ASSESSING THE RELIABILITY OF COMPUTER ADAPTIVE TESTING IN COLLEGE ENTRANCE COMMUNICATION SKILLS EXAMINATIONS IN MALAWI.

MASTER OF EDUCATION (TESTING, MEASUREMENT AND EVALUATION) THESIS

THOKOZANI ELVIS CHISALE

UNIVERSITY OF MALAWI

AUGUST, 2024

ASSESSING THE RELIABILITY OF COMPUTER ADAPTIVE TESTING IN COLLEGE ENTRANCE COMMUNICATION SKILLS EXAMINATIONS IN MALAWI.

MASTER OF EDUCATION (TESTING, MEASUREMENT AND EVALUATION) THESIS

By

THOKOZANI ELVIS CHISALE

Bachelor of Science (Technical Education) –University of Malawi, The Polytechnic

Submitted to the Department of Educational Foundations, School of Education, in partial fulfilment of the requirements for the degree of Master of Education (Testing, Measurement and Evaluation)

UNIVERSITY OF MALAWI

AUGUST, 2024

DECLARATION

I the undersigned hereby declare that the text of this dissertation is my original work which has not been submitted to any institution for similar purposes. Where other people's work has been used, acknowledgements have been made.

THOKOZANI ELVIS CHISALE

_____

Full Legal Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis is the student's own work and effort and has been submitted with our approval.

Signature:_____ Date: _____

Foster Gondwe, PhD (Senior Lecturer)

Main Supervisor

Signature:_____ Date: _____

Yohane Chakasika, (Senior Lecture)

Postgraduate Coordinator (Education Foundation)

DEDICATION

I dedicate this work to my parents Mr. and Mrs. Chisale for their unwavering support, to my wife, Lucy for her understanding during my studies, and lastly to my children Michelle, Wales, and Mtendere for their fortitude and perseverance.

ACKNOWLEDGEMENTS

I sincerely offer my modest gratitude to everyone who encouraged, mentored, and supported me in bringing this research to fulfilment.

First of all, I would like to convey my gratitude to Dr. Foster Gondwe, my supervisor, for his tireless understanding and intellectual support throughout my research. I owe my family a debt of gratitude as well, as they would sleep late only to keep me company while studying. Likewise, I want to express my sincere thanks to Mr Humphrey Kunyenge, my research assistant, for his perpetual help with collecting data.

In addition, I would like to express my gratitude to the head teacher and the entire staff of Ntonda Community Day Secondary School for covering my lessons when I was unable to attend.

Finally, God has been a cornerstone in my thesis. When I thought there was no light at the end of the tunnel, He always shone a light and led me to critical pieces of information to help make this thesis more interesting and relevant.

ABSTRACT

Admission into institutions of higher learning in Malawi is very competitive considering the limited capacity of higher education institutions. Admission procedures vary by institution; some offer entrance examinations, while others do not. Nevertheless, college entrance examinations provide a standardized and objective measure of the academic level of students from diverse educational backgrounds and boost a high predictive validity of student's success in college. In Malawi, the institutions which administer entrance examinations use paper and pencil mode of delivery which seem to have limited reliability among other challenges. On the other hand, computer-adaptive testing (CAT) provide a changed approach and has practical advantages that could be leveraged to overcome the challenges of PBT. However, previous studies have reached different conclusions when comparing the scores from CBT to PBT. Hence, this research assessed the reliability of (CAT) as an alternative to paper-based testing (PBT). The study took a quantitative approach with a quasi-experimental design. Data was collected from a sample of the 2022 DCE and NCE scripts of the paper-based test and the Live CAT administration. The item parameters were examined to determine the paper's quality, then, theta estimates and standard error of measurement were compared. Finally, Pearson moment correlation was used to assess the linear relationship of CAT to PBT estimated scores. The results indicate that Paper-based entrance exams are of moderate quality, CAT estimate scores more precisely than PBT, and with a correlation statistic of 0.717, the scores from CAT and PBT have a positive relationship. The findings suggest that CAT can be relied upon as an alternative to PBT in entrance examinations. Further research must be conducted on the consequences of using a misfit model in computer adaptive tests, understanding the perception of students with the transition from paper-based tests to computer adaptive tests and study on minimum items for calibrating an item bank.

Keywords: Computer adaptive testing, College admission criteria, Item response theory, Test delivery method.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ACRONYMS AND ABBREVIATIONS

BEP	Bayesian Estimating Procedure
CAT	Computer Adaptive Testing
CEE	College Entrance Examination
CTT	Classical Test Theory
CU	Catholic University
DCE	Domasi College of Education
IRT	Item Response Theory
MCHS	Malawi College of Health Sciences
MLE	Maximum Likelihood Estimation.
MSCE	Malawi School Certificate of Education
NCE	Nalikule College of Education.
NEP	National Education Policy
NESIP	National Educational Sector Investment Plan
PBT	Paper-Based Testing (Paper-Pencil Testing)
UEE	University Entrance Examinations
UNIMA	University of Malawi

CHAPTER ONE

INTRODUCTION TO STUDY

1.1 Chapter Overview

This chapter provides background information that brings to light the current college admission testing practice in Malawi. It explains the test delivery mode and measurement theory used (paper-based testing using classical test theory) and the test administration process. It then introduces the alternative test delivery mode and measurement theory (computer-based testing using item response theory) and how the advent of computers has made it possible to integrate IRT in education assessment through computer adaptive testing. The chapter also presents the statement of the problem. It also describes the purpose of the study, the specific questions guiding the research, and the significance of the study.

1.2 Research Background

Technology in teaching and learning has taken centre stage in recent years. Assessment as a part of teaching and learning has not been excused, as digital assessments have blossomed over the years. Different development agendas, as well as policies, have streamlined technology and innovations in education (Ministry of Education [MoE], 2016; National Planning Commission [NPC], 2020; Ministry of Education, Science and Technology. [MoEST], 2020). For example, Malawi Agenda 2063 priority area 2:

industrialization, calls for a redesigned education system to respond to current and future skills in the promotion of research, science, technology, and innovation. This then entails that education institutions will be driven to adopt technological approaches to teaching and learning. The use of technology in assessment is a complex issue, while it offers potential benefits, such as the use of organic data and machine learning algorithms, it also raises concerns about the validity, reliability, measurement bias, and the digital divide of the assessments (Hsu & Liou, 2021). More research therefore is needed to establish the reliability and precision of these assessments before institutions can contemplate the possibility of adopting technology in assessment. This draws the curiosity of the researcher to assess the reliability of technology, specifically computer adaptive testing, in college entrance examinations in Malawi as compared to paper-pencil mode which is the current practice.

Tests are systematic procedures for observing people and describing them and or their abilities with a numerical scale or categorical system. Tests are selected based on their purpose, and to be effective, they must have the following qualities: reliability (the test must produce consistent results), validity (the test must be shown to measure what it is intended to measure), and be unbiased (the test should not place students at a disadvantage because of gender, ethnicity, language, or disability) (Zucker, 2003). Similarly, tests can be delivered in different modes, including oral mode, paper-based mode, and computer-based mode. Comparative studies have raised the issue of mode effects on student performance, i.e., whether the test scores obtained from computer-based tests (CBT) and paper-based tests (PBT) are interchangeable (Oz & Ozturan, 2018). Tests are also supposed to have an underlying measurement theory. Measurement

theory in educational psychology consists of statistical and methodological tools to support inferences about examinees. The theory is commonly called test theory. Test theory is based on a positivist worldview in which latent traits are interpreted in a realist fashion. The common test theories are classical test theory, generalizability theory, and item response theory.

Some tertiary institutions in Malawi administer entrance examinations to admit students to the various programs on offer. Constructs assessed under these examinations are numerical skills, reasoning skills, and communication skills. This study dwells on communication skills paper. Communication skills are considered one of the most critical competencies for academic and career success, as evident in surveys of stakeholders from higher education and the workforce. The structure of the paper differs across test publishers, but the common constructs measured by communication skills paper are parts of speech, question tags, punctuation of sentences, and sentence construction among others. This construct is selected bending to the accession that candidates are more comfortable with language examination in computer-based tests (Oz & Ozturan, 2018). The current testing practice is that all candidates are given the same questions on paper and pencil. They are required to respond to all the items in a specific duration (1 hour) per paper. After the time elapses, they submit the scripts. The examiners then collect all the scripts and mark them. Each question carries one mark and the total number of correct responses is added to give the total test score per paper. The total test scores for all three papers are later added to give the candidate's final score for the test. This final score is ranked and the top students per the number required are selected. The total test score determines the eligibility of a candidate to be admitted to the college.

College entrance examinations are standardized tests that measure an examinee's readiness for tertiary education. These tests measure the projected potential to perform well in future activities rather than knowledge acquired in school. The examination also levels the testing field as candidates who wrote senior secondary examinations from different examination boards are tested on the same constructs.

There are numerous types of entrance examinations and various test publishers. While content and structure may vary between publishers, the skills under assessment remain the same. The most common tests are numerical skills, reasoning skills, and communication skills. Numerical skills tests measure the capacity for dealing with numerical data quickly and accurately, and ability to apply basic arithmetic. Communication skills tests look at the ability to conclude from written information, as well as test vocabulary and language comprehension. Reasoning skills tests are a measure of problem-solving ability and ask a person to identify rules and relationships between abstract sequences (practice aptitude tests, 2020).

The past decade has seen public institutions abolishing college entrance examinations in Malawi. The influencing factor has been the harmonisation drive in admitting students into different public tertiary institutions by the National Council for Higher Education (Singini, 2014). Nevertheless, some public institutions like Domasi College of Education, Nalikule College of Education, Malawi College of Health Sciences and all teacher training colleges still administer college entrance examinations. The expectation is that these institutions will be forced to abolish entrance examinations in the near future based on the said harmonisation drive. However, college entrance examination play a crucial

role in education, serving as a tool for testing knowledge and personality. In China, the college entrance examination has been instrumental in selecting talents and promoting quality education. Jiang and Xuyang (2020) opined that the college entrance examination system in China has been on the road to reform and the function of the college entrance examination in the new era has ushered in a multi-dimensional turn and transformation, that is, the turn of the social function of cultivating diversified talents, the educational function of adaptive examination, and the value function of improving students' comprehensive quality. Hence, instead of abolishing entrance examination, the National Council for Higher Education and tertiary institutions should seek to reform the entrance examination as a selection procedure.

The college entrance examinations using paper-based mode have administrative as well as reliability challenges. *On administrative challenges*, the college entrance examinations are administered on a single day at regional centres. This entails that all candidates must make it without fail on the said day or else they miss their chance for that year. Cheating issues manifest, and scoring is another issue as it takes a lot of time to score and report the scores to examinees (Personal communication, DCE, 2023)**.** *On reliability challenges*, test items are repeated, resulting in item exposure, which affects the measurement of the examinee's ability. The scoring of the entrance examination test also assumes that items are of the same difficulty and discrimination level.

All the institutions that administer college entrance examinations in Malawi use the classical test theory (CTT). While classical test theory has proven very useful in test development, the two statistics that form its cornerstone; indices of item difficulty and

item discrimination are both sample-dependent. In particular, the classical test theory model cannot accommodate tests that target an examinee's aptitude level because it lacks information regarding how an examinee is predicted to perform on a particular item (Hambleton et al, 1991).

On the other hand, item response theory has become an important complement to CTT in the development, interpretation, and evaluation of tests and test items. The interest in IRT grew out of a combination of concerns about the limitations inherent in CTT and the availability of computing systems. IRT has a strong mathematical basis and depends on complex algorithms that are more efficiently solved via computer. It describes the relationship between an examinee's test performance and the traits assumed to underlie such performance on achievement tests as a mathematical function called the item characteristics curve (Hambleton & Swaminathan, 1985). IRT primarily focuses on item-level information, in contrast to the CTT's primary focus on test-level information. Test items are of different difficulty levels, and they discriminate against examinees differently. The issue of guessing also affects the response of the examinee. Raw test responses need to be weighed to come up with estimations of ability using techniques of measurement theory to sift through the factors of item difficulty, discrimination, and guessing. Considering the advantages of IRT and its compatibility with computer adaptive testing, it is worth assessing its reliability to provide precise estimates of ability, standard error of measurement, psychometric properties in comparison with the paper-and-pencil mode. Resultantly, when institutions contemplate the transition, they are aware of the characteristics of alternative test delivery mode from empirical research.

1.3 Statement of the Problem

Admission into tertiary institutions in Malawi is very competitive considering the limited capacity of higher education institutions. Hence, paper-based entrance examinations are administered by several academic institutions to admit deserving students. However, entrance examinations face several challenges among them, administrative and security hindering the reliability of scores obtained from entrance examinations (Yongbo, 2020). This has compelled some institutions to abolish entrance examinations in favour of the Malawi School Certificate of Education (MSCE) or its equivalent as a yardstick for admitting students (University of Malawi (UNIMA), 2015). The challenges with entrance examinations are influenced by test delivery mode among other factors. Scholars have reached different conclusions on the reliability of scores from paper-based test and computer-based tests. Bennett et al. (2008), Piaw (2012), and Kalender and Berberoglu (2017) found that computer-based tests and paper-based tests provide the same estimates of ability, Wang et al. (2008) revealed no comparability between scores obtained from the two testing modes whilst Clariana and Wallace (2002) suggest that it is not necessary that equivalent measures be produced from CBT and PBT. Computer adaptive testing (a component of CBT) provides a changed approach to assessment and with the advancement of technology and its sophistication in the analysis of test items, it could be leveraged as an alternative test delivery mode. However, CAT has primarily been used in developed nations, and its application in developing countries including Malawi is minimally known. Yet, it seems there is insufficiency of research on assessing the reliability of computer adaptive testing in college entrance examinations in Malawi.

