitude towards mathematics learning* and *self-efficacy* positively affect *motivation*. The result shows that students with a more positive *attitude towards mathematics learning* tend to have a higher *motivation* to mathematics. The higher their *attitude towards mathematics learning* is, the more motivated they are to mathematics. Another scale that has a positive direct effect on *motivation* is *self-efficacy*. Students who have more confidence in mathematics learning are more motivated to do mathematics. The negative path coefficient indicates that students whose parents possess a higher level of education are less motivated to study mathematics. Their parental involvement affects the achievement more than the *parents' education level*.

10.4.3 ICT familiarity (ictfam)

A scale with a negative direct effect on *ICT familiarity* is *gender* status (β = -0.192). The result shows that boys tend to access ICT and use ICT more than girls do. Another scale directly affects *ICT familiarity* is their *parents' education level* (β = 0.149). Students whose parents possess a higher level of education are more familiar with ICT.

10.4.4 Attitude towards ICT (attict)

In **Table 10.2**, four scales directly affect the *attitude towards ICT*. These scales are *gender* (β = -0.245), *parents' education level* (β = 0.178), *ICT familiarity* (β = 0.234), and *attitude towards mathematics learning* (β = 0.151). The positive path coefficients indicate that *ICT familiarity, parents' education level,* and *attitude towards mathematics learning* positively affect their *attitude towards ICT*.

This study shows that *ICT familiarity* positively influences the *attitude towards ICT*. The more students are familiar with ICT, the more positive their *attitude to ICT*. Hence, it is suggested that schools need to support students access ICT to become more familiar. In addition, the *attitude towards mathematics learning* indicates a positive path coefficient on the *attitude towards ICT*. This means that students who have a positive *attitude towards mathematics learning* are more likely to have a positive *attitude towards ICT*.

Moreover, students whose parents have higher education levels are more likely to have a positive *attitude towards ICT*. The negative path coefficient indicates that their gender status is a negative path coefficient to their *attitude towards ICT*. This almost implies that the boys tend to have a more positive *attitude towards ICT* than girls do.

10.4.5 Attitude towards formative assessment (attfa)

There are three scales found to have a significant direct effect on the *attitude towards formative assessment*. These scales are the *expected education level* with the path coefficient (β = 0.129), the *self-efficacy* (β = - 0.369) and *motivation* (β = 0.833).

The positive path coefficient indicates that their *expected education level* and *motivation* positively affect their *attitude towards formative assessment*. The result shows that students with the expectation of a higher education level tend to have a more positive *attitude towards formative assessment*. Also, the more motivated they are learning, the higher their *attitude towards formative assessment*.

There is a negative path coefficient of *self-efficacy* influencing *attitude towards formative assessment*. Students who believe in their capacity to execute their attainment tend to have a less positive *attitude towards formative assessment*. However, those students are good at learning mathematics; consequently, they rarely need the help of classroom formative assessment.

10.4.6 Achievement improvement due to PPT (g_ppt)

Table 10.2 and **Figure 10.2** show that three scales directly affect *score improvement due to PPT*. These scales are *parents' education level* (β = - 0.186), *gender* (β = 0.092), *attitude towards PPT* (β = 0.357).

The positive path coefficients indicate that *gender* and *attitude towards PPT* positively impact their improvement in mathematics due to PPT. The direct positive path coefficient of *gender* on the achievement improvement indicates significant differences between boys and girls on the score improvement. The result shows that girls tend to improve more in their achievement due to receiving formative feedback through PPT. Likewise, the positive direct effect of *attitude towards PPT* indicates that the more positive attitudes towards PPT they are likely to, the more likely they improve in mathematics due to receiving formative feedback through PPT. However, this study found *parents' education level* to have a negative path coefficient on the gained WLE in mathematics due to receiving formative feedback through PPT. This means that their *parents with higher education level* negatively affect students' mathematics score improvement due to receiving formative feedback through PPT. One reason for this finding is that parents who have lower level of education tend to have higher value on educational achievement.

10.4.7 Fit indices of Student_PPT model

This section is concerned with the fit indices obtained at the student-level model of improving mathematics achievement for the PPT experiment (Student_PPT Model). **Table 10.3** shows the model fit indices, which identify how well the student_PPT model fits the data. The model obtains Chi-square value (CMIN) of = 442.972 (df = 104) with a corresponding *p*-value of 0.000. The Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and the Standardised Root Mean Residual (SRMR) values are 0.868, 0.908, 0.070, and 0.064, respectively. The final model is relatively good because the TLI and CFI values are close to 1, and the SRMR and the RMSEA are less than 0.080 and are close to 0.050 (Byrne, 2010). Consequently, the final model indicates an acceptable model fit.

Table 10.3

Fit Indices of Student_PPT Model

Model	CMIN	df	CMIN/df	*p*	TLI	CFI	RMSEA	SRMR
Student_PPT model	442.972	104	4.259	0.000	0.868	0.908	0.070	0.064

10.5 Results of Student_LOT model

This study examined the structural equation model at the student-level to explore the factor influencing the achievement improvement due to applying the feedback through LOT. The results in the structural equation model and the main outcomes and their predictors and the model are in **Table 10.4** based on fit indices. The graphical diagram of this structural equation model is in **Figure 10.3**.

Table 10.4

Results of Student_LOT model

Outcome(s)	Predictor(s)	β	se	C.R	p
expedu	gen	0.138	0.035	3.923	0.000
	predu	0.444	0.037	12.046	0.000
moti	predu	-0.138	0.038	-3.607	0.000
	attm	0.469	0.034	13.886	0.000
	selfeffi	0.589	0.030	19.353	0.000
ictfam	gen	-0.200	0.037	-5.346	0.000
	predu	0.193	0.042	4.615	0.000
attict	gen	-0.243	0.040	-6.071	0.000
	predu	0.176	0.052	3.389	0.001
	attm	0.150	0.041	3.693	0.000
	ictfam	0.224	0.043	5.253	0.000
attfa	expedu	0.133	0.052	2.543	0.011
	selfeffi	-0.368	0.139	-2.644	0.008
	moti	0.831	0.212	3.911	0.000
attlot	predu	-0.164	0.052	-3.124	0.002
	attict	0.104	0.050	2.074	0.038
g_lot	attlot	0.283	0.037	7.569	0.000

Note: b = unstandardised estimate; β = standardised estimate

se = standard error; C.R = Critical Ratio (C.R) of standardised estimate and standard error

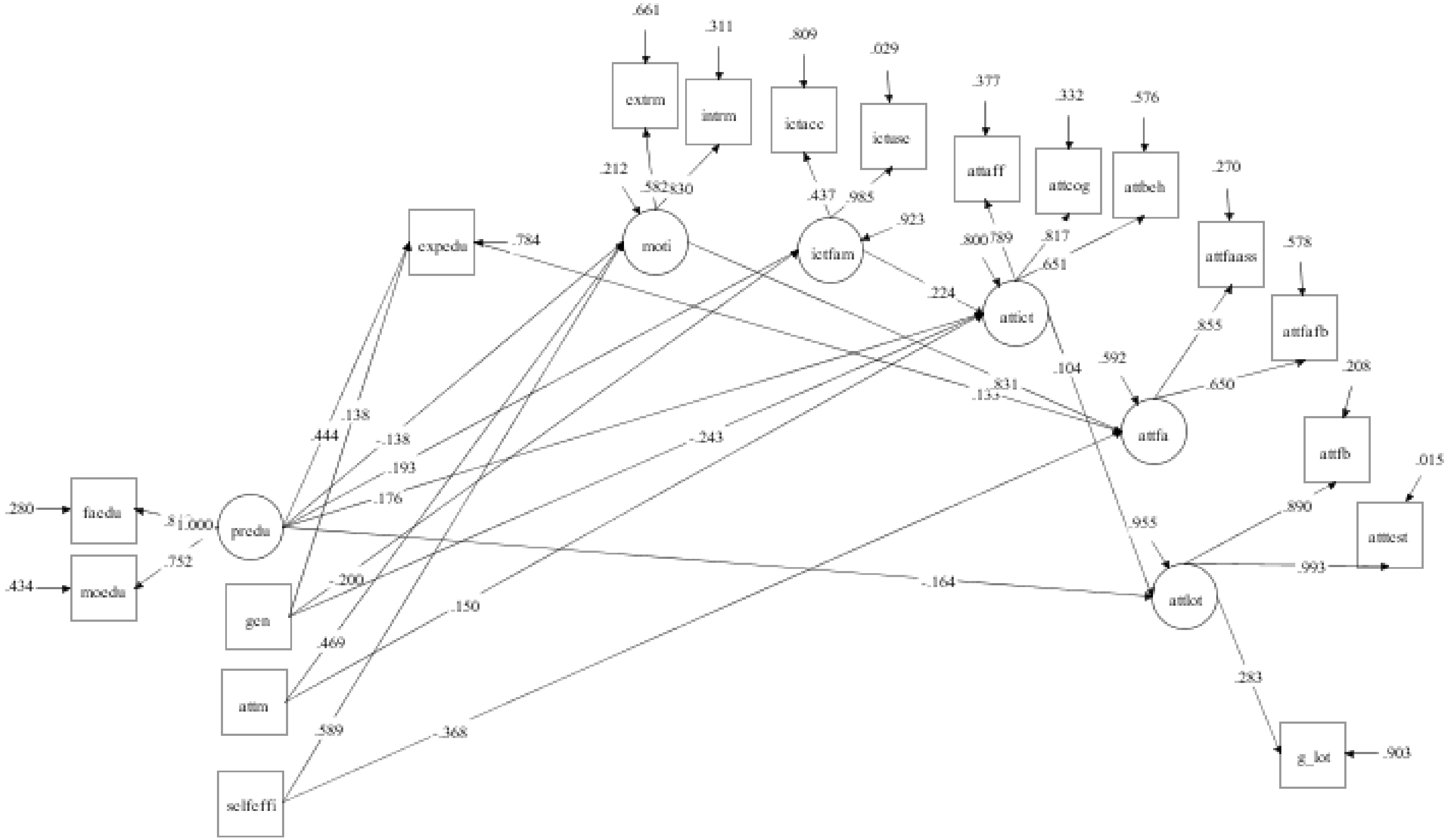

Figure 10.3 Student_LOT model

10.5.1 Factors influencing outcome scales

This Student_LOT Model is analysed whether any predicators, directly and indirectly, affect the seven outcome scales. This Student_LOT Model provides results for the first five outcome scales similar to the student_PPT model. There are two scales on the *expected education level,* which are *gender* status (β = 0.138), and *parents' education level* (β = 0.444). There are three scales on the *motivation*, which are *parents' education level* (β = -0.138), *attitude towards mathematics learning* (β = 0.469), and *self-efficacy* (β = 0.589). Two scales which are *gender* (β = -0.200) and *parents' education level* (β = 0.193) on *ICT familiarity*. On the *attitude towards ICT*, there are four scales which are *gender* (β = -0.243), *parents' education level* (β = 0.176), *attitude towards mathematics learning* (β = 0.150), and *ICT familiarity* (β = 0.224). Three scales which have direct effect on *attitude towards formative assessment* are *expected education level* (β = 0.133), *self-efficacy* (β = -0.368), and *motivation* (β = 0.831). The results of the last two scales: *attitude towards LOT* (attlot) and *achievement improvement due to LOT* (g_lot) are described as follows.

10.5.2 Attitude towards LOT (attlot)

In **Table 10.4**, there are two scales found to have direct effects on the *attitude towards LOT*. These scales are *parents' education level* (β = -0.164), and *attitude towards ICT* (β = 0.104). The path coefficients indicate that their *parents' education level* negatively affects their *attitude towards LOT*. Students whose parents have a lower education level have a more positive *attitude towards LOT*. The *attitude towards ICT* has a positive path coefficient relative to the *attitude towards LOT*. It reveals that students with a more positive *attitude towards ICT* are more likely to have a more positive *attitude towards LOT*.

10.5.3 Achievement improvement due to LOT (g_lot)

This aspect of study found only a scale that directly affect the score improvement due to receiving formative feedback through LOT. The scale is *attitude towards LOT* (β = 0.283). The positive path coefficient indicates that their *attitude towards LOT* positively impacts their mathematics improvement due to receiving formative feedback through LOT. Hence, the more positive *attitude towards LOT*, the more achievement due to receiving formative feedback through LOT.

10.5.4 Fit indices of Student_LOT model

This section is concerned with the fit indices obtained at the student-level model of improving mathematics achievement for the LOT experiment (Student_LOT Model). **Table 10.5** shows the model fit indices, which identify how well the student_LOT model fits the data. This model obtains CMIN of = 399.050 (df = 104) with a corresponding *p*-value of 0.000. The TLI, CFI, RMSEA, and SRMR values are 0.891, 0.925, 0.066, and 0.065, respectively. The final model is relatively good because the TLI and CFI values are close to 1, and the SRMR and the RMSEA are less than 0.080 and are close to 0.050 (Byrne, 2010). Consequently, the final model indicates an acceptable model fit.

Table 10.5

Fit indices of Student_LOT model

Model	CMIN	df	CMIN/df	*p*	TLI	CFI	RMSEA	SRMR
Student_LOT model	399.050	104	3.837	0.000	0.891	0.925	0.066	0.065

10.6 Results of Student_CAT Model

This study examined a structural equation model at the student-level to explore the factor influencing the achievement improvement due to applying the feedback through CAT. The results in the structural equation model and the major outcomes and their predictors and the model are in **Table 10.6** based on fit indices. The graphical diagram of this structural equation model is in **Figure 10.4**.

Table 10.6

Results of Student_CAT model

Outcome(s)	Predictor(s)	β	se	C.R	p
expedu	gen	0.138	0.035	3.923	0.000
	predu	0.442	0.037	11.933	0.000
moti	predu	-0.137	0.038	-3.589	0.000
	attm	0.469	0.034	13.874	0.000
	selfeffi	0.588	0.030	19.335	0.000
ictfam	gen	-0.200	0.037	-5.344	0.000
	predu	0.192	0.042	4.593	0.000
attict	gen	-0.243	0.040	-6.081	0.000
	predu	0.177	0.052	3.424	0.001
	attm	0.147	0.041	3.611	0.000
	ictfam	0.223	0.043	5.248	0.000
attfa	expedu	0.131	0.052	2.530	0.011
	selfeffi	-0.362	0.138	-2.631	0.009
	moti	0.825	0.210	3.926	0.000
attcat	ictfam	0.128	0.043	2.988	0.003
	attict	0.131	0.051	2.585	0.010
g_cat	attfa	0.170	0.070	2.410	0.016
	attict	0.112	0.049	2.304	0.021
	attcat	-0.084	0.040	-2.132	0.033

Note

b = unstandardised estimate; β = standardised estimate

se = standard error; C.R = Critical Ratio (C.R) of standardised estimate and standard error

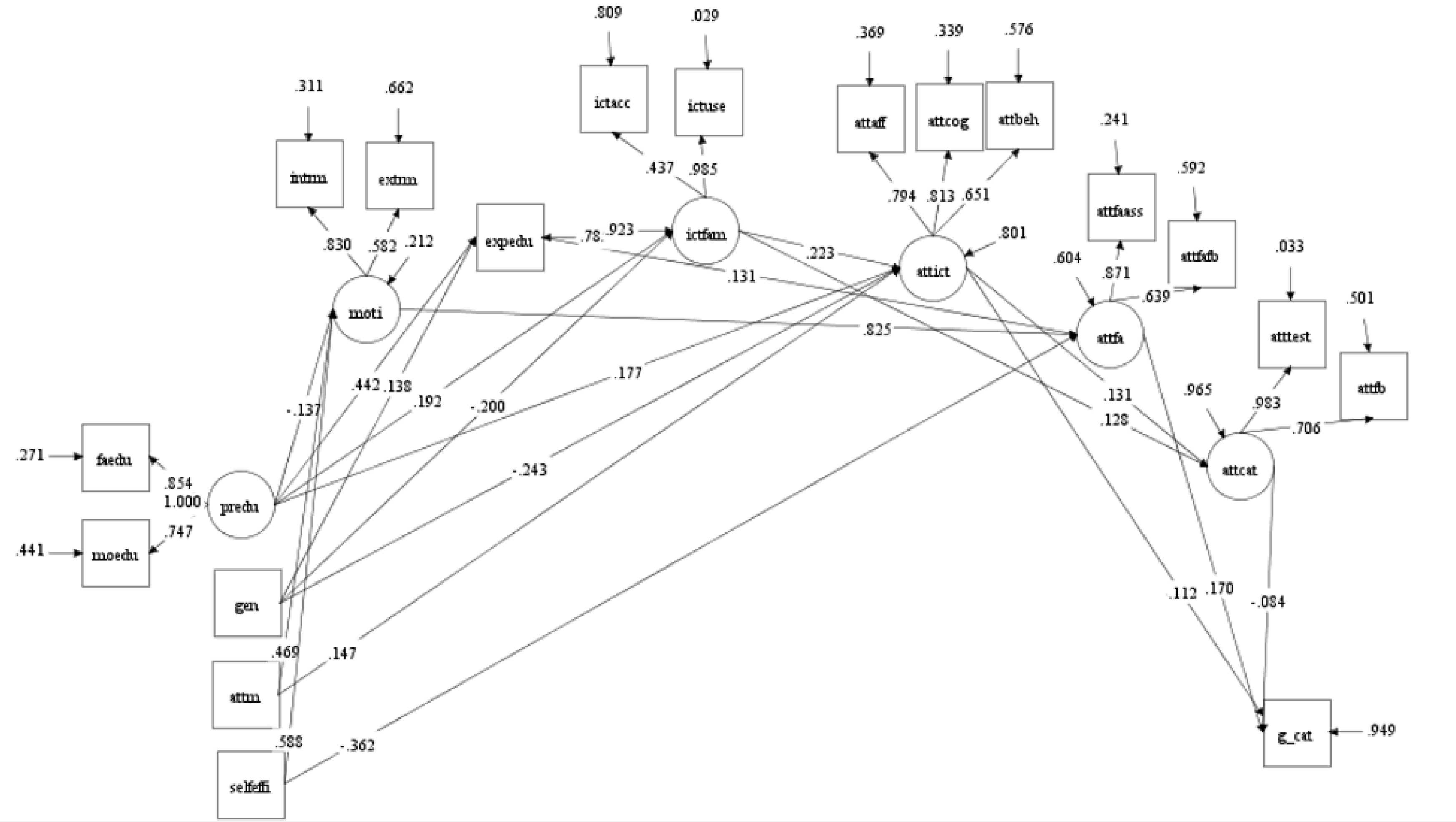

Figure 10.4 Student_CAT model

10.6.1 Factors influencing outcome scales.

This study explored a structural equation model (Student_CAT model) to determine which predictors directly affect seven outcome scales. This Student_CAT model can provide the results for the first five outcome scales similar to Student_PPT model. There are two scales on the *expected education level*, which are *gender* status (β = 0.138), and *parents' education level* (β = 0.442). There are three scales on the *motivation*, which are *parents' education level* (β = -0.137), *attitude towards mathematics learning* (β = 0.469) and *self-efficacy* (β = 0.588). Two scales which are *gender* (β = -0.200) and *parents' education level* (β = 0.192) on *ICT familiarity*. On the *attitude towards ICT*, there are four scales which are *gender* (β = -0.243), *parents' education level* (β = 0.177), *attitude towards mathematics learning* (β = 0.147), and *ICT familiarity* (β = 0.223). Three scales which have direct effect on *attitude towards formative assessment* are *expected education level* (β = 0.131), *self-efficacy* (β = -0.362), and *motivation* (β = 0.825). The last two scales, *attitude towards CAT* (attcat) and score improvement due to CAT (G_CAT), are described.

10.6.2 Attitude towards CAT (attcat)

In **Table 10.6**, two scales directly affect the *attitude towards CAT*. These scales are *ICT familiarity* (β = 0.128), and *attitude towards ICT* (β = 0.131). The positive path coefficient of *ICT familiarity* suggests that students with more *ICT familiarity* have a more positive *attitude towards CAT*. The *attitude towards ICT* has a positive path coefficient on the *attitude towards CAT*. It reveals that a more positive *attitude towards ICT* is more likely to be associated with a more positive *attitude towards CAT*. In addition, students are not familiar with CAT, and the nature of CAT is more complicated.

10.6.3 Achievement improvement due to CAT (g_cat)

Three scales directly affect the score improvement due to receiving formative feedback through CAT. These scales are *attitude towards formative assessment* (β = 0.170), *attitude towards ICT* (β = 0.112) and *attitude towards CAT* (β = -0.084).

The *attitude towards formative assessment* positively affects the achievement due to receiving formative feedback through CAT and shows that students who have a more positive *attitude towards formative assessment* achieve more due to receiving formative feedback through CAT. The *attitude towards ICT* positively impacts the achievement improvement due to CAT. Students with a more positive *attitude towards ICT* achieve more due to receiving formative feedback through CAT. On the other hand, the *attitude towards CAT*, negatively impacts their improvement in mathematics due to receiving formative feedback through CAT. This means that the more positive their *attitude towards CAT*, the less their achievement improvement is due to receiving formative feedback through CAT. By the nature of CAT, students with less performance take a short time to take CAT of classroom formative assessment, and students with more performance take longer to take CAT.

10.6.4 Fit indices of Student_CAT model

This section is concerned with the fit indices obtained at the student-level model of the improvement in mathematics achievement for CAT experiment (Student_CAT model). **Table 10.7** shows the model fit indices, which identify how well the Student_CAT model fits the data. The model obtains CMIN of = 409.997 (df = 104) with a corresponding *p*-value of 0.000. The TLI, CFI, RMSEA, and SRMR values are 0.867, 0.908, 0.067, and 0.065, respectively. The final model is relatively good because the TLI and CFI values are close to 1, and the SRMR and the RMSEA are less than 0.08 and are close to 0.05 (Byrne, 2010). Consequently, the final model indicates an acceptable model fit.

Table 10.7

Fit indices of Student_CAT model

Model	CMIN	df	CMIN/df	*p*	TLI	CFI	RMSEA	SRMR
Student_CAT model	409.997	104	3.9423	0.000	0.867	0.908	0.067	0.065

10.7 Summary

This chapter provides the structural equation modelling for the student-level and examines the relationships of the student-level scales separately using the MPlus (v.8) statistical software package. The student-level structural model shows how students' *gender, parents' education level, expected education, attitudes towards learning mathematics, motivation, self-efficacy, attitude towards formative assessment, ICT familiarity, attitude towards ICT, attitude towards the test modes (PPT, LOT, and CAT)* relate to their achievement improvement due to the test modes. According to the nature of the study, there are three SEM models due to the test mode: Student_PPT model, Student_LOT model, and Student_CAT model.

In all three models, there are same factors that positively or negatively influence on expected education, motivation, ICT familiarity, and attitudes towards ICT and formative assessment. Girls expected higher education level than boys do. Students' parent education levels positively affect their expected education level. Their attitude towards mathematics and self-efficacy positively related to their motivation. Only factor which negatively affect their motivation is their parents' education level. Boys have higher ICT familiarity than girls do in this sample. Parents with higher education level are more likely to encourage their children to be more familiar with ICT. Boys have the more positive attitude towards ICT. Students' parent education level positively relates to attitude towards ICT. The more familiar with ICT and the more positive attitude towards mathematics learning, they have more positive attitude towards ICT. Their self-efficacy negatively relates to their attitude towards formative assessment. Their expected education and motivation positively affect their attitude towards formative assessment.

In the Student_PPT model, the results indicate that there is no factor on their attitude towards PPT. Their gender and *attitude towards PPT* positively influence their achievement improvement due to PPT. However, the *parents' education level* has a negative direct impact on their achievement improvement. According to the results of Student_LOT and Student_CAT model, their attitude towards ICT positively relate to their attitudes towards LOT and CAT. Their parent education level negatively impacts on their attitude towards LOT. The *attitude towards LOT* has a positive direct impact on their achievement improvement due to LOT mode. Their ICT familiarity negatively affects their attitude towards CAT. Further, in Student_CAT model, *attitude towards formative assessment and ICT* positively influences students' achievement improvement due to CAT. However, their *attitude towards CAT* negatively impacts their achievement improvement. In comparison, the findings from the student-level provide essential insights into how the student-level scales interact with one another and impact students' achievement improvements in all these SEM models.

Chapter 11

Impacts on Teacher Attitude towards Test modes

11.1 Introduction

This chapter is concerned with the research question: "What are the teachers' factors that significantly influence their attitude towards the different test modes?" The structural equation model (SEM) is a methodology used to examine the relationships between the scales — measured scales and latent constructs. This study used the SEM to answer the research question mentioned above. There are three targeted outcome measures at the teacher-level, which include their attitude towards (1) PPT, (2) LOT, (3) CAT. For each outcome, the researcher conducted a structural equation model. Consequently, this chapter examined and reported three structural equation models for each outcome measure. These hypothesised models at the teacher-level were analysed using M-plus (v.8). In the investigation, the researcher took an exploratory approach, which, in turn, includes all possible paths. The models were trimmed by removing insignificant paths based on the initial results. For detailed methods and procedures for the single-level models, see **Chapter 4**.

First, the study presents the scales included in each model, followed by the hypothesised teacher-level model. The results of its structural models are presented next, followed by the fit indices for the model. Finally, a summary of this chapter is presented to capture the main points of the analyses. This study applied only the original item responses or raw scores for the demographic scales and used the Weighted Likelihood Estimates (WLE) score for the remaining scales. The Rasch measurement model estimated the WLE score of the scales during the validation process.

11.2 Scales in teacher-level models

The models for teacher-level factors aim to investigate the impact of the hypothesised scales on the size of the different attitudes towards test modes. The attitude towards test modes differs among the three experiments. For this study, teacher participants apply the classroom formative assessment in three test modes (PPT, LOT, and CAT). There are nine manifest (or observed) scales included in each teacher-level SEM model. They are their *highest education level (t_hgedu)*; *qualification in education (t_qedu)*; *class size (t_clsz)*; *multi-subject in an academic year (t_mlsub)*; teachers' *general practices of formative assessment (t_gpfa)*; *specific practices of formative assessment (t_spfa)*; and *attitude towards formative assessment (t_attfa)*.

Two scales related to ICT: their *ICT familiarity (t_ictfam)* and their *attitude towards ICT (t_attict)*; see **Table 11.1**.

Further, there are three main targeted outcome scales. For teachers' *attitude towards applying PPT in formative assessment* is *tch_ppt*; for their *attitude towards applying LOT in formative assessment* is *tch_lot;* for *attitude towards applying CAT in formative assessment* is *tch_cat*. The study used the WLE scores estimated using the Rasch measurement model for the eight latent scales apart from demographic scales. **Table 11.1** describes the list of these latent and manifest scales.

Table 11.1

Manifest scales in the teacher-level model

Manifest (Observed Scales)	Description	Code
t_hgedu	high education level	1= bachelor's degree of science (BSc) 2= bachelor's degree of arts (BA) 3= bachelor's degree of education (BEd)
t_qedu	qualification in education	1= short-term teaching education program 2 = 2-year teacher education program 3= 4-year teacher education program
t_clsz	class size	1 = Class having 61 to 70 students 2 = Class having 51 to 60 students 3 = Class having 41 to 50 students 4 = Class having 31 to 40 students
t_mlsub	multi-subject in an academic year	1 = teaching mathematics and other two subjects 2 = teaching mathematics and another subject 3= teaching only mathematics
t_gpfa	Teachers' general practices of formative assessment	WLE score
t_spfa	Teachers' specific practices of formative assessment	WLE score
t_attfa	Teachers' attitude towards formative assessment	WLE score
t_ictfam	Teachers' ICT familiarity	WLE score
t_attict	Teachers' attitude towards ICT	WLE score
tch_ppt	Teachers' attitude towards PPT	WLE score
tch_lot	Teachers' attitude towards LOT	WLE score
tch_cat	Teachers' attitude towards CAT	WLE score

11.3 Hypothesised model

Chapter 3 described the theoretical framework, which has been thoroughly giving rise to the general hypothesised model. For this chapter, **Figure 11.1** shows the graphical description of the hypothesised model. The hypothesised model has three stages: the presage, process, and product stages. This study hypothesised the scales at the presage stage as exogenous scales as other scales do not influence the model. The other scales in the process and production stages are assumed to be endogenous scales because they may interact in the process and production stages. Then, the scales may influence them in the presage stage.

The main theoretical framework in this study proposed that exogenous scales in the presage are teachers' demographic scales. The scales are their *highest education level (t_hgedu)*; *qualification in education (t_qedu)*; *class size (t_size)* and *multi-subject in an academic year (t_multi)*. The scales from the presage stage affect the scales in the process stage. These scales in the process stage are teachers' *general practices of formative assessment (t_gpfa)*, their *specific practices of formative assessment (t_spfa)*, their *attitude towards formative assessment (t_attfa)*, their *ICT familiarity (t_ictfam)*, and their *attitude towards ICT (t_attict)*. Further, the study viewed the outcome scales in the product stage (*tch_ppt*, *tch_lot*, *tch_cat*) as endogenous scales since scales influence them from the presage and process stages in the models. This study compares effects on attitudes towards different test modes. Consequently, the following sections will explore SEM models.

The SEM examines the strength of the relationships among the latent scales and manifest scales included in this study. In addition, the teacher-level model will examine the casual relationships among the teacher-level factors discussed earlier. **Figure 11.1** shows the graphical description of the hypothesised model. The following sections show the results of the structural models.

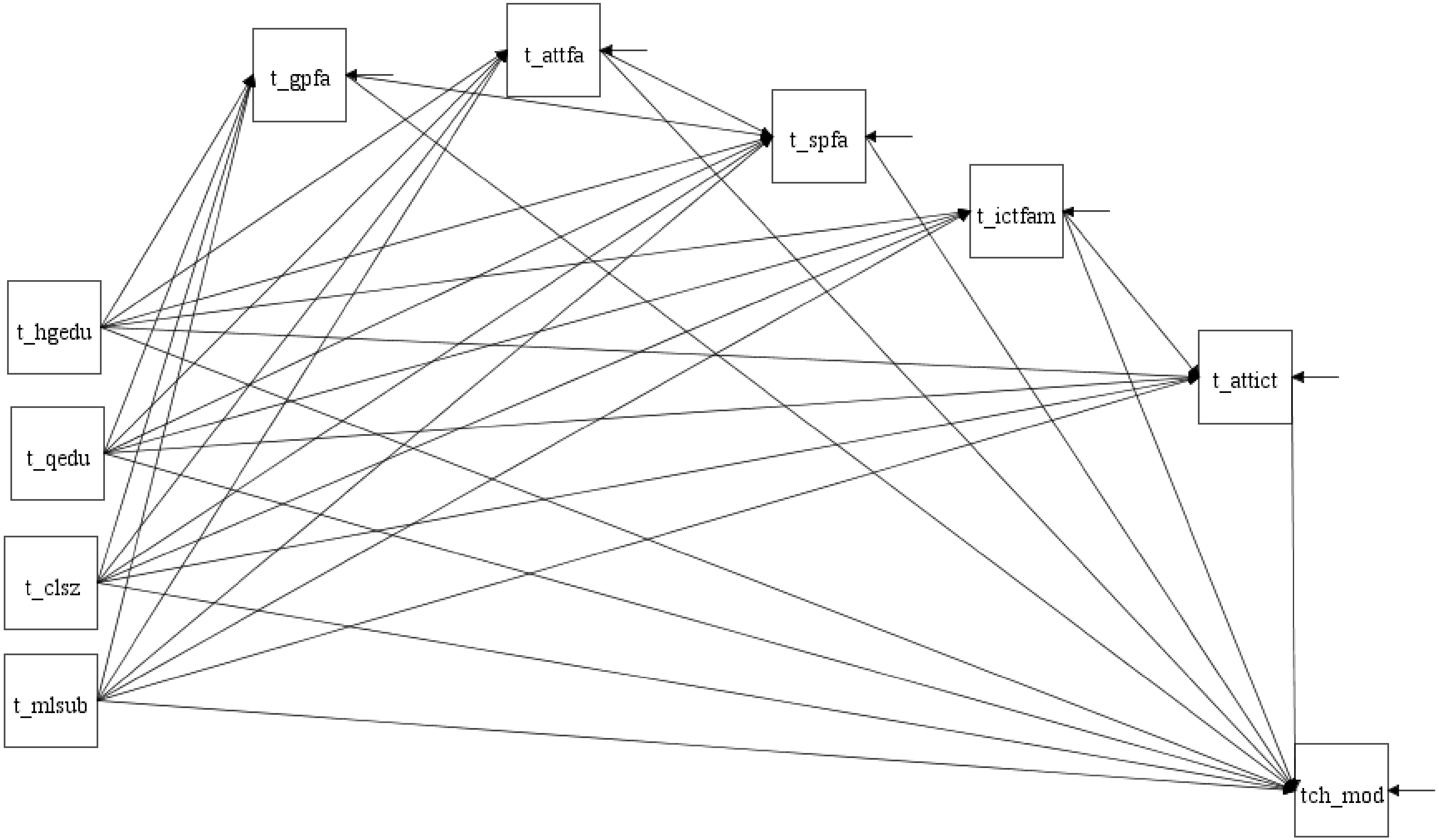

Figure 11.1 Hypothesised model at Teacher-Level

11.4 Results of Teacher_PPT model

The study examined the teacher-level structural equation model to explore the factors influencing teachers' *attitudes towards applying PPT in formative assessment*. **Table 11.2** presents the results for this model, and **Figure 11.2** illustrates its graphical diagram. Overall, there is no significant effect of the demographic scales on any scales of the process and product stages in the Teacher_PPT model.

Table 11.2

Results of the Teacher_PPT model

Outcome(s)	Predictor(s)	β	se	C.R	p
t_spfa	t_gpfa	0.482	0.220	2.188	0.029
	t_attfa	0.493	0.220	2.242	0.025
t_attict	t_ictfam	0.843	0.075	11.285	0.000
tch_ppt	t_spfa	0.826	0.264	3.131	0.002

Note: β = standardised estimate; se = standard error

C.R = Critical Ratio (C.R) of standardised estimate and standard error

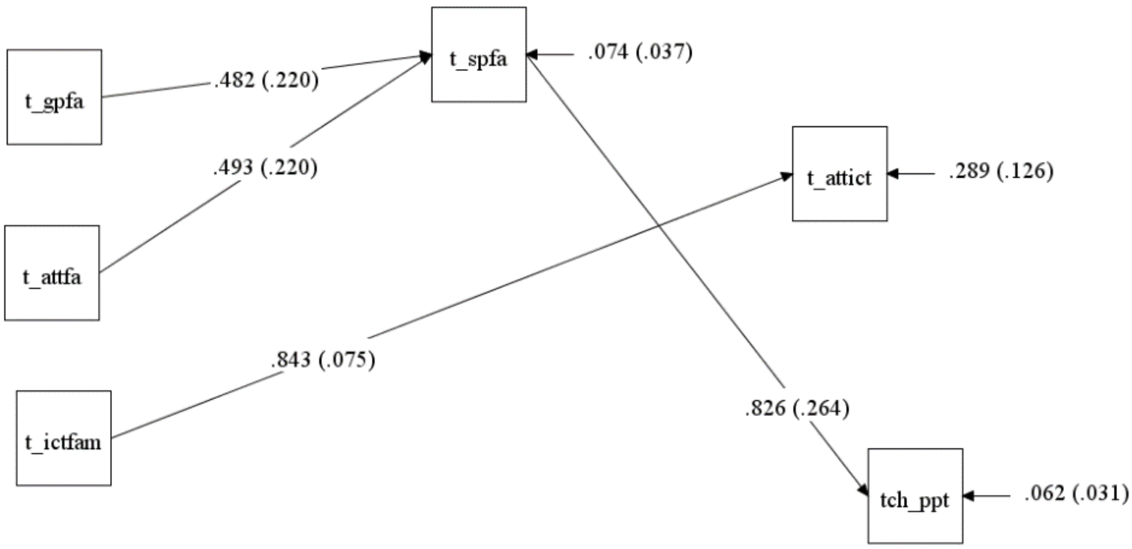

Figure 11.2 Teacher_PPT model

11.4.1 Specific practices of formative assessment (t_spfa)

The scales that have a significant direct effect on teachers' *specific practices of formative assessment (t_spfa)* are their *general practices of formative assessment* (t_gpfa) (β = 0.174) and their *attitude towards formative assessment (t_attfa)* (β = 0.493). The direct positive path coefficient of their *general practices of formative assessment* (*t_gpfa*) indicates that teachers who are more familiar with the general practices apply specific practices of formative assessment more frequently. Teachers who have a positive attitude towards formative assessment use specific practices of formative assessment more frequently. This finding is consistent with previous research findings (see Bernard, 1997; Michael–Chrysanthou, Gagatsis, & Vannini, 2014).

11.4.2 Attitude towards ICT (t_attict)

The scale that directly affects their attitude towards ICT (t_attict) is ICT familiarity — t_ictfam ($\beta = 0.843$), see **Table 11.2**. The positive path coefficient indicates that teachers who are more familiar with ICT have a more positive attitude towards ICT.

11.4.3 Attitude towards PPT (tch_ppt)

Only one scale, teachers' *specific practice of formative assessment* ($\beta = 0.813$), directly affects the *attitude towards PPT* (**Table 11.2**). The direct path coefficient indicates that teachers who use specific practices of formative assessment more frequently have a more positive attitude towards the use of PPT in formative assessment. This finding aligned with previous studies (Bernard, 1997; Michael–Chrysanthou, Gagatsis, & Vannini, 2014).

11.4.4 Fit indices of Teacher_PPT model

This section is concerned with the results of the fit indices obtained at the teacher-level model of the teacher attitude towards PPT (Teacher_PPT model). The model obtains Chi-square value (CMIN) of = 14.991 (df = 5) with a corresponding *p*-value less than 0.005 (**Table 11.3**). The model is still acceptable since the ratio of CMIN by the degrees of freedom (2.998) is between 2.0 and 5.0 (Hooper et al., 2008)—see **Table 11.3**. The Root Mean Square Error of Approximation (RMSEA, 0.206) of the final model is not below 0.08 (Byrne, 2010) because the sample size is only 15 (see **Table 11.3**), smaller. However, other indices that are the Tucker-Lewis Index (TLI, 0.967), and the Comparative Fit Index (CFI, 0.920), and the Root Mean Square Error of Approximation (SRMR, 0.014) values are relatively good because the TLI and CFI value are close to 1 (**Table 11.3**). The SRMR is less than 0.080 and is close to 0.050 (Byrne, 2010) (**Table 11.3**). According to these model fit indices, the analysis identifies how well the Teacher_PPT model fits the data.

Table 11.3

Fit indices of the Teacher_PPT model

Model	CMIN	df	CMIN/df	*p*	TLI	CFI	RMSEA	SRMR
Teacher_PPT model	14.991	5	2.998	0.005	0.967	0.920	0.206	0.014

11.5 Results of Teacher_LOT model

This study examined this structural equation model at the teacher-level to explore the factor influencing the teachers' attitude towards applying LOT in formative assessment. **Table 11.4** shows the results in the structural equation model, and **Figure 11.3** displays the graphical diagram of this structural equation model.

Table 11.4

Results of the Teacher_LOT model

Outcome(s)	Predictor(s)	β	se	C.R	p
t_spfa	t_gpfa	0.482	0.220	2.188	0.029
	t_attfa	0.493	0.220	2.242	0.025
t_attict	t_ictfam	0.843	0.075	11.285	0.000
tch_lot	t_attict	0.582	0.205	2.844	0.004

Note: β = standardised estimate; se = standard error

C.R = Critical Ratio (C.R) of standardised estimate and standard error

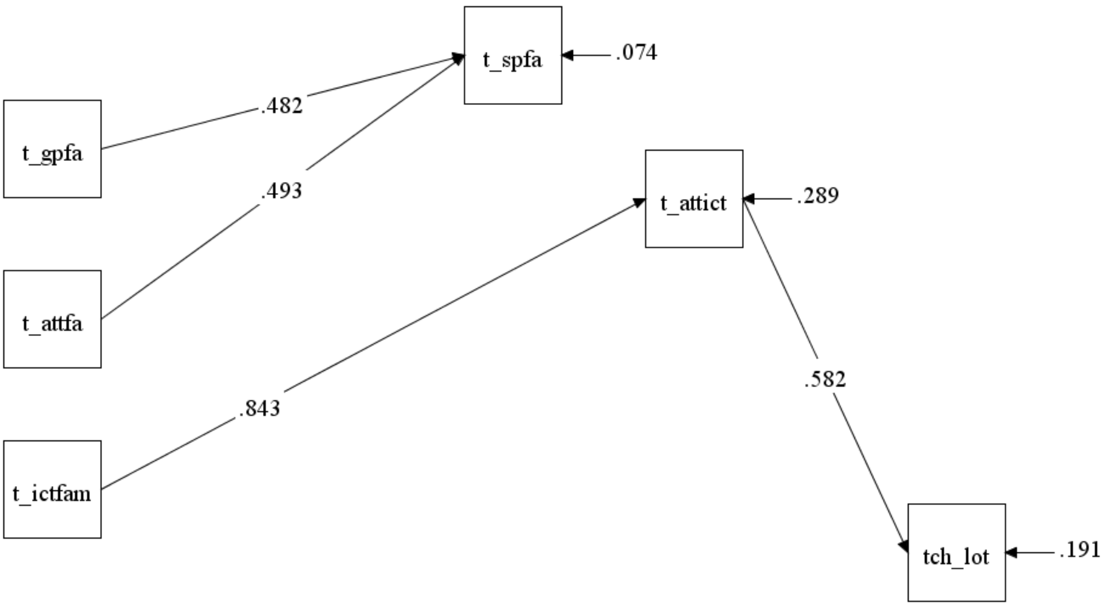

Figure 11.3 Teacher_LOT model

11.5.1 Factors influencing outcome scales

Structural equation modelling (Teacher_LOT model) explores which predicators directly affect the seven outcome scales. Like the previous model (Teacher_PPT model), the demographics scales do not significantly affect any scales of the process and product stages in the model. This Teacher_LOT model provides a similar result for the two outcome scales like the Teacher_PPT model. There are two scales on the teachers' specific practices of formative assessment (β = 0.174) and their attitude towards formative assessment (β = 0.493), see **Table 11.4**. There is only one scale on the teachers' attitude towards ICT, namely ICT Familiarity (β = 0.843).

11.5.2 Teachers' attitude towards LOT (tch_lot)

As shown in **Table 11.4**, only one scale that directly affects the attitude towards LOT is the teachers' attitude towards ICT (β = 0.587). The direct path coefficient indicates that teachers who have a more positive attitude towards ICT have a more positive attitude towards LOT. This result agrees with Herman et al. (2015), who discovered that attitude towards ICT positively affects attitude towards computer-based test mode.

11.5.3 Fit indices of Teacher_LOT model

This section is concerned with the fit indices obtained at the teacher-level model of the teacher attitude towards LOT (Teacher_LOT model). **Table 11.5** shows the model fit indices. The model obtains Chi-square value (CMIN) of = 16.563 (df = 5) with a corresponding *p*-value less than 0.005. The model is still acceptable since the ratio of CMIN by the degrees of freedom (CMIN/df, 3.313) is between 2.0 and 5.0 (Hooper et al., 2008). The Root Mean Square Error of Approximation (RMSEA, 0.393) of the final model is not below 0.08 (Byrne, 2010) because the sample size is small, only 15. However, other indices, including the Tucker-Lewis Index (TLI, 0.859), the Comparative Fit Index (CFI, 0.953), and the Root Mean Square Error of Approximation (SRMR, 0.016) values are relatively good. Because the TLI and CFI values are close to 1, and the SRMR is less than 0.080, which is close to 0.050 (Byrne, 2010), the model indicates an acceptable model fit. Consequently, the analysis proves how well the Teacher_LOT model fits the data.

Table 11.5

Fit indices of the Teacher_LOT model

Model	CMIN	df	CMIN/df	*p*	TLI	CFI	RMSEA	SRMR
Teacher_LOT model	16.563	5	3.3126	0.005	0.859	0.953	0.393	0.016

11.6 Results of Teacher_CAT model

This study examined this structural equation model at the teacher-level to explore the factor influence on the teachers' attitude towards applying CAT in formative assessment. The results are in **Table 11.6**, and the graphical diagram of this model is in **Figure 11.4**.

Table 11.6

Results of the Teacher_CAT model

Outcome(s)	Predictor(s)	β	se	C.R	p
t_spfa	t_gpfa	0.482	0.220	2.188	0.029
	t_attfa	0.493	0.220	2.242	0.025
t_attict	t_ictfam	0.843	0.075	11.285	0.000
tch_cat	t_spfa	0.918	0.319	2.876	0.004
	t_attict	0.365	0.154	2.370	0.018

Note: β = standardised estimate; se = standard error

C.R = Critical Ratio (C.R) of standardised estimate and standard error

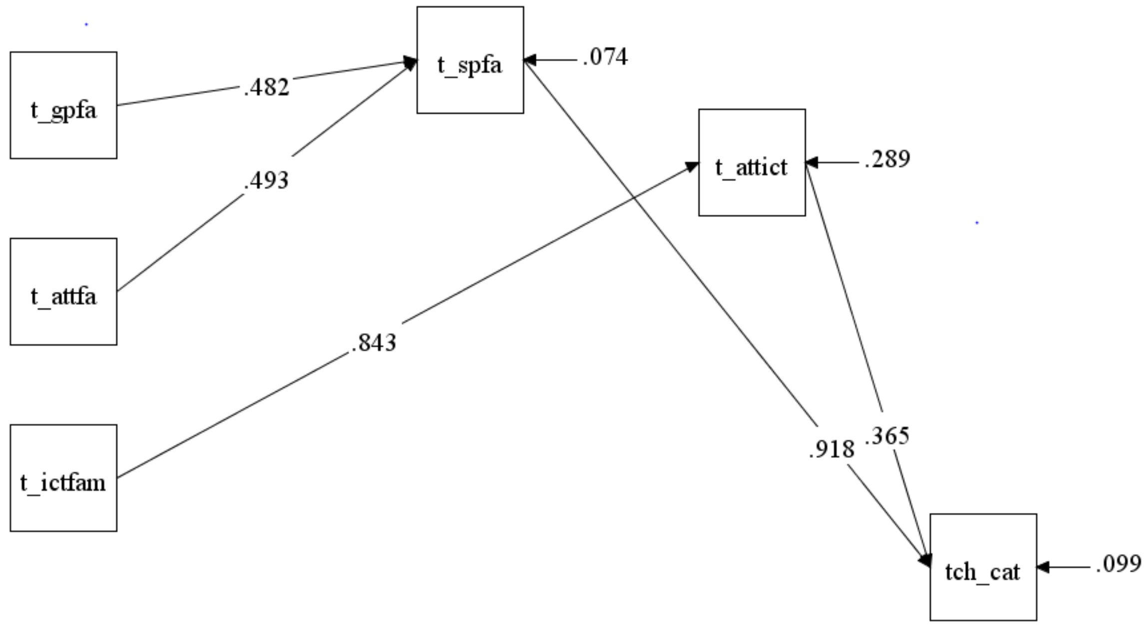

Figure 11.4 Teacher_CAT model

11.6.1 Factors influencing outcome scales

This structural equation modelling (Teacher_CAT model) explores which predicators directly affect the seven outcome scales. Like the Teacher_LOT model, the demographics scales do not significantly affect any scales of the process and product stages in the model. This SEM_CAT model provides a similar result for the two outcome scales, like the Teacher_LOT model. There are two scales on the teachers' *specific practices of formative assessment* (β = 0.174) and their *attitude towards formative assessment* (β = 0.493). There is only one scale on the teachers' *attitude towards ICT*: their *ICT Familiarity* (β = 0.843), see **Table 11.6**.

11.6.2 Attitude towards CAT (tch_cat)

Two scales have direct effects on the *attitude towards CAT*. These scales are the teachers' *specific practice of formative assessment* (β = 0.925) and the teachers' *attitude towards ICT* (β = 0.368). The direct path coefficient indicates that teachers who use more practices that are specific have a more positive attitude towards CAT. Likewise, teachers with a more positive attitude towards ICT have a more positive attitude towards CATs. These findings are consistent with findings from other studies (Michael–Chrysanthou, Gagatsis, & Vannini, 2014; Herman, Osmundson, Dai, Ringstaff, and Timms, 2015).

11.6.3 Fit indices of Teacher_CAT model

This section discusses the fit indices obtained at the teacher-level model of the teacher attitude towards CAT (Teacher_CAT model). **Table 11.7** shows its model fit indices. The model obtains Chi-square value (CMIN) of = 14.986 (df = 5) with a corresponding *p*-value less than 0.004. The model is still acceptable since the ratio of CMIN by the degrees of freedom (CMIN/df, 2.997) is between 2.0 and 5.0 (Hooper et al., 2008). The Root Mean Square Error of Approximation (RMSEA, 0.365) of the final model is not below 0.08 (Byrne, 2010) because the sample size is small, only 15. However, other indices, that is, the Tucker-Lewis Index (TLI, 0.852), the Comparative Fit Index (CFI, 0.957), and the Root Mean Square Error of Approximation (SRMR, 0.015) values are relatively good because the TLI and CFI value are close to 1 and the SRMR is less than 0.080, and are relatively close to 0.050 (Byrne, 2010). Consequently, the model indicates an acceptable model fit. According to these model fit indices, the analysis shows how well the Teacher_CAT model fits the data.

Table 11.7

Fit indices of the Teacher_CAT model

Model	CMIN	df	CMIN/df	*p*	TLI	CFI	RMSEA	SRMR
Teacher_CAT model	14.986	5	2.997	0.004	0.852	0.957	0.365	0.015

11.7 Summary

This chapter provides an analysis of the teacher-level models. The MPlus (v.8) statistical software package can provide the different relationships of the teacher-level scales. The teacher-level structural model shows how teachers' *general and specific practices of formative assessment, attitude towards formative assessment, ICT familiarity, attitude towards ICT, relate to their attitude towards the test modes (PPT, LOT, CAT)*. According to the nature of the study, there are three SEM models due to the test mode: Teacher_PPT model, Teacher _LOT model, and Teacher _CAT model.

In all three models, there are same factors that positively influence on teachers' specific practices of formative assessment and attitude towards ICT. teachers' general practices of formative assessment and their attitude towards formative assessment positively affect their specific practices of formative assessment. Their ICT familiarity positive relate to their attitude towards ICT.

In the Teacher_PPT model and Teacher _CAT model, the results indicate that their specific practice of formative assessment has a positive direct impact on their attitude towards PPT mode and LOT mode. In the Teacher _LOT model and Teacher _CAT model, their attitude towards ICT positively influences their attitude towards LOT mode and CAT mode. Moreover, the findings from the teacher-level provide essential insights into how the teacher-level scales interact with one another and impact teachers' attitude test mode in all these SEM models.

Chapter 12

Hierarchical Linear Models for Achievement Improvement

12.1 Introduction

Chapters 10 and 11 analysed student-level and teacher-level structural equation models (SEMs) in which the factors are influencing score improvement of the students due to applying each of the three test modes in formative assessment. The researcher of this study constructed SEMs for both student-level scales and teacher-level scales. These single-level SEMs provide insights into the direct and mediation effect of the scales but cannot estimate the cross-level interaction or moderation effect in single-level SEMs.

Further, as discussed in **Chapter 5**, bias could be introduced if data from two levels are aggregated or disaggregated. Therefore, this study employed the Hierarchical Linear Modelling (HLM) or multilevel regression analysis technique to overcome bias caused by aggregation and/or disaggregation and thus minimise the bias. In addition, the HLM models provide the estimate for the direct and the cross-level interaction, i.e., student- and teacher-levels. This is because data collected from classes are hierarchical, and students are sampled within classes.

Chapter 4 described the analysis procedure of HLM. The purpose of this chapter is to discuss the findings from the two-level HLM analyses on the student- and teacher-level scales and their impact on students' score improvement according to the different test modes. This chapter begins with its scales and its hypothesised two-level HLMs on students' score improvement. There are three experiments in three test modes: PPT, LOT, and CAT. Then, the chapter will describe, compare, and contrast the findings of three HLMs: PPT_HLM model, LOT_HLM model, and CAT_HLM model, and their respective null models. Finally, a summary is provided for this chapter. Further, this chapter supports the aims of the study and integrates the qualitative responses from teacher and student participants with the quantitative findings to help enrich the interpretation of the quantitative results on the experiment of the effectiveness of different test modes. This chapter carried out all analyses underpinned by a mixed-methods design, following the theoretical framework presented in Chapter 3.

12.2 Scales in two-level model

This study uses HLM because the scales are entwined at the student- and teacher-level and can display a hierarchical structure. The HLM is more complex than the ordinary multiple regression model. There are scales of HLM in which the assumptions of normality, linearity, and homogeneity of variance can be violated (Hox, 2010). However, these assumptions are not violated for all scales in this model.

Two sets of scales at student- and teacher-levels are specified to carry out the HLM analysis. First, as mentioned in Chapter 11, all scales except the demographic scales are subjected to Rasch analysis. Then, the Weighted Likelihood Estimate (WLE) scores are obtained to use in subsequent analyses such as the analyses of student- and teacher-level models as well as the hierarchical model.

Moreover, the scales that confirmed the one-factor structure are represented by a single set of WLE scores used in this analysis. On the other hand, scales with a hierarchical structure, which refer to latent scales with multiple WLE scores indicators, are simplified into scales with a one-factor structure. Finally, as HLM does not allow the formation of latent scales, the principal component scores are calculated for all latent scales that had manifest scales using the IBM SPSS (v.26).

Teachers' factors, including contextual and psychological factors, directly affect students' score improvement. Similarly, students' contextual and psychological factors, directly affect their score improvement. Teachers' factors also have the cross-level interaction effects on the slope of students' factors and their score improvement. The conceptual model described the direct effect and cross-level interaction effect. In addition, **Figure 12.1** shows conceptual model for the two-level model of factors influencing students' score improvement.

The researcher used the terms level-1, student-level, and micro-level interchangeably in this section. Similarly, level-2, teacher-level, and macro-level are used interchangeably. The sample of this study involved 659 secondary students and 15 teachers in 5 schools. Data from student-level (Level-1) are obtained from students' questionnaires and their achievement tests. For Level-1, this study applied information related to student contextual and attitudinal factors collected from student questionnaires to observe their effect on the outcome scale. Data from the teacher-level (Level-2) are obtained from the teacher questionnaires. In Level-2, the study observed contextual and attitudinal factors collected from the teacher questionnaire, whether their effects are on the outcome scale.

Apart from demographic scales, the study obtained the WLE scores of all scales after the raw scores were subjected to Rasch analysis for subsequent analyses. Therefore, score improvement due to applying test modes in classroom formative assessment is determined as an outcome scale in this study. **Table 12.1** describes the names, codes, and descriptions of the predictor scales tested from the Level-1 (micro-level) and the Level-2 (macro-level) for each level.

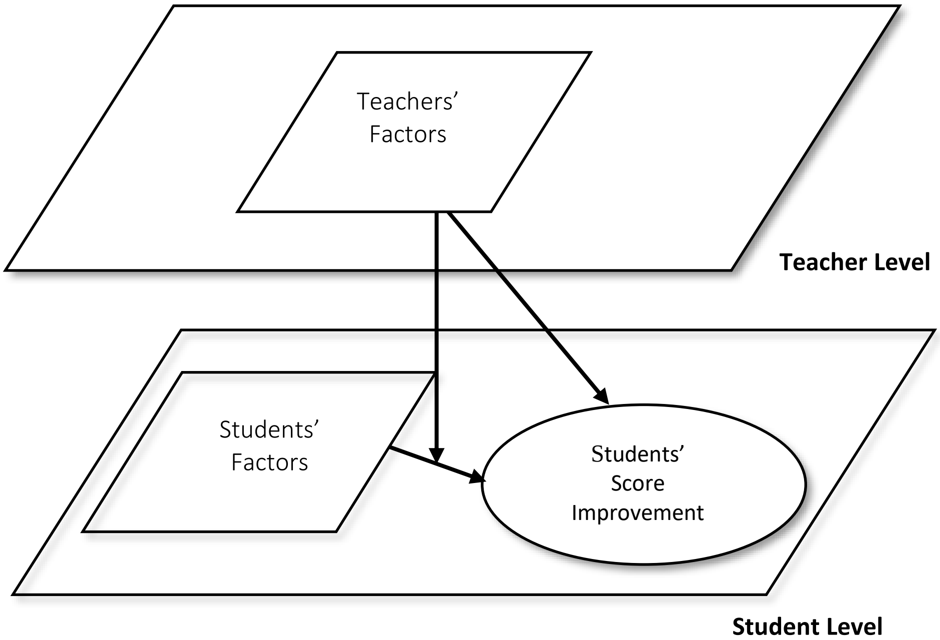

Figure 12.1 Conceptual model of two-level scales on score improvement

Table 12.1
List of scales tested at the two-level HLM

Scale name	Scale code	Scale description	Code
Level-2		(Teacher-Level Scale)	
	t_hgedu	high education level	0= bachelor's degree of science (BSc) 1= bachelor's degree of arts (BA) 2= bachelor's degree of education (BEd)
	t_qedu	qualification in education	0 = short-term teaching education program 1 = 2-year teacher education program 2 = 4-year teacher education program
	t_clsz	class size	0 = Class having 61 to 70 students 1 = Class having 51 to 60 students 2 = Class having 41 to 50 students 3 = Class having 31 to 40 students

Scale name	Scale code	Scale description	Code
	t_mlsub	multi-subject in an academic year	0 = teaching mathematics and other two subjects 1 = teaching mathematics and another subject 2= teaching only mathematics
	t_gpfa	Teachers' general practices of formative assessment	WLE score
	t_spfa	Teachers' specific practices of formative assessment	WLE score
	t_attfa	Teachers' attitude towards formative assessment	WLE score
	t_ictfam	Teachers' familiarity with ICT	WLE score
	t_attict	Teachers' attitude towards ICT	WLE score
	tch_mod	Teachers' attitude towards Test mode	WLE score
	tch_ppt	Teachers' attitude towards PPT Mode	WLE score
	tch_lot	Teachers' attitude towards LOT Mode	WLE score
	tch_cat	Teachers' attitude towards CAT Mode	WLE score
Level-1		(Student- Level Scale)	
Student background (N= 659)	gen	Gender	0 = Male 1 = Female
	expedu	Expected Education Level	0 = High School Graduate 1 = Diploma's Degree 2 = Bachelors' Degree 3 = Masters' Degree 4 = Doctoral Degree
	predu	Parent Education Level	0= Less than High School 1 = Some High School 2 = High School Graduate 3 = Diploma's Degree 4 = Bachelor's Degree 5 = Masters' Degree 6 = Doctoral Degree
	attm	Attitude towards mathematics learning	WLE scores
	selfeffi	Self-Efficacy	WLE scores
	moti	Motivation	WLE scores
	attfa	Attitude towards Formative Assessment	WLE scores

Scale name	Scale code	Scale description	Code
	ictfam	Familiarity with ICT	WLE scores
	attict	Attitude towards ICT	WLE scores
	attmod	Attitude towards Test mode	WLE scores
	attppt	Attitude towards PPT Mode	WLE scores
	attlot	Attitude towards LOT Mode	WLE scores
	attcat	Attitude towards CAT Mode	WLE scores
Outcome	gain	Score Improvement due to applying test mode in formative assessment	WLE scores
	g_ppt	Achievement improvement due to PPT Mode	WLE scores
	g_lot	Achievement improvement due to LOT Mode	WLE scores
	g_cat	Achievement improvement due to CAT Mode	WLE scores

12.3 Hypothesised two-level model of achievement improvement

This study derived a hypothesised two-level model of students' score improvement from the theoretical framework presented in Chapter 5. The theoretical framework proposed that scales in the presage stage are four teacher-level scales and three student-level scales. These teacher-level scales include teachers' highest education level (t_hgedu), their qualification in education (t_qedu), their class size (t_clsz), and their multi-subject teaching (t_mlsub). These student-level scales include their gender status (gen), their expected education (expedu), and their parent education level (predu).

The proposed scales in the process phase include six teacher-level scales and seven student-level scales in the process stage. These teacher-level scales in the process phase are the teachers' general practices of formative assessment (t_gpfa), their specific practices of formative assessment (t_spfa), attitude towards formative assessment (t_attfa), their familiarity with ICT (t_ictfam), their attitude towards ICT (t_attict), and their attitude towards test mode (tch_mod). This study substituted teachers' attitude towards PPT (tch_ppt) for PPT_HLM, or their attitude towards LOT (tch_lot) for LOT_HLM, attitude towards CAT (tchcat) for CAT_HLM instead of tch_mod.

These student-level scales in the process phase are their attitude towards mathematics learning (attm), their self-efficacy (selfeffi), their motivation (moti), their attitude towards formative assessment (attfa), their familiarity with ICT (ictfam), their attitude towards ICT (attict), their attitude towards test mode (attmod). This study substituted their attitude towards PPT (attppt) for PPT_HLM, or their attitude towards LOT (attlot) for LOT_HLM, attitude towards CAT (attcat) for CAT_HLM instead of attmod.

The proposed students' level scales in the product stage is their achievement improvement due to applying Test mode in formative assessment (gain). For PPT_HLM, the scale in the product stage is g_ppt. Likewise, the scale in the product stage in LOT_HLM is g_lot, and that of CAT_HLM is g_cat.

According to the theoretical framework, all presage and process factors are related to the factors in the product. Consequently, **Figure 12.2** indicates only a few hypothesised cross-level interaction effects (blue arrows) of Level-2 scales as the interaction effects between the predictors and outcome scales in Level-1 to simplify the presentation of the model.

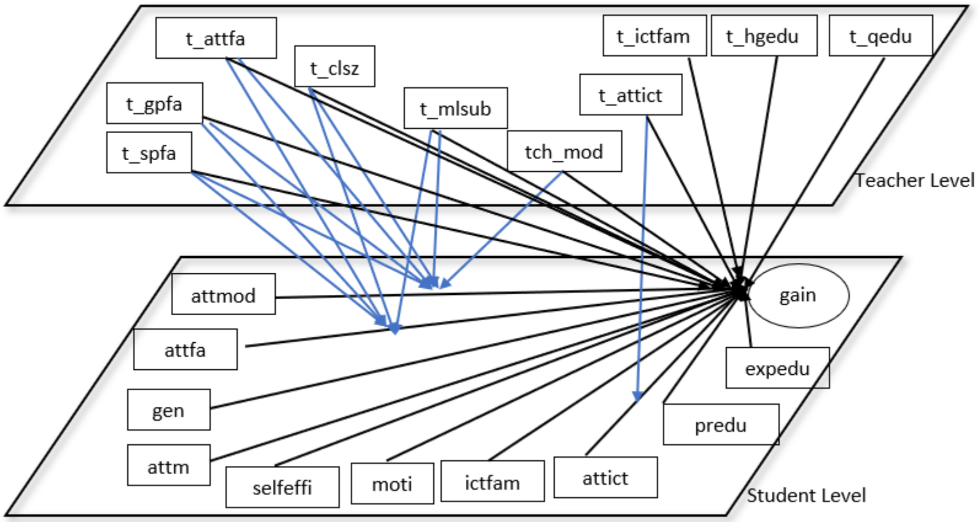

Figure 12.2 Hypothesised two-level HLM of score improvement

12.4 Null models for score improvement in PPT_HLM; LOT_HLM and CAT_HLM Model

The study carried out two-level HLM analyses to examine the relationships between level-1 and level-2 predictors and the outcome scales in the models of PPT_HLM; LOT_HLM; CAT_HLM. The HLM 6.0 is used to analyse hierarchical linear modelling (Raudenbush, Bryk, & Congdon, 2009), while the data are organized using the SPSS software (v.26). Firstly, the null (or fully unconditional) models are analysed to examine the amount of variance available at each level of the hierarchy (Bryk and Raudenbush, 1992). The null models are to estimate whether there are random effects or not by the result of reliability (Raudenbush & Bryk, 1992). Then, the researcher conducted follow-up analyses to build up the final HLM models if there were random effects. In addition, the ICC is one of the indicators used to decide whether it is worthwhile to build up the final HLM models.