1.4 Purpose of the Study

The purpose of the study was to assess the reliability of computer adaptive testing as an alternative to paper-based testing in college communication skills entrance examination in Malawi, using the Domasi College of Education 2022 communication skills paper as an example.

1.5 Research Questions

1.5.1 Main Research Question

The main research question is, what is the reliability of computer adaptive testing as an alternative to paper-based testing in college communication skills entrance examinations in Malawi?

1.5.2 Specific Research Questions

The following research questions guided the research.

1. What is the quality of the communication skills entrance examination paper?
2. How comparable is the frequency of correct responses to items in CAT and PBT?
3. How comparable are the candidate's ability measurements in CAT to PBT?
4. What is the relationship between candidates' scores in CAT and PBT?

1.6 Significance of the Study

The study will inform tertiary institutions of test delivery methods that provide reliable estimates of ability. It is envisaged to establish the reliability of computer adaptive testing

so that tertiary institutions are informed in the event of a decision to transform from a conventional mode of testing to a computer mode in entrance examinations.

This study provides empirical evidence on how test delivery modes influence the estimation of ability in college entrance examinations in Malawi, as Chulu (2013, p. 3) emphasized that an effective assessment system requires the availability of a comprehensive policy framework, structures, and scientific data to support the system.

The study confirmed that the assumptions of item response theory which is the framework for computer adaptive testing can reliably be used to admit students into college. Test items which have known parameters and are adapted to each examinee can precisely measure the ability of candidates and rank them for merit selection into the college programs.

1.7 Operational Definitions

Classical Test Theory (CTT) – is a theory based on the assumption that an examinee has an observed score, a true score, and an error score. Spearman's model envisioned observed test scores to be a composite of two hypothetical components, a true score and a random error component (Crocker and Algina, 1989).

College Entrance Examination (CEE) - also referred to as the University Entrance Examination (UEE) is a test used to determine an individual's skill or propensity to succeed in tertiary education.

Computer –Based Testing (CBT) – is a test delivery method using a computer.

Computer Adaptive Testing (CAT) - is a distinct approach to the assessment of latent traits through a computer, where the test item is specifically matched to the ability of each examinee (Davey & Pitoniak, 2006).

Item Response Theory (IRT) – Item response theory is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test (Hambleton and Jones, 1993).

Paper-Based Testing (PBT) – is a test delivery method through paper and pencil.

Test - is a systematic procedure for observing persons and describing them with a numerical scale or categorical system (Zucker, 2003).

Test Reliability – is a measure of the consistency in the estimation of the examinee's ability informed by the item quality, test information, and minimal errors.

1.8 Chapter Summary

This chapter provided a Background to the study. College admission testing practice in Malawi, test delivery mode and issues of mode effect, measurement theories, and the need to conduct reliability studies in different contexts were explained. The statement of the problem, the purpose of the study, the research questions, and the significance of the study were also presented. The literature review in the next chapter builds on this background to provide relevant knowledge in the field.

CHAPTER TWO

LITERATURE REVIEW

2.1 Chapter Overview

This chapter reviews relevant literature pertaining to the study. It begins with an explanation of college entrance examinations, the qualities of effective tests, test delivery modes and the process of computer adaptive testing. It further expounds on estimations of ability and standard error of measurement, comparability and reliability studies on CAT, and the landscape of digital technology in assessment in Malawi.

2.2 Qualities of an Effective Test.

Tests are selected based on their purpose and to be effective they must have the following qualities: Reliable (The test must produce consistent results), Valid (The test must be shown to measure what it is intended to measure, and Unbiased (The test should not place students at a disadvantage because of gender, ethnicity, language, or disability) (Zucker, 2003).

This paper focuses on test reliability issues. Among the several definitions of test reliability, the fundamental idea has been the precision and consistency of test scores, with various methods used to estimate it (Cronbach, 1947; Mead, 2005; Wells 2013). Several terms associated with the concept of test reliability include: “true score,” “error of measurement,” “alternate-forms reliability,” “interrater reliability,” “internal

consistency," "reliability coefficient," "standard error of measurement," "classification consistency," and "classification accuracy." (Livingstone, 2018). This study looks at test reliability as the extent to which measurements resulting from a test are characteristics of those being measured. In a technical sense, the theoretical definition of test reliability is the proportion of score variance (measurement error) that is caused by systematic variation in the population of test takers.

Test reliability is a joint characteristic of a test and examinee group, not just a characteristic of the test. There are three major sources of error that affect test reliability: factors in the test itself, factors in the students taking the test, and scoring factors. Most tests contain a collection of items that represent particular skills. Error, however may be introduced by the selection of particular items to represent the skills and domains. The particular cross-section of test content that is included in the specific items on the test will vary with each test form, introducing sampling error and limiting the dependability of the test, since we are generalizing to unobserved data, namely; ability across all items that could have been on the test. Other sources of test error include the effectiveness of the distractors (wrong options) in multiple-choice tests, partially correct distractors, multiple correct answers, and difficulty of the items relative to the student's ability. Test takers are not always consistent and also introduce errors into the testing process. Whether a test is intended to measure typical or optimal student performance, changes in such things as student's attitudes, health, and sleep may affect the quality of their efforts and thus their test-taking consistency. For example, test takers may make careless errors, misinterpret test instructions, forget test instructions, inadvertently omit test sections, or misread test items. Lastly, scoring errors are a third potential source of error. On

objective tests, the scoring is mechanical, and the scoring error should thus be minimal. On constructed-response items, sources of error include clarity of the scoring rubrics, clarity of what is expected of the student, and a host of rater errors.

To improve test reliability there is a need to develop better tests with less random measurement error than simply documenting the amount of error. Measurement error is reduced by writing items clearly, making the instructions easy to understand, adhering to proper test administration, and providing consistent scoring. Because a test is a sample of the desired skills and behaviours, some scholars have proposed longer tests as they generally yield more reliable scores in educational and psychological measurement (Aday, 2018; Murphy, 2022). On the other hand some scholars have proposed that longer tests might not be feasible in most cases, as such with fewer test items adaptive tests can be more reliable (Weiss, 1982; Hambleton et al., 1991; Wang et al., 2010).

This piece of literature directed this study to look at reliability of a test as a multifaceted variable. It is a composite of measurement error. The magnitude of the error foretells the reliability of the test. Instead of just reporting the error, test administrators should ask question on what multifaceted methodologies can minimise the error? And how feasible are such methodologies. The knowledge has informed this study to look at how Computer adaptive testing as an alternative to Paper-based testing can minimise the error from the test takers, the test and scoring procedures.

2.3 Measurement Theories in Educational Psychology

Measurement theory in educational psychology consists of statistical and methodological tools to support inferences about examinees (Mislevy, 1995). The theory is commonly

called Test Theory. Test theory is based on a positivist worldview in which latent traits are interpreted in a realist fashion. The common test theories are Classical test theory, Generalizability theory, and Item response theory.

2.3.1 Classical Test Theory

This theory is based on the assumption that an examinee has an observed score and a true score. Spearman's model explains that an observed test score is a composite of two hypothetical components, a true score and a random error component-expressed in the form

$$X = T + E$$

Where **X** represents the observed test score; **T** is the individual's true score, and **E** is a random error component (Crocker & Algina, 1995 p. 106). While CTT has proven very useful in test development, the two statistics that form its cornerstones; indices of item difficulty and item discrimination are both sample-dependent. In particular, the classical test theory model cannot accommodate tests that target an examinee's aptitude level because it lacks information regarding how an examinee is predicted to perform on a particular item (Hambleton et al., 1991).

2.3.2 Generalizability (G) Theory

This is a psychometric theory based on a statistical sampling approach that partitions scores into their underlying multiple sources of variation (Li et al., 2015). This theory is done in two phases: a generalizability (G) study and a decision (D) study. In generalizability theory, a set of measurement conditions is called a facet. A facet may be

treated as fixed or random. This uses analysis of variance (ANOVA) to estimate reliability coefficients and errors of measurement.

2.3.3 Item Response Theory

This is a measurement theory that postulates examinees' ability and the latent trait on the same continuum. It describes the relationship between an examinee's test performance and the traits assumed to underlie such performance on achievement tests as a mathematical function called the item characteristics curve (Hambleton & Swaminathan, 1985). IRT primarily focuses on item-level information. The relationship between the examinee's ability and performance on an item is described by one or more parameters depending on which IRT model is used.

The popular IRT models are

One Parameter Logistic model (1PL),

$$P(x_i = 1 | \theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Two Parameter Logistic model (2PL),

$$P(x_i = 1 | \theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Three Parameter Logistic model (3PL).

$$P(x_i = 1 | \theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

The parameters are known as **a**-parameter (discrimination), **b**-parameter (difficulty), and **c**-parameter (Pseudo-Guessing).

The measurement theory supports the conclusions drawn from examinee scores, and if the measurement theory is incorrect, the measurement process is compromised. This study used item response theory for item analysis, item bank calibration, and Live CAT administration. Item response theory (IRT) has grown in popularity as a methodological framework for modelling response data from educational and health tests, yet it is not widely used by educational psychologists in Malawi. This research intends to provide an instructive application of IRT and highlight some of its benefits for psychological test construction. The use of IRT in test development has several advantages over other measurement theories, mainly because IRT produces person parameter invariance when the model fit is present, and test information functions provide the amount of information or "measurement precision" captured by the test on the scale measuring the construct of interest and other features (Zanon et al, 2016).

2.4 Test Delivery Modes

Tests can be delivered in different modes among them: Oral mode, Paper-based mode, and Computer-based mode. Tests should be developed and delivered in a way that allows the participation of the widest possible range of students and results in reliable and valid inferences about performance for all students who participate in the assessment (Thompson et al., 2002, p. 5). Thus, the delivery mode must resemble the environment and enhancements of learning. With the inclusion of Technology as instruction media at

colleges, it is conspicuous to have empirical data on whether having entrance examinations in computer mode of delivery could provide precise ability estimates.

In today's digital age, tests are increasingly being delivered on computers. Many of these computer-based tests (CBTs) have been adapted from paper-based tests (PBTs). However, this change in the mode of test administration has the potential to introduce construct-irrelevant variance, affecting the validity of score interpretations. Because of this, when scores from a CBT are to be interpreted in the same way as a PBT, evidence is needed to support the reliability and validity of these scores (American Educational Research Association [AERA], 2014). The Standards for Educational and Psychological Testing (AERA, 2014) state that a rationale is needed for adapting a test to a new mode of administration. Given the COVID-19 safety concerns, there has been a very good reason to adapt paper-and-pencil tests for computer administration.

Lynch (2022) highlights that the benefits of CBTs over PBTs are also a major motivation for the transition. Computerized tests are seen to be more efficient than their paper-based counterparts because scoring is automated, enabling faster reporting and feedback; administration is better controlled, improving standardization and test security; and more data can be gathered, permitting more sophisticated psychometric analyses (Way & Robin, 2016; Wise, 2018). Some of these benefits such as the absence of errors in scoring, test automation, and simulations improve reliability. Candidates themselves seem to be motivated by computer-delivered tests

Wang and Kolen (2001) cautioned against assuming the interchangeability of scores from a PBT and adapted CBT without evidence. To address these comparability concerns, *the Standards for Educational and Psychological Testing* (AERA et al., 2014) and *the International Guidelines on Computer-Based and Internet-Delivered Testing* (International Test Commission [ITC], 2005) outline best practices when such adaptations are made. Both publications underscore the need to demonstrate the comparability of the two test modes and minimize sources of construct-irrelevant variance. The ITC guidelines (2005) are specific in their recommendations, stating that a PBT and CBT should be comparable in terms of their reliabilities, means, and standard deviations; the two versions should be correlated, and should correlate with similar measures; and a CBT should be designed to minimize sources of construct-irrelevant variance. AERA et al. (2014) provide more general advice, stating that empirical evidence supporting the validity of interpretations and the reliability of test scores of a CBT adapted from a PBT should be warranted. However, studies often reveal mixed results regarding the comparability issues of CBT and PBT.

Several studies have explored the comparability of Computer-Based Tests (CBT) and Paper-Pencil Tests (PBT). Wang et al. (2008) and Hakim (2017) both found that CBT can be more efficient and effective, with immediate scoring and reporting of results, and improved test performance. However, Yao (2019) noted differences in test takers' performance across different CEFR levels, with only the CEFR A2 level showing a statistically significant difference between CBT and PBT. This suggests that while CBT may offer advantages, it is important to consider the specific context and test takers' characteristics.

Computer-based tests can be categorized into linear computer-based tests and non-linear computer-based tests. Linear CBT are tests delivered through a computer but the candidates answer the same number of tests whilst non-linear CBT are tests delivered through a computer where the candidates write different items. The items could be administered randomly or tailored to the candidate's ability based on the responses they provide in real-time. The current study concentrates on Computer adaptive tests (a type of non-linear computer-based test) compared to paper and pencil tests.

2.5 Use of Technology for Assessment Purposes.

The use of technology in assessment has transformed the field, offering new opportunities and tools for administrators. This transformation is driven by advances in cognitive and measurement science, which have the potential to fundamentally change assessment in areas such as test design, item generation, and scoring. Mobile technologies, including PDAs and mobile phones, have been particularly influential in assessment, offering anytime, anyplace data collection and multimedia capabilities (Sandars & Dearnley (2008). These tools have been adopted to improve the quality, timeliness, and cost efficiency of assessment and evaluation processes.

It seems that technology will minimize some of the challenges of entrance examinations in Malawi. Technology innovates and enhances assessments in terms of item and test design, methods of test delivery, data collection and analysis, and the reporting of test results. Testing process like item writing, pretesting (using actual examinees or

simulations) to have the parameters of the items, security and scoring will be without errors, hence improving test reliability.

Whilst there are huge potential benefits of technology-enhanced assessments there should be recognition that some practices may make assessment more accessible and comfortable; others may be divisive or exclusionary and there are possibilities of new divides emerging with new practices of technology-enhanced assessment (Grant and Villalobos, 2008). This foretells that if technology-enhanced assessments are to be used, some care must be taken to avoid obtaining undesired scores, hence diminishing test reliability.

2.6 Computer Adaptive Testing

Computer-adaptive testing is a distinct approach to the assessment of latent traits through a computer, where the test item is specifically matched to the ability of each examinee (Davey & Pitoniak, 2006). Development and administration of CAT take 5 stages: item bank calibration, starting rule, item selection rule, scoring rule, and stopping rule. The test items are selected depending on the answer of the previous item: If the Previous item is answered correctly, the next item will be more difficult. If the item is answered incorrectly, then the next item will be less difficult. CATs are more efficient than paper based tests (Weiss and Kingsbury, 1984), their efficiency measure range from generally reducing test length by 50% or more, controlling measurement precision, adding security (candidates are administered a different set of test items), frequent retesting, and immediate results scoring and reporting, to having more item formats**.**

Computer adaptive tests (CAT) offer several advantages over traditional linear tests. They can reduce test anxiety and length while increasing the precision of ability estimates (Stepanek, 2020). CAT also provides a balance of accuracy and efficiency in knowledge evaluation. The use of item response theory in CAT ensures comparability of test scores and allows for immediate judgment of response quality. Furthermore, CAT can be implemented in computer-assisted learning and e-learning, providing more efficient test administration and intelligent learning evaluation.

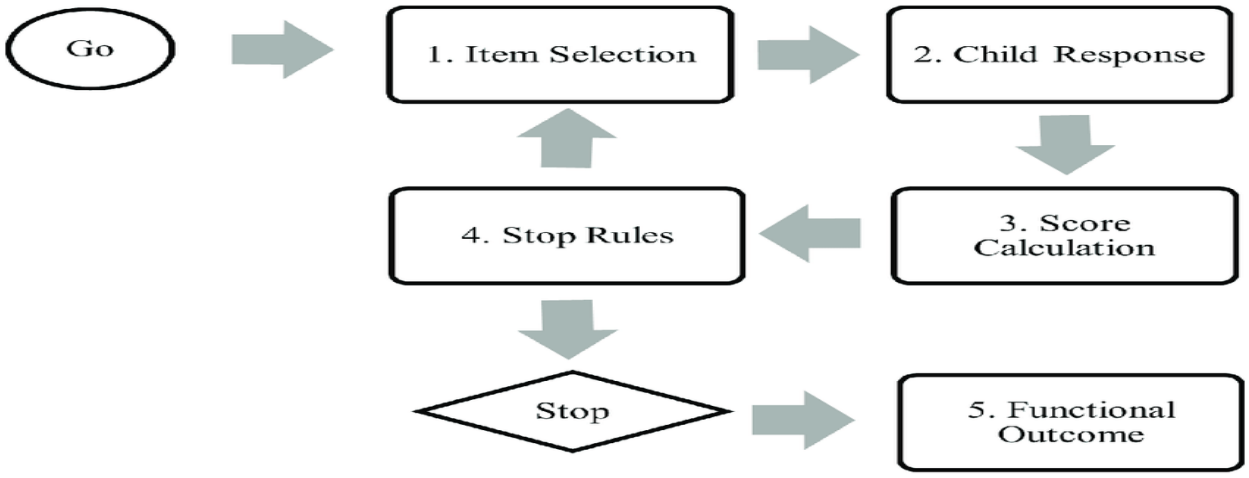


Figure 1: Illustration of the steps of computer adaptive testing CAT

Note. The steps of a CAT in assessment. From "Item-saving assessment of self-care performance in children with developmental disabilities: A prospective caregiver-report computerized adaptive test," by Chen C.T, Chen Y.L, Lin Y.C, Hsieh C.L, Tzeng J.Y, and Chen K-L, 2018, PLoS ONE 13(3): e0193936. p. 3 (https://doi.org/10.1371/journal.pone.0193936). Copyright 2018 by Creative Commons Attribution License.

2.6.1 Item Bank Calibration

Computer adaptive testing requires a calibrated item bank. Since each examinee has different traits and as such will be administered a different set of items, the item bank

must contain items that will cater for a wide range of examinees and which can measure the ability of learners in extremes. A classic recommendation is 12–16 times the adaptive test length (Stocking, 1994), based on considerations of content constraints, item exposure, and test overlap. Larger item banks are needed when examinees' latent traits cover a broader range so that there are adequate items to match the examinees. For example, an item bank of 1000 to 2000 items has been used where the test length averages 20 to 30 items (Kingsbury & Houser, 1999; He & Min, 2017). In simulations, 100 to 500 items for a variable-length test have been used (Han, 2012; Sahin and Ozbasi, 2017, Sahin and Weiss, 2015, Rudner and Guo, 2011). Studies which used data from the Paper-based version of the test have used a limited number of test items on average 30 to 60 items bank (Kalender and Berberoglu, 2017, Kaya, 2021). The data collection instrument in this study has 30 items for item bank calibration.

2.6.2 Starting Rule and Item Selection Rule

There are multiple rules to choose from when deciding on the first item to be administered. Given that the program uses information about the examinee to choose an item, the first starting rule could be to use some prior information on the examinee, prior information from the tested population when there is no prior information for the examinee, use an item of average difficulty level (b = 0.0), an item from within the initial difficulty range of -0.5 to 0.5, and giving easy items initially to the examinees to help reduce test anxiety, though not based on psychometric properties but is rather based on old habit by test developers. There are many issues to consider when choosing a starting rule. One such issue is whether different proficiency estimates and different items adversely affect final estimates. The method of initial item selection does not adversely

affect final score estimates when using a likelihood-based estimator but could affect the estimates when using a Bayesian method (Thissen & Mislevy, 2000). It has been shown that the longer the test, however, the less the initial item will affect the final estimate of proficiency level (Lord, 1980).

After the examinee responds to the first item, the adaptive algorithm begins. Using the parameters of the IRT model, the computer now administers items based on the examinee's previous pattern of correct/incorrect responses. Two decisions to be made are: how to score the responses and how to choose the next item for administration (Embretson & Reise, 2000). There are multiple psychometric methods for item selection, with the common ones being maximum Fisher information and Kullback-Leibner information-based selection. However, all of the methods choose the single "best" item at each stage for administration. Some of the item selection constraints that should be considered include not administering any of the items twice, minimizing item exposure as well as item content. Eggen (2012) expressed that the practice of a computerized adaptive learning system needs to have better continuously updated estimates of the individual's ability and it is recommended to combine the item selection method with better estimation methods during test administration.

2.6.3 Scoring Rule and Stopping Rule

According to van der Linden and Pashley (2000), in CAT, there are three stages of ability estimation: estimates to start the item selection, estimates during the test to adapt item selection to the examinee's ability, and estimates at the end of the test to have a final theta score of the examinee.

The estimates to start item selection includes prior examinees estimates, average difficulty and random easy items (explained under starting rule). The various methods for estimating ability to adapt item selection include Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP), and Expected A Posteriori (EAP). The scoring can either be on one item for unidimensional computer adaptive tests or per section for multidimensional computer adaptive tests. The last scoring is estimation at the end of the test commonly referred to as the stopping rule or termination rule.

There are numerous stopping rules (test termination criteria) for use in CAT. The broader categories are fixed length and variable length stopping rules. Fixed-length CATs terminate when the set maximum number of items have been administered whilst variable-length CATs terminate using Standard Error of Measurement minimum information (MI), standard error (SE), change in theta, generalized likelihood ratio (GLR), predicted standard error reduction (PSER), minimum determinant rule (D-rule), minimum eigenvalue rule (E-rule), and maximum trace rule (T-rule)., among other rules. Research on termination criteria in computer adaptive tests (CATs) has highlighted the importance of variable-length CATs for efficient and effective measurement (Babcock & Weiss, 2012). However, the use of fixed-length CATs is not necessarily inferior, as they can perform comparably to variable-length CATs (Babcock & Weiss, 2009). In practice, some have combined both fixed and variable length as well as time to terminate computer adaptive tests. Kalender and Berberoglu (2017) and Kaya (2021) used SE and test length as stopping rules. Babcock and Weiss (2012) in their study on used different variable length termination criteria in combination with different minimum numbers of items.

Based on the findings, standard error of measurement in combination with a minimum of 15 items may be ideal, depending on the precision needs of the test user and the discriminations of the items in the bank.

2.7 Studies on Comparability of CAT and PBT

Score comparability between the two administration techniques becomes the main problem when the existing PBT method is replaced by CAT administration or when the methods are used interchangeably (Wang & Kolen, 2000). As a result, research is required to provide evidence on the reliability and comparability of the scores. Wang and Shin (2010) explains that CAT introduces a new testing paradigm in which test items are not the same for each examinee, as they are in conventional tests. CAT offers notable variances in the testing framework, comparability between various test modes cannot be naively assumed. Instead, it should be investigated. It is important to take into account how the PBT and CAT variants of the same test's scores might be compared, as well as the impact of changing the medium of administration and overall paradigm.

Wang and Shin (2010) state that administration mode is an influential factor needing investigation across CAT and PBT, also pointing out that score comparability from different administrations of a test should be fully satisfied. The study made a comparison between CAT and PBT by reporting descriptive statistics, such as central tendency and dispersion measures, rank orders, and the validity and reliability evidence for the test scores.

It is evident that studies produce contrasting views on the comparability of CAT and PBT scores. Vispoel (2000), Paek (2005), and Wang et al (2008) in their study on adaptive testing administered in K-12 and college-level testing programs concluded that scores obtained from the two testing modes are not comparable. Similarly, Schaeffer et al. (1998) on GRE scores found no comparability of the scores. On the contrary, Kalender and Berberoglu (2017) in their study on CAT in admission of students in Turkey found the correlation of ability estimates between PBT and CAT to be .764 ($p < .05$). This showed that the scores are comparable. Likewise, Kaya (2021) reported no significant difference observed between PBT and CATs terminated with SE threshold and fixed-item stopping rules. Babcock and Weiss (2012) found that CATs produced comparable ability estimates with their PBT counterparts regardless of the test termination method.

Due to defensibility and accountability concerns, as well as professional testing standards, empirical evidence demonstrating comparability of test scores obtained in the two administration modes is necessary whenever paper- and computer-based assessments of the same subject matter are conducted. In the absence of such proof, CAT cannot be suggested as a credible substitute for an exam's PBT equivalent (Kaya, 2022).

2.8 Studies on Reliability of CAT

Reliability studies are seen to be classical ways of data analysis. However, their computations are still present as well as very useful to date. There are several measures of the reliability of computer adaptive tests among them the standard error of measurement, and Pearson moment correlation. The standard error of measurement (SEM) provides a way of summarising the amount of error or inconsistency in test scores. The SEM is a

function of two values: the standard deviation of the test and the reliability of the test. The higher the reliability of the test, the smaller the SEM, and the more precise the test is.

$$SEM = \sigma_X \sqrt{1 - \rho^2},$$

Figure 2: Standard error of measurement equation

Culligan (2008) explained that in item response theory, item information functions (IIF) are vital to the calculation of the standard error of measurement. Item information function is the proportion of the square root of the differentiation of item characteristics to its variance. It shows the amount of information each item provides and it is calculated by multiplying the probability of endorsing a correct response multiplied by the probability of answering incorrectly.

$$SEM = \frac{1}{\sqrt{I(\theta)}}$$

Figure 3: The equation that links the information function with the standard error of measurement

Graphically the item information function will be a curve on a probability of correct response and latent trait continuum plane (see figure 4). The item information functions can then be summed into a test information function. Lastly, the test information function is often inverted into the conditional standard error of measurement function, which is extremely useful in test design and evaluation.