The null models estimate the partition of the variance and the Intraclass Correlation (ICC). The ICC calculates the ratio between the group variance to the total variance. It represents the percent of the variance in score improvement available between classrooms relative to the total variance. For the HLM, the initial ANOVA provides the amount of variance within a classroom (σ^2) and the amount of variance between the classroom (τ_B). There is no cut-off point; however, if the ICC is very low, the HLM analyses may not yield different results from a traditional analysis (Woltman, Feldstain, MacKay & Rocchi, 2012).

In the Level-1 equations of the three models, the level of score improvement of student i under teacher j is equivalent to the teacher mean plus a random error. For this thesis, the researcher assumed that respective Level-1 error, R_{ij}^{MOD}, is normally distributed with a mean of zero, and the specific constant Level-1 variance, σ^2, refers to student-level variability (Raudenbush & Bryk, 1992).

In the Level-2 equation, the students' average score improvement for a teacher is equivalent to the overall mean plus a random error. The variability between teachers is represented as τ_B . The null model allows the estimation of the proportions of variation within teachers among schools. For the Level-1, ICC (1) = σ^2/ (σ^2 + τ_B) is the proportion of variance within teachers; and for the level-2, ICC (2) = τ_B / (σ^2 + τ_B) is the proportion of variance among teachers.

In the null PPT_HLM, σ^2 (0.651) parameter represents student-level variability, while the τ_B (0.570) parameter represents teacher-level variability (**Table 12.2**). The proportions of variation for Level-1 are 0.533 or 53% of the variance. Level-2 is either 0.467 or 46.7% in the outcome measure of students' score improvement in PPT.

In the null LOT_HLM, σ^2 (1.226) parameter represents student-level variability, while the τ_B (0.334) parameter represents teacher-level variability (**Table 12.2**). The results estimated the proportions of variation for Level-1 to be 0.786 or 78.6% of the variance and Level-2 is either 0.214 or 21.4% in the outcome measure of students' score improvement in LOT.

In null CAT_HLM, σ^2 (1.384) parameter represents student-level variability, while the τ_B (1.090) parameter represents teacher-level variability (**Table 12.2**). The proportions of variation for Level-1 equal 0.556 or 55.6% of the variance. Level-2 is equal to 0.441 or 4.41% in the outcome measure of students' score improvement in CAT.

Table 12.2
Null Models' Results of PPT_HLM, LOT_HLM and CAT_HLM Models

Final estimation of variance components:							
Null Models	Random Effect	Reliability	Standard Deviation	Variance Component	df	Chi-square	P-value
PPT_HLM	INTRCPT1, U_{0j}^{PPT}	0.974	0.755	0.570	14	530.281	0.000
	Level-1, R_{ij}^{PPT}		0.807	0.651			
LOT_HLM	INTRCPT1, U_{0j}^{LOT}	0.921	0.578	0.334	14	194.909	0.000
	Level-1, R_{ij}^{LOT}		1.107	1.226			
CAT_HLM	INTRCPT1, U_{0j}^{CAT}	0.982	1.174	1.378	14	836.092	0.000
	Level-1, R_{ij}^{CAT}		1.044	1.090			

Null Models	σ^2	τ_B	Proportion of variance at Level-1 (n = 659)	Proportion of variance at Level-2 (n = 15)
PPT_HLM	0.651	0.570	0.651/ (0.0651 + 0.570) = 0.533	0.570/ (0.0651 + 0.570) = 0.467
LOT_HLM	1.226	0.334	1.226/ (1.226+0.334) = 0.786	0.334/ (1.226+0.334) = 0.214
CAT_HLM	1.384	1.090	1.384/ (1.384+1.090) = 0.559	1.090/ (1.384+1.090) = 0.441

In summarise, because the ICC (2) values are not very low the HLM analyses may provide different results from a traditional analysis (Woltman et al. 2012). Apart from the ICC values of each model, another critical value is reliability. Moreover, Raudenbush and Byrk (1992) note that when the reliability values in a model is less than 0.05, it possible to assume that there is no random effect for such coefficient. However, in **Table 12.2,** the reliability coefficients for Level-2 estimates in three models are more than 0.05, indicating that there may be a random effect for students' score improvement. In other words, more studied are required to examine factors influencing students' score improvement. Consequently, the researcher conducted follow-up analyses to examine conditional models and to build up the final models.

This study conducted hierarchical linear models to examine the effects of scales from student and teacher-level on the score improvement due to applying PPT, LOT, and CAT modes as classroom formative assessment. The general null model is MOD_HLM. Hence, they are called PPT_HLM, LOT_HLM and CAT_HLM, respectively. In all these null models MOD_HLM, without any student- and teacher-level predictors, the study specified outcome scales at student-level.

In the MOD_HLM model, the equations formed are as follows:

Level-1 Model

$Y_{ij}^{MOD} = B_{0j}^{MOD} + R_{ij}^{MOD}$

Level-2 Model

$B_{0j}^{MOD} = G_{00}^{MOD} + U_{0j}^{MOD}$

where,

Y_{ij}^{MOD} is the score improvement in test mode of student *i* under teacher *j*

B_{0j}^{MOD} is the intercept or mean score improvement in test mode of students under teacher *j* and

R_{ij}^{MOD} is a random 'student' effect or error (i.e., the deviation of student *i's score* from the mean score of students under teacher *j)*

G_{00}^{MOD} is the intercept or grand mean score improvement in test mode and

U_{0j}^{MOD} is a random 'teacher' effect or error (i.e., the deviation of teacher *j's mean* score from the grand mean of the score improvement in test mode mode)

12.5 Final models of score improvement in PPT_HLM, LOT_HLM and CAT_HLM

The study examined the direct effects of ten student-level scales to estimate a Level-1 model. At this stage, the researcher followed a step-down approach to examine which scales significantly influence an outcome scale at a 95% confident interval and removed insignificant scales from the Level-1 model. Among those, the estimated coefficients of the three scales are significant.

In order to estimate a Level-2 model, the researcher added ten Level-2 predictors into a model using the step-down strategy. As a result, three scales in Level-2 are significantly included in the model for the intercept. This process builds up the final model with significant teacher-level predictors and the significant student-level predictors for the outcome scales.

For the final PPT_HLM, the Level-1 and Level-2 equations are as follows:

Level-1 model

$$Y_{ij}^{PPT} = B_{0j}^{PPT} + B_{1j}^{PPT} * (\text{gen}) + B_{2j}^{PPT} * (\text{predu}) + B_{3j}^{PPT} * (\text{attppt}) + R_{ij}^{PPT}$$

Level-2 model

$$B_{0j}^{PPT} = G_{00}^{PPT} + G_{01}^{PPT} * (\text{t\_attfa}) + G_{02}^{PPT} * (\text{tch\_ppt}) + U_{0j}^{PPT}$$

$$B_{1j}^{PPT} = G_{10}^{PPT} + U_{1j}^{PPT}$$

$$B_{2j}^{PPT} = G_{20}^{PPT} + U_{2j}^{PPT}$$

$$B_{3j}^{PPT} = G_{30}^{PPT} + G_{31}^{PPT} * (t\_spfa) + U_{3j}^{PPT}$$

By substituting the Level-2 equations into the Level-1 equation, the equation below details a final model equation:

$$\begin{aligned} Y_{ij}^{PPT} = {} & G_{00}^{PPT} + G_{01}^{PPT} * (\text{t}_{\text{attfa}}) + G_{02}^{PPT} * \left(\text{tch}_{\text{ppt}}\right) + U_{0j}^{PPT} + G_{20}^{PPT} * (\text{gen}) + U_{2j}^{PPT} \\ & * (\text{gen}) + G_{20}^{PPT} * (\text{predu}) + U_{2j}^{PPT} * (\text{predu}) + G_{30}^{PPT} * (\text{attppt}) + G_{31}^{PPT} \\ & * (t\_spfa) * (\text{attppt}) + U_{3j}^{PPT} * (\text{attppt}) + R_{ij}^{PPT} \end{aligned}$$

For the final LOT_HLM, the Level-1 and Level-2 equations are as follows:

Level-1 model

$$Y_{ij}^{LOT} = B_{0j}^{LOT} + B_{1j}^{LOT} * (\text{attict}) + B_{2j}^{LOT} * (\text{attlot}) + R_{ij}^{LOT}$$

Level-2 model

$$B_{0j}^{LOT} = G_{00}^{LOT} + G_{01}^{LOT} * (t\_attfa) + G_{02}^{LOT} * (\text{t\_attict}) + U_{0j}^{LOT}$$

$$B_{1j}^{LOT} = G_{10}^{LOT} + U_{1j}^{LOT}$$

$$B_{2j}^{LOT} = G_{20}^{LOT} + G_{21}^{LOT} * (t\_spfa) + U_{2j}^{LOT}$$

The equation below details a final model equation of LOT_HLM:

$$\begin{aligned} Y_{ij}^{LOT} = {} & G_{00}^{LOT} + G_{01}^{LOT} * (t\_attfa) + G_{02}^{LOT} * (\text{t\_attict}) + U_{0j}^{LOT} + G_{10}^{LOT} * (\text{attict}) + U_{1j}^{LOT} \\ & * (\text{attict}) + G_{20}^{LOT} * (\text{attlot}) + G_{21}^{LOT} * (t\_spfa) * (\text{attlot}) + U_{2j}^{LOT} * (\text{attlot}) \\ & + R_{ij}^{LOT} \end{aligned}$$

For the final CAT_HLM, the Level-1 and Level-2 equations are as follows:

Level-1 model

$$Y_{ij}^{CAT} = B_{0j}^{CAT} + B_{1j}^{CAT} * (\text{attict}) + B_{2j}^{CAT} * (\text{attcat}) + R_{ij}^{CAT}$$

Level-2 model

$$B_{0j}^{CAT} = G_{00}^{CAT} + G_{01}^{CAT} * (t\_attict) + G_{02}^{CAT} * (\text{tch\_cat}) + U_{0j}^{CAT}$$

$$B_{1j}^{CAT} = G_{10}^{CAT} + U_{1j}^{CAT}$$

$$B_{2j}^{CAT} = G_{20}^{CAT} + G_{21}^{CAT} * (t\_spfa) + U_{2j}^{CAT}$$

The equation below details a final model equation for CAT_HLM:

$$Y_{ij}^{CAT} = G_{00}^{CAT} + G_{01}^{CAT} * (t\_attict) + G_{02}^{CAT} * (\text{tch\_cat}) + U_{0j}^{CAT} + G_{10}^{CAT} * (\text{attict}) + U_{1j}^{CAT} * (\text{attict}) + G_{20}^{CAT} * (\text{attcat}) + G_{21}^{CAT} * (t\_spfa) * (\text{attcat}) + U_{2j}^{CAT} * (\text{attcat}) + R_{ij}^{CAT}$$

Table 12.3 shows the final PPT_HLM, LOT_HLM and CAT_HLM results with the direct and cross-sectional effects from the predictors and the outcome scales are in. The graphical representation of the direct and interaction effects for PPT_HLM, LOT_HLM, and CAT_HLM are in **Figures 12.3, 12.4, and 12.5**.

Table 12.3
Final Model Results in PPT_HLM, LOT_HLM and CAT_HLM

	PPT_HLM			LOT_HLM			CAT_HLM		
Fixed Effect	Coe	se	*p*	Coe	se	*p*	Coe	se	*p*
For INTRCPT1, B_{0j}									
INTRCPT2, G_{00}	0.469	0.183	0.025	0.229	0.103	0.000	0.463	0.227	0.029
t_attfa, G_{01}	0.319	0.131	0.007	0.460	0.226	0.013			
tch_ppt, G_{02}	0.278	0.127	0.027						
tch_cat, G_{02}							0.106	0.046	0.027
t_attict, G_{02}				0.490	0.245	0.021	0.352	0.168	0.035
For attppt slope, B_{3j}									
INTRCPT2, G_{30}	0.229	0.080	0.038						
t_spfa, G_{31}	0.363	0.151	0.025						
For attlot slope, B_{2j}									
INTRCPT2, G_{20}				0.083	0.035	0.026			
t_spfa, G_{21}				0.086	0.037	0.041			
For attcat slope, B_{2j}									
INTRCPT2, G_{20}							-0.192	0.078	0.038
t_spfa, G_{21}							0.265	0.159	0.011
For attict slope, B_{1j}									
INTRCPT2, G_{10}				0.614	0.292	0.045	0.386	0.105	0.003
For gen slope, B_{1j}									
INTRCPT2, G_{10}	0.005	0.002	0.044						
For predu slope, B_{2j}									
INTRCPT2, G_{20}	-0.021	0.009	0.045						

Note: Coe, coefficient; se, standard error; *p*, statistical significance

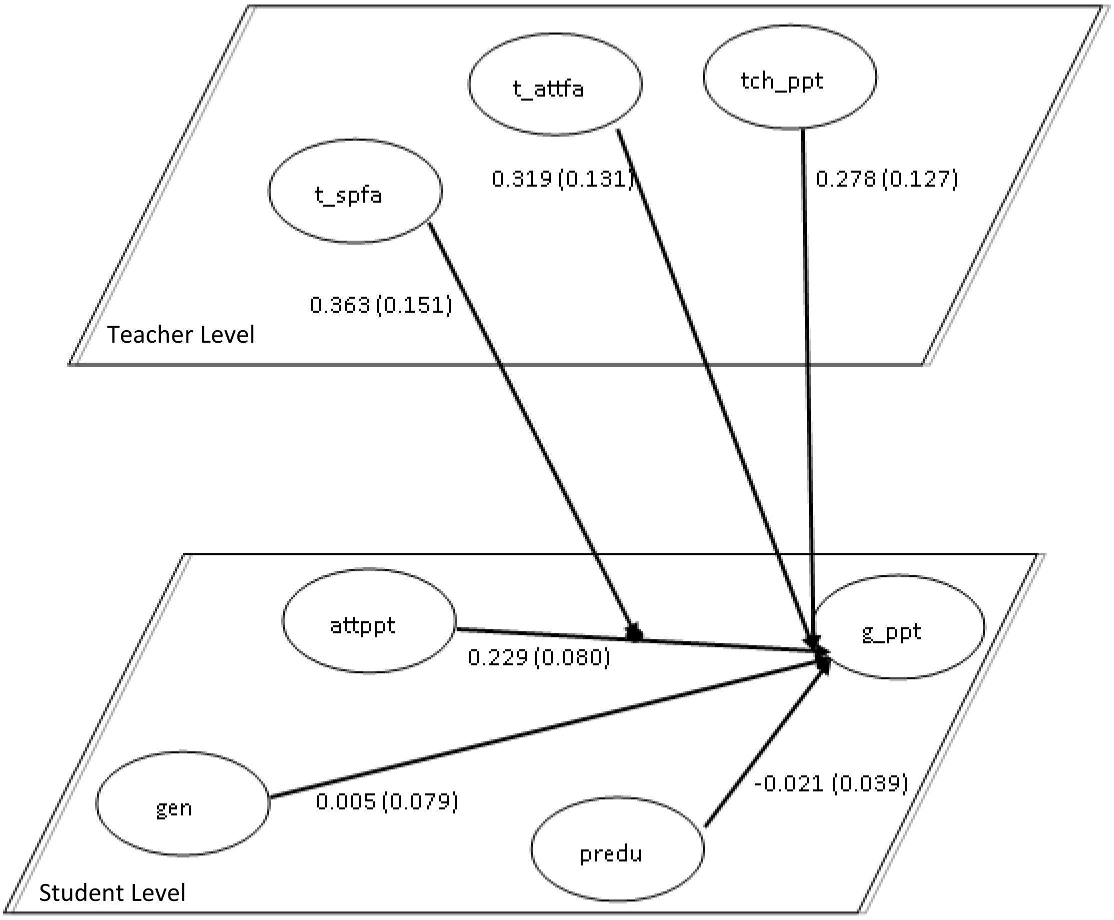

Figure **12.3** Final PPT_HLM model

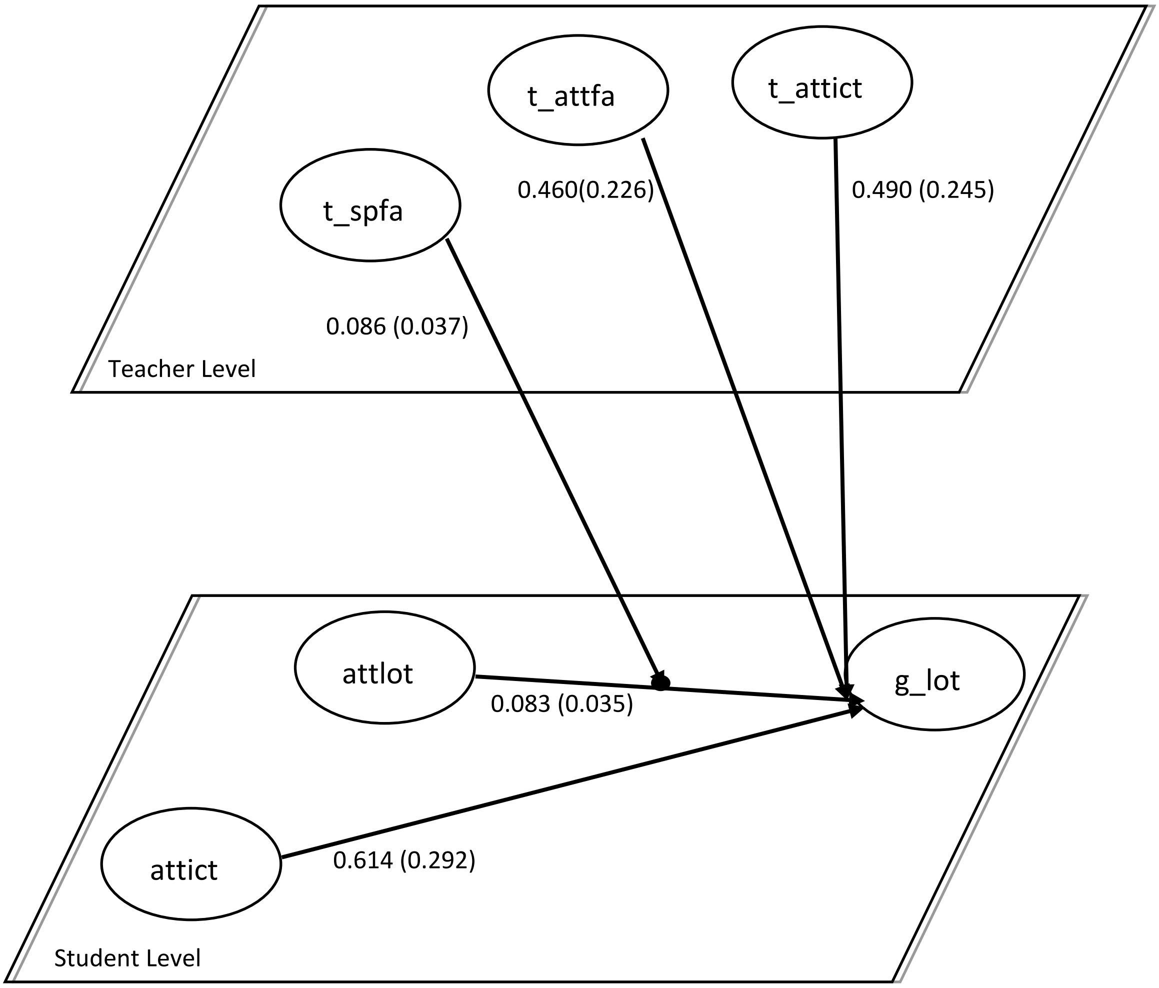

***Figure* 12.4** Final LOT_HLM model

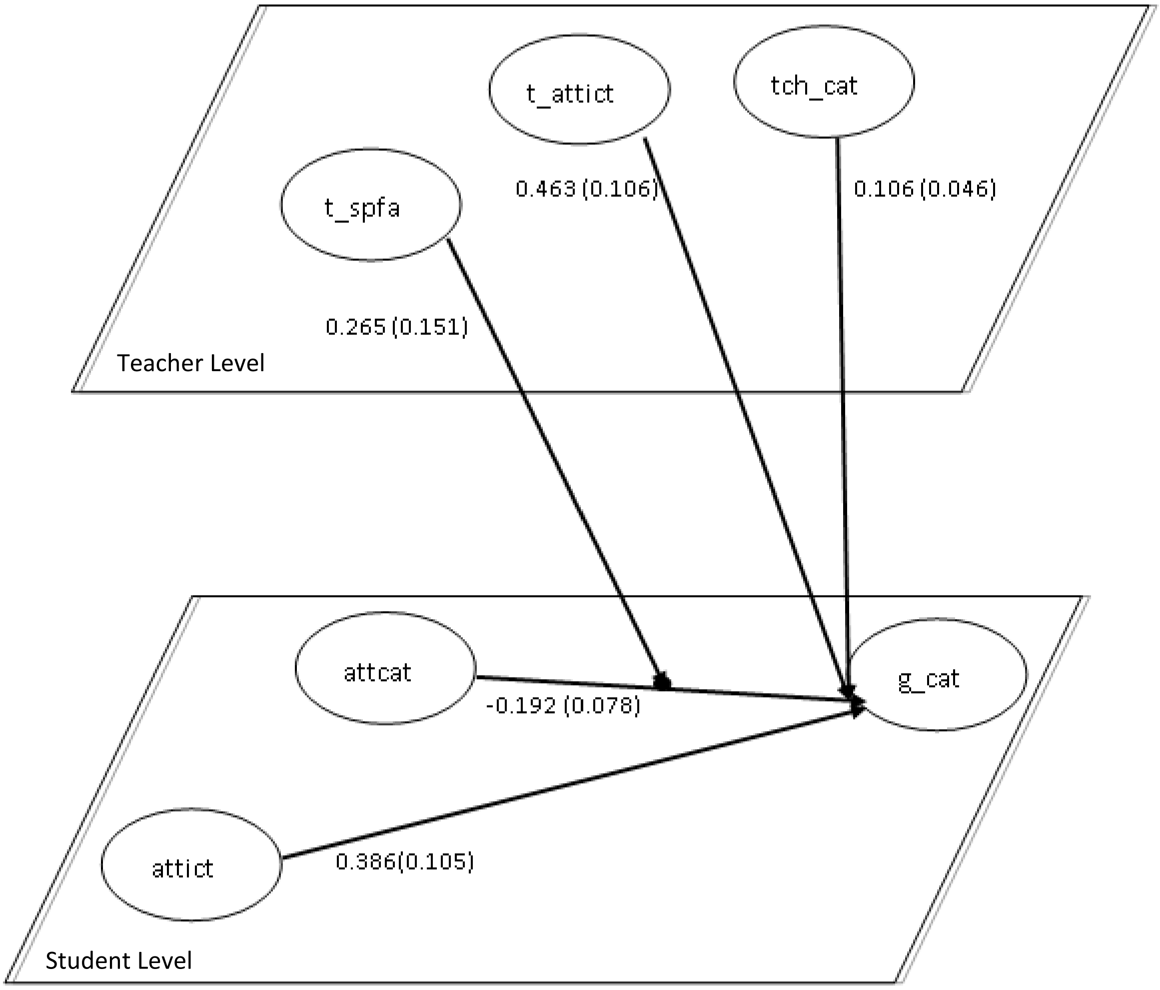

***Figure* 12.5** Final CAT_HLM model

In PPT_HLM, there are five scales with a main direct effect on students' mathematics achievement improvement at Level-1 and Level-2. There are direct effects of gender (gen), parental education (predu) and attitude towards PPT mode (attppt) at Level-1 and teachers' attitude towards formative assessment (t_attfa), and teachers' attitude towards PPT mode (tch_ppt) at Leve-2. At Level-1, there are positive direct effects of gender (G_{10} = 0.005) and attitude towards PPT mode (G_{30} = 0.229) on improving mathematics achievement in PPTs. Female students improve their mathematics scores more than the male students do. Likewise, the more positive attitude towards PPT, the more their mathematics score improves in PPT. However, the negative direct effect of parental education (G_{20} = - 0.021) implies that students with the parents' better education level improve less in mathematics achievement in PPT. Most parents with higher education levels do not prefer PPTs (Sánchez, Reyes, & Singh, 2006).

For Level-2, teachers' attitude towards formative assessment and teachers' attitude towards PPT positively affect the improvement of mathematics achievement in PPT. Positive direct effect by teachers' attitude towards formative assessment (t_attfa) (G_{01} = 0.319) suggests that the more positive attitude towards formative assessment, the more improvement of mathematics score in PPT. The positive direct effect of the teachers' attitude towards PPT (G_{02} = 0.278) implies that the teachers' positive attitude towards PPT schools tends to improve students' achievement in PPT.

In LOT_HLM, the four main effects include the direct effects from an attitude towards ICT (attict) and attitude towards LOT mode (attlot) in the Level-1 and from teachers' attitude towards formative assessment (t_attfa) and teachers' attitude towards ICT (t_attict). At Level-1, there are positive direct effects of attitude towards ICT (G_{10} = 0.614) and attitude towards LOT mode (G_{30} = 0.083) on the improvement of mathematics achievement in LOT mode. The more positive attitude towards ICT students has, the more their mathematics score improves in LOT mode that applies ICT. Likewise, the more positive their attitude towards LOT mode, the more their mathematics score improves in LOT mode. For Level-2, teachers' attitude towards formative assessment and teachers' attitude towards ICT provide positive direct effects on the improvement of mathematics achievement in LOTs. There is a positive direct effect by teachers' attitude towards formative assessment (t_attfa) (G_{01} = 0.460) and a positive direct effect by the teachers' attitude towards ICT (G_{02} = 0.490). Students under teachers with more positive attitudes towards formative assessment tend to have bigger improvements. Likewise, their teachers with more positive attitudes towards LOT mode tend to improve students' achievement in LOT mode.

In CAT_HLM, there are four scales with direct effect on students' mathematics achievement improvement at the Level-1 and Level-2. At Level-1, there are direct effects of attitude towards ICT (G_{10} = 0.386) and attitude towards CAT (G_{30} = -0.192) on improving mathematics achievement in CATs. The more positive attitude towards ICT they have, the more they improve in mathematics achievement in CAT. However, the more positive attitude towards CAT mode, the less their mathematics score improves in CAT. For Level-2, the positive direct effect by the teachers' attitude towards ICT (G_{01} = 0.463) implies that the teachers' positive attitude towards ICT tends to improve students' achievement in CAT. Likewise, teachers' attitude towards CAT mode and teachers' attitude towards ICT provide positive direct effects on the improvement of mathematics achievement in CAT. Positive direct effect by teachers' attitude towards CAT mode (tch_cat) (G_{02} = 0.106) suggests that the more positive attitude towards CAT, the more improvement of mathematics score in CAT.

To summarise the results of direct effects, there is a direct effect from students' attitude towards respective test mode on their achievement improvement in all three HLMs. Their attitude towards the test modes of PPT and LOT positively affects their achievement improvement but their attitude towards CAT has a negative effect. This may be because students are familiar with PPT and LOT modes, but they are not familiar with the concepts of CAT, and such concept is itself complicated. Consequently, the study required further investigation to explore the negative effect of the attitude towards CAT. Moreover, students' attitude towards ICT directly affects their achievement improvement in LOT_HLM and CAT_HLM. However, there is no direct effect of their attitude towards ICT in PPT_HLM. This is because LOT and CATs that are ICT-related test mode are related directly to the attitude towards ICT.

Each of PPT_HLM, LOT_HLM, and CAT_HLM has only one significant cross-level interaction effect between Level-2 and Level-1 predictors and the outcome scales at Level 1. The cross-level interaction is between students' attitudes towards test mode (for example, attppt, attlot, attcat) and the teachers' specific formative assessment practices (t_spfa). In PPT_HLM, the significant cross-level interaction effect indicates that the teachers' specific practices of formative assessment interact with students' attitudes towards PPTs (attppt) with an interaction effect coefficient of 0.363. Likewise, in LOT_HLM, the significant cross-level interaction effect indicates that the teachers' specific formative assessment interacts practices with students' attitudes towards LOT (attlot) with an interaction effect coefficient of 0.086. Similarly, in CAT_HLM, the significant cross-level interaction effect indicates that the teachers' specific formative assessment practices interact with students' attitudes towards CAT (attcat) with an interaction effect coefficient of 0.265. Consequently, these findings suggest that the teachers' specific practices of formative assessment have a positive effect on the slope of students' attitude towards test mode (PPT, LOT, and CAT) on their improvement in mathematics achievement. Therefore, the teachers' specific classroom formative assessment practices are considered important factors from the teacher-level, and school principals and educational leaders support teachers to have more effective specific practices of classroom formative assessment.

This study calculated the coordinates for the graphs of the cross-level interaction using the procedure of Aiken and West (1991). The researcher of this thesis applied the final model equation to calculate the coordinates of the cross-level interaction and kept other scales and the direct effect of the scale as constant or zero.

In PPT_HLM, the equation below shows the part of the final model equation involving the cross-level interaction effect of t_spfa and attppt.

$$Y_{ij}^{PPT} = G_{00}^{PPT} + G_{30}^{PPT} * (\text{attppt}) + G_{31}^{PPT} * (t\_spfa) * (\text{attppt}) + R_{ij}^{PPT}$$

G_{00} represents the average students' mathematics achievement across classroom. In this two-level PPT_HLM model, $G_{00} = 0.469$, $G_{30} = 0.229$. and $G_{31} = 0.363$.

Hence, these results in the following equation:

$$Y_{ij}^{PPT} = 0.469 + 0.229 \text{ (attppt)} + 0.363 \text{ (t\_spfa) (attppt)} + R_{ij}$$

The study used the above equation to calculate teacher-level coordinates to obtain a graphical representation of the cross-level interaction effect. The mean and standard deviation for attppt is zero and one, respectively. Likewise, the mean and standard deviation for t_spfa are zero and one, respectively. For example, teachers' specific practices of formative assessment with the most frequency (High) by the high level of students' positive attitude towards PPT (t_spfa = 1; attppt = 1); Y (Students' improvement of mathematics achievement in PPT) = 0.469 + 0.229 (1) + 0.363 (1) (1) = 1.061. Consequently, **Table 12.4** shows the coordinates and **Figure 12.6** depicts their graphical representation.

Under teachers with the average level of specific formative assessment practice, their students with more positive attitudes towards PPT had a bigger improvement in their mathematics achievement. In addition, the effect of student's attitudes towards PPT on score improvement is stronger for students under teachers with a higher level of specific practices of formative assessment and vice versa.