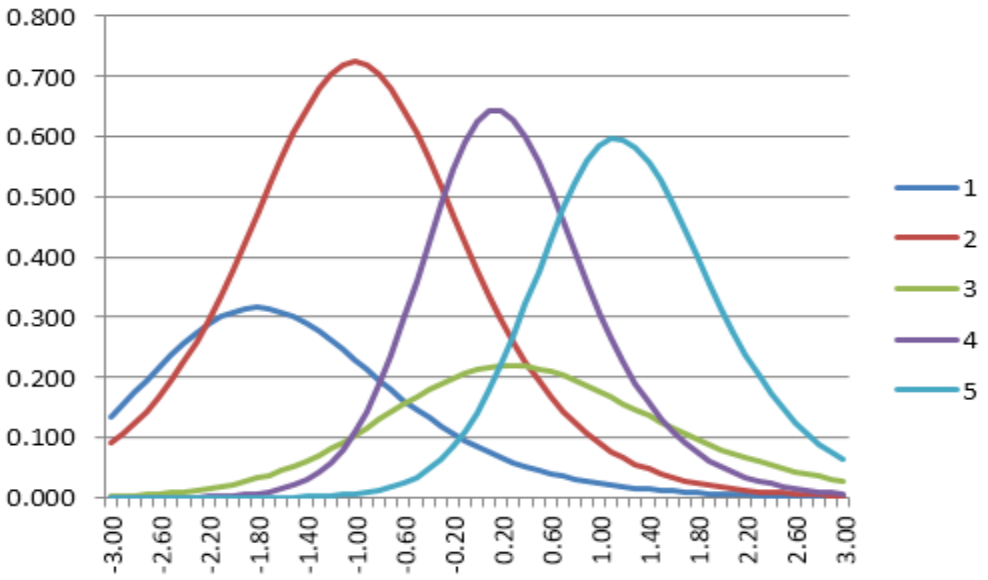


Figure 4: Item information function for 5 items

Note. Item 2 has more information for candidates with ability level of -1.40, item 5 is difficult, so it is not very useful for examinees in the bottom half of ability. From "IRT test information function," by Nathan Thompson, 2021. Copyright 2024 by Assessment Systems.

In addition, Reliability of computer adaptive tests have been computed using Pearson moment correlation. The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (see figure 5). The Pearson product moment correlation coefficient is a widely used measure of the strength and direction of a linear relationship between two variables (Franke, 2010; Wilson, 2018). The coefficient ranges from -1 to +1, with values less than 0.3 indicating weak correlation, values between 0.3 to 0.7 indicating moderate correlation and values above 0.7 indicating a stronger linear

relationship, the negative sign indicate the direction of the relationship (Wilson, 2018). The coefficient can be interpreted as the predicted standard deviation difference in the dependent variable for a one standard deviation difference in the independent variable.

Kalender and Berberoglu (2017) used Pearson moment correlation to check the reliability of CAT. The Correlations were obtained from the full test length, with the mathematics subtest scores as an external criterion. On average, for the full test length, a correlation of $r > .91$ was found at $SE = .30$ whilst for correlation with the mathematics subtest, a correlation of $r > .71$ at $SE=.30$ was found. Even though slightly lower correlations were obtained with the mathematics subtest scores compared to full test length, the correlations were not lower than .65 for any of the different test termination rules. The results indicated that CAT reduces the number of items used for all termination rules. Likewise, Kaya (2022) analysed the reliability of CAT and PBT using Pearson Moment correlation. Two termination criteria were used; thus standard error and test length. Using standard errors from .50 to .10 an average correlation of .767 was found whilst using test length between the ranges of 10 test items to 50 an average correlation of .748 was found. Overall, all the correlations are higher than .70. All the correlation values were significant at the alpha level of .01. However, the correlations between the ability estimations terminated with standard error threshold rule are slightly higher than the correlations based on fixed-length CATs.

2.9 Chapter Summary

The literature research has revealed three key findings about assessing the reliability of computer adaptive testing as an alternative to paper-based assessment in Malawi. Firstly, it is apparent that there is a variance of findings on the comparability of test delivery modes in estimating ability and their corresponding standard error of measurement. As a result, the comparability of PBT and CAT should not be assumed but investigated. Secondly, as one of the pillars for assessing the reliability of CAT as an alternative to other test delivery methods defined by Wang and Kolen (2001), is that to check the information functions of the tests and how the scores from CAT correlate with scores obtained in PBT administration. The estimates obtained in CAT must correlate highly with other delivery modes. Lastly, technology-enhanced assessment must minimize the challenges of the paper-based delivery mode whilst not being divisive considering socio-economic status and computer literacy levels in Malawi. Hence, there is a need to have empirical evidence that digital assessment approaches like CAT which utilize the IRT framework can provide precise estimates of ability, SEM, and psychometric properties like the paper versions of the Test.

Using the DCE and NCE 2022 communication skills paper as an instrument, this study investigated the reliability of computer adaptive testing as an alternative to paper-based testing in the college communication skills entrance examination in Malawi. Based on the literature review, methodologies for conducting the study are explained in the next chapter.

CHAPTER 3.

METHODOLOGY

3.1 Chapter Overview

This chapter describes the methodology of the study. The research paradigm, approach, design, sampling, and data collection procedures are explained. Likewise, the data management methods, data dissemination strategy, data analysis techniques, limitations of the study, and ethical considerations are described.

3.2 Research Paradigm and Approach

This research employed a positivist paradigm. The student responses collected were objective in nature and it was a conviction of the researcher that there is one objective reality. A quantitative approach was be used in this study. This approach was selected so that the findings' applicability could be generalized to a larger population. Johnson et. al (2007) explained that data collected through a quantitative approach could be generalized to a larger degree, with which the data for the same issue with different social contexts is collected.

3.3 Research Design

A Quasi- experimental design was used in this study. This design was selected because of its capability to demonstrate correlation of variables. The researcher compared the ability estimates, standard error and psychometric properties of the items obtained from two different test delivery methods. Brennan (2001) records that a research design provides a framework for the collection and analysis of data; in other words, a good choice of a research design reflects decisions about the priority being given to the range of dimensions of the research process.

3.4 Study Population and Sampling

The population of the study was 4360 scripts for DCE communication skills paper of which a sample size of 1005 was used for item bank calibration. For the Live CAT administration, the population of the study was 2204 first-year students at Catholic University (CU), Domasi College of Education (DCE), Malawi College of Health Science (MCHS), and Nalikule College of Education (NCE) of which a sample size of 546 students was used. A sample of above 500 is a requirement of IRT techniques as recommended by Ree and Jensen (1983) for statistically meaningful results. The researcher stratified the sample students into colleges. A student sample was 150 from CU, 150 from DCE, 150 from MCHS, and 96 from NCE. The participants were notified that they had been sampled to participate in the study. Those who voluntarily accepted formed part of the final sample for the study. The four colleges have been purposively selected since they administer entrance examinations in Malawi.

Participants included in this study were first-year students in the four colleges. The first-year students were the recent cohort to have sat for the entrance examinations which is a selection criterion to admit students in the colleges. It is assumed that these students had comparable ability to those who sat the scripts used for item bank calibration. The study excluded students in the other levels of study.

3.5 Study Sites

The study was conducted in four tertiary institutions Malawi which were Catholic University, Domasi College of Education, Malawi College of Health Sciences, and Nalikule College of Education. Two of the study sites are in the southern region of Malawi, while one is in the eastern region and another is in the central region of Malawi. These institutions were purposively selected because they administer entrance examinations to select students for various programs that they offer.

3.6 Study Period

The study was conducted in 19 months. It began in September 2022 and concluded in March 2024. The first six months were for the development of the research proposal, defence of the proposal, and submission to the University of Malawi Research and Ethics Committee (UNIMAREC) for approval. The next eleven months were for data collection and analysis, a major portion was spent on software search, since, commercial software like FastTest was costly and open-source software like Concerto could not run with the specifications of the computer the researcher was using, and the last two months being for addressing comments and submission to the education foundations department.

3.7 Instrumentation and Data Collection Tools

DCE and NCE 2022 Communication skills entrance examination paper was used for this study. Data was collected in two folds. Firstly, 1005 communication skills scripts for the 2022 DCE and NCE entrance examinations were recorded for the calibration of the instrument. These scripts were systematically sampled. Every 4^{th} script was sampled to be part of the study. The total was 1009 but after a data audit 4 cases were dropped. Correct responses were assigned a 1 whilst incorrect responses and missing responses were assigned a 0 in SPSS. The SPSS file was then saved in tab-delimited format which is a format that X-calibre software recognizes. The first six columns were for candidate ID (000001 to 001005) whilst from column seven to thirty-six were the candidate's responses (10100100010010011111000011010O). This was compiled for the 1005 scripts and formed the data matrix file. A control file was coded with 30 items, having a single response, and dichotomously scored. X-calibre was then run to produce the item parameters, a process called item bank calibration. The Item characteristics were used to adaptively select items based on the candidate's previous response. Secondly, a live CAT version was administered after item bank calibration. In this delivery mode candidate's latent traits and item characteristics were measured concurrently. The instrument had 30 items and according to Kalender and Berberoglu (2017), tests of 30 to 60 items produce robust results. The data collection instrument was purposively selected because it was the most recent paper administered.

3.8 Data Analysis

3.8.1 Preliminary Analysis

Firstly, IRT Assumptions of Local independence and unidimensionality were checked. KMO and Bartlett's test was used to test for local independence, whilst principal factor analysis was used to test for unidimensionality. A test is unidimensional when a single latent trait accounts for all the common variance among item responses (Morizot et al., 2007, p. 413). Eigenvalues was computed and first factor value compared with the other factors. A scree plot was schemed as well (see figure 5). The assumption of unidimensionality and local independence were obtained. On the other hand other IRT assumptions of monotonicity and item invariance were not checked because the data was not grouped.

A model-data fit assessment was done to choose among the three IRT models to be used for the study. 3 Parameter Logistic model (3PL) fitted the data and was used in this study. Maximum Likelihood Estimation (MLE) was used for the estimation of the candidate's ability. Maximum Likelihood Estimation (MLE) is a widely used method for estimating parameters of statistical models due to its desirable properties, such as unbiasedness, small variance, and ease of approximation. MLE also offers a way to devise estimators of unknown population parameters without the need for calculating expected values (González et al., 2016).

3.8.2 Comparability Analysis

The study compared item administration frequency, percentage of correct responses to items, average testing time, and estimates of the ability of examinees. Using frequency statistics and Independent sample t-test analysis the decision on the comparability of the two different test delivery modes was made at 95% confidence level. The null hypothesis (H_0) was that there is no difference between average estimates of ability and standard error of measurement in CAT and PBT.

3.8.3 Correlation Analysis

The Standard error of measurement (SEM) and Pearson moment correlation were used to assess the correlation of CAT as an alternative to PBT. The correlation $r < 0.3$ means a low positive correlation, $0.3 < r > 0.7$ shows a moderate positive correlation and $r > 0.7$ shows a high positive correlation. To show the significance of the correlation value at alpha of 0.05 (95% confidence level) the P-value should be less than the alpha value (α=.05). This enabled the researcher to deduce the linear relationship of CAT estimates of theta to paper versions of the test.

3.9 Data Management Methods

This research used objective data from candidate scores. The researcher ensured that all the collected data reflected the realities by using a standardized instrument, following data collection procedures, and checking the completeness and consistency of participant's responses in the data collection tool. Data collected was processed to Nominal form. Where 0 represents an incorrect response and 1 represents a correct response. Input files (Data matrix file and item control file) of the calibrated item bank

were created for use in X-calibre 4.2.2 software. Data for the study was stored securely using a password electronically in Google Drive. This is so because electronic storage requires little space, can be easily accessed and is simple to back up. In the event of data sharing and transmitting to other parties, the data would be encrypted. Before data is shared to other institutions the researcher would ensure that confidentiality and ownership agreements are made through a memorandum of understanding. Data quality was key to having authentic and scientific data and therefore the researcher gave utmost significance by recording the data promptly, legibly and accurately (TDR-IR Toolkit, 2023).

3.10 Limitations of the Study

This study was conducted in an area where CAT has never been used. There was an extensive need to explain to participants the concept that by using different sets of calibrated test items, the examinee's ability can be estimated fairly, and this is against the background of conventional tests where the test items are the same for all examinees.

This study assumed that candidates had an above-average computer proficiency level since they were college students. It also did not take into account the user's perception of the transformation from convention to computer-delivered tests and the paradigm of CAT. Furthermore, the validity of CAT is not taken into account.

For Live CAT administration, the item bank of around 30 items could not be sufficient to precisely provide ability estimates of extreme candidates. However, the data used for item bank calibration was from a real test. Research on computerized adaptive testing (CAT) has shown that the number of items required for item bank calibration can vary.

Kalender and Berberoglu (2017), and Kaya (2021) suggest that 30 to 45 items from real testing data can be valid for item bank calibration in CAT, but further research is needed to explore the optimal number of items for different test lengths and conditions.

3.11 Ethical Considerations

Ethics in research deals with making sure that participants or respondents are safe from any harm and are protected from unnecessary stress (Cacciattolo, 2015). The research followed the UNIMAREC guidelines to comply with ethical issues. Participation in the study was voluntary. All participants were required to give informed consent to participate, and they were allowed to withdraw at any point (see Appendix 5). Before giving consent, the researcher explained the aims, methodology, and potential risks of participating in the research.

Privacy and confidentiality of information and participants were taken into account as well. The data collected was used for research purposes only and participants were given random numbers or pseudonyms. Data was not disclosed or shared without the consent or authorization of research participants or the relevant authorities. The rights and preferences of research participants were respected when collecting, storing, using, and sharing their data. Kitchin (2007) provided for respect for autonomy, justice, and beneficence as general principles that guide ethical practice in research involving technology. This was operationalized by obtaining informed consent from participants.

The principle of beneficence requires researchers to evaluate all physical, social, psychological or medical harms or risks that their participants may face by virtue of being

in the project, and make every possible attempt to minimize these harms and maximize safety (Kitchin, 2007). Within the context of research using technology, the risk of harm arises when there is a disclosure of participant's identity or any other sensitive information that may expose them to the risk of embarrassment, reputational damage, or legal prosecution (Townsend & Wallace, 2016). The equipment's used should not in any way cause harm as well. The researcher isolated the eminent risks and ways of managing and avoiding them (see table 1)

Table 1: Research risks and ways of managing them

SN	RISK	WAYS OF MANAGING
1	Hacking of information/Misuse of information by third parties.	Personal data collected was protected using data encryption and passwords.
2	Feeling embarrassed by not being able to operate the computers.	A research assistant (see CV appendix 8) was available to help such participants to operate the computers.
3	Software used damaging the ICT infrastructure.	The software (s) used were obtained from credible suppliers and were uninstalled after the research.
4	Participants being kept for many hours waiting to participate in the study.	The researcher developed a time plan for test administration and informed the participants about the duration of their participation.
5	Psychological discomfort due to variable length of tests to students.	The researcher explained to the participants that the test will be of varied items and the examination ends when a termination criterion is satisfied.

The researcher sought approvals from UNIMAREC, DCE, NCE, MCHS, and CU to conduct the study. The study was within the limits of approved activities only. The activities done were documented and reported to the institutions' authorities.

3.12 Validity and Reliability of the Study

The purpose of establishing reliability and validity in research is essentially to enhance the believability and trustworthiness of the research findings especially if the study is repeated by different investigators under the same conditions or with different research instruments measuring the same construct. The evidence of validity and reliability are prerequisites to assure the integrity and quality of the research process and have a great bearing on the results (Kimberlin & Winterstein, 2008). To ensure this, the research took several measures. Firstly, the research questions were clear and SMART (Specific, Measurable, Attainable, Realistic, and Time-bound). For example, when doing an independent sample t-test the research question being addressed was: what is the significant difference in estimated scores between computer-adaptive tests and paper-based tests? Secondly, the version of the software used was recent and obtained from a reliable source. In addition, a pilot study was conducted to check if the research instruments were providing the required data. The pilot was done at Domasi College of Education, which was was closer to the researcher's base. 50 participants took part in the pilot. Two key issues emerged during the pilot: (1) there was a time delay when the algorithm was selecting the next question and (2) when network was poor the software could log out the candidates. The researcher used this information to advise participants prior to taking the test to be patient when the algorithm was selecting questions. The time

delay was 30 seconds and on the other hand candidates were told to keep their log in credentials with them so that when the system has logged them out they can log in once more to complete the test. Pilot studies play a crucial role in research, particularly in testing the validity and reliability of research instruments (Mocorro, 2017). They also provide a practical understanding of the research domain, allowing for meaningful revisions to the theoretical framework and methodology. The researcher also made sure that the data collection instrument was free of errors when transforming the items from paper to computer. The researcher proofread the items after the conversion. Furthermore, a data audit was conducted by the researcher to check that the data collected was free of error and no omissions were present. In the audit, four cases were dropped for analysis and item bank calibration because some scores were omitted.

3.13 Chapter Summary

This chapter has explained the methodologies employed to get the required data and how it was analysed to satisfactorily test the hypothesis of the study. It first describes the research paradigm, approach and design. Then it explains how the preliminary, comparability and reliability analysis was done as well as data management methods and research dissemination strategies. Lastly, the limitations, ethical issues considered, and measures to ensure reliability and validity were explained. The next chapter outlines the results and discussion of what such findings tell.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Chapter Overview

This chapter presents the findings of the study and their discussion in the quest to respond to the research questions. To begin with, the assumptions of IRT were tested to ascertain if the test functioned independently and unidimensional. After that, a model data fit analysis was conducted to enable the selection of a suitable mode among the IRT models. Subsequently, item parameters were analysed to provide an overview of the quality of test items used in the test and calibrate the item bank. Finally, an analysis was carried out to compare and show a relationship between scores in PBT and CAT to ascertain if CAT is a reliable alternative to PBT.

4.2 Testing IRT Assumptions

4.2.1 Checking for Local Independence.

The assumption of local independence is usually checked by computing x^2 statistic of items at all levels of ability. A 2 x 2 contingent table is then constructed:

		Item 2		
		Incorrect	Correct	Total
Item 1	Incorrect	A	B	A+B
	Correct	C	D	C+D
	Total	A+C	B+D	

Using the formula,

$$x^2 = \frac{N(AD - BC)^2}{(A + B)(B + D)(D + C)(C + A)}$$

We compute x^2 then compare it with the critical value of x^2 (at α=0.05 and one degree of freedom) and when computed x^2 is greater than x^2 critical, the hypothesis is rejected. Thus, local independence does not hold at that level. Then we compute the x^2 statistics at different ability levels and then sum them across the levels. This procedure is rigorous and requires a lot of computations. However, in this study, a KMO and Bartlett's Test were used to check the sampling adequacy and independent functioning (local independence) of items in the data. With a KMO of 0.701, the data was suitable for factor analysis. Kaiser and Rice (1974) stipulated that KMO test values should be greater than 0.6 for an acceptable analysis, greater than 0.7 for a good analysis, greater than 0.8 for a very good analysis, and greater than 0.9 for an excellent analysis. Bartlett's test with a significance value of 0.000 which is less than 0.05 indicated that the variables in the correlation matrix are not interrelated and that they could be used in factor analysis (Bartlett, 1954). This explains that the items in the data set function independently, hence local independence is obtained. Table 2 summarises the KMO and Bartlett's Test

Table 2: Results of KMO and Bartlett's Test

KM0 and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.701
Bartlett's Test of Sphericity	Approx. Chi-Square	1294.040
	df	435
	Sig.	.000

4.2.2 Checking for Unidimensionality

Unidimensionality in item response theory (IRT) refers to the assumption that the latent trait driving item responses is one-dimensional (Junker, 1990). However, the degree of unidimensionality can vary, and this can impact IRT calibration and scoring. Unidimensionality can be checked through eigenvalue plots of the inter-item correlation matrix to determine whether a dominant first factor exists, comparing the eigenvalue plots obtained using real data and simulated data, DIMTEST procedure, or Multidimensional Scaling (MDS) procedures. This study used eigenvalue plots of the inter-item correlation matrix with a Varimax rotation to extract the components. To have a dominant component the first factor should account for more than 20% of variability in the data or the first eigenvalue should be four times larger than the second eigenvalue. Figure 5 shows the scree plot that was obtained.

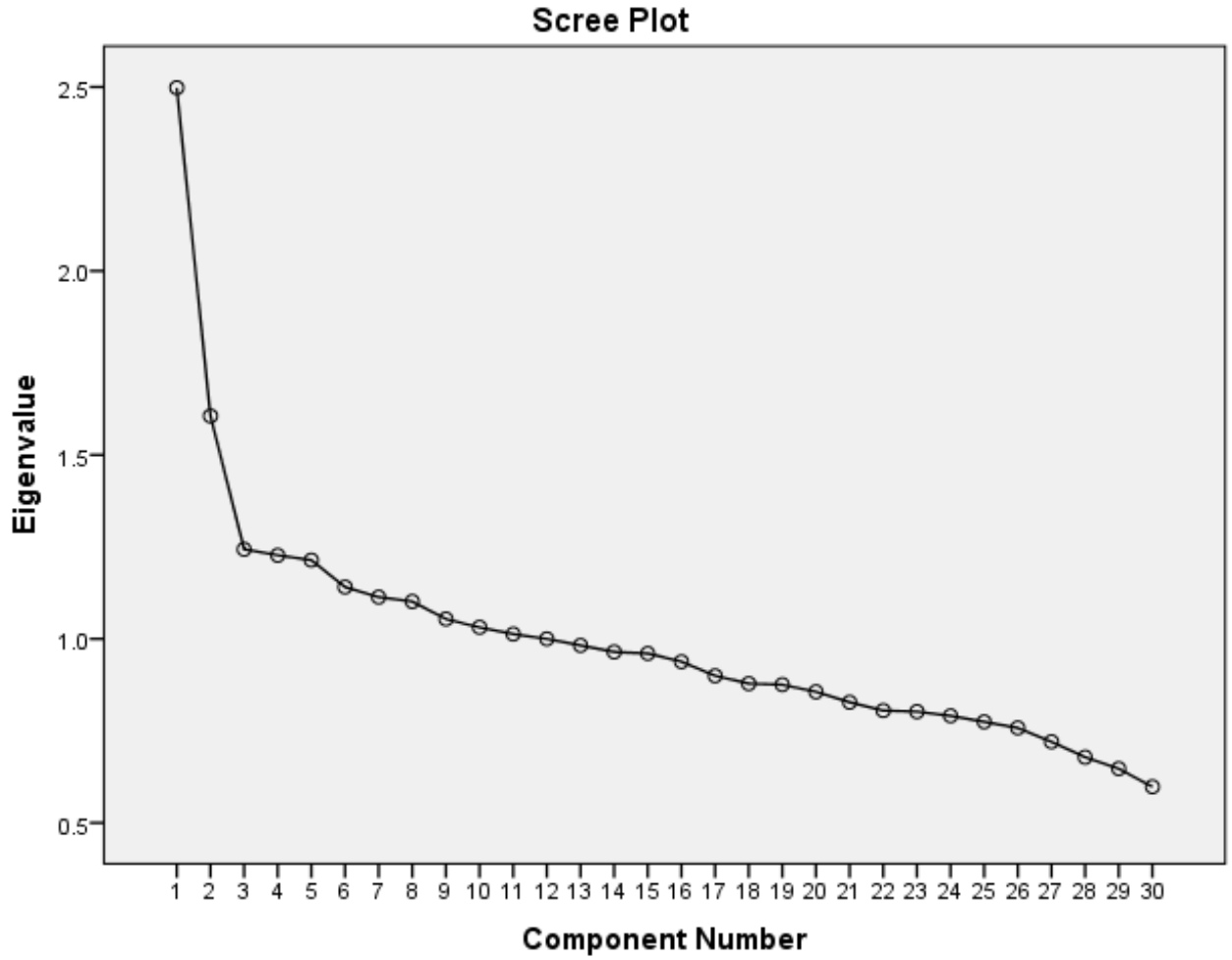


Figure 5: A scree plot showing the extracted Eigenvalues

The Principal component analysis extracted 11 components which explains the variability of examinees. This shows that 11 items provide the differences of examinees and cumulatively provide 47.48% explanation of the test. Component 1 explains 8.327% of the total variability, which suggested a dominant first factor and the assumption of unidimensionality holds. However, it is a weak dominant first factor as the variability is less than 20%. Since this holds then the Local independence assumption holds as well.

Table 3: A summary of the extracted loadings and eigenvalues

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.498	8.327	8.327	2.498	8.327	8.327
2	1.606	5.353	13.680	1.606	5.353	13.680
3	1.243	4.145	17.824	1.243	4.145	17.824
4	1.227	4.091	21.915	1.227	4.091	21.915
5	1.214	4.047	25.962	1.214	4.047	25.962
6	1.141	3.803	29.765	1.141	3.803	29.765
7	1.113	3.711	33.477	1.113	3.711	33.477
8	1.102	3.672	37.149	1.102	3.672	37.149
9	1.054	3.513	40.662	1.054	3.513	40.662
10	1.031	3.438	44.100	1.031	3.438	44.100
11	1.013	3.377	47.477	1.013	3.377	47.477
12	1.000	3.333	50.810			
13	.982	3.274	54.084			
14	.965	3.215	57.299			
15	.960	3.201	60.500			
16	.938	3.127	63.627			
17	.900	2.999	66.626			
18	.878	2.928	69.554			
19	.875	2.918	72.472			
20	.856	2.854	75.326			
21	.828	2.759	78.085			
22	.805	2.683	80.769			
23	.802	2.674	83.443			
24	.791	2.636	86.079			
25	.775	2.582	88.661			
26	.758	2.526	91.187			
27	.720	2.401	93.588			
28	.679	2.262	95.850			
29	.647	2.157	98.007			
30	.598	1.993	100.000			

Table 3 shows a summary of the extracted loadings and eigenvalues. Since the local independence and unidimensionality assumptions were obtained then we can be confident that the test items are independent of each other, such that a response to one item does not depend on the response to another item. Likewise, we can conclude that the test measures one construct; in this case communication skills. In any test, it is important to ensure that test items are often derived from the same construct. This rational method of testing is meant to ensure that all items capture the construct aimed at and only this construct. A test that conforms to the assumption of unidimensionality is believed to be testing a single construct. Since the two assumptions of unidimensionality and local independence hold, we can proceed to apply the item response theory framework in data analysis.