Table 12.4
Coordinates due to cross-level interaction on the achievement improvement in PPT

	t_spfa = -1	t_spfa = 0	t_spfa = 1
att_ppt = -1	0.335	0.240	-0.123
att_ppt = 0	0.469	0.463	0.469
att_ppt = 1	0.603	0.698	1.061

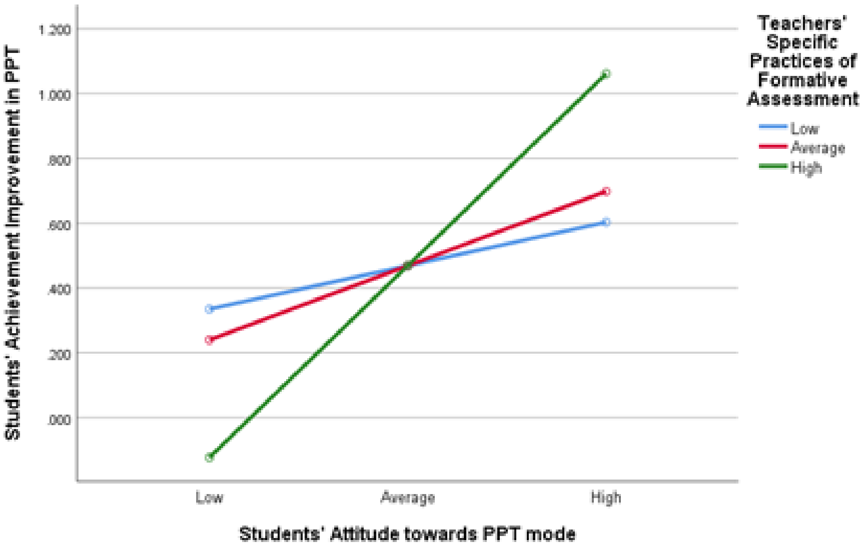

Figure 12.6 Cross-level interaction effect of PPT_HLM

In LOT_HLM, the equation below shows part of the final model equation involving the cross-level interaction effect of t_spfa and attlot:

$$Y_{ij}^{LOT} = G_{00}^{LOT} + G_{20}^{LOT} * (\text{attlot}) + G_{21}^{LOT} * (t\_spfa) * (\text{attlot}) + R_{ij}^{LOT}$$

G_{00} represents the average students' mathematics achievement across classroom. In this analysis, $G_{00} = 0.229$, $G_{02} = 0.460$ and $G_{21} = 0.086$. These results in the following equation:

$$Y_{ij}^{LOT} = 0.229 + 0.460\ (\text{attlot}) + 0.086\ (\text{t\_spfa})\ (\text{attlot}) + R_{ij}$$

The mean and standard deviation for attlot is zero and one, respectively. Likewise, the mean and standard deviation for t_spfa are zero and one, respectively. Consequently, **Table 12.5** shows the coordinates using the abovementioned equation, and **Figure 12.7** exhibits their graphical representation.

Under teachers with the average level of the specific practice of formative assessment, their students with more positive attitudes towards LOT improved their mathematics achievement to a higher level. In addition, the effect of students' attitudes towards LOT on score improvement is stronger for students under teachers with a higher level of specific practices of formative assessment and vice versa.

Table 12.5

Coordinates due to cross-level interaction on the achievement improvement in LOT

	t_spfa = -1	t_spfa = 0	t_spfa = 1
att_lot = -1	-0.145	-0.231	-0.231
att_lot = 0	0.229	0.229	0.229
att_lot = 1	0.603	0.689	0.775

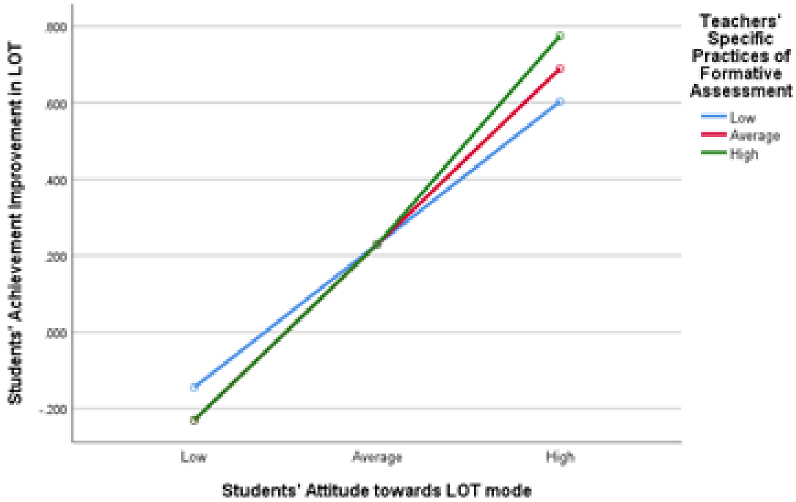

Figure 12.7 Cross-level interaction effect of LOT_HLM

In CAT_HLM, the equation below details the part of the final model equation involving the cross-level interaction effect of t_spfa and attcat.

$$Y_{ij}^{CAT} = G_{00}^{CAT} + G_{20}^{CAT} * (\text{attcat}) + G_{21}^{CAT} * (t\_spfa) * (\text{attcat}) + R_{ij}^{CAT}$$

G_{00} represents the average students' mathematics achievement across classroom. In this analysis, $G_{00} = 0.463$, $G_{20} = -0.192$, and $G_{21} = 0.363$. These results in the following equation:

Y_{ij}^{CAT} = 0.463 – 0.192 (attcat) + 0.363 (t_spfa) (attcat) + R_{ij}

The mean and standard deviation for *attcat* are zero and one, respectively. Likewise, the mean and standard deviation for *t_spfa* are zero and one, respectively. Consequently, **Table 12.6** shows the coordinates using the abovementioned equation, and **Figure 12.8** displays their graphical representation.

Under teachers with the average level of the specific formative assessment practice, their students with less positive attitudes towards CAT improved higher in their mathematics achievement. The effect of students' attitudes towards CAT on score improvement is directly impacted by the teachers' level of specific formative assessment practices. The effect of students' attitudes towards CAT became more negative on their score improvement in CAT under teachers with a lower level of specific formative assessment practices. The higher level of specific formative assessment practices inversely affects the direct relation with their students' attitudes towards CAT and their score improvement.

Table 12.6
Coordinates due to cross-level interaction on the achievement improvement in CAT

	t_spfa = -1	t_spfa = 0	t_spfa = 1
att_cat = -1	1.018	0.655	0.292
att_cat = 0	0.463	0.463	0.463
att_cat = 1	-0.092	0.271	0.634

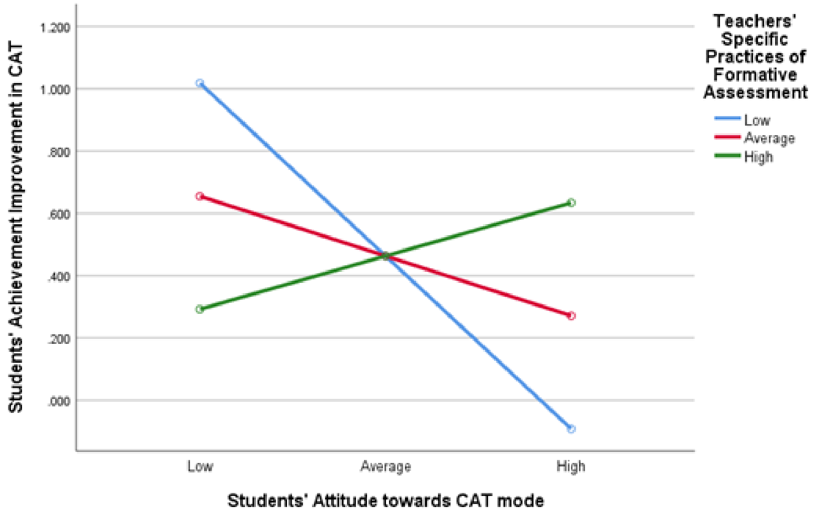

Figure 12.8 Cross-level interaction effect of CAT_HLM

12.6 Variance available in PPT_HLM, LOT_HLM and CAT_HLM

The hierarchical linear modelling provides additional information related to the variance components of the model. The study examined the proportion of variance available in all the levels, and **Table 12.7** shows the estimated variance components for the outcome scale in PPT_HLM, LOT_HLM, and CAT_HLM.

Table 12.7
Estimation of variance components in PPT_HLM, LOT_HLM and CAT-HLM

Model	PPT		LOT		CAT	
	Estimation of variance components					
	Between Students	Between teachers	Between Students	Between teachers	Between Students	Between teachers
Fully unconditional model	0.651	0.57	1.226	0.334	1.384	1.09
Final model	0.637	0.56	1.198	0.312	1.378	1.007
	Proportion of variance available at each level at the fully unconditional model					
Between Students	0.651/ (0.651+0.570) = 0.533 = 53%		1.226/ (1.226+0.334) = 0.786 = 79%		1.384/ (1.384+1.090) = 0.559 = 56%	
Between teachers	0.570/ (0.651+0.570) = 0.467 = 47%		0.334/ (1.226+0.334) = 0.214 = 21%		1.090/ (1.384+1.090) = 0.441 = 44%	
	Proportion of variance explained at each level by final model					
Between Students	(0.651 - 0.637)/0.651 = 0.022 = 2.2%		(1.226 - 1.198)/1.226 = 0.023 = 2.3%		(1.384 - 1.378)/1.384 = 0.004 = 0.4%	
Between teachers	(0.570 - 0.560)/0.570 = 0.017 = 1.7%		(0.334 - 0.312)/0.334 = 0.064 = 6.4%		(1.090 - 1.007)/1.090 = 0.076 = 8%	
	Proportion of total variance explained by final model					
	(0.022 x 0.533) + (0.017 x 0.467) = 0.019 = 1.9%		(0.023 x 0.786) + (0.064 x 0.214) = 0.032 = 3.2%		(0.004 x 0.559) + (0.076 x 0.441) = 0.036 = 3.6%	
	Statistics for covariance components model					
	Deviance	No of estimated parameters	Deviance	No of estimated parameters	Deviance	No of estimated parameters
Fully unconditional model	1986.899	2	2343.436	2	1986.899	2
Final model	1343.009	11	2038.969	7	1784.721	7

In null PPT_HLM, the proportion of variance available at student-level is 53% and at teacher-level is 47%. Likewise, in null LOT_HLM, the proportion of variance available at the student-level is 79%, and 21% is available at the teacher-level. In addition, in null CAT_HLM, 56 % of the variance is available at the student-level, and 44% at the teacher-level.

In PPT_HLM, compared to the null model, the final model explained about 1.7% of variance available at the teacher-level and 2.2% at the student-level. In the final model, this study shows about 6.3 % of the variance at the teacher-level and 2.3 % at the student-level in

LOT_HLM. Relative to the null model, in the final model, the study depicted about 8% of the variance at the teacher-level, and 0.4% at the student-level in CAT_HLM.

However, in LOT_HLM and CAT_HLM, there are more proportions of total variance available by their final model, 3.2% and 3.6%, respectively, than PPT_HLM, 1.9%. Consequently, the large proportion of unexplained variance in each model represents that there can be other significant factors at the student-, teacher- and school-levels. Further, future studies will explore the unavailable variance in student-, teacher- and school-levels.

Further, in PPT_HLM, the deviance value of the final model is reduced by 305.924 compared with the deviance of the null model with 9 additional degrees of freedom. In LOT_HLM, the deviance value of the final model is reduced by 304.469 compared with the deviance of the null model with 5 additional degrees of freedom. Finally, in CAT_HLM, the deviance value of the final model is reduced by 202.178 compared with the deviance of the null model with 5 additional degrees of freedom. In all three models, since the ratio of the decrease of deviance increasing degrees of freedom is greater than one, this study considered a final model better than the null model (Brykl & Raudenbush, 1992).

12.7 Summary

Based on the results obtained from this study, factors such as students' *gender, parental education, attitudes towards ICT,* and *attitudes towards the test modes*, and teachers' *attitudes towards test mode*, *attitudes towards formative assessment*, and their *specific practices of formative assessment* are found to be significant factors at the student- and teacher-levels. This study has clearly explained how these factors interact and influence students' achievement improvement due to the test mode. Due to the three test modes, there are three hierarchical linear modelling (HLM): PPT_HLM, LOT_HLM, and CAT_HLM.

This study compared the results among PPT_HLM, LOT_HLM, and CAT_HLM. At the student-level, the researcher found demographic factors affecting students' achievement improvement, which are *gender* and *parental educational levels*, in PPT_HLM. Attitude towards ICT at the student-level directly affects students' achievement improvement due to the ICT-integrated test modes: LOT, and CAT, see LOT_HLM and CAT_HLM models. Attitude towards test mode directly affects students' achievement improvement due to their respective test mode. In PPT_HLM, the attitude towards PPT at the student-level directly affects the achievement improvement of PPT. In LOT_HLM, the attitude towards LOT at the student-level directly affects achievement improvement due to LOT. In CATT_HLM, the attitude towards CAT at the student-level directly affect the achievement improvement due to CAT. The teacher-level attitude towards formative assessment also directly affects the outcome scales, the achievement improvement due to PPT and LOT, in PPT_HLM and LOT_HLM.

Only one cross-level interaction effect from teachers' specific practices of formative assessment is on the slope of attitude towards test mode (PPT, LOT, CAT) and the achievement improvement due to test mode (PPT, LOT, CAT) respectively, see PPT_HLM, LOT_HLM, and CAT_HLM. The attitude towards ICT from the teacher-level also directly affects that outcome scales, the achievement improvement due to ICT-integrated test modes, LOT and CAT, in LOT_HLM and CAT_HLM. Only one cross-level interaction effect of teachers' specific formative assessment practices is on the slope of attitude towards LOT and achievement improvement.

On identifying student- and teacher-level factors affecting students' achievement improvement due to the test mode, this study has contributed to the growing theories and knowledge in the field of innovation of technology-based test mode for formative assessment and mathematics education. In addition, the use of the HLM approach for the high school education context allows the analyst to model relationships that are, theoretically and empirically, complex. As a result, this study provides results that are more valid and confirmed it further to model for teaching and learning. Overall, the findings of this study provide a meaningful understanding of what student- and teacher-level factors influence students' achievement improvement due to the test modes. Moreover, the findings provide how they affect students' achievement improvement in high school education, and from this study, a comprehensive model of the achievement improvement in test modes in high school education can emerge.

Chapter 13

Discussion and Conclusion

13.1 Introduction

This study explores the effectiveness of paper-and-pencil test (PPT), linear-online test (LOT), and computer-adaptive test (CAT). For this thesis, the researcher used these modes to test classroom formative assessment for mathematics achievement of Grade-10 students in Myanmar. The study integrates two ICT-based test modes: LOT and CAT, in five high schools in Myanmar. The study aims to compare the significant effect of test modes: PPT, LOT and CAT on students' achievement improvement in learning mathematics. This study determines which test mode best helps students improve their mathematics achievement. In addition, it identifies the contextual factors from students and teachers' levels that influence students' achievement improvement due to different test modes.

Further, **Chapter 1** reflects the objectives of the research questions stated. This study adopts a theoretical framework (**Chapter 2**) and employs methods to analyse the resulting data from the responses of teacher and student participants (**Chapter 4**) to answer the research questions. Finally, the study uses the data analysis results to draw conclusions on findings and implications. Specifically, this concluding chapter highlights the design of the study and provides a summary of the findings, relevant implications, and limitations of the study.

13.2 Design of the study

This research study is generally concerned with the introduction of ICT-related test modes for a classroom formative assessment and how the test modes contribute to academic achievement. ICT-related test modes which are LOT and CAT were constructed by the application of Monkey survey and Concerto platform respectively. For the ICT-related test modes, item banks for three concepts (function, remainder and factor) were constructed by the calibration of item in Rasch dichotomous model (see **Chapter 5**). The sixty items for the item-bank of *function* content domain, forty-two items for the *remainder* content domain, and thirty-nine items for the *factor* content domain were assembled in the item banks.

This study poses a number of questions that involve scales or factors at the teacher and student levels to address this. The scales include teachers' formative assessment practices, ICT familiarity and attitudes towards formative assessment, ICT, and test modes at the teacher level. In addition, demographic factors such as age, academic qualification, professional qualification, years of teaching experience, multi-class teaching, and class size have also been included. At the student level, the factors include students' motivation, attitude, and self-efficacy towards mathematics, ICT familiarity and attitudes towards ICT, test modes, formative assessment, and achievement. Demographic factors such as gender, parental education, and expected education levels have also been included. The study investigates the factors and the proposed relationships among them using the responses of 15 teachers, and 687 Grade-10 students from five high schools in Myanmar. The researcher of this study collected the responses during the 2019-2020 Academic Year.

The study gauges the factors in the students and teachers survey through careful selection, modification, development, and scale validation instruments. This thesis bases the appropriateness of the scales on the study objectives and the research questions. Using ConQuest 4.0, the Rasch models establishes the validity and reliability of the scales (Adam, Wu & Wilson, 2015) and confirmatory factor analysis (CFA) using AMOS. Moreover, this study uses the data gathered from the participant teachers and students to establish the psychometric utilities of the scales.

The data from the validated scales are utilised to examine the factors considered in this study and answer the research questions. In the investigation process, the researcher employed descriptive and inferential statistical techniques and corresponding specialised software and used specifically the frequency, mean, standard deviation, and percentage to describe the levels of the attitude towards formative assessment, ICT and test modes, and the distributions of the demographic factors.

As a check on the data collected for this study, the researcher used analysis of variance (ANOVA) using SPSS 26 to determine any significant difference in the improvement of students' achievement levels due to the different test modes. Using MPlus software (v 8), structural equation modelling (SEM) is utilised to examine the directional relations among the scales at each teacher and student level. Further, SEM helps determine the directional influence of teachers' attitudinal and demographic scales as well as students' attitudinal and demographic scales through the intervening scales at the teacher and student-level in the three test mode models (PPT_HLM, LOT_HLM, and CAT_HLM). The study uses HLM 6.08 (Raudenbush, Bryk, & Congdon, 2009) to run three HLM analyses and employes interview questions to gather the qualitative data.

The qualitative responses from teacher and student participants are integrated with the quantitative findings to help enrich the interpretation of the quantitative results on the experiment of the effectiveness of different test modes. The study conducted a semi-structured interview to explore some quantitative findings qualitatively, and the researcher interviewed five male and female students and five teachers each. This study employs mixed methods designs, following the theoretical framework in **Chapter 2.**

13.3 Discussion of findings

This section discusses the key findings from the data analysed. The findings answer the research questions presented in **Chapter 3** and contribute to the Myanmar education system and, more broadly, to the assessment literature. According to the scales and the specific questions, these findings are discussed below.

13.3.1 Effectiveness of test modes on achievement improvement

Research Question 1*: Which test mode (PPT, LOT, or CAT) makes students' achievement improve more when the test mode is applied as classroom formative assessments in Myanmar high schools?*

According to research question 1, the study demonstrates which test modes mainly affect the achievement improvement in mathematics achievement by the ANOVA analyses. After the experiment in which one month of formative assessment using each of the different modes, computer-based test modes make students' achievement improve more than traditional test mode does. As evidenced by the findings from the ANOVA results, the average score improvement for a student on a test marked out of 100 is +30% for CAT, +27% for LOT, and +8% for PPT, and this across each of the content areas. This is because LOT and CAT can provide immediate and specific scores and feedback, but PPT is with the delayed score and feedback. Immediate and specific scores and feedback helps students have more time for learning. Of the two computer-based assessment methods, CAT is the most effective for formative assessments in the participant high schools in Myanmar. Achievement improvement in mathematics due to LOT is slightly lower than the results obtained for CAT. The three-way interaction effect among the independent factors has the smallest. The two-way interaction effect size between the groups and the mathematics-concepts has the most significant effect, followed by the main effect of test modes. This is because each group of students has its strength and weakness in each concept.

As stated in **Chapter 2** (literature review), researchers examined the effectiveness of test modes and compared the traditional test modes and the computer-based or technology-based test modes to show mixed results. Some studies proved that there is no significant impact of test mode on learning progress in LOT and CAT modes (Hattie and Timperley, 2007; Shute, 2008 and Stobart, 2008). Other studies revealed that the computer-based test modes, especially LOT and CAT, cannot improve student achievement than PPT (Nicol and Macfarlane-Dick, 2007; Hattie and Timperley, 2007; and Shute, 2008). However, this study shows that computer-based test modes offer many advantages over PPT and help students improve their achievement (Csapó et al., 2009; and Meijer, 2009). The computer-based test modes can provide immediate feedback and scores. Specifically, while LOT and CAT substitute/replace PPT, LOT can enhance the learning achievement because of prompt feedback (Hartnell-Young, Harrison, Crook, Davies, Fisher, Pemberton, & Smallwood, 2007).

13.3.2 Impact of tested factors on students' achievement improvement

There are two research questions of the exploration of impact of test factors on students' achievement improvement:

2.1: What are students' contextual factors and attitudinal factors that significantly influence their achievement improvement due to the different test mode (PPT, LOT, and CAT)?

2.2: What are the teachers' contextual factors and attitudinal factors, which significantly influence their students' achievement improvement due to the different test mode (PPT, LOT and CAT)?

The study conducted three student-level structural equation models, teacher-level structural equation models and three hierarchical linear models to explore the impact of test factors from student-level and teacher-level on students' achievement improvement. Then, semi-structured interviews to students and teachers were conducted. The findings from structural equation models from student-level and teacher-level explore direct effect and mediating effect or indirect effect on students' achievement improvement. The results of hierarchical linear models of PPT, LOT, and CAT provide the direct and cross interaction effect from the tested factor from students and teachers' levels on the targeted scale by the HLM analyses. Also, the qualitative results from semi-structured interviews are applied to support these quantitative findings.

The above research question number (2.1) is a specific inquiry and relates to the findings from test factors of student level that influence students' achievement improvement in term of test modes. Likewise, research question number (2.2) resides at the core of systematic investigation and relates to the findings from test factors of teacher level. This study discussed their findings according to test factors.

Student-level Factors

According to Research Question 2.1, students' contextual factors and attitudinal factors which significantly influence their achievement improvement due to test modes are found through SEM and HLM analyses. Their findings are discussed as follows:

Influence of students' demographic scales on the achievement improvements in PPT, LOT and CAT models

Among the demographic scales of Grade-10 students, gender (β = 0.090), parental education (β = -0.178) affect their achievement improvement due to PPT. The female students improve their mathematics scores more than male students do. Likewise, other studies mentioned that common potential bias related to gender is small due to different assessment modes (Thorndike, and Thorndike-Christ, 2010; Leeson, 2006; and Hensley, 2015). Gender influences the difference in gained cognitive achievement due to PPT and CBT. However, there is no achievement improvement difference for gender due to the different test modes (Clariana & Wallace, 2002; Horkay, Bennett, Allen, Kaplan & Yan, 2006; Poggio & McJunkin, 2012). Students with the parents' better education level improve less in mathematics achievement in PPT for the parental education level. Most parents with higher education levels do not prefer PPT. There is no impact of their demographic scales on their gained mathematics achievement due to the respective test modes in LOT and CAT models. Students' parent education level significantly and directly affects more on the achievement improvement by the PPT than CAT (Clariana, & Wallace, 2002; Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008). Their expected education level does not significantly influence their learning progress due to these three test modes. However, a few studies found this scale to be a direct factor in learning progress due to different test modes (Horkay et al., 2006, Poggio & McJunkin, 2012; Hensley, 2015).

Influence of students' motivation, self-efficacy, and attitude towards mathematics on the achievement improvements in PPT, LOT and CAT models

The significant impact of motivation, self-efficacy, or attitude towards mathematics on gained mathematics achievement could not be found in all three models of SEM and HLM. Students' gender status and their parents' education levels positively affect their self-efficacy from the result of SEMs in three test modes. Their motivation is positively affected by their attitude towards mathematics learning and self-efficacy and negatively affected by their parents' education levels. The results related to gained mathematics achievement of this study contradict the previous literature review. Researchers such as Ak and Sayil (2006), Newton and Mwisukha (2009) and Geddes, Murrell and Bauguss (2010) indicated that there is a significant relationship between the concept of attitude towards learning mathematics and academic achievement due to any test modes. According to previous studies by Candeias, Rebelo and Oliveira (2010), and Verešová and Malá (2016), higher academic achievement is associated with a positive attitude of a student towards learning due to any test modes. Schunk (1995) has highlighted that motivation and self-efficacy are key factors in students; achievement improvement because their level of motivation and self-efficacy are enhanced when students achieve more due to the effect of any test modes. In addition, Liu and Koriala (2009) study provides that mathematics self-efficacy was a significantly positive predictor of mathematics achievement.

Influence of students' ICT familiarity and attitude towards ICT on the achievement improvements in PPT, LOT and CAT models

Students' ICT familiarity does not affect the achievement improvements due to the respective test mode in all three test mode models of SEM and HLM analyses. According to SEMs at student level, their ICT familiarity and attitude towards ICT is positively affected by their parents' education levels. Boys are more familiar with ICT than girls are. Also, boys have more positive attitude towards ICT than girl does. Their attitude towards learning mathematics and ICT familiarity positively impacts on their attitude towards ICT. In PPT HLM model, attitude towards ICT does not affect the achievement improvements due to PPT. Their attitude towards ICT ($\beta = 0.614$) in LOT and attitude towards ICT ($\beta = 0.386$) in CAT has positive impacts on the achievement improvements due to the respective test mode in respective HLM analyses. Previous literature shows that ICT familiarity and attitude towards ICT affect students' achievement improvement due to CBT. The more ICT familiarity students have and the more positive attitude towards ICT they have, the more gained in their achievement (Leeson, 2006; Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008).

Influence of students' attitude towards formative assessment on the achievement improvements in PPT, LOT and CAT models

The study by Quyen and Khairani (2017) reveals significant relationships between students' attitude towards formative assessment and their improvement in achievement in traditional test modes and technology-based ones. However, in this study, no significance in the relationships of students' attitudes towards formative assessment and their improvement in achievement in test modes, apart from CAT at student level. Participant students' attitude towards formative assessment (β = 0.170) positively affect their gained mathematics achievement due to CAT. Their expected education and motivation positively affect their attitude towards formative assessment, but their self-efficacy negatively affect it. Findings from HLM and SEM analyses contradict their interview responses. This study conducted a semi-structured interview to explore their attitude towards classroom formative assessment qualitatively. The researcher of this thesis interviewed five male students and five female students and analysed the responses using thematic analysis. Five interviewees have positive attitudes towards formative assessment, and others have negative ones with their reasons. Those who have a positive attitude towards formative assessment believe that formative assessments motivate their improvement in learning. Examples of such responses are as follows: *"I like classroom formative assessments because they cover an only small amount of content. They make me study more. They help me to improve achievement in learning. I can know the weak point in new concepts. (Student-female interviewee No. 4 (Sf4))".* Students who have a negative attitude towards formative assessment think that formative assessment does not help their learning progress and improve their test anxiety. An example of such a response is as follows: *"I do not like classroom formative assessments because I am always busy with tests. My test anxiety increased because of them. There is not enough time to study for a concept whole, especially for slow learners like me. The contents were forgotten after testing. (Student-male interviewee No. 2 (Sm2))"*

Influence of students' attitude towards respective test modes on the achievement improvements in PPT, LOT and CAT models

Through SEM analyses and HLM analyses, the attitude towards test mode is significant test factor on students' achievement improvement due to respective test mode. Their attitude towards PPT (β = 0.359), attitudes towards LOT (β = 0.083) and attitudes towards CAT (β = -0.192) significantly influence their achievement improvement due to PPT, LOT and CAT respectively. The more positive attitude towards PPT, the more their mathematics score improves due to PPT mode. Likewise, the more positive their attitude towards LOT mode, the more their mathematics score improves due to LOT mode. Additionally, the more positive attitude towards CAT mode, the less their mathematics score improves due to CAT. This is because a more positive attitude towards CAT, the less is their improvement in achievement.

Students' attitude towards ICT positively affects their attitude towards LOT and CAT from the SEM analyses. However, previous literature shows a positive relationship between students' attitude towards test modes and their achievement improvement. Especially in computer-based test modes, their attitude towards test modes positively influences their achievement improvement (see Agarwal & Prasad, 1999; Hu, Chau, Sheng, & Tam, 1999; Venkatesh, 1999; Ong & Lai, 2006; Lee, 2006; Van Raaij & Schepers, 2008; Way, Davis, & Fitzpatrick, 2006; Penuel, 2006). This is because students who believe that computer-based test modes will improve their knowledge, comprehension, and performance will achieve more in their learning.

For this study, the researcher conducted qualitative semi-structured interviews to explore the findings related to students' attitude towards test modes. Some students believe that PPT is more familiar and easier to take. They prefer paper-based feedback and teacher feedback. On the other hand, some students like computer-based tests, but they prefer to take PPT. This is because they do not have enough knowledge to use computers and the internet, and they do not access ICT at home and at school. Their responses pinpoint that their choice is due to the lack of ICT familiarity and accessibility. Students who do not trust the ability of ICT will choose PPT, which is easier for them. The results from qualitative analysis support the findings from quantitative analyses that are SEMs and HLMs. Models of PPT show no impact of attitude towards ICT in achievement improvement.

Some students who prefer PPT believe that PPT is more familiar and easier to take. They prefer paper-based feedback and teacher feedback. Some students like a computer-based test, but they prefer to take PPT. This is because they do not have enough knowledge to use computers and the internet, and they do not access ICT at home and school. Their responses pinpoint that their choice is due to the lack of ICT familiarity and accessibility. Students who do not trust the ability of ICT will choose PPT, which is easier for them. An example of such a response is "*I like computer-based test because we get immediate score and feedback which makes me study more, know which is my error or mistakes and motivate in learning. In addition, I am not familiar with computer and ICT devices. Wait for a good connection. I prefer to choose PPT because I like teacher feedback. I want a teacher to explain my error rather than machine feedback. I like hardcopy (print) feedback to electronic feedback. This is because I am not familiar with the computer, and I have less access to ICT.* (*Sf1)"*

It appears that some students would prefer to have more participation in the computer-based test modes. This is because they believe that they have enough ICT familiarity and accessibility and a positive attitude towards ICT. The preference of LOT and CAT is associated with immediate score and feedback. Students who prefer LOT to CAT think that LOT is simpler than CAT. Otherwise, students who like CAT more than LOT mentioned that CAT has shorter test taken time. The responses related to choosing either LOT or CAT as their favourite are as follows: "*I prefer LOT. I like immediate feedback and scoring and its simplicity. LOT is simpler than CAT. The reason why I choose LOT as my like is that I know how to use computer and internet well. Mostly I access the internet, computer at home, and school. In addition, I trust the utility of ICT in the future education and work. (Sf5)" and "I like CAT. It is because I like the short test time and its immediate score and feedback from CAT. Their feedback is clear. I like the use of computer and the display of one question in one screen because no need to focus on other questions. (Sm1)"*

Teacher-level Factors

According to research question number 2.2, there are some contextual factors and attitudinal factors, from teacher questionnaire, which directly and indirectly influence their achievement improvement due to test modes through SEM and HLM analyses. Their findings are discussed as follows:

Influence of teachers' demographic scales on the achievement improvements in PPT, LOT and CAT models

Participant teachers' demographic scales are their highest education levels and their highest qualification level in the education field. In all three test mode models, there is no main direct effect on their demographic scales on their achievement improvement due to the respective test modes. However, the study by Brown (2004), and Jamil, Tariq, and Shami (2012) reveal that some external factors, which are their highest education levels and their highest qualification level in education, positively influence their student achievement improvement due to the effect of any test mode.

Influence of teachers' workload in their classroom on the achievement improvements in PPT, LOT, and CAT models

This workload scale in their teachers' classrooms is measured in two scales: multi-subject teaching and class size. These scales can hinder classroom teaching and assessment. There is no effect of their workload scales on their achievement improvement due to the respective test modes in all three test mode models. However, in previous studies by Burghof, 2001; Eggen, 2007; Rashad, and his colleagues 2008 reveal that class size and multi-subject teaching can increase their work loading related to test development and administration and negatively affect students' learning (Burghof, 2001; Eggen, 2007; Rashad, et.al, 2008).

Influence of teachers' formative assessment practices and their attitude towards formative assessment on the achievement improvements in PPT, LOT and CAT models

There are two scales in teachers' formative assessment practices: *general practices and specific practices* in formative assessment. There is no direct effect of their general practices and specific practices in formative assessment on students' achievement improvement due to the respective test modes in all three test mode models of HLM analyses. However, their general practices of formative assessment (β = 0.482) and attitude towards formative assessment (β = 0.493) positively impact on their specific practices of formative assessment in their three SEM models due to three test modes at teacher level. Previous literature proved the direct relationship between teachers' practices of formative assessments and their students' achievement improvement due to any test modes. Ellis, Loewen and Erlam (2006) described that teachers' practices of formative assessments facilitate students to achieve better, inform them on how well they are doing, help them to improve specific points, encourages them to work hard and guide them what they need to focus on when they are having difficulty. In addition, teachers' practices in formative assessment positively influence their students' achievement progress in any test modes (Quyen and Khairani, 2017). However, this study reveals a cross-level interaction effect from the teachers' specific practices of formative assessment on the relationship between the attitude towards test mode and the achievement improvement due to the respective test mode in the test mode models.

The teachers' attitudes towards formative assessment (β = 0.319) and (β = 0.460) directly affect students' achievement improvement due to PPT and LOT. However, there is no direct effect of teachers' attitude towards formative assessment on students' achievement improvement in CAT model. However, the literature provides mixed findings. Nesa (2014) no relationship between teachers' attitude towards formative assessment and their students' achievement improvement in any test modes. On the other hand, Pinchok and Brandt (2009) mentioned that teachers' attitudes towards formative assessment positively influence their students' achievement improvement in any test modes. This is because teachers believed that formative assessment helps students to progress in their learning and the feedback helps them improve specific points or help plan their learning. Other studies support the results of this study (Opre, 2010; Quyen and Khairani, 2017) in that there is a direct effect of teachers' attitude towards formative assessment of their students' achievement progress in any test modes.

This study conducted a qualitative interview conducted to explore the findings related to teachers' attitude towards formative assessment. All participant teachers in the interview session believed that formative assessment is necessary and effective in students' gained achievement. The results from qualitative analysis support the findings from quantitative analyses, which are HLMs. Their PPT and LOT models show no impact of teachers' attitude towards the formative assessment in the achievement improvement. The interview result cannot reveal that there no significant impact of teachers' attitude towards the formative assessment in the achievement improvement in the CAT model.

All participant teachers in the interview session believed that it does not matter which test mode is used, the nature of formative assessment is necessary and has effect on students' gained achievement. They believed that formative assessment makes students learn more, helps to catch up with every lesson, and points out their strength and weakness. See an example of a response: "*I totally agree with the use of FA, if teacher knows which mistake students did, teacher can do the remedial and help their teaching. However, teacher need more time for scoring. I cannot explain their individual mistake. I cannot do tutorial for every concept. And I do not have time to score and tell the mistake to individual. (Teacher interviewee No. 1 (T1))"*

Influence of teachers' ICT familiarity and attitude towards ICT on the achievement improvements in PPT, LOT and CAT models

Due to the respective test mode in all three test mode models of HLM analyses, there is no direct or cross interaction effect of teachers' ICT familiarity on students' achievement improvement. However, their ICT familiarity ($\beta = 0.843$) positively impact on their attitude towards ICT. Their attitude towards ICT does not directly affect students' achievement improvement due to PPT. There is no direct effect of the teachers' attitude towards ICT ($\beta = 0.490$) and ($\beta = 0.463$) on students' achievement improvement in LOT and CAT respectively. According to SEM models at teacher level, teachers' attitude towards ICT is positively affected by their ICT familiarity. However, present findings contract the study of Quyen and Khairani (2017). Their finding is that teachers' ICT familiarity and attitude towards ICT positively influence their students' achievement in any test mode.

Influence of teachers' attitude towards test mode on the achievement improvements in PPT, LOT and CAT models

According to SEM analyses at teacher level, teachers' attitude towards LOT and CAT is positively affected by their attitude towards ICT. Their specific practices of formative assessment (β = 0.826) and (β = 0.918) positively impact on their attitude towards PPT and CAT. Through HLM analyses, teachers' attitude towards PPT (β = 0.278) and CAT (β = 0.106) directly affects students' achievement improvement. There is no direct effect of teachers' attitude towards LOT on students' achievement improvement. According to the responses of the interview, four teachers prefer LOT, and two teachers prefer CATs. Only one teacher in the interview session prefers PPT. All teachers in the interview session believe that the effect of LOT and CAT has a positive impact on students' achievement improvement. However, previous literature shows a relationship between teachers' attitude towards test modes and their students' achievement improvement. Teachers' attitude towards test modes, particularly computer-based test modes, have a positive impact on their students' scores (Agarwal & Prasad, 1999; Hu, Chau, Sheng, & Tam, 1999; Venkatesh, 1999; Ong & Lai, 2006; Lee, 2006; Van Raaij & Schepers, 2008, Khoshsima, 2019).

According to their interview responses, there are different preferences in test mode with different reasons. Teachers who prefer PPT believe that schools are not ready to use LOT and CAT. This is because they need enough computers and ICT access. In addition, LOT and CAT need more effort to construct LOT and CAT. See the example of such a response: ***"**Currently there are not enough computers for all students in classroom. So, I think computer-based tests are not ready to use in the classroom currently. So, I chose the paper based as the best applicable. There are advantages of CBT, because they do not need for scoring, and they are able to provide feedback immediately. In PPT, students need to wait for teacher scoring and feedback. (T1)".* However, some teachers respond that LOT and CAT are their preference because these test modes make them save time and effort and teachers can easily trace the students' learning progress. A response who prefers LOT and CAT to PPT: *"I like LOT and CAT because it scores automatically and provide feedback immediately. Trace individual improvement. Security is good. Everything restores in computer. And I am so familiar with computer, I understand the advantage of technology. I prefer LOT and CAT to PPT. I spend one time for make questions and save and applied multiple time. Especially I like item-bank. Sacrifice for a year, and then apply multiple times. Some necessary will be added because student will change, contents do not change, and teacher or instructor does not change. If instructor change, he can pass it on next one. But I consider for other teachers who are lack of knowledge of computer. They will face more challenges. For students, CAT and LOT is good. (T5)".*

13.3.3 Perception towards PPT, LOT and CAT as classroom formative assessment

This study explores students' and teachers' perception towards PPT, LOT, and CAT to provide a future innovation in LOT and CAT for classroom formative assessment. Below is a research question for such exploration:

2.3: What are students' and teachers' perception towards PPT, LOT, and CAT as classroom formative assessment?

For this study, the researcher conducted semi-structured interviews with students and teachers. Five teachers, five male students, and five female students from five participating schools, were participated in these interviews. There is an interview question, in students' semi-structured interview session, which examines their perception towards PPT, LOT, and CAT is "*Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?*" Because of students', interview responses, students who do not have enough ICT familiarity and accessibility and a positive attitude towards ICT choose PPT as their favourite. On the other hand, students chose LOT and CAT because of their immediate scoring and feedback. The preference of LOT and CAT is associated with ICT familiarity, accessibility, attitude towards ICT, and immediate scoring and feedback. However, the major challenge in taking LOT and CAT is a poor internet connection. The example of responses related to choosing LOT and CAT as their favourite is as follows:

"I like LOT and CAT because I like to study or test on screen. I believe that the use of ICT makes me to improve my confidence in learning. I prefer to choose LOT because I get immediate scores and feedback according to my answers in detail. No challenges, apart from sometimes, a poor internet connection. I prefer testing in the computer classroom because they turn on the soothing music. The computer classroom is a better place to take a test. I prefer computer feedback to teacher feedback. (Sf3)". The response of choosing PPT as favourite is "*I prefer PPT to two other computer-based test. This is because I do not have enough knowledge of computer use. I am not familiar with the internet and computer. I have no confidence in using the computer. I prefer paper-based feedback. That is the best for us. We can study it again. I prefer the print of feedback sheet to the electronic one. Print one can be saved a long time because the electron one easily to lose and I still do not know how to save it electronically or soft copy, because I am not familiarly with a computer. (Sm4)".*

An interview question that explored teachers' perception towards PPT, LOT, and CAT is "*Which test mode do you prefer, paper-based or linear-online or computer-adaptive test* *mode?*" According to interview a response, teachers have perceptions towards different test modes with their challenges. Teachers who prefer PPT highlight the challenges of LOT and CAT. This is because LOT and CAT need enough computers, ICT access, and more test-administers. For example, "*I chose PPT as the best applicable because it does not need extra support. Not enough computers for all students in a classroom. For LOT and CAT, there are extra teachers and periods because there are not enough computers in the classroom. (T2)"*. However, teachers who choose LOT and CAT as their favourite pinpoint the advantage of LOT, and CAT overtakes the disadvantage of PPT. PPT makes teachers have more workload question making, test security, administration, score, and feedback. There is an example of such a response: *"I like LOT. I can easily measure the ability and provide scores and feedback. Teachers save time for question making (item-bank), test security, testing administration (because students cannot copy other answers, scoring and providing feedback. Class teachers can trace their abilities. Students can get to know their abilities immediately. (T3)"*

After conducting these interview sessions, the researcher explored students and teachers' perception towards PPT, LOT, and CAT as well as their challenges in conducting PPT, LOT, and CAT. There are more advantages of LOT and CAT than PPT for teachers and students because LOT and CAT overtake challenges of PPT, for example, overtime for scoring. For students, LOT and CAT can provide immediate scoring and feedback. LOT and CAT can easily trace students' learning progress and so teacher can save more time and effort for classroom instruction. However, LOT and CAT need enough computers, ICT access, and ICT familiarity for teachers and students.

To sum up, students and teachers without enough ICT familiarity and attitude towards ICT are more likely to choose PPT as their favourites. Students and teachers with ICT accessibility, ICT familiarity and positive attitude towards ICT prefer to participate in LOT and CAT test modes.

13.4 Limitations and recommendation

Although the present study contributes significantly to the development of knowledge in classroom assessment practices, mathematics assessment, and learning, it has several limitations. First, this study focuses on students and mathematics teachers in high schools of Yangon Region, Myanmar. Consequently, the results of this study may be restricted to other states and areas such as the rural areas and students from other primary and secondary levels.

Consequently, further studies focus on other subjects of the primary, secondary or university level in other states.

Second, data collection had also posed some challenges in this study. There are many high schools and students in Yangon Region, a developed and commercial city in Myanmar. Due to the nature of the study, sample schools needed to have already installed internet access and computer classrooms. Even though most high schools in Yangon Region are ready to conduct this study according to national educational data, upon visiting the schools, the researcher found that only a very limited number were, in fact, ICT-ready. This led to delays in the data collection, and the data cannot represent all high school students in Yangon Region as expected in a large-scale study.

Third, student- and teacher-level factors included in this study do not represent all the factors that may affect or influence students' mathematics achievement improvement due to test modes. There are many other factors, which may contribute to students' mathematics achievement, such as mathematics-related resources from home and school, anxiety in mathematics learning, previous year achievement, and test anxiety. There are also other factors related to teacher level such as their working experience, ICT anxiety, Moreover, this study did not include school-level factors such as principals' attitude towards new test modes or other related factors. Consequently, it demands future studies that explore other factors from student-, teacher- and school-levels to get a better picture of assessment innovation in education.

In addition, this study modified and utilised some instruments developed in other countries. Although the instruments were validated and found to have acceptable measurement properties. This study suggested developing new relevant instruments that are more appropriate for Myanmar students and teachers to obtain more results that are meaningful. Then, administration of the survey instruments or scales should be made consistent (i.e., distribution and collection and time allotted for completing the instruments or scales) as much as possible throughout the data collection. Such survey instruments will reduce the additional facets or biases that need to be considered in data analysis. It is also crucial to revise interview questions to elicit more information about the tested scales and provide in-depth interpretation of the quantitative findings should further research in the same area be undertaken.

Additionally, according to the nature of this study, the researcher should conduct a longitudinal study. Due to the time and financial limitations, researcher will conduct a longitudinal study in the nearest future. Due to some limitations, suggestions are consequently advanced for further studies. Samples from all schools and more classes representative of the target population are needed, and proper sampling methods such as multistage random sampling should be employed, considering the hierarchal nature of the data. As a result, in the future, more large-scale studies and longitudinal studies will collect data from more schools and explore the deeper information related to the innovation of technological test modes. Findings in the nearest future will provide thorough information for a national educational system. This study strongly suggests longitudinal studies to examine the interrelationships among student- and teacher-level factors and their impact on students' mathematics achievement. In addition, longitudinal studies allow researchers to trace changes of participants' characteristics, such as students' attitudes and motivation, which may provide a more in-depth understanding of the problems faced by the Myanmar High schools.

13.5 Methodological implications

The research questions advanced in this study address which test mode is more effective for students' achievement improvement and how many factors affect students' different test modes and their attitude towards test modes (see **Chapter 2**). Review of related literature facilitated the research design employing both quantitative and qualitative methods for obtaining data. As the research methodology, mixed method with multilevel analyses is helpful in this study to draw meaningful findings. As the quantitative method, the pre-test and post-test were applied for measuring students' improvement in mathematics achievement due to the three test modes, and questionnaires were applied to measure the factors influencing their improvement. The teacher-level factors and student-level factors were included in questionnaires. The qualitative methods employed interviews. This study is composed of various statistical data analyses for answering the targeting research questions. The comparison of students' achievement improvement, their significant factors in teacher-level and student-level, the in-depth interpretation from interview response among three test modes provide the in better interpretation and application of theoretical framework for education innovators or researchers.

Firstly, it is important to subject the questionnaire to rigorous validation processes to ensure dependable data and achieve a desirable degree of objectivity. This requires the use of the appropriate techniques when gauging any measuring instrument. This study has applied confirmatory factor analysis (CFA) and Rasch measurement models to validate and verify the instruments or scales. The CFA provides information of confirm the structure of the instruments and scales. The Rasch measurement models have special properties of item and person independence and unidimensionality, and its characteristics of being mathematically sound provide strength and ensure the objectivity measurement. Hence, the use of the CFA and Rasch model is promising, especially in the context of Myanmar or other developing countries, where the CFA and Rasch model are not widely employed and where educational research in the form of surveys is common practice.

Then, this study transformed raw scores into logit scores to measure uniformity for more valid interpretation of the results. The study employed weighted likelihood estimation (WLE) logit score through the Rasch model of ConQuest program to transform scores to measures. This technique has the advantage of a minimised estimation bias compared to similar transformation techniques. Several studies in the literature employed a WEL technique as part of the data analyses in recent large-scale studies such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). Consequently, the use of the estimation of WLE through the Rasch model of ConQuest program will be more widely employed in other social research studies.

There is a web of educational scales operating at different levels. As such, educational data are nested in nature. It is important to capture this characteristic of educational data to untangle the web of relationships among educational factors. The data collected for this study is hierarchical in nature. It is always interesting to examine the combination of different dataset hierarchies into a single level, even when problems arise. Problems of analysing multilevel data using single-level techniques usually have something to do with the effect of introducing bias and over- or underestimation of the magnitude of the effects. **Chapters 4** and **12** detailed this discussion. In this study, the researcher carried out single-level analyses using the MPlus (v 8). The resulting models form the basis for further analysis using single level techniques. This study carried out further analyses employing a multilevel technique to acknowledge the problems associated with single-level techniques. This either eliminates or minimises issues inherent within single-level modelling techniques. Analysing the relationships and interactions between teacher-level scales and student-level scales and the influence of all scales from the two levels on the outcome scales requires a proper technique. Furthermore, such a technique also provides for the estimation of the cross-level interaction effects that may be present between the teacher-level factors and student-level factors. Consequently, Hierarchical Linear Modelling (HLM), the multilevel modelling technique, was employed in this study and the researcher applied HLM (v.6) for the HLM analysis. Hence, appropriate techniques such as

SEM and HLM in the analysis of this kind of data should be useful for further social research studies.

Additionally, so much information needs to be unpacked in the educational context. Although the quantitative data is common in use and achieves a high level of objectivity and generalisation, the qualitative data can provide more information and deeper interpretation about educational phenomena should be meaningful. The researcher collected qualitative data from selected teacher and student participants through semi-structured interviews to support and enrich the interpretation of the quantitative results. This is because qualitative data can provide more profound information about the phenomena. This implies that it is possible to explore other factors that are not included in this study. Analysis of interview responses was undertaken by identifying the common themes, which are based on the scales from quantitative studies. Consequently, this study implied the implication of triangulation in mixed methods designs.

Consequently, the current developments in validation and verification by Rasch measurement models and confirmatory factor analysis (CFA), moderation effects, and mediation effects by the multilevel analysis and mixed methods analysis are even more promising in understanding the complex educational phenomena. In Myanmar or other developing countries, where mixed methods design, the validation and verification of Rasch models, and the CFA analyses, the single level and multilevel analysis by SEM and HLM are not widespread, local educational researchers should find these techniques more advantageous.

13.6 Theoretical, and practical implications

The trend of ICT application in classroom education for over the last three decades has been the centre of attention for many education researchers. Because of the utility of ICT and the accessibility of ICT in classroom teaching and learning in all levels of education, the increase in the application of ICT in teaching, learning and assessment has aroused the interest of educational researchers, who attempt to find answers to the basic and practical questions of *'How effective is ICT integrated teaching, learning and assessment?'* or *'Which technique is the best and most suitable for the targeted group?'* Many of these researchers have examined numerous factors that may contribute to the effectiveness of test modes in teaching, learning, and assessment. Further, many researchers have found that these factors related to the examination and comparison of test modes appear to affect students' achievement improvement. However, it is impossible to generalize or apply these findings in the school setting in some locations.

One reason is that most research on this problem has only been carried out where the education systems are well equipped, mostly in well-developed countries. There is very little similar research carried out in developing countries such as Myanmar, where they are only just taking their first steps to update their education systems. Consequently, the underlying assumptions about the factors playing a role in shaping students' achievement improvement may not be applicable. This was why the present study was conceptualised. This study established an initial conceptual framework by reviewing existing literature on different factors found to have a significant influence on students' achievement improvement due to different test modes. The researcher presents this framework in **Chapter 2** (**Figure 2.2)** and included teacher- and student-level factors that are found by numerous studies to affect achievement improvement as the effect of test modes.

These findings lead to some theoretical and practical implications. First, this study assembled and calibrated items for LOT and CAT item banks. Through Rasch dichotomous model, items for Function concepts, Remainer concept, and Factor concept are assembled and calibrated for item banks. This study provides three item banks of Function, Remainder, and Factor for Grade 10 mathematics teachers. Also, it technically contributes to the psychometricians, test-makers, and classroom teachers on how to construct future item banks for various subjects that could be used in schools and institutions.

One of the main purposes of this study is to determine which test mode is more effective for formative assessment in terms of students' achievement improvement. Due to the ANOVA analysis, ICT-related test modes (LOT and CAT) help to improve students' achievement in mathematics more than the PPT mode does. LOT is also much easier to utilise for both teachers and students than CAT. Therefore, in Myanmar or other developing countries that are ready to update their education system, the use of ICT-related test modes is advantageous in students' learning progress. Among the ICT-related test modes, the use of LOT in classroom formative assessment is much easier than the use of CAT which is very new for teachers and students from developing countries.

Additionally, in this comparative study, students have different prior abilities in mathematics achievement. In addition, some students achieve higher in different concepts due to the level of concepts' difficulty. For example, in this study, students improve their achievement in function concepts than in remainder and factor concepts. For further studies, it is better to do comparative studies in test mode on the same level of concept difficulties or to check the participants' prior knowledge or ability.

Second, this study provides empirically based analytical procedures for testing and extending the existing research frameworks and models and how different factors affect students' achievement improvement in mathematics. The research theoretical framework developed for this study takes into account the three stages of teaching and learning processes. The framework also contributes evidence on what factors from teachers and students influence students' achievement improvement in mathematics due to test modes in their use for formative assessment. Particularly, comparing student- and teacher-level factors in three test modes is conducted to determine factors contributing to enhancing students' achievement improvement due to test modes. Because of employing different statistical techniques, student- and teacher-level structural equation models (SEM) and the hierarchical linear models (HLM) accounting for students' achievement improvement in the Myanmar High Schools have emerged. As evidenced by the findings from this study, student- and teacher-level factors influence one another and positively affect student achievement improvement in mathematics due to PPT, LOT and CAT. These analyses provide a better understanding of what factors, directly and indirectly, influence Myanmar High School students' achievement improvement.

Based on the results obtained from the SEM and HLM analyses, direct and indirect factors influence students' achievement improvement at student level and teacher level. In all test modes (PPT, LOT and CAT), **students' and teachers' attitude towards test modes** significantly impacts on students' achievement improvement. In addition, students' and teachers' attitude towards ICT directly affects students' achievement improvement due to computer-based test modes, especially LOT and CAT. Consequently, this study provides implications of the importance of participant's attitudinal factors related to test modes in innovation in classroom assessment in either paper-based or ICT-assisted. Students' and teachers' attitude towards ICT is direct factor to their achievement improvement due to computer-based test modes (LOT and CAT). Education innovators need to examine the participant's attitudinal factors such as attitude towards test modes and ICT for the readiness of installing new test modes into classroom assessment.

Students' attitude towards computer-based test modes, LOT and CAT, is positively influenced by their ICT familiarity and attitude towards ICT. Also, teachers' attitude towards computer-based test modes (LOT and CAT) is positively affected by their attitude towards

ICT. Consequently, important factors for updating the education system with ICT are participants' ICT familiarity and attitude towards ICT.

Teachers' attitude towards ICT is positively affected by their ICT familiarity. Ministry of education encourages in-service teachers to get ICT access and improve ICT familiarity. Students' attitude towards ICT is positively influenced by their ICT familiarity and their attitude towards mathematics learning. So, schools encourage students to improve their ICT familiarity and their attitude towards mathematics learning,

Teachers' attitude towards either paper-based or computer-based (PPT or CAT) test modes is positively affected by their specific practices of formative assessment, which is affected by their general practices of formative assessment and their attitude towards formative assessment. Also, teachers' attitude towards formative assessment directly affects students' achievement improvement due to test modes, either paper-based or computer-based modes, especially PPT and LOT. Through HLM analyses, teachers' specific practices of formative assessment positively affect the slope of students' attitude towards test mode (PPT, LOT, and CAT) on their improvement in mathematics achievement. Thus, it suggests that teachers' specific classroom formative assessment practices are considered important factors from the teacher-level, and school principals and educational leaders support teachers to have more effective specific practices of classroom formative assessment. Ministry of education encourages and supports in-service teachers to use general practices and specific practices formative assessment. Hence, pre-service teachers need more training or programme that provide the information about the use of classroom formative assessment.

There are two demographics factors, which indirectly affects students' achievement improvement due to test modes. They are gender and expected education levels. Gender has negative impacts on the ICT familiarity and attitude towards ICT. Boys had more ICT familiarity and more positive attitude towards ICT than girls did. So there need more programmes for upgrading girls' ICT familiarity to reduce the gender gap. Then, gender positively affects the expected education level. Girls are expected to accomplish a higher level of education in their future than boys do. Thus, more studies need to explore the gender gap concerning the expected education level.

Students' attitude towards formative assessment, which is indirectly related to their achievement improvement due to test modes, is positively related to their expected education levels and motivation. Students who have a higher expectation in their future education are more likely to have a positive attitude towards formative assessment and prefer more formative assessment for their learning progress. However, self-efficacy negatively affects attitude towards formative assessment. This means that students who believe in their capacity to execute their attainment rarely need the help of classroom formative assessment. Motivation in learning mathematics is positively affected by self-efficacy and attitude towards learning mathematics. Students who believe in their ability to achieve have a more positive attitude towards learning mathematics have higher motivation to learn mathematics. Thus, the main attitudinal factors and other related attitudinal factors should be considered for future studies related to the installation of test modes.

One of the factors concerning socioeconomic status is parents' education level. In this study, through SEM analyses, parent education positively impacts ICT familiarity and attitude towards ICT. This meant that their parents with higher education levels give their children the ICT access and encourages their children to get more ICT familiarity and have a more positive attitude towards ICT. Then, parent education levels positively affect the expected education level. A student whose parents have a higher education level expects to achieve a higher education level. However, parent education levels negatively affect motivation. Students whose parents possess a higher level of education are less motivated to study mathematics. Therefore, their parental involvement affects the achievement more than the *parents' education level*. This suggests that parents' education level is considered one of the key factors for studies relating students' achievement improvement.

Additionally, teachers and students' diverse perception towards PPT, LOT, and CAT is important through the analysis of interview responses. The qualitative data give researcher the insight relating to benefit and challenges of using PPT, LOT and CAT. Teachers and students perceived the benefits of LOT and CAT to help teachers to overtake challenges of classroom formative assessment in PPT, especially for large class size. For example, overtime for scoring, providing prompt feedback individually, making question, having high-test security, administering test, and tracing students' learning progress. As a result, a teacher can save more time and effort, so they can update their classroom teaching. Students prefer LOT and CAT to PPT, because of immediate feedback and scoring, which helps their learning progress and know their strengths and deficiencies in specific content areas. However, LOT and CAT need ICT accessibility, especially enough computers and a good internet connection, and ICT familiarity through teachers' and students' responses. Therefore, it suggests that school management should encourage teachers to use computer-based test modes in their classroom assessment. In addition, schools should be equipped with enough computers and access to good internet connection. In addition, computer technicians should be employed to educate computer and ICT literacy and to maintain and fix computers and the internet.

This study provides implications for innovation in classroom assessment in computer-based or ICT-related test modes, national assessment systems, and professional development for mathematics teachers in high schools in Myanmar. Consequently, Myanmar educators can formulate better strategies for policies and curricula that can enhance the mathematics achievement of their high school students. National assessment systems and professional development for mathematics teachers in high schools in Myanmar. Consequently, Myanmar educators can formulate better strategies for policies and curricula that can enhance the mathematics achievement of their high school students.

13.7 Conclusions

Nowadays, computer-based test modes are being used, rather than paper-based test modes, for classroom formative assessment, summative assessment, and national examinations. Researchers are working in this research area to help educational institutions to have a successful implementation of computer-based test modes.

To conclude, this study proposes an acceptance model for computer-based test modes and introduces two new computer-based test modes, LOT and CAT, for the use of classroom formative assessments. LOT and CAT are more effective with immediate scoring and feedback on students' mathematics learning progress than a traditional test mode (PPT) which generally delays feedback and students' score. Further, the study aims to explore and demonstrate other related scales from student-level and teacher-level impacting on the effects and determine its possible relationships with other scales, especially their influence on students' improvement in mathematics achievement due to the different test modes. In addition, the study generally adds information related to participants' demographic status, in-service teachers' assessment practices and techniques, and students' attitudinal factors towards specific subjects to the available literature by providing more findings on the achievement improvement due to experiencing test modes and on its link with other education scales as implied in the literature.

This study provides a first step towards the use of computer-based test modes for classroom formative assessment in Myanmar, contrary to the literature that mainly focused on the e-learning environment or computer-based test modes for summative assessment. As a result, it suggests the application of two new computer-based test modes that are very important for classroom formative assessment in developing countries.

In terms of installing new test modes in the classroom formative assessment instead of the traditional PPT, this study is deemed successful in providing additional findings on experiences of in-service teachers to new test modes. In fact, it is the first study to provide empirical evidence on the assessment practices of Myanmar teachers, as no study of this kind has been conducted in Myanmar. In addition, this study, to the best of my knowledge, is the first to provide evidence on the relationship of ICT-related assessment practices with relevant education scales. While this finding is far from conclusive and warrants further investigations, this study provides initial data for other educational researchers to confirm or refuse and develop new frameworks to advance the study of assessment practices.

With the likely proliferation of ICT-based devices and automatic proctoring technologies, there will likely be a substantial increase in ICT related or computer-based test modes. However, it is also true that it is beyond the abilities of most teachers, school principals, educational innovators, and in fact, even most instructional designers, and instructional technologists, to design advanced test modes because they do not have the skills nor the time to craft and extensively pilot their test mode for classroom formative assessment. As a result, instructors, innovators, and instructional designers must invest additional time and effort to learn how to construct computer-based test modes for classroom formative assessment and summative assessment as well as national examinations and design high-quality test items for the use of computer-based test modes. It is important to construct a national item bank with good psychometric values accessible for all schools countrywide. Further, there is a need for psychometric studies to compare LOT and CAT. Ministries of education or institutions in developing countries should spend the time, cost, and effort to use widely computer-based test modes for classroom better education systems.

Moreover, when incorporating the idea of computer-based test modes into existing classroom assessment practices, teachers can search for creative and effective test modes and evaluate systems in addition to the traditional multiple choice. For instance, they can include web-based test modes or e-portfolio, or computer simulations for hands-on performance assessments in classroom formative assessment for students' learning progress. Thus, there are more programmes, short training, or courses upgrading ICT literacy for teachers and students to apply well the application of computer-based test modes.

School principals have crucial roles to play for successful practices of computer-based test modes in classroom formative assessment. It is possible to argue that many schools, especially in developing countries, do not have the technology infrastructure and/or the budgets to support the integration of computer-based test modes into classroom assessments. Government and Ministry of Education should support teachers to access free ICT devices, internet, and online resources for constructing computer-based test modes. According to Dede (2003), the fundamental barriers to employing these technologies effectively for learning are not technical or economic, but psychological, organizational, political, and cultural. This is because there is still a real challenge to overcome the fear, suspicion, and doubt found in many schools about importance of such efforts.

Further, another point to remember is that most teachers and students in this study prefer CBTs because they have already known their benefit even though they were more familiar with the PPT in all assessments. Consequently, now is the right time to install CBTs in the classroom formative assessment. From that movement, CBT in classroom assessments or large-scale national studies could be greater – however, further studies are required before innovation.

Another contribution that this study provides is its methodological approaches to address the objectives or the research questions. For example, the use of mixed-methods design allowed more information and more profound interpretation of some analysis results. Moreover, the use of single-level (SEM) and multilevel (HLM) analysis techniques provided the strength in data handling and analysis and the validity of the results because this study addressed the issues associated with the ordinary statistical techniques (i.e., the loss of the information, erroneous estimation). These methods are considered beneficial in education research, especially given the complexity of educational phenomena for which appropriate procedures are needed to help obtain proper inferences.

In this study, the use of the SEM and HLM approaches has provided more results that are valid and further confirmed the classroom assessment model in the Myanmar context. Overall, the findings of this study provide a meaningful understanding of what students-, and teacher-level factors influence their mathematics improvement, interrelationship, and how they affect their mathematics improvement in Myanmar High Schools. From this beginning, a comprehensive model of mathematics learning and assessment, which is applicable for the Myanmar context and other developing countries, can emerge.