4.2.3 Testing for Monotonicity and Item Invariance

Monotonicity is best displayed on a graph as a curve called the item characteristics curve (ICC), which is assumed to reflect the true relationship between the trait and the responses to the item. For example, in an educational setting, what we see is that as the ability level increases, the probability of getting the item correct increases monotonically. Within a health setting, that would mean that as the ability level increases, the participant is more likely to endorse a higher response option for that item. Monotonicity can be tested using methods like Mokken analysis, spearman rank order correlation and checking Kendaull's tau-b. On the other hand, the assumption of invariance is best understood as the characteristics of the item parameters and latent trait being independent of the sample characteristics within a population. That means, for an item, the item parameters estimated by an IRT model would not change even if the characteristics of the

candidate, such as age or gender, changes. Under IRT, the ability of a candidate under measure does not change due to sample characteristics. Differential Item Functioning (DIF) analysis is often used to evaluate if this assumption if violated.To test for both monotonicity and item invariance, the data must be grouped. This study did not group the data, hence monotonicity and item invariance assumptions were not tested.

4.3 Model Data Fit

4.3.1 Assessing Model Data Fit

IRT has great potential for solving many measurement problems. However, the advantages of IRT models can be obtained only when the fit between the model and the test data of interest is satisfactory. A poorly fitting IRT model will not yield invariant item and ability parameters. Therefore, there was a need to assess the fit of the model to the data before employing a particular model.

Hambleton and Swaminathan (1985) suggested collecting three types of evidence to help one decide which model fits the data: (1) Validity of the assumptions of the model for the data. (2) The extent to which the expected properties of the model (e.g., invariance of parameters) are obtained, and (3) Accuracy of model predictions using real and simulated data. In this study, the assumptions of local independence and undimensionality were obtained, and the entrance examination was not a speeded test hence the validity of the assumptions was obtained. For the second type of evidence: invariance property, this study did not do any analysis to check this evidence because the test data used was not grouped. This property is checked when the data is grouped either by gender,

geographical location, or other type of participant grouping. The third type of evidence was tested using empirical proportions. To check item fit the empirical proportions were superimposed on an ICC. The expectation was that if the predicted ICC follows closely the empirical trace line implied by the proportions, an item is assumed to have a satisfactory fit. The graph below shows the empirical proportion for 1PL, 2PL, and 3PL models respectively.

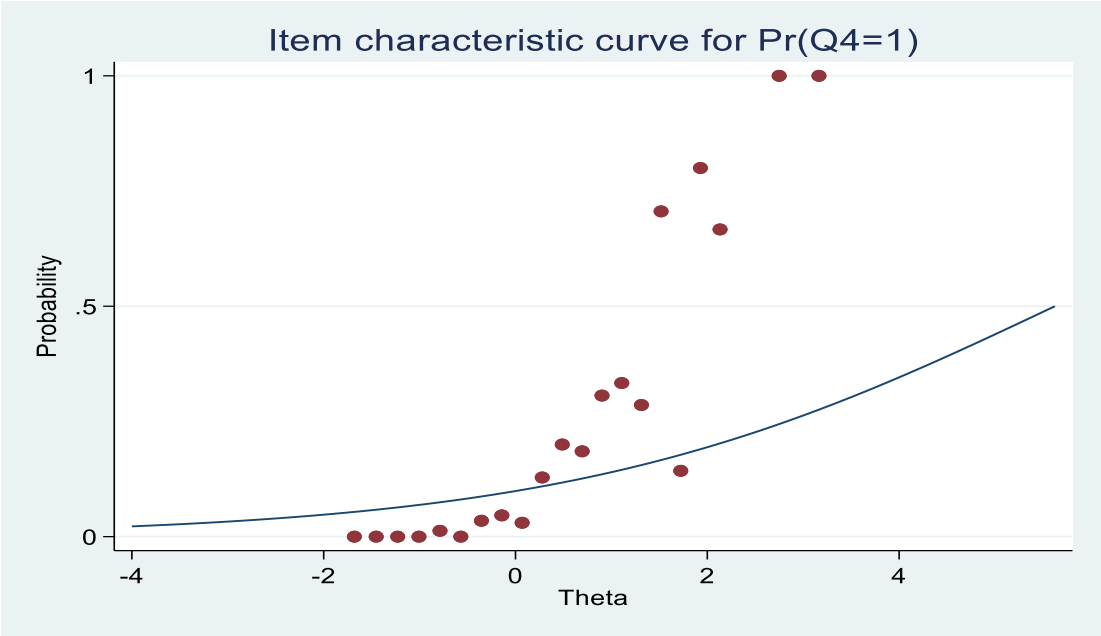


Figure 6: IRT ICC for item 4 in 1-pl model

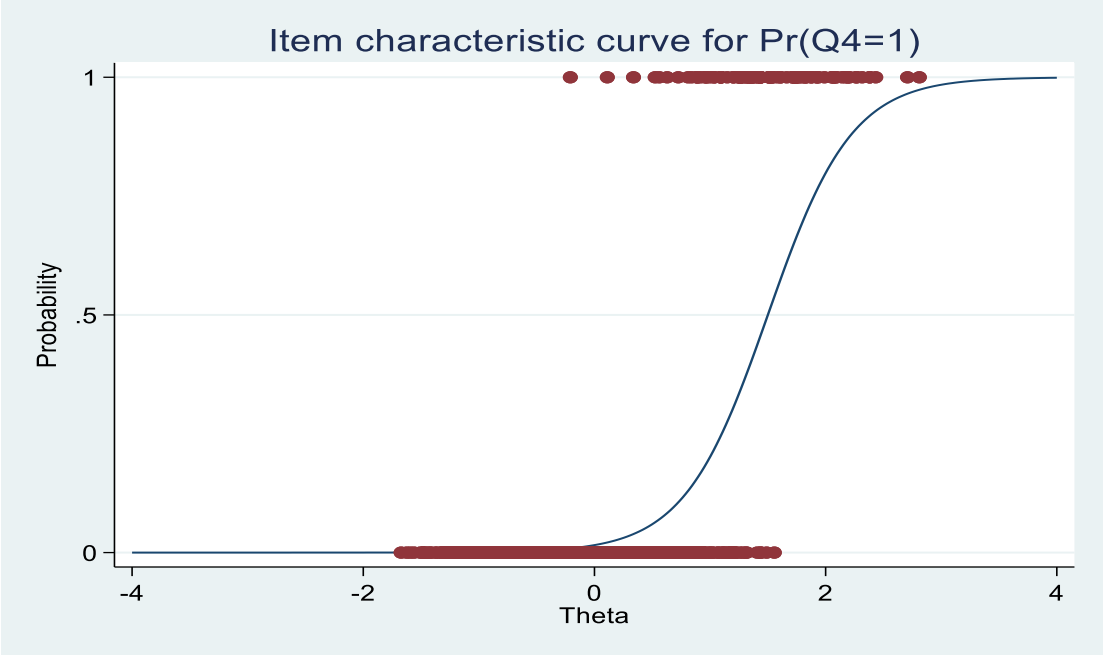


Figure 7: IRT ICC for item 4 in 2-pl model.

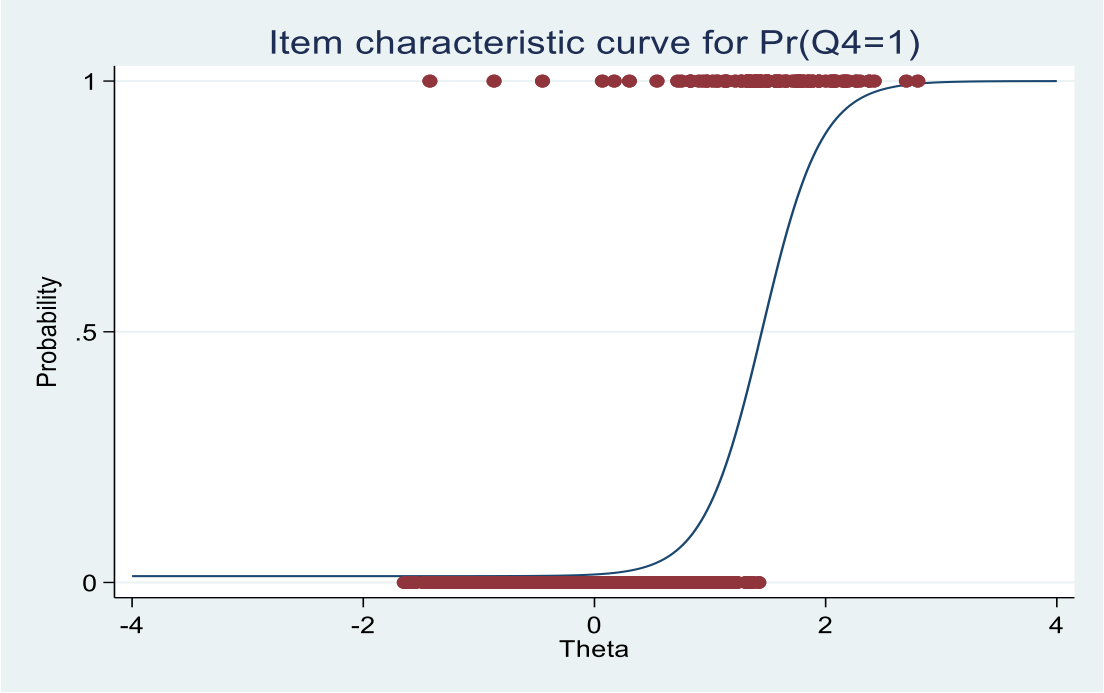


Figure 8: IRT ICC for item 4 in 3-pl model.

4.3.2 Model Data Fit Prediction.

Model-data fit should be verified as a prerequisite to using IRT models. The crucial benefits of IRT models are realized to the degree that the data fit the different models. Model-data fit is a major concern when applying item response theory (IRT) models to real test data. Although there is an argument that the evaluation of fit in IRT modelling has been challenging, the use of item response theory model checking and item fit statistics serve as crucial factors to effective IRT use in psychometrics for information on items and model selections (Essen et al, 2017). Looking at Figures 6, 7 and 8, it is evident that the 3-PL Model (Figure 8) fits the data as the probability against theta plots runs through the trace lines. However, it is worth noting that the fit is not perfect. This study employed the 3-PL Model which fitted the data in its analysis to compute the item parameters for item bank calibration.

4.4 Quality of the Test Items

4.4.1 Item Parameters

The three-parameter logistic model has 3 parameters: discrimination (a), difficulty (b), and guessing (c). The parameters explain the relationship between the examinees' latent traits and the items. They also explain the quality of the items based on the test purpose. Scholars have classified the ranges of the parameter values to give meaning and set boundaries for the selection of items (Hambleton & Swaminathan, 1985; Adedoyin & Mokobi, 2013: Baker, 2001). This study used Baker’s item parameters classification framework to categorize the items that made up the 2022 DCE PBT as shown in the table

4. To obtain the parameters, a logistic approximation of the parameter estimates was run using a maximum likelihood estimation procedure with 40 quadrature points and 100 iteration loops using X-calibre software. Table 5 shows the parameter values for the communication skills paper delivered through Paper and Pencil.

Table 4: Baker's Classification of Discrimination and Difficulty parameter.

Discrimination Parameter Classification				Difficulty Parameter Classification		
Verbal Label	Range of Values			Verbal Label	Range of Values	
None	0			Very easy	below -2	
Very low	0.01	0.34		Easy	-2	-0.49
Low	0.35	0.64		Medium	-0.5	0.5
Moderate	0.65	1.34		Hard	0.51	2
High	1.35	1.69		Very hard	Above 2	
Very High	above 1.7					
Perfect	Infinity					

Table 5: Item parameter for 2022 DCE and NCE communication skills paper.

Item ID	Max Info	Theta at Max	Discrimination (a)	Difficulty (b)	Guessing (c)
Q1	0.007	2.64	0.1251	1.1617	0.2522
Q2	0.0664	3.28	0.3683	2.8469	0.2035
Q3	0.2304	1	0.7824	0.7225	0.336
Q4	2.7565	1.56	2.2178	1.5034	0.1319
Q5	0.6727	2.24	1.2697	2.0788	0.285
Q6	0.1584	1.1	0.6066	0.7811	0.2693
Q7	1.9168	1.24	1.9503	1.1591	0.1879
Q8	1.194	1.96	1.5619	1.854	0.2032
Q9	1.5797	2.7	1.6938	2.6343	0.1412
Q10	1.2296	0.96	1.6351	0.8445	0.2354
Q11	1.4158	1.62	1.6476	1.5274	0.1698
Q12	0.6304	0.78	1.2015	0.6285	0.262
Q13	0.0623	-0.4	0.3768	-0.9096	0.26
Q14	0.5058	0.86	1.076	0.691	0.2617
Q15	1.0721	3.06	1.4302	2.9692	0.1672
Q16	0.1039	0.02	0.4872	-0.3709	0.2607
Q17	0.0457	2.34	0.332	1.7347	0.2884
Q18	0.1118	1.46	0.5094	1.0907	0.2687
Q19	0.0947	0.12	0.464	-0.2935	0.2581
Q20	0.4747	1.04	1.0369	0.8517	0.2564
Q21	1.0328	2.56	1.4429	2.4439	0.196
Q22	0.9171	2.22	1.3417	2.104	0.1822
Q23	1.3176	3.12	1.5128	3.0534	0.1177
Q24	0.8097	2.18	1.321	2.0391	0.2308
Q25	0.1536	1.56	0.5943	1.2487	0.2642
Q26	0.268	1.72	0.8043	1.465	0.2885
Q27	1.6761	2.06	1.7326	1.9839	0.1338
Q28	0.0538	-0.1	0.351	-0.6422	0.2624
Q29	1.3929	2.88	1.5743	2.807	0.1305
Q30	0.9512	1.88	1.3575	1.7723	0.1754

4.4.2 Classification of item discrimination parameter

The test items had discrimination values ranging from 0.1251 for Q1 to 2.2178 for Q4. No item was classified as non-discriminatory, Two items (Q1 and Q17) provided very low discrimination, Eight items (Q28, Q2, Q13, Q19, Q16, Q18, Q25, Q6) provided low discrimination, another eight items (Q3, Q26, Q20, Q14, Q12, Q5, Q24, Q22) provided moderate discrimination, nine items (Q30, Q15, Q21, Q23, Q8, Q29, Q10, Q11, Q9) provided high discrimination, and three items (Q27, Q7, Q4) provided very high discrimination. Figure 9 shows a pie chart of the percentage classification of the items based on discrimination parameter.