On identifying student- and teacher-level scales influencing students' achievement improvement in high schools of Myanmar, this study has contributed to the growing theories and knowledge in mathematics education and innovation in assessment. This study found multiple interrelationships among student-level scales in the student-level model and teacher-level scales in the teacher-level model. This thesis has clearly explained how the scales interact and influence students' achievement improvement among three test modes. Based on the results obtained, factors such as students' ICT familiarity, their attitude towards ICT, and formative assessment, as well as teachers' specific practices of formative assessment and attitude towards formative assessment, ICT and test modes are significant on the effect of the test mode on students' achievement improvement.

Assessment practice and techniques have key roles to play in improving quality education in any country. This study has likewise provided findings based on empirical evidence that could help guide future development efforts in integrating ICT in assessment in developing countries. In Myanmar or other developing countries, there is an urgent need for a national strategy of educational innovation in teacher assessment practices that the government should focus to attain a proper height in education sector. It is possible to replicate this study at the national level to identify specific needs of teachers to improve and access more ICT and ICT-based assessment techniques through professional development programs.

In conclusion, this study has provided additional knowledge that helps advance the understanding of assessment techniques or test modes, particularly with ICT-related test modes, their roles in fostering students' learning improvement, supporting teachers' effectiveness of formative assessment, and their paramount importance in education, training, and practice. Through this study, the researcher strongly recommends the Myanmar government or governments in developing countries to support ICT access and have enough ICT devices or facilities and technicians in every school, especially free Internet and electricity access through its Ministry of Education. In addition, in-service teachers need to be equipped with professional development programs to enhance their knowledge with assessment and measurement, especially how to construct item banks, a material support for innovative assessment techniques.

References

Acharya, N., & Joshi, S. (2009). Influence of parents' education on achievement motivation of adolescents. Indian Journal Social Science Research, 6(1), 72-79.

Adams, R., & Khoo, S. T. (1996). *Quest.* Australia Council for Educational Research. Melbourne, Australia.

Adams, R., Wilson, M., & Wang, W. C. (1997), The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), pp. 1-23.

Adams, R. & Wu, M. (2010). *Multidimensional models.* Australian Council for Educational Research.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

Afrassa, T. M. (1999). Monitoring mathematics achievement over time. In G. N. Masters & J. P. Keeves (1999) (Eds) Advances in Measurement in Educational Research and Assessment (pp. 115-137). Amsterdam: PERGAMON.

Agarwal, R., & Prasad, J. (1999). Are Individual Differences Germane to the Acceptance of New Information Technologies? *Decision Sciences*, *30*(2), 361-391.

Aiken, L.S. and West, S.G. (1991) Multiple Regression: Testing and Interpreting Interactions. Sage Publications, Inc., Thousand Oaks.

Ajogbeje, O.J. (2014). Effect of formative testing on students' achievement in junior secondary school Mathematics. European Scientific Journal, 8.8

Ak, L., & Sayil, M. (2006). Three different types of elementary school students' school achievements, perceived social support, school attitudes and behavior-adjustment problems. Educational Sciences: Theory & Practice, 6, 293-300.

Akyüz, G. (2014). The effects of student and school factors on mathematics achievement in TIMSS 2011. *Education and Science: Large-Scale Assessment Special Issue.* 39(172), 150-162.

Alaga, N. A. C. (2016). Motivation and attitude of students toward learning English language. In The International Conference on Research in Social Science, Humanities, and Education, Cebu, Philipines.

Al-Shebri, K. D. (2008). Teachers' and students' attitudes towards formative assessment and feedback in teaching English for specific purposes ESP.

Ak, L., & Sayil, M. (2006). Three different types of elementary school students' school achievements, perceived social support, school atti- tudes and behavior-adjustment problems. Educational Sciences: Theory & Practice, 6, 293-300.

Anderson, J. O., Lin, H. S., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education*, *5*(4), 591-614.

Andrich, D. (1999). *Rating scale analysis. In G. N. Masters & J. P. Keeves (Eds). Issues in educational measurement* (pp. 110-121). Amsterdam: Pergamon.

Australian Council for Educational Research (2016). ACER ConQuest, Version 4. Retrieved from https://www.acer.edu.au/conquest/acer-conquest1

Backes, B., & Cowan, J. (2018). Is the Pen Mightier than the Keyboard? The Effect of Online Testing on Measured Student Achievement. Working Paper 190. *National Center for Analysis of Longitudinal Data in Education Research (CALDER).*

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment*, *6*(9).

Bennett, R. E. (2010) Technology for large-scale assessment, in: P. PETERSON, E. BAKER&B. MCGAW (Eds) *International Encyclopedia of Education* (3rd ed., Vol. 8, pp. 48–55) (Oxford, Elsevier).

Bennett, R. E. (2011). Formative assessment: A critical review. Assessment in education: principles, policy & practice, 18(1), 5-25.

Bernard E. W. Jr. (1997). Gender Differences in Computer-Related Attitudes and Behavior: A Meta-Analysis, Computers in Human Behavior, 13(1), 1-22

Biggs, J. B. (1989). Approaches to the enhancement of tertiary teaching. *Higher Education Research & Development*, 8(1), 7-25.

Blais, J.-G., Grondin, J., Loye, N., & Raiche, G. (2011) Rasch model's contribution to the study of items and item response scales formulation in opinion/perception questionnaires. In N.J.S. Brown, B. Duckor, K. Draney. & M. Wilson (Eds.), *Advances in Rach measurement volume 2* (pp. 183-202). Maple Grove, Minnesota: JAM Press.

Bloom, B. S. (1968). Learning for Mastery. Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Evaluation comment*, *1*(2), n2.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York and London: Routledge.

Boone, W. J., & Scantlebury, K. (2005). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests, Wiley InterScience, pp. 253-268, DOI 10.1002/sce.20106

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, Edward Arnold (Publishers) Ltd , vol 3, pp. 77-101

Bridgeman, B. (2009) Experiences from large-scale computer-based testing in the USA, in: F. Scheuermann & J. Bjornsson (Eds) *The Transition to Computer-Based Assessment* (Luxembourg, Office for Official Publications of the European Communities).

Broughton, S. J. (2017). *An evaluation of the effectiveness of a computer-aided assessment system for mathematics and engineering students* (Doctoral thesis, © Stephen James Broughton).

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, London, New Delhi: SAGE Publications, Inc.

Bunderson, V. C., Inouye, D. K. & Olsen, J. B. (1989) The four generations of computerized educational measurement, in: R. L. LINN (Ed) *Educational Measurement* (Third ed., pp. 367–407) (NewYork, Macmillan).

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education*, *28*(3), 282-290.

Burghof, K. L. (2001). Assembling an item-bank for computerised linear and adaptive testing in Geography. International Educational Journal, 2(4), pp. 74-83.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1988). The four generations of computerized educational measurement. *ETS Research Report Series*, *1988*(1), i-148.

Burghes, D. (2011). International comparative study in mathematics teacher training. London.

Byrne, B. M. (2010). *Structural equation modelling with AMOS: Basic Concepts applications, and Programming* (2nd ed.). New York and London: Routledge.

Candeias, A. A., Rebelo, N., & Oliveira, M. (2010). *Student' Attitudes Toward Learning and School – Study of Exploratory Models about the Effects of Socio-demographics and Personal Attributes*. Retrieved January 10, 2019 from http://www.projectored.uevora.pt/documentos/LICE.pdf.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Sage University Paper 17. Beverly Hills: Sage Publications.

Chua, Y.P. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behaviour, 28*(5), 1580–1586. http://dx.doi.org/10.1016/j.chb.2012.03.020.

Chua, Y.P., & Don, Z.M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behaviour*, *29*(5), 1889–1895. http://dx.doi.org/10.1016/j.chb.2013.03.008.

Clariana, R. B., Wagner, D., & Murphy, L. C. R. (2000). Applying a connectionist description of feedback timing. Educational Technology Research and Development, 48(3), 5–21.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with test mode effect. *British Journal of Educational Technology*, *33*(5), 593-602.

Cohen, J. (1969). Statistical power analysis for the behavioural sciences. New York: Academic Press.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. (6th ed.). New York: Routledge.

Cohen, R. A., Sparling-Cohen, Y. A., & O'Donnell, B. F. (1993). *The neuropsychology of attention*. New York: Plenum Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Creswell, J. W., (2009). Research Design: Qualitative, Quantitative and Mixed Methods Approaches (3rd ed.). London: SAGE Publications, Inc.

Creswell, J.W. (2014). Research Design. (4th ed). London: SAGE.

Creswell, J.W. (2015). A concise introduction to mixed methods research. London: SAGE.

Csapó, B., Molnár, G., & Tóth, K. R. (2009). Comparing Paper-and-Pencil and Online Assessment of Reasoning Skills: A Pilot Study for Introducing TAO in Large-scale

Assessment in Hungary. In F. Scheuermann, & J. Björnsson, (2009). The transition to computer-based assessment. Luxembourg: Office for official publications of the European communities. Pp. 120-126.

Csapó, B., Ainley, J., Bennett, R., Latour, T. & Law, N. (2010) *Technological Issues for Computer-Based Assessment*. http://atc21s.org/wp-content/uploads/2011/11/3-Technological-Issues.pdf

Curtis, D. D. (2005) Comparing classical and contemporary analyses and Rasch measurement. In S. Alagumalai, D. D. Curtis, & N. Hungi, (2005). *Applied Rasch Measurement: A Book of Exemplars*. p. 179-195, The Netherlands: Springer.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. Exceptional Children, 52, 219–232.

Department of Basic Education (2016). Statistical data for Basic Education High Schools. Ministry of Education (MOE), Myanmar.

Department of Basic Education (2018). Myanmar Matriculation Exam Statistics Handbook. Ministry of Education (MOE), Myanmar.

Drasgow, F (2002). The work ahead: A psychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J.J. Fremer, & W.C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 67–88). Hillsdale, NJ: Lawrence Erlbaum.

Duque, M., & Schools, B. C. P. (2016). Is there a PARCC mode effect? *Strategic Data Project Fellowship*.

Ebel, R. L., & Frisbie, D. A., (1986). Essentials of Education Measurement, New Jersey: Prentice Hall, Englewood Cliffs.

Education Policy and Data Center (EPDC) (2012). SACMEQ and PASEC viewed on https://www.epdc.org/data-about-epdc-data-epdc-learning-outcomes-data/sacmeq-and-pasec

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved October 5, 2018, from www.psych.umn.edu/psylabs/CATCentral/

Eggen, T. J. H. M. & Straetmans, G. J. J. M. (2009) Computerized adaptive testing of arithmetic at the entrance of primary school teacher training college, in: F. Scheuermann & J.

Bjornsson (Eds) *The Transition to Computer-Based Assessment* (Luxembourg, Office for Official Publications of the European Communities).

Ellis, R., Loewen, S. and Erlam, R. (2006). *Implicit and explicit corrective feedback and the acquisition of L2 grammar studies of second language acquisition*, 28, 339 – 368.

Fageeh, A. (2011). EFL students' readiness for e-learning: factors influencing e-learners acceptance of the Blackboard in a Saudi university. *JALTCALL Journal*, 7 (1), 19–42.

Fakeye, D. O. (2016). Secondary school teachers' and students' attitudes towards formative assessment and corrective feedback in English language in Ibadan metropolis. *Journal of Educational and Social Research*, *6*(2), 141.

Fishman, B., Riconscente, M., Snider, R., Tsai, T., & Plass, J. (2014). *Empowering Educators: Supporting Student Progress in the Classroom with Digital Games*. Ann Arbor: University of Michigan. gamesandlearning.umich.edu/agames

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2011). *How to design and evaluate research in education*. New York: McGraw-Hill Humanities/Social Sciences/Languages.

Gay, L .R., Mills, G, E., & Airasian, P. (2012). Educational Research: Competencies for analysis and applications. (3rd ed.). New York: Pearson.

Geddes, J. D., Murrell, A. R., & Bauguss, J. (2010). Childhood Learning: An Examination of Ability and Attitudes toward School. *Creative Education*, 1(3), 170-183.

Goldberg, A., & Pedulla, J.J. (2002). Performance differences according to test mode and computer familiarity on a practice GRE. Educational and Psychological Measurement, 62(6), 1053-1067.

Granello, D. (2001). 'Promoting cognitive complexity in graduate written work: Using bloom's taxonomy as a0p pedagogical tool to improve literature reviews', *Counselor Education and Supervision*, Vol. 40, No. 4, pp. 292-308.

Gravemeijer, K., Stephan, M., Julie, C., Lin, F-L., & Ohtani, M. (2017). What Mathematics Education May Prepare Students for the Society of the Future? International Journal of Science and Mathematics Education, 15(1), pp. S105–S123 DOI 10.1007/s10763-017-9814-6.

Greenwood, L., Cole, U.M., McBride, F.V., Morrison, H., Cowan, P. & Lee, M. (2000) Can the same results be obtained using computer-mediated tests as for paper-based tests for

National Curriculum assessment? *Proceedings of the International Conference on Mathematics/Science, Education and Technology*. Vol. 2000 (1) 179-84.

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

Gustafsson, J. E. (2018). International Large Scale Assessments: Current Status and Ways Forward. Vol. 62 (3) pp. 328-332

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). USA: Pearson Prentice Hall.

Hardman, F., Stoff, C., Aung, W., & Elliott, L. (2014). Developing pedagogical practices in Myanmar primary schools: possibilities and constraints, Asia Pacific Journal of Education, DOI: 10.1080/02188791.2014.906387.

Harks, B., Rakoczy, K., Hattie, J., Besser, M & Klieme, E. (2014). The effects of feedback on achievement, interest, and self-evaluation: the role of feedback's perceived usefulness, Educational Psychology, 34:3, 269-290, DOI:10.1080/01443410.2013.785384

Hartnell-Young, E., Harrison, C., Crook, C., Davies, L., Fisher, T., Pemberton, R., & Smallwood, A. 2007, *Impact study of e-portfolios on learning*, British Educational Communications Technology Agency, Coventry. http://partners.becta.org.uk/index.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Hayden, M., & Martin, R. (2013). Recovery of the education system in Myanmar. *Journal of International and Comparative Education (JICE)*, 47-57.

Hensley, K.K. (2015). *Examining the effects of paper-based and computer-based modes of assessment of mathematics curriculum-based measurement*. Unpublished PhD thesis, University of Iowa, Iowa.

Herman, J., Osmundson, E., Dai, Y., Ringstaff, C., & Timms, M. (2015). Investigating the dynamics of formative assessment: Relationships between teacher knowledge, assessment practice and learning. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 344-367.

Hogan, T. (2014). Using a computer-adaptive test simulation to investigate test coordinators' perceptions of a high-stakes computer-based testing program.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. The Electronic Journal of Business Research Methods, 6, 53-60.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does It Matter if I Take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2), n2.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York and Hove: Routledge Taylor and Francis Group.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modelling: A Multidisciplinary Journal, 6, 1-55.

Hu, P. J., Chau, P. Y. K., Sheng, O. R. L., & Tam, K. Y. (1999). Examining the technology acceptance model using physician acceptance of telemedicine technology. Journal of Management Information Systems, 16(2), 91–112.

IBM Corp. (2019). *IBM SPSS Statistics for Windows*, Version 26.0. Armonk, NY: IBM Corp. Retrieved from https://hadoop.apache.org

International association -IAE (2015). Third International Mathematics and Science Study (TIMSS) background questionnaire, viewed on https://timssandpirls.bc.edu/timss1999i/pdf/BM2_TeacherS.pdf

Jacobsen, J., Ackermann, R., Egüez, J., Ganguli, D., Rickard, P., & Taylor, L. (2011). Design of a computer-adaptive test to measure English literacy and numeracy in the Singapore workforce: Considerations, benefits, and implications. *Journal of Applied Testing Technology*, *12*(1).

Jamil, M., Topping, K. J., & Tariq, R. H. (2012). Perceptions of university students regarding computer assisted assessment. The Turkish Online Journal of Educational Technology, 11(3), 267–277.

Jones, T., & Clarke, V. A. (1994). A computer attitude scale for secondary students. *Computers & Education*, *22*(4), 315-318.

June, Z., & Eamoraphan, S. (2019). The relationship of attitudes toward mathematics and mathematics self-efficacy with mathematics achievement of grade 10 students at a Len Bum IDPs high school in Kachin State, Myanmar. *Scholar: Human Sciences*, *11*(2), 180-180.

Karim, B. H. (2015). The Impact of Teachers' Beliefs and Perceptions about Formative Assessment in the University ESL Class. *International Journal of Humanities, Social Sciences and Education*, *2*(3), 108-115.

Keeves, J.P., and Alagumalai, S. (1999). 'New Approaches to Measurement'. In Masters, G. N., and J. P. Keeves (Eds), *Advances in Measurement in Educational Research and Assessment*, Pergamon, pp. 23-48.

Ker, H. W. (2016). The impacts of student-, teacher-and school-level factors on mathematics achievement: an exploratory comparative investigation of Singaporean students and the USA students. *Educational Psychology*, *36*(2), 254-276.

Khoshsima, H., Hashemi Toroujeni, S. M., Thompson, N., & Reza Ebrahimi, M. (2019). Computer- Based (CBT) vs. Paper-Based (PBT) Testing: Mode Effect, Relationship between Computer Familiarity, Attitudes, Aversion and Mode Preference with CBT Test Scores in an Asian Private EFL Context. *Teaching English with Technology*, *19*(1), 86-101.Kingston, N. (2010). Computerized adaptive testing. In N. J. Salkind (Ed.), Encyclopedia of research design (pp. 203-207). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412961288.n65

Kline, R. B. (2011). *Principles and practice of structural equation modelling (3rd ed.).* New York and London: The Guilford Press.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, *119*(2), 254.

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research*, *43*(4), 489-510.

Kozlowski, S. W., & Klein, K. J. (2000). A multi-level approach to theory and research in organizations: contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), Multi-level theory, research, methods in organizations (pp. 3–90). San Francisco: Jossey-Bass

Kusaeri, K., Aditomo, A., Ridho, A., & Fuad, A. (2018). Socioeconomic status, parental involvement in learning and student'mathematics achievement in Indonesian senior high school. *Cakrawala Pendidikan Jurnal Ilmiah Pendidikan*, *37*(3), 333-344.

Lazendic, G. (2015). National Assessment and Surveys Online Program – 2013 Tailored test design study, NSW: Australian Curriculum, Assessment and Reporting Authority

Lee, J. S. (2006). Exploring the relationship between electronic literacy and heritage language maintenance. *Language Learning & Technology, 10* (2), 93–113.

Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. Educational Technology Research and Development, 58, 1–20. doi:10.1007/s11423-010-9153-6.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing*, *6*(1), 1-24.

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied psychological measurement*, *41*(7), 495-511.

Liu, X., & Koirala, H. (2009). The effect of mathematics self-efficacy on mathematics achievement of high school students.

Lopez, L. (2009). Effects of delayed and immediate feedback in the computer-based testing environment [Electronic version]. Doctoral dissertation, Department of Curriculum, Instructional, and Media Technology, Indiana State University. Available from http://gradworks.umi.com/33/58/3358462.html.

Luecht, R. M., & Clauser, B. E. (2002). Test models for complex computer-based testing. In C. N. Mills, M. T. Potenza, J. J. Fremer, &W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 67-88). Hillsdale, NJ: Lawrence Erlbaum.

Maamin, M., Maat, S. M., & Iksan, Z. H. (2021). Analysis of the factors that influence mathematics achievement in the ASEAN countries. *Cypriot Journal of Educational Sciences*, *16*(1), 371-388.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modelling in psychological research. Annual Review of Psychology, 51, 201-226.

Martin, R. (2008) New possibilities and challenges for assessment through the use of technology; in: F. Schuermann & A. G. Pereira (Eds) *Towards a Research Agenda on Computer-Based Assessment* (Luxembourg: Office for Official Publications of the European Communities).

Martin, A. J., & Lazendic, G. (2017). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Education Psychology*. DOI: 10.1037/edu0000205

McGahee, T. W. & Ball, J. (2009). How to read and really use an item analysis. Nurse Educator, 34, 166- 171.

Meijer, R. (2009). Transition to Computer-based Assessment: Motivations and Considerations. In F. Scheuermann, & J. Björnsson, (2009). The transition to computer-based assessment. Luxembourg: Office for official publications of the European communities. Pp. 104-107.

Michael–Chrysanthou, P., Gagatsis, A., & Vannini, I. (2014). Formative assessment in mathematics: a theoretical model. *Acta Didactica Universitatis Comenianae–Mathematics*, *14*, 43-70.

Miller, T. (2009). Formative computer-based assessment in higher education: The effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education*, *34*(2), 181-192.

Ministry of Education (MOE) (2004). Development of education in Myanmar. The Government of the Union of Myanmar.

Moe, E. (2009) Introducing large-scale computerised assessment lessons learned and future challenges, in: F. Scheuermann & J. Bjornsson (Eds) *The Transition to Computer-Based Assessment* (Luxembourg, Office for Official Publications of the European Communities).

Morgan, C., & O'Reilly, M. (2001). Innovations in online assessment. In F. Lockwood & A. Gooley (Eds.), *Innovation in open and distance learning: Successful development o*f *online and web-based learning* (pp.179–188). London: Kogan Page.

Mory, E. H. (2004). Feedback research revisited. In D. Jonassen (Ed.), Handbook of research on educational communications and technology (pp. 745–783). Mahwah: Erlbaum.

Mullis, I. V. S., & Martin, M. O. (2017). TIMSS 2019 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/

Myanmar Examination Board, (2018). Question Collection of Myanmar National Matriculation examination (2011-2018). Ministry of Education. Myanmar: Yangon.

Nesa S. (2014). What are the advantages and disadvantages of formative assessment? Retrieved on 6th August, 2019 from *everyday life.globalpost.com/advantages-disadvantages-formativeassessment-28407.html*

Newton, M. A., & Mwisukha, A. (2009). Relationship between peer attitudes towards school, selected peer group activities and academic achievement of secondary school students in Nairobi. *Journal of Educational Research and Development*, 4(1), 99-104.

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modelling of educational data.* Charlotte, NC: Information Age.

Oermann, H. M. & Gaberson, B. K. (2006). Evaluation and Testing in Nursing Education. 2nd ed. Springer. New York - USA.

Ofsted. (2012). Mathematics: made to measure. (pp. 3-18). Manchester.

Oliver, R. (2002). The role of ICT in higher education for the 21st century: ICT as a change agent for education. *Retrieved April*, *14*, 2007.

Olson (2004). Computerised adaptive testing: Effective measurement for all students. The Journal.

Opre, D. (2015). Teachers' conceptions of assessment. *Procedia-Social and Behavioral Sciences*, *209*, 229-233.

Organisation for Economic Co-operation and Development (OECD), (2016). Programme for International Student Assessment (PISA), Student Questionnaire. Paris: OECD Publishing.

Oxford Business Group, (2017). Changes to Myanmar's education sector needed, Oxford Business Group, retrieved in October 20, 2018, from https://www.oxfordbusinessgroup.com/overview/learning-curve-changes-are-needed-revitalise-country’s-schools.

Özyurt, H. & Özyurt, Ö. (2015). Ability Level Estimation of Students on Probability Unit via Computerized Adaptive Testing. Eurasian Journal of Educational Research, 58, 27-44. http://dx.doi.org/10.14689/ejer.2015.58.7

Pallant, J. (2011). *SPSS survival manual (4th ed.).* NSW Australia: Allen& Unwin.

Parshall, C. G., & Kromrey, J. D. (1993). Computer testing versus paper-and-pencil: an analysis of examinee characteristics associated with mode effect. In *A paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA, April (Educational Resources Document Reproduction Service (ERIC)# ED363272)*.

Pinchok, N. and Brandt, W.C. (2009). Connecting formative assessment research to practice: An introductory guide for educators. *Learning Point Associates*, retrieved 4th November, 2020 from http://www.learningpt.org/rel/events.php

Penn, B. K. (2009). Test item development and analysis. Presented at Creighton University School of Nursing Faculty Retreat, Omaha, NE.

Poggio. J. & McJunkin, L. (2012). History, current practice, perspectives and what the future holds for computer based assessment in K-12 education. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments* (pp. 25-53). Charlotte, NC; Information Age Publishing, Inc.

Public Broadcasting Service (PBS) (2019) Mathematics in everyday life, viewed at https://www.pbs.org/parents/mathematics-in-everyday-life/maths

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2009). HLM 6.08 [computer software]. *Scientific Software International, Chicago*.

Quellmalz, E. S. & Pellegrino, J. W. (2009). Technology and testing. *Science*, *323,* 75-78.

Quyen, N. T. D., & Khairani, A. Z. (2017). Reviewing the challenges of implementing formative assessment in Asia: The need for a professional development program. *Journal of Social Science Studies*, *4*(1), 160-177.

Rashad A.M., Youssif A.A.A., Abdel-Ghafar R.A., Labib A.E. (2008) E-Assessment Tool: A Course Assessment Tool Integrated into Knowledge Assessment. In: Iskander M. (eds) Innovative Techniques in Instruction Technology, E-learning, E-assessment, and Education. Springer, Dordrecht

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: applications and data analysis methods (2nd ed.). Thousand Oaks: Sage.

Redecker, C., & Johannessen, Ø. (2013). Changing assessment—Towards a new assessment paradigm using ICT. *European Journal of Education*, *48*(1), 79-96.

Rezaie, M., & Golshan, M. (2015). Computer Adaptive Test (CAT): Advantages and Limitations, *International Journal of Educational Investigations*, 2(5), pp. 128-137, 2015 (May) ISSN: 2410-3446

Richardson, J.T.E. (2011). Eta squared and partial eta squared as measurements of effect size in educational research. Educational Research Review, 6, 135-147.

Rickards, J. P., & Friedman, F. (1978). The encoding versus the external storage hypothesis in note taking. *Contemporary Educational Psychology, 3*, 136–143.

Ricketts, C., Filmore, P., Lowry, R., & Wilks, S. (2003). *How should we measure the costs of computer aided assessment? Proceedings from the 7th Computer Assisted Assessment Conference, Loughborough University, UK.*

Ripley, M. (2009). Transformational Computer-based Testing. In F. Scheuermann, & J. Björnsson, (2009). The transition to computer-based assessment. Luxembourg: Office for official publications of the European communities. Pp. 92-98.

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 1155.

Rudner, L.M. (1998). Item banking. *Practical Assessment, Research & Evaluation,* Vol. 6, No. 4. Available at http://ericae.net/pare/getvn.asp?v=6&n=4

Sánchez, S., Reyes, O., & Singh, J. (2006). Makin' it in college: The value of significant individuals in the lives of Mexican American adolescents. Journal of Hispanic Higher Education, 5, 48-67.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. American Psychological Association. Washington, DC, USA.

Shin, N., Sutherland, L. M., Norris, C. A., & Soloway, E. (2012). Effects of game technology on elementary student learning in mathematics. *British journal of educational technology*, *43*(4), 540-560.

Schreiber, J. B. (2008). Core reporting practices in structural equation modeling. *Research in Social and Administrative Pharmacy*, *4,* 83–97. https://doi.org/10.1016/j.sapharm.2007.04.003

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2016). Reporting structural equation modelling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99(6),* 323-338.

Schunk, D. H. (1995). Self-efficacy, motivation, and performance. *Journal of applied sport psychology*, *7*(2), 112-137.

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78, 153–189. doi:10.3102/0034654307313795.

Slavin, R. E. (2008). Perspectives on evidence-based research in education. What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practices,* 18(1), 23-27.

Stobart, G. (2008). Testing times: The uses and abuses of assessment. London: Routledge.

Suurtamm, C., Thompson, D. R., Kim, Moreno, L. D., Sayac, N., Schukajlow, S., Silver, E., Ufer, S., Vos, P., (2016). Assessment in mathematics education: large-scale assessment and classroom assessment. Hamburg: Springer Open.

Tabachnick, B. G. & Fidell, L. S. (2013). Using multivariate statistics (6th ed.). USA: Pearson Education, Inc.

Terzis, V., & Economides, A. A. (2011a). The acceptance and use of computer-based assessment. *Computers & Education*, *56*(4), 1032-1044.

Terzis, V., & Economides, A. A. (2011b). Computer based assessment: Gender differences in perceptions and acceptance. Computers in human behavior, 27(6), 2108-2122.

Thein, M. M. M., & Thein, N. N. (2018). The correlation between middle school students' attitudes and achievement in mathematics. *Journal of Myanmar Academy Arts and Science*. 16-9A, 259-289.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*.

Thorndike, R. M. & Thorndike-Christ, T. (2010). Measurement and Evaluation in Psychology and Education (8th ed.), New York: Pearson.

Timmers, C. F., Braber-Van Den Broek, J., & Van Den Berg, S. M. (2013). Motivational beliefs, student effort, and feedback behaviour in computer-based formative assessment. *Computers & education*, *60*(1), 25-31.

Timmers, C. F., & Veldkamp, B. P. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. Computers & Education, 56, 923–930. doi: 10.1016/j.compedu.2010.11.007.

Torres, C. M. D. P., Lopes, A. P. F. F., Azevedo, J. M. M. L., & Babo, M. D. L. (2009). Developing multiple-choice questions in mathematics. *International conference of education, research and innovation (ICERI2009)*, pp. 6218-6229. Madrid: IATED.

Ullman, J. B. (2013). Structual equation modelling. In B. G. Tabachnick & L. S. Fidell (Eds), *Using multivariate statistics* (6th ed., pp. 681-785). USA: Pearson Education, Inc.

Umar, J. (1999). 'Item Banking', In Masters, G. N., and J. P. Keeves (Eds), *Advances in Measurement in Educational Research and Assessment*, Pergamon, pp. 207-219.

Van Der Linden, W. J., & Glas, C. A.W. (2000). *CATing: Theory and practice*. Boston: Kluwer Academic.

Van Lent, G. (2009). Risks and Benefits of CBT versus PBT in High-Stakes Testing: Introducing key concerns and decision-making aspects for educational authorities. In F. Scheuermann, & J. Björnsson, (2009). The transition to computer-based assessment. Luxembourg: Office for official publications of the European communities. Pp. 83-91.

Van Raaij, E.M. and Schepers, J.J.L. (2008) The Acceptance and Use of a Virtual Learning Environment in China. Computers & Education, 50, 838-852. http://dx.doi.org/10.1016/j.compedu.2006.09.001

Velasco, M. M. (2017). The importance of item bank in the teaching-learning process, Press Reader, retrieved from https://www.pressreader.com/philippines/sunstarpampanga/20170729/281754154394965

Venkatesh, V. (1999). Creation of favourable user perceptions: Exploring the role of intrinsic motivation. MIS Quarterly, 23, 239–260.

Verešová, M., & Malá, D. (2016). Attitude toward school and learning and academic achievement of adolescents. In *7th International Conference on Education and Educational Psychology, Published by Future Academy*.

Vorderman, C., Porkess, R., Budd, C., Dunne, R., & Rahman-hart, P. (2011). A world-class mathematics education for all our young people. London.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (Eds.). (2000). *Computerized adaptive testing: Aprimer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Wandall, J. (2009) National tests in Denmark — CAT as a pedagogic tool, in: F. Scheuermann & J. Björnsson (Eds) *The Transition to Computer-Based Assessment* (Luxembourg, Office for Official Publications of the European Communities).

Waters, J. K. (2012). Resolving the Formative Assessment Catch-22: Teachers Often Have a Hard Time Embedding Assessment in Their Instruction, but Some Technologies Are Making It Easier. *THE Journal (Technological Horizons in Education)*, *39*(7), 8.

Watson, D. B. (2001). Key factors affecting conceptual gains from CAL materials. *British Journal of Educational Technology*, *32*(5), 587-593.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills. In *annual meeting of the National Council on Measurement in Education, San Francisco, CA*.

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006). Practical questions in introducing computerized adaptive testing for K-12 assessments. *Retrieved October 1*, 2009.