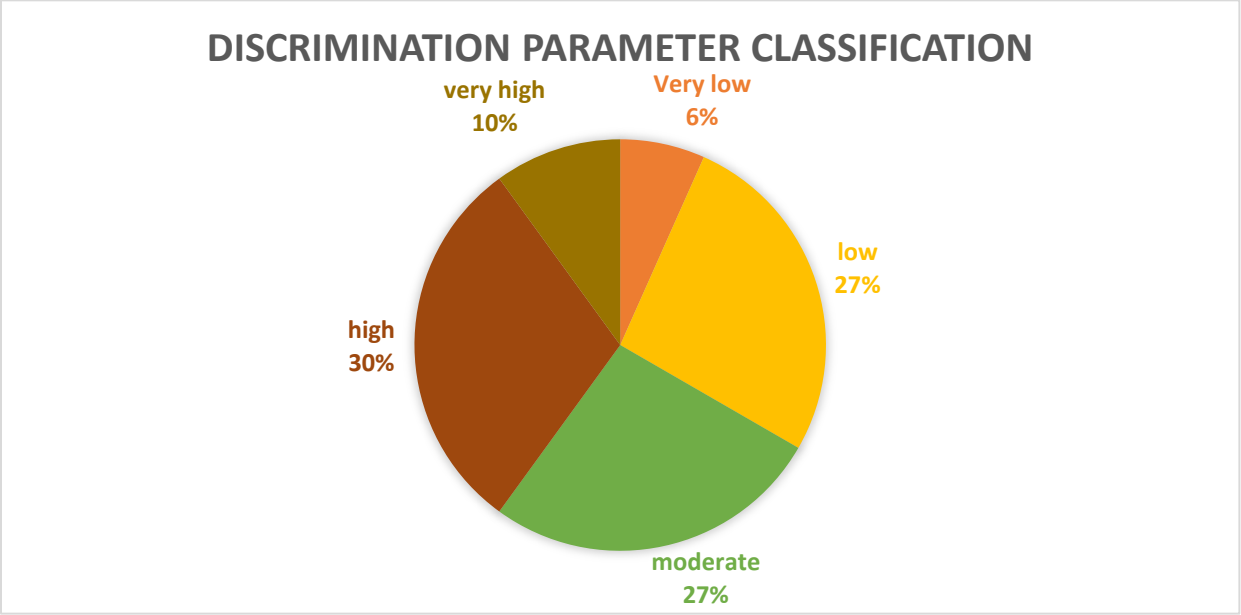


Figure 9: Pie chart of percentage classification of the items based on discrimination parameter

The test was for selection purposes into the first year of the college, hence only items with moderate values and above were supposed to be part of the test. Thus, only 20 (66.6%) of the items in the 2022 DCE communication skills paper could discriminate appropriately

between candidates with low ability and those with the required ability to succeed in college education. Item discrimination parameter is an important tool in standardized tests to ensure that questions are varied enough to discriminate between high-ability and low-ability examinees. It helps test developers in selecting the items that will make up the test. With items that discriminate the candidates appropriately based on their abilities, there will be consistency in the measurement of students' abilities. Likewise, understanding and accurately estimating the discrimination parameter is crucial in the development, calibration, and interpretation of tests based on Item Response Theory. It provides valuable insights into the quality and effectiveness of test items in measuring the intended construct or trait. Based on the results of this study, the test developer could be compelled to select only 20 items whose discrimination parameter is above 0.65, it conversely means that the other 10 items require replacing or rephrasing since they have weaker discrimination, indicating that the item may not effectively discriminate between individuals with different trait levels.

4.4.3 Classification of item difficulty parameter

The test items had difficulty levels ranging from -0.9096 for Q13 to 3.0534 for Q23. There was no item which was very easy in terms of difficulty level, two items (Q13, Q28) were easy, another two items (Q16, Q19) were of medium difficulty, seventeen items (Q12, Q14, Q3, Q6, Q10, Q20, Q18, Q7, Q1, Q25, Q26, Q4, Q11, Q17, Q30, Q8, Q27) were hard, nine items (Q24, Q5, Q22, Q21, Q9, Q29, Q2, Q15, Q23) were very hard. Overall, the test was comprised of difficult items. Figure 10 is a pie chart showing percentage distribution of items according to difficulty level. The pie chart explains that 37% of the

test items (easy + very hard) were not of appropriate quality. Such items should not form part of a test. 63% of the test items (medium + hard) are of good quality and appropriate difficulty level. Item difficulty parameter estimation is a crucial aspect of item response theory (IRT) and it shows how difficult the item is, or the construct level at which we would expect examinees to have a probability of 0.50 (assuming no guessing) of providing the correct response to the item. Item difficulty parameter enables the test developer to predict how examinees would fare on different items. Since every test has a purpose, then the test developer must choose items appropriate to the purpose. Bulut (2015) found that higher difficulty levels of the items and higher omitted response rates affect the estimation of guessing parameter as well as the selection of students for Graduate Studies. It thus corresponds with the findings of this study that items with high-difficulty parameter should not be part of the test for selecting students for college. Using such items increases guessing and cases of omitting responses. The difficulty level parameter is also important in computer adaptive tests as it allows for the classification of items and enables the algorithm in the selection of items. Based on the difficulty parameter values, we can probably say that some candidates provided guessed responses in the 9 very hard items and that all examinees provided correct responses in 2 easy items. This diminishes the reliability of the test and scores obtained in such tests may not be consistent.

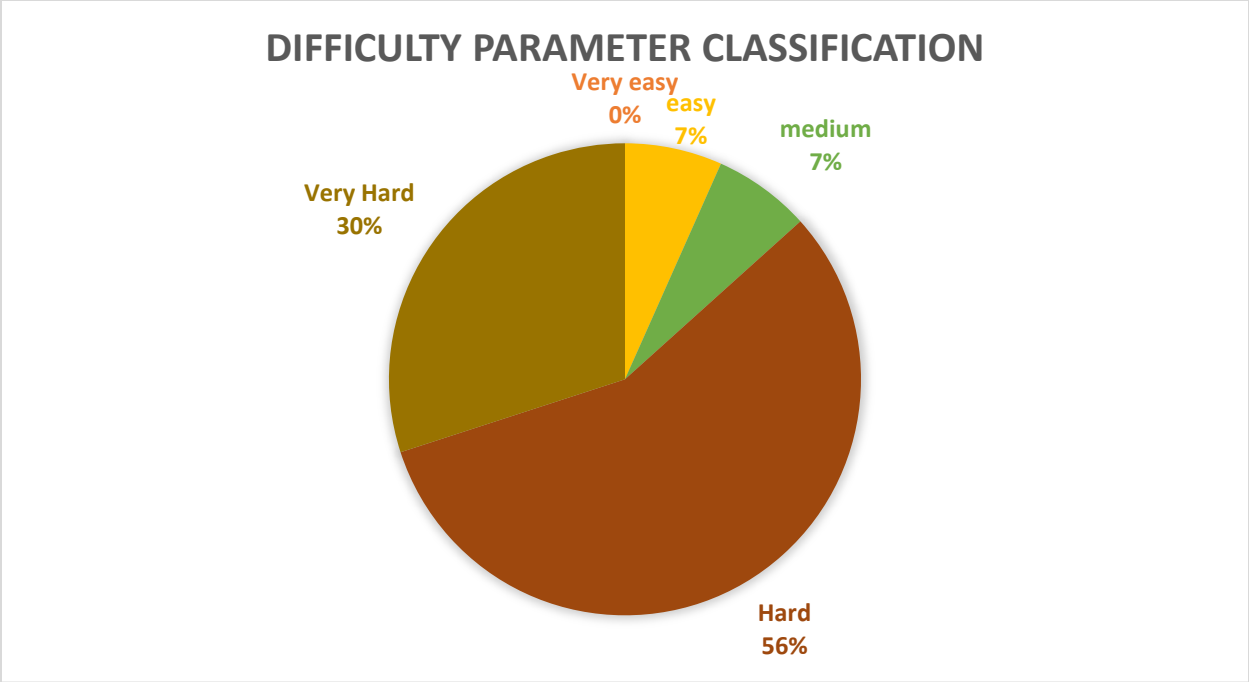


Figure 10: Pie chart of percentage distribution of items according to difficulty level.

4.4.4 Item Guessing Parameter.

The three-parameter model incorporates the possibility of guessing. This parameter expresses the probability that an examinee with low ability can be able to get an item correctly and, therefore, has a greater-than-zero probability of answering an item correctly in a test. The guessing parameter c is the lowest value that an ICC attains. The results show that guessing parameter ranged from 0.1177 for Q23 to 0.336 for Q3 with a mean of 0.22267. This explains that on average an examinee had 22.27% of getting a correct response through guessing. With dichotomous items in a paper-based test, this parameter is important as examinees resort to guesswork when they do not have an idea. Guessing parameter is a critical factor in various fields, including test reliability, and parameter estimation. It is worth noting, that the value of *c* does not vary as a function of the trait/ability level, i.e. examinees with high and low ability levels have the same probability

of responding correctly by guessing. Theoretically, the guessing parameter ranges between 0 and 1, but practically values above 0.35 are considered unacceptable, hence the range $0 < c < 0.35$ is applied. A value higher than $1/k$, where k is the number of options, often indicates that a distractor is not performing. In this case with an average guessing parameter of 0.22267, we can conclude that the guessing is below the unacceptable level.

A critical analysis of the items shows that only four items (Q8, Q10, Q11, Q30) could be included in the test if the inclusion criteria were that an item should satisfy both item discrimination and item difficulty appropriate levels.

4.4.5 Item Characteristics and Test Characteristics.

Item characteristic curve graphically depicts the relationship between an examinee's ability and the probability of a correct response to test items. Hence, the higher the individual's ability, the higher the probability of a correct response. Figure 11 shows the ICCs for all the items in this study.

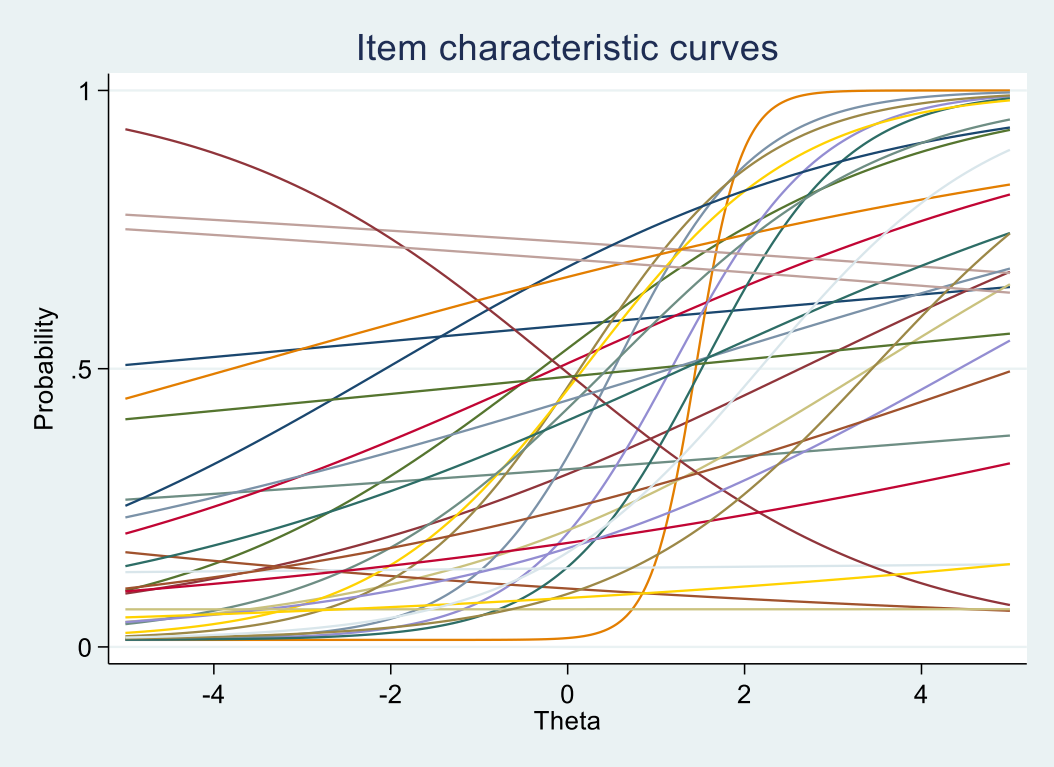


Figure 11: ICC for all the items in the study.