Weaver, A. J., & Raptis, H. (2001). Gender differences in introductory atmospheric and oceanic science exams: multiple choice versus constructed response questions. *Journal of Science Education and Technology, 10*(2), 115-126.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems, Journal of Educational Measurement, 21 (4), pp. 361-375

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counselling and education. Measurement and Evaluation in Counselling and Development, 37, pp. 70-84.

Weiss, D. J. (2011). Item banking, test development, and test delivery. *The APA handbook on testing and assessment*, 185-200.

Wood, R., & Skurnik, L.S., (1969). *Item Banking*, National Foundation for Educational Research in England and Wales, Slough.

World Bank (2020). Myanmar overview. Available at http: www.worldbank.org/en/country/Myanmar/overview. [accessed 1 June 2020].

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values, Rasch Measurement Transaction, 8:3, pp.370.

Wright, B. D. & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wilmington, Delaware: Wide Range, Inc.

Wu, M. L., & Adams, R. J. (2007). Applying the Rasch model to psycho-social measurement: A practical approach. Melbourne. Australia: Educational Measurement Solutions

Yuan, R. (1999). Chinese language learning and the Rasch model. In G. N. Masters & J. P. Keeves (1999) (Eds) Advances in Measurement in Educational Research and Assessment, pp. 115-137. Amsterdam: PERGAMON.

Appendices

Appendix A: Ethics Approval

THE UNIVERSITY of ADELAIDE

RESEARCH SERVICES
OFFICE OF RESEARCH ETHICS, COMPLIANCE AND INTEGRITY
THE UNIVERSITY OF ADELAIDE

LEVEL 4, RUNDLE MALL PLAZA
50 RUNDLE MALL
ADELAIDE SA 5000 AUSTRALIA

TELEPHONE +61 8 8313 5137
FACSIMILE +61 8 8313 3700
EMAIL hrec@adelaide.edu.au

CRICOS Provider Number 00123M

Our reference 33386

06 March 2019

Dr Igusti Darmawan
School of Education

Dear Dr Darmawan

ETHICS APPROVAL No: H-2019-039
PROJECT TITLE: An Exploratory Study of Assessment of High School Mathematics Students in Myanmar

The ethics application for the above project has been reviewed by the Low Risk Human Research Ethics Review Group (Faculty of Arts and Faculty of the Professions) and is deemed to meet the requirements of the *National Statement on Ethical Conduct in Human Research (2007)* involving no more than low risk for research participants.

You are authorised to commence your research on: 06/03/2019
The ethics expiry date for this project is: 31/03/2022

NAMED INVESTIGATORS:

Chief Investigator:	Dr Igusti Darmawan
Student - Postgraduate Doctorate by Research (PhD):	Ms Hnin Nwe Nwe Tun
Associate Investigator:	Associate Professor Nicholas Buchdahl
Associate Investigator:	Associate Professor Sivakumar Alagumalai

CONDITIONS OF APPROVAL: Thank you for your responses to the matters raised. The revised application provided on 06/03/19 has been approved.

Ethics approval is granted for three years and is subject to satisfactory annual reporting. The form titled Annual Report on Project Status is to be used when reporting annual progress and project completion and can be downloaded at http://www.adelaide.edu.au/research-services/oreci/human/reporting/. Prior to expiry, ethics approval may be extended for a further period.

Participants in the study are to be given a copy of the information sheet and the signed consent form to retain. It is also a condition of approval that you immediately report anything which might warrant review of ethical approval including:

- serious or unexpected adverse effects on participants,

- previously unforeseen events which might affect continued ethical acceptability of the project,
- proposed changes to the protocol or project investigators; and
- the project is discontinued before the expected date of completion.

Yours sincerely,

Dr Anna Olijnyk
Convenor

Dr Jungho Suh
Convenor

The University of Adelaide

Appendix B: Student Questionnaire

An Exploratory Study of Assessment of High School Mathematics Students in Myanmar

(Student Questionnaire)

Information about this Questionnaire

This questionnaire is intended for Grade-10 students in Yangon, Myanmar. It contains items that asks for students' background information, their attitude toward mathematics, their attitude toward computer, computer familiarity and their attitude towards the three test methods (Paper-and-Pencil Test (PPT); Linear-Online Test (LOT) and Computerised-Adaptive Test (CAT)) and feedback from these test methods.

Your response to the questionnaires contributes the evaluation of the effectiveness and efficiency of these test modes (PPT, LOT and CAT) for classroom formative assessments in this study and to value other studies related to the development of technology-based test modes in Myanmar. It is vital that you respond to each of the items very carefully so that the information provided reflects your situation. All your responses and identity will be kept strictly confidential.

General instructions to student participants

1. Please read each item carefully and answer them accurately as you can. For every section of the question, specific instructions of how to answer the items are given.
2. The questionnaire must be returned to the researcher as soon as it has been completed.
3. Complete the questionnaire during your non-contact period in the school or at your home. The researcher will help to distribute and explain the instructions any time. Please contact the researcher.
4. Time allocated to answer to the questions is 20 minutes
5. School Name, Teacher name, and student name will be replaced by codes by the researchers.

Cut here

✂………✂

Please fill in your details

School Name: …………………………………………

Teacher Name: …………………………………………

Student Name: …………………………………………

For Researchers Use Only

School Code # ………………………………………

Teacher Code # ………………………………………

Student Code # ………………………………………

Instruction: Please fill in spaces with the words or sentences that correspond to your answer. For the items are multiple choice questions, check ☒ boxes that correspond to your answer in the table.

General information about participants

1. Are you female or male?

□ Female □ Male □ Other

2. How old are you? ………………………………

3. What is the highest education level completed by your father?

□ Less than High School	□ Bachelor's Degree
□ Some High School	□ Master Degree
□ High School Graduate	□ Doctoral Degree
□ Diploma's Degree	□ Other …………………………………..

4. What is the highest education level completed by your mother?

□ Less than High School	□ Bachelor's Degree
□ Some High School	□ Master Degree
□ High School Graduate	□ Doctoral Degree
□ Diploma's Degree	□ Other …………………………………..

5. How far in your education do you aspire?

□ High School Graduate	□ Master Degree
□ Diploma's Degree	□ Doctoral Degree
□ Bachelor's Degree	□ Other …………………………………..

Mathematics

How much do you agree with these statements about learning mathematics?

		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	I enjoy learning mathematics.	☐	☐	☐	☐
b.	I wish I did not have to study mathematics.	☐	☐	☐	☐
c.	Mathematics is boring.	☐	☐	☐	☐
d.	I learn many interesting things in mathematics.	☐	☐	☐	☐
e.	I like mathematics.	☐	☐	☐	☐
f.	I like to solve mathematics problems	☐	☐	☐	☐
g.	I look forward to mathematics class.	☐	☐	☐	☐
h.	Mathematics is one of my favourite subjects	☐	☐	☐	☐

How much do you agree with these statements about your mathematics lessons?

		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	I know what my teacher expects me to do	☐	☐	☐	☐
b.	My teacher gives me interesting things to do	☐	☐	☐	☐
c.	My teacher has clear answers to my questions	☐	☐	☐	☐
d.	My teacher is good at explaining mathematics	☐	☐	☐	☐
e.	My teacher does a variety of things to help us learn	☐	☐	☐	☐
f.	My teacher tells me how to do better when I make a mistake	☐	☐	☐	☐
g.	My teacher links new lessons to what I already know	☐	☐	☐	☐
h.	My teacher explains a topic again when we don’t understand	☐	☐	☐	☐

How much do you agree with these statements about mathematics?

		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	I usually do well in mathematics	☐	☐	☐	☐
b.	Mathematics is more difficult for me than for many of my classmates	☐	☐	☐	☐
c.	Mathematics is not one of my strengths	☐	☐	☐	☐
d.	I learn things quickly in mathematics	☐	☐	☐	☐
e.	Mathematics makes me nervous	☐	☐	☐	☐
f.	I am good at working out difficult mathematics problems	☐	☐	☐	☐
g.	My teacher tells me I am good at mathematics	☐	☐	☐	☐

h.	Mathematics is harder for me than any other subject	☐	☐	☐	☐

How much do you agree with these statements about mathematics?

		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	I think learning mathematics will help me in my daily life.	☐	☐	☐	☐
b.	I need mathematics to learn other school subjects.	☐	☐	☐	☐
c.	I need to do well in mathematics to get into the college or university of my choice.	☐	☐	☐	☐
d.	I need to do well in mathematics to get the job I want.	☐	☐	☐	☐
e.	It is important to learn about mathematics to get ahead in the world.	☐	☐	☐	☐
f.	Learning mathematics will give me more job opportunities when I am an adult.	☐	☐	☐	☐
g.	My parents think that it is important that I do well in mathematics.	☐	☐	☐	☐

Students' Attitude towards Formative Assessment					
How much do you agree with these statements?					
		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	The use of formative assessment improves my performance.	☐	☐	☐	☐
b.	Formative assessment makes me to be actively involved in learning process.	☐	☐	☐	☐
c.	I enjoy my teacher asking questions during lesson.	☐	☐	☐	☐
d.	Asking me questions when the lesson is going on distracts my attention	☐	☐	☐	☐
e.	Formative assessment is time consuming.	☐	☐	☐	☐
f.	Corrective feedback enhances my learning.	☐	☐	☐	☐
g.	I adopt a deeper approach to learning whenever I am corrected.	☐	☐	☐	☐
h.	Corrective feedback helps me to know where I am lacking after each feedback	☐	☐	☐	☐
i.	I like it when my teacher points out my mistakes	☐	☐	☐	☐

Information and Communications Technology (ICT)			
		Yes	No
a.	Do you have a computer, or laptop or tablet or smart phone for your own use?	☐	☐
b.	Do you have personal email address?	☐	☐
c.	Do you have Wi-Fi or internet access at home?	☐	☐
d.	Have you taken the training on how to use the internet and email?	☐	☐
e.	I often use ICT for playing on a computer, online games	☐	☐
f.	I often use ICT for playing instructional programs, tutorial, remedial, or mastery learning	☐	☐
g.	I often use ICT for communicating by email, or social network such as Facebook, WhatsApp, Viber, etc.	☐	☐
h.	I often use ICT for participating social network such as Facebook, WhatsApp, Viber, etc.	☐	☐
i.	I often use ICT for downloading learning material, music, film, games, software.	☐	☐
j.	I often use ICT for uploading your own created music, poetry, videos.	☐	☐
k.	In a typical week, how many hours do you spend on using ICT devices? ---------------------- hours		
l.	In a typical week, how many hours do you spend on internet surfing? ---------------------- hours		

How much do you agree with these statements?

		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	ICT devices do not scare me at all.	☐	☐	☐	☐
b.	Working with ICT devices would make me very nervous.	☐	☐	☐	☐
c.	ICT devices make me feel uncomfortable.	☐	☐	☐	☐
d.	I would feel at ease in the ICT class	☐	☐	☐	☐
e.	I would feel comfortable working with a computer.	☐	☐	☐	☐
f.	ICT devices bore me.	☐	☐	☐	☐
g.	ICT devices are difficult to use	☐	☐	☐	☐
h.	Learning about ICT devices is a waste of time.	☐	☐	☐	☐
i.	People that use of ICT devices are seen as being more important than those who don’t.	☐	☐	☐	☐
j.	People who work with ICT devices make really good money.	☐	☐	☐	☐

k.	I learn new tasks of use of ICT by trial and error.	☐	☐	☐	☐
l.	When I have a problem with ICT devices, I will usually solve it on my own.	☐	☐	☐	☐
m.	Using the ICT devices has increased my interaction with other students.	☐	☐	☐	☐
n.	I develop short cuts, and more efficient ways to use ICT devices.	☐	☐	☐	☐
o.	I would like to learn more about ICT devices.	☐	☐	☐	☐
p.	If I need ICT skills for my career choice, I will develop them.	☐	☐	☐	☐

Attitude towards Paper-and-Pencil Test (PPT) and its feedback					
11. How much do you agree with these statements about attitude toward PPT test mode and its feedback?					
		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	PPT helps me to identify my weak areas	☐	☐	☐	☐
b.	PPT helps me to build my confidence.	☐	☐	☐	☐
c.	PPT helps me to improve mathematics learning	☐	☐	☐	☐
d.	In PPT, I was very focused on understanding the questions and tasks	☐	☐	☐	☐
e.	I persisted in PPT mode even when it was challenging or difficult.	☐	☐	☐	☐
f.	I was anxious in PPT.	☐	☐	☐	☐
g.	The feedback from PPT helps me reach my learning goal.	☐	☐	☐	☐
h.	The feedback from PPT helps me recognize where I can improve.	☐	☐	☐	☐
i.	The feedback from PPT lets me know which types of tasks I should practice	☐	☐	☐	☐
j.	The feedback from PPT lets me know whether I should/have to prepare myself better.	☐	☐	☐	☐
k.	After receiving the feedback from PPT, I make more effort.	☐	☐	☐	☐

Linear-Online Test (LOT) and its feedback					
12. How much do you agree with these statements about attitude toward LOT test mode and its feedback?					
		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	LOT helps me to identify my weak areas	☐	☐	☐	☐
b.	LOT helps me to build my confidence.	☐	☐	☐	☐
c.	LOT helps me to improve mathematics learning	☐	☐	☐	☐
d.	In LOT, I was very focused on understanding the questions and tasks	☐	☐	☐	☐
e.	I persisted in LOT mode even when it was challenging or difficult.	☐	☐	☐	☐
f.	I was anxious in LOT.	☐	☐	☐	☐
g.	The feedback from LOT helps me reach my learning goal.	☐	☐	☐	☐
h.	The feedback from LOT helps me recognize where I can improve.	☐	☐	☐	☐
i.	The feedback from LOT lets me know which types of tasks I should practice	☐	☐	☐	☐
j.	The feedback from LOT lets me know whether I should/have to prepare myself better.	☐	☐	☐	☐
k.	After receiving the feedback from LOT, I make more effort.	☐	☐	☐	☐

Computer-Adaptive Test (CAT) modes and its feedback					
13. How much do you agree with these statements about attitude toward CAT test mode and its feedback?					
		Disagree a lot	Disagree a little	Agree a little	Agree a lot
a.	CAT helps me to identify my weak areas	☐	☐	☐	☐
b.	CAT helps me to build my confidence.	☐	☐	☐	☐
c.	CAT helps me to improve mathematics learning	☐	☐	☐	☐
d.	In CAT, I was very focused on understanding the questions and tasks	☐	☐	☐	☐
e.	I persisted in CAT mode even when it was challenging or difficult.	☐	☐	☐	☐
f.	I was anxious in CAT.	☐	☐	☐	☐
g.	The feedback from CAT helps me reach my learning goal.	☐	☐	☐	☐
h.	The feedback from CAT helps me recognize where I can improve.	☐	☐	☐	☐
i.	The feedback from CAT lets me know which types of tasks I should practice	☐	☐	☐	☐
j.	The feedback from CAT lets me know whether I should/have to prepare myself better.	☐	☐	☐	☐
k.	After receiving the feedback from CAT, I make more effort.	☐	☐	☐	☐

Thank you for completing this questionnaire

Please return it to your teacher

Appendix C: Teacher Questionnaire

An Exploratory Study of Assessment of High School Mathematics Students in Myanmar

(Teacher Questionnaire)

Information about this Questionnaire

This questionnaire is intended for Grade-10 mathematics teachers in Yangon, Myanmar. It contains items that asks for background information, practices and the challenges of classroom formative assessment, attitude toward formative assessment and feedback for students' learning progress, and attitude and perceptions of computer-based test modes for classroom formative assessments.

Your response to the questionnaires contributes the evaluation of the effectiveness and efficiency of these test modes (Paper-and-Pencil Test (PPT); Linear-Online Test (LOT) and Computer-Adaptive Test (CAT)) for classroom formative assessments in this study and to value other studies related to the development of technology-based test modes in Myanmar. It is vital that you respond to each of the items very carefully so that the information provided reflects your situation. All your responses and identity will be kept strictly confidential.

General instructions to teacher participants

1. Please read each item carefully and answer them accurately as you can. For every section of the question, specific instructions of how to answer the items are given.
2. The questionnaire must be returned to the researcher as soon as it has been completed.
3. Complete the questionnaire during your non-contact period in the school or at your home. The researcher will help to distribute and explain the instructions any time. Please contact the researcher.
4. Time allocated to answer to the questions is 30 minutes.
5. School Name, Teacher name, and the number of students in your class will be replaced by codes by the researchers.

Cut here

✂……………………………………………………………………………………………...………………✂

Please fill in your details

School Name: …………………………………………

Teacher Name: …………………………………………

The number of students in your class: ………………….

For Researchers Use Only

School Code # …………………………………………

Teacher Code # …………………………………………

Class Size Code # ……………………………………..

Instruction: Please fill in spaces with the words or sentences that correspond to your answer. For the items are multiple choice questions, check ☒ boxes that correspond to your answer in the table.

General information about Participants
1. What is your highest level of education you have completed? ☐ Bachelor’s Degree in Education ☐ Bachelor’s Degree in Arts ☐ Bachelor’s Degree in Science ☐ Other …………………….
2. Which teaching qualification did you complete? ☐ Teacher education degree (4-year program) ☐ Teacher education degree (4-year program) ☐ Short term program in teaching education
1. How many students are there in this class? ……………………. Students
2. How many subjects do you teach in this academic year? ……………………. subjects

Formative assessment				
How frequent do you do the following assessment?	Never	Sometimes	Regularly	All the time
a. I use classroom discussion as general practices of formative assessment.	☐	☐	☐	☐
b. I use classroom observation as general practices of formative assessment.	☐	☐	☐	☐
c. I use whole class oral question-and-answer as general practices of formative assessment.	☐	☐	☐	☐
d. I use oral question-and-answer with individual student as general practices of formative assessment.	☐	☐	☐	☐
How frequent do you do the following assessment?	Never	Sometimes	Regularly	All the time
a. I provide your students score on their formative assessments.	☐	☐	☐	☐
b. I describe the results/scores of formative assessments.	☐	☐	☐	☐
c. I provide the written feedback to students.	☐	☐	☐	☐
d. I provide the feedback individual student.	☐	☐	☐	☐
e. I provide students' weaknesses and strengthens as contents of feedback.	☐	☐	☐	☐
f. I provide your students score on their formative assessments.	☐	☐	☐	☐
How much do you agree with these statements?	Disagree a lot	Disagree a little	Agree a little	Agree a lot
a. I rarely use formative assessment to evaluate my students' achievement.	☐	☐	☐	☐
b. Formative assessment is interesting for my class.	☐	☐	☐	☐
c. I do not like asking questions while the lesson is going on.	☐	☐	☐	☐
d. Formative assessment makes my class boring.	☐	☐	☐	☐
e. Formative assessment provides the useful information your students for enhancing students learning progress.	☐	☐	☐	☐
f. Formative assessment conceived as feedback loop to close the gap between students' current learning statues and intended learning outcomes.	☐	☐	☐	☐
g. Formative assessment provides a valuable learning experience for students.	☐	☐	☐	☐

Section 3: Information and Communications Technology (ICT)		
	Yes	No
m. I often use ICT for application of Microsoft Word Document	☐	☐
n. I often use ICT for application of Microsoft Excel Worksheet	☐	☐
o. I often use ICT for application of Microsoft PowerPoint presentation	☐	☐
p. I often use ICT for communicating colleagues, students and students' parents by email or social network (such as Facebook, WhatsApp, Viber)	☐	☐
q. I often use ICT for downloading teaching and learning material, music, film, games, or software from the internet	☐	☐
r. In a typical week, how many hours do you spend on using ICT devices? ---------------------- hours		
s. In a typical week, how many hours do you spend on internet surfing? ---------------------- hours		

How much do you agree with these statements?

	Disagree a lot	Disagree a little	Agree a little	Agree a lot
q. ICT has the capacity to strongly enhance classroom teaching and learning.	☐	☐	☐	☐
r. ICT provides valuable resources and tools to support student learning.	☐	☐	☐	☐
s. ICT provides students with efficient presentation and communication tools	☐	☐	☐	☐
t. I like challenge of exploring new technology and software.	☐	☐	☐	☐
u. I feel apprehensive about using ICT in my classroom teaching.	☐	☐	☐	☐
v. It scares me to think that I could cause the computer to destroy a large amount of information by hitting the wrong key.	☐	☐	☐	☐
w. I hesitate to use ICT tools and equipment for fear of making mistakes that I can't correct.	☐	☐	☐	☐
x. Computer and internet technologies are somewhat intimating to me.	☐	☐	☐	☐
y. I can effectively use ICT as instructional tool.	☐	☐	☐	☐
z. I can effectively manage my classroom when students are using ICT	☐	☐	☐	☐
aa. I can extend my classroom teaching by using computer and internet.	☐	☐	☐	☐
bb. I can learn to use ICT for my teaching and learning process.	☐	☐	☐	☐

PPT mode and its feedback				
1. How much do you agree with these statements? I prefer to use PPT with my students because ………				
	Disagree a lot	Disagree a little	Agree a little	Agree a lot
a. PPT examines students' ability to carry out mathematical procedures and methods.	☐	☐	☐	☐
b. PPT examines students' deeper understanding of mathematical concepts.	☐	☐	☐	☐
c. Students immediately receive good quality feedback in PPT.	☐	☐	☐	☐
d. PPT provides students with the opportunities to progress their learning.	☐	☐	☐	☐
e. PPT provides students with the motivation to mathematics learning.	☐	☐	☐	☐

LOT mode and its feedback				
1. How much do you agree with these statements? I prefer to use LOT with my students because ………				
	Disagree a lot	Disagree a little	Agree a little	Agree a lot
a. LOT examines students' ability to carry out mathematical procedures and methods.	☐	☐	☐	☐
b. LOT examines students' deeper understanding of mathematical concepts.	☐	☐	☐	☐
c. Students immediately receive good quality feedback in LOT.	☐	☐	☐	☐
d. LOT provides students with the opportunities to progress their learning.	☐	☐	☐	☐
e. LOT provides students with the motivation to mathematics learning.	☐	☐	☐	☐

CAT mode and its feedback				
1. How much do you agree with these statements? I prefer to use CAT with my students because ………				
	Disagree a lot	Disagree a little	Agree a little	Agree a lot
a. CAT examines students' ability to carry out mathematical procedures and methods.	☐	☐	☐	☐
b. CAT examines students' deeper understanding of mathematical concepts.	☐	☐	☐	☐
c. Students immediately receive good quality feedback in CAT.	☐	☐	☐	☐
d. CAT provides students with the opportunities to progress their learning.	☐	☐	☐	☐
e. CAT provides students with the motivation to mathematics learning.	☐	☐	☐	☐

Thank you for completing this questionnaire

Please return it to the researcher

Appendix D: Pre-Test (1) for Function

	Name: Gender: ☐ Female ☐ Male ☐ Other	
	(Answer **ALL** questions. Choose the correct or the most appropriate answer for each question. Write the letter of the correct or the most appropriate answer.) Time Allowed: (20) Minutes	
No	**Questions**	**Your Answer**
	Sample Question: 2+2 = **A.** 4 **B.** 2 **C.** 0 **D.** – 2 **E.** – 4	**A**
1	Let f : R → R, g : R → R be given by $f(x) = 2^x$, $g(x) = x^2$. Then (g.f) (1) = A. 4 B. 2 C. 1 D. 0 E. None of these	
2	f : R → R and g : R → R are defined by $f(x) = x^2 + 3$ and g(x) = 3x. If (g.f) (x) = 21, then x = A. ± 1±1 B. -1 C. ±2 D. -2 E. 1	
3	Let f : R → R and g : R →R be defined by $f(x) = 3 + 2x^2$ and g(x) = 2x + 1 If g(f(x)) = 35, then x = A. 7 B. 3 C. $\pm\sqrt{7}$ D. $\pm\sqrt{3}$ E. ±3	
4	Given f : R → R , $f(x) = 2^x$. If (f . f) (a) = 256, then a = A. 1 B. 2 C. 3 D. 0 E. -1	

5	$f : R \to R$ and $g : R \to R$ are given by $f(x) = 3^x + 1$, $g(x) = x – 1$. Then the value of $a \in R$ such that $(g.f)(a) = 27$ is A. 3 B. 9 C. 2 D. $\frac{1}{3}$ E. $\frac{1}{9}$	
6	$f : R \to R$ and $g: R\to R$ are given by $f(x) = x + 1$ and $g(x) = 2^x - 1$. Then the value of $a \in R$ such that $(f.g)(a) = 128$ is A. 5 B. 6 C. 7 D. 8 E. 9	
7	Let $f : R \to R$ and $g : R\to R$ be defined by $f(x) = px + 3$ and $g(x) = qx – 4$. If $(g . f)(x) =10x + 11$, p = A. - 2 B. 2 C. 3 D. -5 E. 5	
8	Given $g(x) = 5x + 4$, $(g . f)(x) = 10x – 11$, then f(x) = **A**. $2x + 3$ **B**. $2x – 3$ **C**. $5x + 3$ **D**. $5x – 3$ **E**. none of these	
9	If $g(x) = x + 1$ and $(g . f)(x) = x^2 + 3$, then f(x) = **A**. x^2-1 **B**. $x^2 + 2$ **C**. $x^2 + 1$ **D**. x^2 **E**. $3 – x^2$	
10	Given $f(x) = 2x – 1$ and $(g . f)(x) = 4x^2 – 2x – 3$, then g(x) = **A**. $2x – 3$ **B**. $4x – 1$ **C**. $x^2 + x – 3$ **D**. $x^2 + x + 3$ **E**. $x^2 – 3x –1$	
11	If $f(x) = x^2 + 3$ and $(g . f)(x) = 2x^2 + 3$, then g(x) = **A**. $2x$ **B**. $2x^2$ **C**. $2x + 3$ **D**. $2x – 3$ **E**. $3 – 2x^2$	
12	If $f(x) = x + 1$ and $(g . f)(x) = x^2 + 5x + 5$, then g(x) =	

	A. $x^2 + 3x + 1$ **B**. $x^2 + 3x$ **C**. $x^2 + 1$ **D**. $x – 1$ **E**. $x + 1$	
13	Given that $f : R \to R$ is defined by $f(x) = 1 – x$, then $f^{-1}(x) =$ **A**. $x – 1$ **B**. $1 + x$ **C**. $1 – x$ **D**. $\frac{1}{1-x}$ **E**. $\frac{1}{1+x}$	
14	Given that $h : R \to R$ is defined by $h(x) = 2x – 3$, then $h^{-1}(x) =$ **A**. $\frac{x+3}{2}$ **B**. $\frac{x-3}{2}$ **C**. $2x + 3$ **D**. $3 – 2x$ **E**. $2x – 3$	
15	Given that $g : R \to R$ is defined by $g(x) = \frac{5-3x}{7+2x}$, $x \neq -\frac{7}{2}$ then $g^{-1}(x) =$ **A**. $\frac{7x-5}{2x+3}$ **B**. $\frac{7x+5}{2x+3}$ **C**. $\frac{5+7x}{3+2x}$ **D**. $\frac{5-7x}{3+2x}$ **E**. $\frac{5-7x}{3-2x}$	
16	Given that $f : R \to R$ is defined by $f(x) = 2 + \frac{3}{x}$, then $f^{-1}(5) =$ **A**. -1 **B**. -2 **C**. 0 **D**. 1 **E**. 5	
17	A function f is defined by $f(x) = \frac{x + 3}{x - 2}$, $x \neq 2$. Then $f^{-1}(-4) =$ **A**. 1 **B**. -4 **C**. 5 **D**. -1 **E**. -5	
18	If $f(x) = \frac{x+a}{x-2}$, $x \neq 2$ and $f(7) = 2$, then $f^{-1}(-4) =$ **A**. 1 **B**. 2 **C**. 3 **D**. 4 **E**. 5	
19	If $f(x) = 3x + k$ and $f(3) = 11$, then $f^{-1}(65) =$ A. 15 B. 17 C. 19 D. 21 E. 23	
20	A function $f: R \longrightarrow R$ is given by $f(x) = 7 – kx$ and $f^{-1}(5) = 1$. Then k =	

	A. - 1 B.2 C. 1 D. -2 E. 4	

Appendix E: Pre-Test (2) for Remainder

	Name: Gender: ☐ Female ☐ Male ☐ Other	
	(Answer **ALL** questions. Choose the correct or the most appropriate answer for each question. Write the letter of the correct or the most appropriate answer.) Time Allowed: (20) Minutes	
No	**Questions**	**Your Answer**
	Sample Question: 2+2 = **A.** 4 **B.** 2 **C.** 0 **D.** – 2 **E.** – 4	**A**
1	If $px^3 + 4x^2 – 5x + 1$ has a remainder – 5 which divided by x + 2, then the value of p is A. -4 B. -1 C. 1 D. 2 E. 4	
2	If the polynomial $x^3 + ax^2 – x – 3$ is divided by x – 2, the remainder is 27, then a = A.-3 B. 2 C. 3 D. 6 E. -6	
3	If the remainder when $4x^2 – 3x + k$ is divided by x – 2 is 8, then k + 6 = A. -16 B. -4 C. 0 D. 4 E. 12	
4	If $x^3 + kx^2 – 10$ is divided by x – 3, remainder is 26. Then k = A. 3 B. -2 C. 2 D. -1 E. 1	

5	If $x^3 - 2x^2 - 3x + 4$ and $x^3 + 5x^2 - 7x - k$ have the same remainder when divided by $x + 1$, then k = A.5 B. 6 C. 7 D. 8 E. 9	
6	If $x^3 - 3x^2 + 5$ and $x^3 + 5x^2 + p$ have same remainder when divided by $x + 1$, p = A. 7 B. -7 C. 13 D. -3 E. 3	
7	If $x^3 - 4x^2 + 2xk + 6$ and $x^3 - 5x + 3k$ have the same remainder when divided by $x + 3$, then k = A. 3 B. - 4 C. 5 D. - 5 E. none of these	
8	$x^2 + ax + 9$ has the same remainder when it is divided by $x - 2$ or $x - 1$, then a = A. - 3 B. 3 C. 2 D. 0 E. -2	
9	When f(x) be a polynomial when f(x) is divisible by 3x, **A.** f(3) = 0 **B**. f(0) = 0 **C**. $f(\frac{1}{3}) = 0$ **D**. $f(-\frac{1}{3}) = 0$ **E**. none of these	
10	When f(x) be a polynomial when f(x) is divisible by $x - 2$, **A**. f(2) = 0 **B**. f(0) = 0 **C**. $f(\frac{1}{2}) = 0$ **D**. f(-2) = 0 **E**. none of these	
11	When f(x) be a polynomial when f(x) is divisible by $3x - 1$, **A**. f(1) = 0 **B**. f(-1) = 0 **C**. f(0) = 0 **D**. $f(\frac{1}{3}) = 0$ **E**. $f(-\frac{1}{3}) = 0$	
12	When the polynomial $g(x) = x^3$ is divisible by 5x,	

	A. $g(-\frac{1}{5}) = 0$ **B**. $g(\frac{1}{5}) = 0$ **C**. $g(5) = 0$ **D**. $g(-5) = 0$ **E**. $g(0) = 0$	
13	When the polynomial $f(x)$ is divisible by $3x + 3$, **A**. $f(\frac{1}{3}) = 0$ **B**. $f(-\frac{1}{3}) = 0$ **C**. $f(3) = 0$ **D**. $f(1) = 0$ **E**. $f(0) = 0$	
14	When $f(x) = x^3 – x^2 + kx + 2$ is exactly divisible by $(x + 1)$, the value of k is A. -3 B. -5 C. 3 D.5 E. 0	
15	If $2x^2 + 5x – k$ is divisible by $x – 2$, then k = A. 26 B. 2 C. -2 D. 18 E. -18	
16	If $4x^3 – 4x + c$ is divisible by $-\frac{1}{2}x + 1$ then the value of c is A. 10 B. 12 C. 24 D. -12 E. -24	
17	If $ax^3 – 9x – 2$ is divisible by $x + 2$, then a = A. 8 B. 2 C. 1 D. 16 E. 6	
18	If the remainder when $2x^3 + kx^2 +7$ is divided by $x – 2$ is half the remainder when the same expression is divided by $2x – 1$, the value of k is A. 2 B. 3 C. 4 D. 5 E. 6	
19	If the remainder when $x^4 + 3x^2 – 2x + 2$ is divided by $x + a$, is square of the remainder when $x^2 - 3$ is divided by $x + a$, the value of $9a^2 + 2a$ is A. 1 B. 5 C. 7 D. 11 E. 40	

20	If $kx^2 + 5x - 6$ is divisible by $2x + 3$, the remainder when it is divided by $3x - 2$, is A. 0 B. 6 C. $\frac{2}{3}$ D. $\frac{3}{2}$ E. $-\frac{3}{2}$	

Appendix F: Pre-Test (3) for Factor

	Name: Gender: ☐ Female ☐ Male ☐ Other	
	(Answer **ALL** questions. Choose the correct or the most appropriate answer for each question. Write the letter of the correct or the most appropriate answer.) Time Allowed: (20) Minutes	
No	**Questions**	**Your Answer**
	Sample Question: 2+2 = **A.** 4 **B.** 2 **C.** 0 **D.** – 2 **E.** – 4	**A**
1	If 2x – 1 is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(1) = 0 **B.** f(-1) = 0 **C.** $f(-\frac{1}{2}) = 0$ **D.** f(-2) = 0 **E.** $f(\frac{1}{2}) = 0$	
2	If x – 1 is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(0) = 0 **B.** f(1) = 0 **C.** f (0) = 1 **D.** f(-1) = 0 **E.** f(0)= -1	
3	If 3x + 4 is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(4) = 0 **B.** f(-4) = 0 **C.** $f(-\frac{4}{3}) = 0$ **D.** $f(-\frac{3}{4}) = 0$ **E.** $f(\frac{3}{4}) = 0$	

4	If 4x is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(4) = 0 **B.** f(-4) = 0 **C.** $f(-\frac{1}{4}) = 0$ **D.** $f(\frac{1}{4}) = 0$ **E.** f(0) = 0	
5	If x – 4 is a factor of a polynomial g(x), which of the following is certainly true? **A.** f(4) = 0 **B.** f(-4) = 0 **C.** g(- 4) = 0 **D.** g(4) = 0 **E.** g(0) = – 4	
6	If $x^2 - 5x + 6$ is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(6) = 0 **B.** f(-3) = 0 **C.** f(1) = 0 **D.** f(-2) = 0 **E.** f(3) = 0	
7	A factor of $x^3 + 9x^2 + 6x - 16$ is **A.** x + 1 **B.** x + 7 **C.** x – 8 **D.** x - 2 **E.** x – 1	
8	The expression $2x^3 - 13x^2 + 23x - 12$ has a factor **A.** 2x + 1 **B.** 3x – 2 **C.** 2x + 3 **D.** 2x – 3 **E.** x + 4	

9	The expression $2x^3 - 13x^2 + 23x - 12$ has a factor **A.** $x + 1$ **B.** $3x - 2$ **C.** $2x + 3$ **D.** $x - 4$ **E.** $x + 4$	
10	Which of the following is a factor of $2x^3 - 3x^2 - 11x + 6$? **A.** $x - 1$ **B.** $x - 2$ **C.** $x + 1$ **D.** $x + 2$ **E.** $x + 3$	
11	Which of the following is a factor of $2x^3 + x^2 + 5x - 3$? **A.** $2x + 1$ **B.** $2x - 1$ **C.** $x + 1$ **D.** $x - 1$ **E.** $x + 3$	
12	If $(x - k)$ is a factor of $f(x) = 4x^3 - (3k + 2)x^2 - (k^2 - 1)x + 3$, then k = **A.** 1 or -1 **B.** 1 or $\frac{3}{2}$ **C.** -1 or $\frac{3}{2}$ **D.** -1 or $-\frac{3}{2}$ **E.** $\frac{2}{2}$ or $-\frac{3}{2}$	
13	If $(x - p)$ is a factor of $4x^3 - (3p + 2)x^2 - (p^2 - 1)x + 3$, then p = **A.** $-\frac{1}{2}$ or 3 **B.** $\frac{1}{2}$ or -3 **C.** -1 or $\frac{3}{2}$ **D.** 1 or $-\frac{3}{2}$ **E.** -1 or $\frac{2}{3}$	

14	If $(x + 2)$ is a factor of $(x + 1)^7 + (2x + k)^3$, then the value of k is **A.** 2 **B.** 1 **C.**4 **D.**3 **E.** 5	
15	If $(x + 2)$ is a factor of $10 + 5x – 4x^2 – ax^3$ then a = **A.** 0 **B.** 11 **C.** 2 or 11 **D.** – 2 **E.** 2	
16	If $x – 1$ is a factor of $f(x) = x^3 – 6x^2 + px – 6$, then $p^2 – 1$ = **A.** 11 **B.** 21 **C.** 121 **D.** 120 **E.** 10	
17	If $x – 3$ is a factor of $f(x) = x^3 – 6x^2 + ax – 6$, then $a^2 – 1$ is A. 105 B. 11 C.121 D.120 E. 10	
18	If $x – 3$ is a factor of $x^3 – 6x^2 + ax – 6$, then a + 4 is A.22 B.15 C.12 D. 11 E. 5	
19	If $x + 2$ is a factor of $f(x) = x^3 – 3x^2 – ax + 2$, then the value of a is	

	A. 9 B. – 9 C. 1 D. – 11 E. – 1	
20	If x + 3 is a factor of $x^3 + 6x^2 + ax + 12$ then $a^2 - 19 =$ A. 14 B. 15 C. 150 D. 169 E. 196	

Appendix G: Post-Test (1) for Function

<table>
<tr><td></td><td>Name:

Gender: ☐ Female ☐ Male ☐ Other</td><td></td></tr>
<tr><td></td><td>(Answer ALL questions. Choose the correct or the most appropriate answer for each question. Write the letter of the correct or the most appropriate answer.)
Time Allowed: (20) Minutes</td><td></td></tr>
<tr><td>No</td><td>Questions</td><td>Your Answer</td></tr>
<tr><td></td><td>Sample Question: 2+2 =
A. 4 B. 2 C. 0 D. – 2 E. – 4</td><td>A</td></tr>
<tr><td>1</td><td>Let f : R → R, g : R → R be given by $f(x) = 2^x$, $g(x) = x^2$. Then (g.f) (1) =
A. 4 B. 2 C. 1 D. 0 E. None of these</td><td></td></tr>
<tr><td>2</td><td>f : R → R and g : R → R are defined by $f(x) = x^2 + 3$ and g(x) = 3x. If (g.f) (x) = 21, then x =
A. ± 1 B. -1 C. ±2 D. -2 E. 1</td><td></td></tr>
<tr><td>3</td><td>Let f : R → R and g : R →R be defined by $f(x) = 3 + 2x^2$ and g(x) = 2x + 1 If g(f(x)) = 35, then x =
A. 7 B. 3 C. $\pm\sqrt{7}$ D. $\pm\sqrt{3}$ E. ±3</td><td></td></tr>
<tr><td>4</td><td>Given f : R → R , $f(x) = 2^x$. If (f . f) (a) = 256, then a =
A. 1 B. 2 C. 3 D. 0 E. -1</td><td></td></tr>
</table>

5	f : R $\rightarrow$ R and g : R $\rightarrow$ R are given by $f(x) = 3^x + 1$, $g(x) = x - 1$. Then the value of $a \in R$ such that (g.f) (a) = 27 is A. 3 B. 9 C. 2 D. $\frac{1}{3}$ E. $\frac{1}{9}$	
6	f : R $\rightarrow$ R and g: R$\rightarrow$ R are given by f(x) = x + 1 and $g(x) = 2^x - 1$. Then the value of $a \in R$ such that (f.g) (a) = 128 is A. 5 B. 6 C. 7 D. 8 E. 9	
7	Let f : R $\rightarrow$ R and g : R$\rightarrow$ R be defined by f(x) = px + 3 and g(x) = qx – 4 . If (g . f) (x) =10x + 11, p = A. - 2 B. 2 C. 3 D. -5 E. 5	
8	Given g(x) = 5x + 4, (g . f) (x) = 10x – 11, then f(x) = **A**. 2x + 3 **B**. 2x – 3 **C**. 5x + 3 **D**. 5x – 3 **E**. none of these	
9	If g(x) = x + 1 and (g . f) (x) = $x^2 + 3$, then f(x) = **A**. x^2-1 **B**. $x^2 + 2$ **C**. $x^2 + 1$ **D**. x^2 **E**. $3 - x^2$	
10	Given f(x) = 2x – 1 and (g . f) (x) = $4x^2 - 2x - 3$, then g(x) = **A**. 2x – 3 **B**. 4x – 1 **C**. $x^2 + x - 3$ **D**. $x^2 + x + 3$ **E**. $x^2 - 3x - 1$	
11	If f(x) = $x^2 + 3$ and (g . f) (x) = $2x^2 + 3$, then g(x) = **A**. 2x **B**. $2x^2$ **C**. 2x + 3 **D**. 2x – 3 **E**. $3 - 2x^2$	
12	If f(x) = x + 1 and (g . f) (x) = $x^2 + 5x + 5$, then g(x) =	

	A. $x^2 + 3x + 1$ **B**. $x^2 + 3x$ **C**. $x^2 + 1$ **D**. $x - 1$ **E**. $x + 1$	
13	Given that $f : R \to R$ is defined by $f(x) = 1 - x$, then $f^{-1}(x) =$ **A**. $x - 1$ **B**. $1 + x$ **C**. $1 - x$ **D**. $\frac{1}{1-x}$ **E**. $\frac{1}{1+x}$	
14	Given that $h : R \to R$ is defined by $h(x) = 2x - 3$, then $h^{-1}(x) =$ **A**. $\frac{x+3}{2}$ **B**. $\frac{x-3}{2}$ **C**. $2x + 3$ **D**. $3 - 2x$ **E**. $2x - 3$	
15	Given that $g : R \to R$ is defined by $g(x) = \frac{5-3x}{7+2x}$, $x \neq -\frac{7}{2}$ then $g^{-1}(x) =$ **A**. $\frac{7x-5}{2x+3}$ **B**. $\frac{7x+5}{2x+3}$ **C**. $\frac{5+7x}{3+2x}$ **D**. $\frac{5-7x}{3+2x}$ **E**. $\frac{5-7x}{3-2x}$	
16	Given that $f : R \to R$ is defined by $f(x) = 2 + \frac{3}{x}$, then $f^{-1}(5) =$ **A**. -1 **B**. -2 **C**. 0 **D**. 1 **E**. 5	
17	A function f is defined by $f(x) = \frac{x+3}{x-2}$, $x \neq 2$. Then $f^{-1}(-4) =$ **A**. 1 **B**. -4 **C**. 5 **D**. -1 **E**. -5	
18	If $f(x) = \frac{x+a}{x-2}$, $x \neq 2$ and $f(7) = 2$, then $f^{-1}(-4) =$ **A**. 1 **B**. 2 **C**. 3 **D**. 4 **E**. 5	
19	If $f(x) = 3x + k$ and $f(3) = 11$, then $f^{-1}(65) =$ A. 15 B. 17 C. 19 D. 21 E. 23	

20	A function f: R $\longrightarrow$ R is given by $f(x) = 7 - kx$ and $f^{-1}(5) = 1$. Then k = A. - 1 B.2 C. 1 D. -2 E. 4	

Appendix H: Post-Test (2) for Remainder

	Name: Gender: ☐ Female ☐ Male ☐ Other	
	(Answer **ALL** questions. Choose the correct or the most appropriate answer for each question. Write the letter of the correct or the most appropriate answer.) Time Allowed: (20) Minutes	
No	**Questions**	**Your Answer**
	Sample Question: 2+2 = **A.** 4 **B.** 2 **C.** 0 **D.** – 2 **E.** – 4	**A**
1	If $px^3 + 4x^2 - 5x + 1$ has a remainder – 5 which divided by x + 2, then the value of p is A. -4 B. -1 C. 1 D. 2 E. 4	
2	If the polynomial $x^3 + ax^2 - x - 3$ is divided by x – 2, the remainder is 27, then a = A.-3 B. 2 C. 3 D. 6 E. -6	
3	If the remainder when $4x^2 - 3x + k$ is divided by x – 2 is 8, then k + 6 = A. -16 B. -4 C. 0 D. 4 E. 12	
4	If $x^3 + kx^2 - 10$ is divided by x – 3, remainder is 26. Then k = A. 3 B. -2 C. 2 D. -1 E. 1	

5	If $x^3 - 2x^2 - 3x + 4$ and $x^3 + 5x^2 - 7x - k$ have the same remainder when divided by $x + 1$, then k = A.5 B. 6 C. 7 D. 8 E. 9	
6	If $x^3 - 3x^2 + 5$ and $x^3 + 5x^2 + p$ have same remainder when divided by x + 1, p = A. 7 B. -7 C. 13 D. -3 E. 3	
7	If $x^3 - 4x^2 + 2xk + 6$ and $x^3 - 5x + 3k$ have the same remainder when divided by x + 3, then k = A. 3 B. - 4 C. 5 D. - 5 E. none of these	
8	$x^2 + ax + 9$ has the same remainder when it is divided by x – 2 or x – 1 , then a = A. - 3 B. 3 C. 2 D. 0 E. -2	
9	When f(x) be a polynomial when f(x) is divisible by 3x, **A.** f(3) = 0 **B**. f(0) = 0 **C**. $f(\frac{1}{3}) = 0$ **D**. $f(-\frac{1}{3}) = 0$ **E**. none of these	
10	When f(x) be a polynomial when f(x) is divisible by x – 2, **A**. f(2) = 0 **B**. f(0) = 0 **C**. $f(\frac{1}{2}) = 0$ **D**. f(-2) = 0 **E**. none of these	
11	When f(x) be a polynomial when f(x) is divisible by 3x – 1, **A**. f(1) = 0 **B**. f(-1) = 0 **C**. f(0) = 0 **D**. $f(\frac{1}{3}) = 0$ **E**. $f(-\frac{1}{3}) = 0$	
12	When the polynomial $g(x) = x^3$ is divisible by 5x,	

	A. $g(-\frac{1}{5}) = 0$ **B**. $g(\frac{1}{5}) = 0$ **C**. $g(5) = 0$ **D**. $g(-5) = 0$ **E**. $g(0) = 0$	
13	When the polynomial f(x) is divisible by $3x + 3$, **A**. $f(\frac{1}{3}) = 0$ **B**. $f(-\frac{1}{3}) = 0$ **C**. $f(3) = 0$ **D**. $f(1) = 0$ **E**. $f(0) = 0$	
14	When $f(x) = x^3 - x^2 + kx + 2$ is exactly divisible by $(x + 1)$, the value of k is A. -3 B. -5 C. 3 D.5 E. 0	
15	If $2x^2 + 5x - k$ is divisible by $x - 2$, then k = A. 26 B. 2 C. -2 D. 18 E. -18	
16	If $4x^3 - 4x + c$ is divisible by $-\frac{1}{2}x + 1$ then the value of c is A. 10 B. 12 C. 24 D. -12 E. -24	
17	If $ax^3 - 9x - 2$ is divisible by $x + 2$, then a = A. 8 B. 2 C. 1 D. 16 E. 6	
18	If the remainder when $2x^3 + kx^2 + 7$ is divided by $x - 2$ is half the remainder when the same expression is divided by $2x - 1$, the value of k is A. 2 B. 3 C. 4 D. 5 E. 6	
19	If the remainder when $x^4 + 3x^2 - 2x + 2$ is divided by $x + a$, is square of the remainder when $x^2 - 3$ is divided by $x + a$, the value of $9a^2 + 2a$ is A. 1 B. 5 C. 7 D. 11 E. 40	

20	If $kx^2 + 5x - 6$ is divisible by $2x + 3$, the remainder when it is divided by $3x - 2$, is A. 0 B. 6 C. $\frac{2}{3}$ D. $\frac{3}{2}$ E. $-\frac{3}{2}$	

Appendix I: Post-Test (3) for Factor

	Name: Gender: ☐ Female ☐ Male ☐ Other	
	(Answer **ALL** questions. Choose the correct or the most appropriate answer for each question. Write the letter of the correct or the most appropriate answer.) Time Allowed: (20) Minutes	
No	**Questions**	**Your Answer**
	Sample Question: 2+2 = **A.** 4 **B.** 2 **C.** 0 **D.** – 2 **E.** – 4	**A**
1	If 2x – 1 is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(1) = 0 **B.** f(-1) = 0 **C.** $f(-\frac{1}{2}) = 0$ **D.** f(-2) = 0 **E.** $f(\frac{1}{2}) = 0$	
2	If x – 1 is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(0) = 0 **B.** f(1) = 0 **C.** f (0) = 1 **D.** f(-1) = 0 **E.** f(0)= -1	
3	If 3x + 4 is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(4) = 0 **B.** f(-4) = 0 **C.** $f(-\frac{4}{3}) = 0$ **D.** $f(-\frac{3}{4}) = 0$ **E.** $f(\frac{3}{4}) = 0$	

4	If 4x is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(4) = 0 **B.** f(-4) = 0 **C.** $f(-\frac{1}{4}) = 0$ **D.** $f(\frac{1}{4}) = 0$ **E.** f(0) = 0	
5	If x – 4 is a factor of a polynomial g(x), which of the following is certainly true? **A.** f(4) = 0 **B.** f(-4) = 0 **C.** g(- 4) = 0 **D.** g(4) = 0 **E.** g(0) = – 4	
6	If $x^2 - 5x + 6$ is a factor of a polynomial f(x), which of the following is certainly true? **A.** f(6) = 0 **B.** f(-3) = 0 **C.** f(1) = 0 **D.** f(-2) = 0 **E.** f(3) = 0	
7	A factor of $x^3 + 9x^2 + 6x - 16$ is **A.** x + 1 **B.** x + 7 **C.** x – 8 **D.** x - 2 **E.** x – 1	
8	The expression $2x^3 - 13x^2 + 23x - 12$ has a factor **A.** 2x + 1 **B.** 3x – 2 **C.** 2x + 3 **D.** 2x – 3 **E.** x + 4	

9	The expression $2x^3 – 13x^2 + 23x – 12$ has a factor **A.** $x + 1$ **B.** $3x – 2$ **C.** $2x + 3$ **D.** $x – 4$ **E.** $x + 4$	
10	Which of the following is a factor of $2x^3 – 3x^2 – 11x + 6$? **A.** $x – 1$ **B.** $x – 2$ **C.** $x + 1$ **D.** $x + 2$ **E.** $x + 3$	
11	Which of the following is a factor of $2x^3 + x^2 + 5x – 3$? **A.** $2x + 1$ **B.** $2x – 1$ **C.** $x + 1$ **D.** $x – 1$ **E.** $x + 3$	
12	If $(x – k)$ is a factor of $f(x) = 4x^3 – (3k + 2)x^2 – (k^2 – 1) x + 3$, then k = **A.** 1 or -1 **B.** 1 or $\frac{3}{2}$ **C.** -1 or $\frac{3}{2}$ **D.** -1 or $-\frac{3}{2}$ **E.** $\frac{2}{2}$ or $-\frac{3}{2}$	
13	If $(x – p)$ is a factor of $4x^3 – (3p + 2)x^2 – (p^2 – 1) x + 3$, then p = **A.** $-\frac{1}{2}$ or 3 **B.** $\frac{1}{2}$ or -3 **C.** -1 or $\frac{3}{2}$ **D.** 1 or $-\frac{3}{2}$ **E.** $–1$ or $\frac{2}{3}$	

14	If $(x + 2)$ is a factor of $(x + 1)^7 + (2x + k)^3$, then the value of k is **A.** 2 **B.** 1 **C.**4 **D.**3 **E.** 5	
15	If $(x + 2)$ is a factor of $10 + 5x – 4x^2 – ax^3$ then a = **A.** 0 **B.** 11 **C.** 2 or 11 **D.** – 2 **E.** 2	
16	If $x – 1$ is a factor of $f(x) = x^3 – 6x^2 + px – 6$, then $p^2 – 1$ = **A.** 11 **B.** 21 **C.** 121 **D.** 120 **E.** 10	
17	If $x – 3$ is a factor of $f(x) = x^3 – 6x^2 + ax – 6$, then $a^2 – 1$ is A. 105 B. 11 C.121 D.120 E. 10	
18	If $x – 3$ is a factor of $x^3 – 6x^2 + ax – 6$, then a + 4 is A.22 B.15 C.12 D. 11 E. 5	
19	If $x + 2$ is a factor of $f(x) = x^3 – 3x^2 – ax + 2$, then the value of a is	

	A. 9 B. – 9 C. 1 D. – 11 E. – 1	
20	If x + 3 is a factor of $x^3 + 6x^2 + ax + 12$ then $a^2 - 19 =$ A. 14 B. 15 C. 150 D. 169 E. 196	

Appendix J: Semi-structure interview

Students' Interview

Student-male interviewee No. 1 (Sm1)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sm1: *Yes, I like classroom formative assessment because it lets me know my current ability.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sm1: *I like CAT. It is because I like the short test time and its immediate score and feedback from CAT. Their feedback is clear. I like the use of computer and the display of one question in one screen because no need to focus on other questions. I like receiving immediate score and feedback from CAT.*

Student-male interviewee No. 2 (Sm2)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sm2: *I do not like classroom formative assessments because I am always busy with tests. My test anxiety is increased because of them. There is not enough time to study for a concept whole, especially for slow learners like me. The contents were forgotten after testing.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sm2: *Currently, PPT is easier to take than LOT and CAT (CBT). However, I like LOT the most because I received the immediate score and feedback. The feedback makes me to get more understanding. CBT improves the computer literacy and familiarity. There are no challenges in CBT apart from the poor connection.*

Student-male interviewee No. 3 (Sm3)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sm3: *I like classroom formative assessments because the assessments help us to get deeper understanding of new concepts. It makes me have more study time. It makes me know which subjects are needed to be more focused.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sm3: *I like computer-based test because immediate score and feedback. I like anything on computer. However, if I pressed wrong answer and forgot to check back, I can lose one. In poor connection and lack of electricity, if computer don't save it, we need to start again. In PPT, we can do multiple checking. I prefer to PPT test and its feedback because electronic feedback is not unreliable and feedback form teachers. We hardly access the devices. If we have owned devices, I prefer electronic feedback.*

Student-male interviewee No. 4 (Sm4)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sm4: *I am not happy to take classroom formative assessments because the number of tests is too frequent.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sm4: *I prefer PPT to two other computer-based test. This is because I do not have enough knowledge of the use of computer. I am not familiarity with internet and computer. I have no confidence of using the computer. I prefer paper-based feedback. That is the best for us. We can study it again. I prefer the print of feedback sheet to the electronic one. Print one can be saved long time because the electron one easily to lose and I still do not know how to save it electronically or soft copy, because I am not familiarly with computer.*

Student-male interviewee No. 5 (Sm5)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sm5: *There are no challenges, but formative assessment makes me a bit busy. And it makes me more motivate learning. Increase motivation, and more focus on the study and improved my achievement. If I got high score in formative assessment, it makes me happy and joyful in learning. If lose mark, figure out my weakness.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sm5: *I like anything on computer screen. ICT can give the efficient ways. I like both LOT and CAT. I want to choose CAT as my favourite because I get short test time, immediate score, and feedback. Especially receiving feedback highlight me which is correct or incorrect in short time. It measures my mathematics skills in short time.*

Student-female interviewee No. 1 (Sf1)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sf1: *I like to take classroom formative assessment. It makes us to more study and more improve in learning, reduce test anxiety because of frequency. They can help us to know which subject needs to have more emphasis.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sf1: *I like computer-based test because we get immediate score and feedback which makes me study more, know which is my error or mistakes and motivate in learning. Also, I am not familiar with computer and ICT devices. Wait for good connection. I prefer to choose PPT because I like teacher feedback. I want teacher to explain my error rather than machine feedback. I like hardcopy (print) feedback to electronic feedback. This is because I am not familiar with the computer, and I have less access ICT.*

Student-female interviewee No. 2 (Sf2)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sf2: *I like classroom formative assessments because I can know which are my mistakes and weaknesses. I can correct myself and I can study more in my weak subject. The classroom formative assessments make me to reduce test anxiety also because I am so used to taking test.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sf2: *I want to choose LOT, as my favourite, among three test modes. I am familiar with ICT to some extent because, at home and school, I also access ICT. So, I am confident with the use of ICT. Then, I was getting immediate score and feedback after computer-based test, so I prefer LOT.*

Student-female interviewee No. 3 (Sf3)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sf3: *I do not like taking classroom formative assessment frequently. Tests make me busy, there is no time for study which I want. I do not have relaxation due to the frequent tests. I* *cannot focus on the weak subject because I hardly get result from tests. The amount of study contents is gained from year to year. Good points: we can know our ability.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sf3: *I like LOT and CAT because I like study or test on screen. I believe that the use of ICT makes me to improve my confidence in learning. I prefer to choose LOT because I get the immediate score and feedback according to my answers in details. No challenges, apart from sometimes poor internet connection. I prefer testing in the computer classroom because they turn on the soothing music. The computer classroom is better place to take test. I prefer computer feedback to teacher feedback.*

Student-female interviewee No. 4 (Sf4)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sf4: *I like classroom formative assessments because they cover only small amount of content. They make me to do study more. They help me to Improve achievement in learning. I can know the weak point in new concepts.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sf4: *I am more familiar PPT, but I like CBT. I want to choose LOT as my favourite because of its immediate feedback, and score. LOT can give us the news skill, in the further studies like TOFEL, which is one of CBTs. CBT provide us the familiarity with future education and work opportunity. I prefer computer feedback and the CBT immediately shows all my mistakes. Teacher cannot give the feedback on each mathematics question.*

Student-female interviewee No. 5 (Sf5)

Researcher: *What are your attitude/perceptions of classroom formative assessment?*

Sf5: *I don't like classroom formative assessments which makes me busy and stressful. There is no time to study which I want, and to relax. I cannot also focus on the weak subject because the amount of study contents is improved from year to year. The good points is that we can know our ability through formative assessment.*

Researcher: *Which test mode do you like among PPT, LOT, or CAT? Why do you choose it as your preference?"*

Sf5: *I prefer LOT. I like the immediate feedback and scoring and its simplicity. LOT test is simpler than CAT. The reason why I choose LOT as my like is because I know how to use computer and internet well. Mostly I access the internet and computer at home and school. Also, I trust the utility of ICT in the future education and work.*

Teachers' Interview

Teacher interviewee No. 1 (T1)

Researcher: *What is the teachers' perception of formative assessment?*

T1: *I totally agree with the use of FA, if teacher knows which mistake students did, teacher can do the remedial and help their teaching. However, teacher need more time for scoring. I cannot explain their individual mistake. I cannot do tutorial for every concept. And I do not have time to score and tell the mistake to individual.*

Researcher: *Which test mode do you prefer, paper-based or linear online or computer-adaptive test mode?*

T1: *Currently there are not enough computers for all students in classroom. So, I think computer-based tests are not ready to use in the classroom currently. So, I chose the paper based as the best applicable. There are advantages of CBT, because they do not need for scoring, and they are able to provide feedback immediately. In PPT, students need to wait for teacher scoring and feedback.*

Teacher interviewee No. 2 (T2)

Researcher: *What is the teachers' perception of formative assessment?*

T2: *I like FAs, so students continuously study always. I make tutorial according to the number of concepts in each chapter. Tutorial make me busy mostly. I believe that tutorial is good type of formative assessment. It helps student in learning progress. Sometimes printing the questions is costly. Teachers pay for it themselves. The students enjoy taking tutorial. For teacher, there are disadvantage in FA: time consuming and heavy workload in making question, administering classroom tests and making marking scheme and scoring all answer sheets of all students who they are teaching. This cause to reduce the classroom teaching time. Teachers reduce the contents administered and select only focused or targeted content for each class. They do not have enough classroom teaching time to cover all content areas.*

Researcher**:** *Which test mode do you prefer, paper-based or linear online or computer-adaptive test mode?*

T2: *I chose PPT as the best applicable because it does not need extra support. Not enough computers for all students in a classroom. For LOT and CAT, there are extra teachers and periods because there are not enough computers in the classroom.*

Teacher interviewee No. 3 (T3)

Researcher: *What is the teachers' perception of formative assessment?*

T3: *FAs are good for students, so they can study well every concepts. At least there is two FA test in a month, so they are studying well. The student covers the whole content small by small. Every concept there is a test, so they study every day, now they won't. For the average or lower achievers, the previous monthly test is the best. Therefore, I did tutorial myself after finishing a concept. Tutorial make students be harder. Testing makes students improve learning.*

Researcher**:** *Which test mode do you prefer, paper-based or linear online or computer-adaptive test mode?*

T3: *I like LOT. I can easily measure the ability and provide scores and feedback. Teachers save time for question making (item-bank), test security, testing administration (because students cannot copy other answers, scoring and providing feedback. Class teachers can trace their abilities. Students can get to know their abilities immediately.*

Teacher interviewee No. 4 (T4)

Researcher: *What is the teachers' perception of formative assessment?*

T4: *FA positively impact on students' achievement. Tutorials are used as classroom formative assessment. Tutorial makes students study more. I see tutorials have some better points for their achievement than summative assessment. If I can do tutorial twice in a month according to each concept, I find their achievement improvement.*

Researcher: *Which test mode do you prefer, paper-based or linear online or computer-adaptive test mode?*

T4: *I like LOT. I can save time. I prefer item banking, scoring, and providing feedback. I can save their track of achievement. I receive more student interest. Test security. Save more time for scoring.*

Teacher interviewee No. 5 (T5)

Researcher: *What is the teachers' perception of formative assessment?*

T5: *I believe the FA is the most suitable and necessary for student learning improvement. The FA as tutorial is good for students. Students can learn well and catch up every lesson. Measure their ability and achievement. However, making FA too frequently takes over all class time and teachers and student are so busy with FAs.*

Researcher: *Which test mode do you prefer, paper-based or linear online or computer-adaptive test mode?*

T5: *I like LOT and CAT because it scores automatically and provide feedback immediately. Trace individual improvement. Security is good. Everything restores in computer. And I am so familiar with computer, I understand the advantage of technology. I prefer LOT and CAT to PPT. I spend one time for make questions and save and applied multiple time. Especially I like item-bank. Sacrifice for a year, and then apply multiple times. Some necessary will be added because student will change, contents do not change, and teacher or instructor does not change. If instructor change, he can pass it on next one. But I consider for other teachers who are lack of knowledge of computer. They will face more challenges. For students, CAT and LOT is good.*

Psalm 23

1 The LORD is my shepherd; I shall not want.

2 He makes me to lie down in green pastures:

He leads me beside the still waters.

3 He restores my soul:

He leads me in the paths of righteousness for His name's sake.

4 Yea, though I walk through the valley of the shadow of death, I will fear no evil:

for You are with me;

Your rod and Your staff they comfort me.

5 You prepare a table before me in the presence of my enemies:

You anoint my head with oil; my cup runs over.

6 Surely goodness and mercy shall follow me all the days of my life: and

I will dwell in the house of the LORD forever and ever more, Amen