The probability of providing a correct response to items concentrates from theta above zero. This explains that candidates with theta below 0 have a low probability of providing correct responses. In general, we can say the items were of higher difficulty level. Examinees with average and low ability could not provide correct responses to the majority of the items. To specifically illustrate this assertion figure 12 shows the ICCs of five items (Q4, Q12, Q16, Q22, and Q30). The items cover a wide difficulty level spectrum from -1.99 to 4.5. This explains that item 22 is more difficult than other items. Hence an examinee with a theta of zero will be expected to correctly respond to item 16 and provide incorrect responses to the other items.

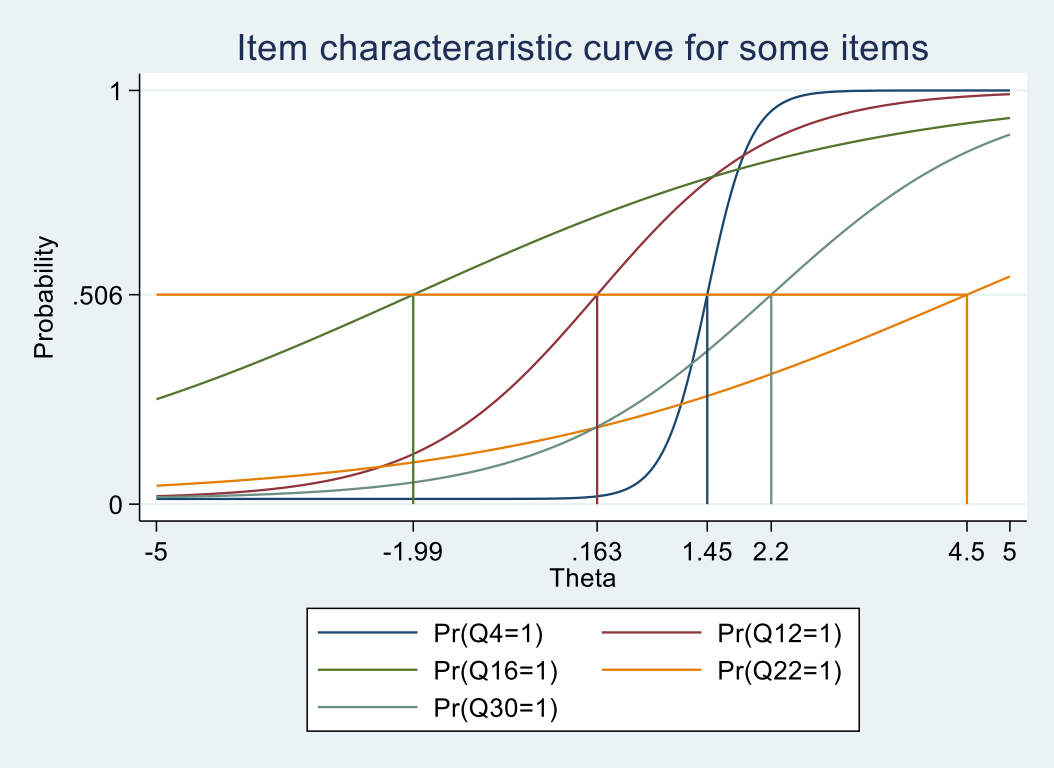


Figure 12: ICC of Items 4, 12, 16, 22 and 30 and the corresponding difficulty level values.

The sum of the ICCs gives us the expected score on the whole test called a test characteristic curve (TCC). A person with an ability level ($\theta = 0$) could possibly respond correctly to 10 items. Likewise, it is observed that to correctly respond to half of the total items it requires an examinee with ability level ($\theta \geq 2$). It is worth noting that within the range of -4 to +4 ability estimates the candidates obtained scores of 6 to 18. A quality test, which follows the normal distribution should have equal magnitude differences from the half mark which is 15 in this case. Figure 13 show the test characteristic curve.

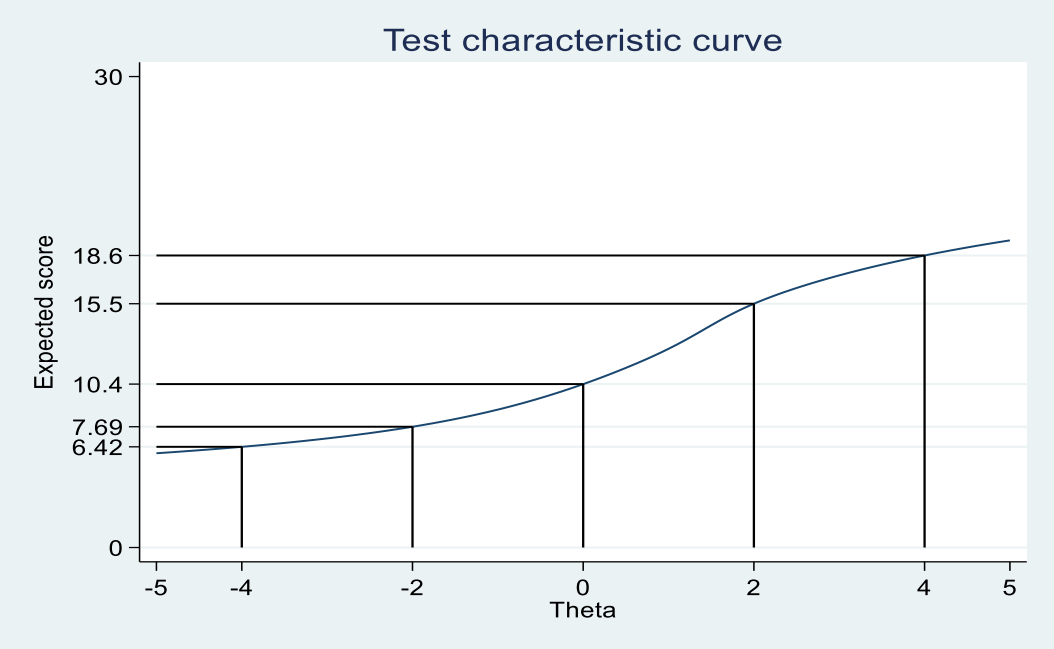


Figure 13: TCC for expected scores at different ability levels.

4.4.6 Item Information and Test Information.

The information function of an item for a given ability level can be defined as the proportion of the square root of the differentiation of item characteristics to its variance (Hambleton et al, 1991). The IIF tells us how individuals, in terms of ability, are distinguished best by the items. In this study, it is observed that most items except for one, had little information (Figure 14)

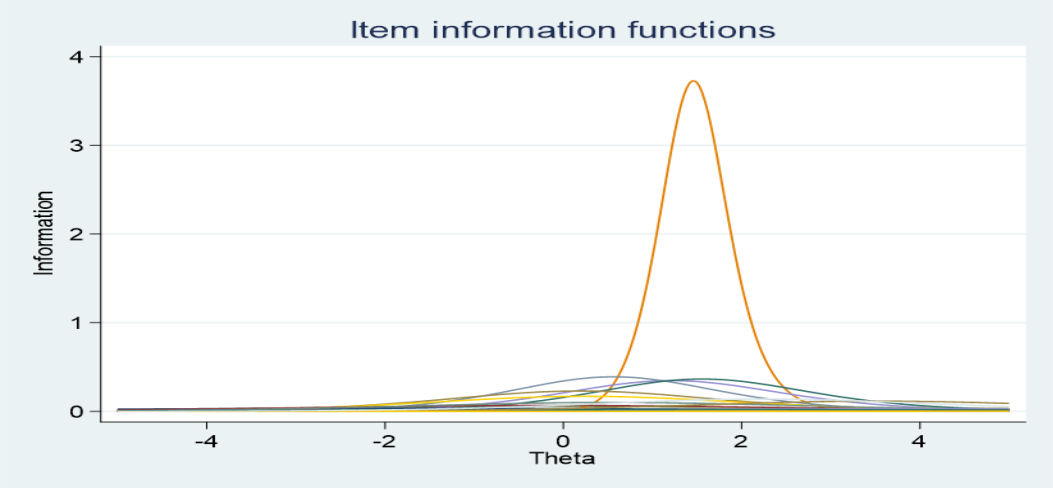


Figure 14: Item information functions for all items

While an information function can be obtained for each item in a test, the amount of information yielded by each item is rather small, and mostly, the examinee's abilities are not estimated with a single item. Consequently, the amount of test information at an ability level and the test information function are of primary interest. Alan and Yen (1979) explained that when the slope is steep and the variance is low, the information function would be larger; however, when the slope is not that steep and the variance is great, the information function would be less.

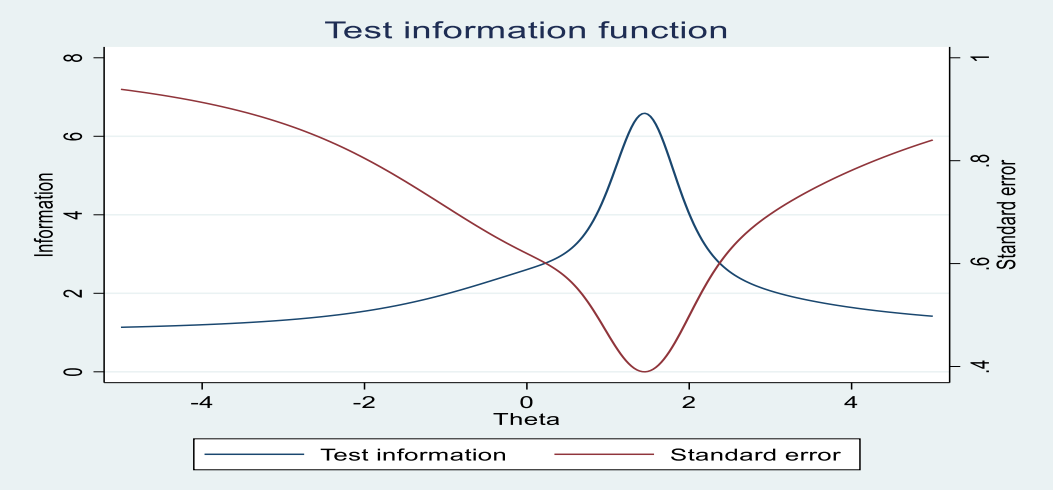


Figure 15: TIF curve and SE curve.

The Test Information Function shows the maximum amount of information that was available for the 2022 DCE communication skills paper. From the curve above, we observe that the test explained 6 to 8 items and peaked at theta (θ) = 1.75 with a standard error of 0.755. This explained that the test would be best for estimating the ability of examinees whose abilities were 1.0 to 2.5. This could inform the institution in deciding whether to administer the test or not depending on the purpose of the test. It also shows that the information obtained in the test could be achieved with only 6 to 8 items as opposed to 30 items.

4.5 Calibrated Item Bank and Live CAT Administrations

Item bank calibration was the first step in Live CAT administration. The common software for CAT administration is FastTest, Concerto, CATIRT, and CAT Korea. Some of these are commercial whilst others are open-source software. This study used CAT Korea to administer the Live CAT. A request was made through email at sales@thecatkorea.com to use the software.

To calibrate the item bank, firstly, items were registered in the system. The category of the Test is cognitive, with multiple choice questions, having 4 responses with 1 correct answer. Secondly, an item bank was registered with the 30 items having a theta range of -5 to +5. The item bank used a 3PL dichotomous IRT model with a maximum information of 3.989. The difficulty parameter ranged from -0.91 to 3.05 with the discrimination parameter ranging from 0.13 to 2.22. Figure 16 shows the aspects of the item bank.

Thereafter, the test was registered. The starting rule was a random theta to minimize the exposure of the item. The test used the maximum likelihood procedure to estimate the ability of examinees. Item selection was based on the maximum information of the items remaining in the item bank. The termination rule employed in this test was a standard error (SE) of measurement and fixed test length. A SE of 0.35 and a test length of 20 items was used. CAT Korea has three options for terminating the test: SE, test length or a combination of SE and test length. This study used a combination of SE and test length in unison with Kalender & Berberoglu (2017) and Kaya (2021). For test length, the developer has liberty to input the specific number of items whilst for SE there are pre-set three values SE of 0.3, 0.35 and 0.4. This study used the moderate value of 0.35 in the quest to measure candidates with a broad theta continuum. The scores were published publicly, and examinees were able to see the results after completion of the test. The scores reported were theta values and t-score conversion. A list of candidates, with identification and passwords was uploaded in the system for identification of test takers. Pseudo names (Malawian candidate 1 to Malawian candidate 1200) were used with identification numbers 300990001 to 300991200 and a password 1234 as log in credentials. The platform then generated a URL: https://medmevcat.livecat-assessment.com/taker/taker-login/22/1813 to be used for test administration. The link took test takers to the login page (figure 17) to enter their credentials ready for examination

Question area		Number of questions	Difficulty distribution range	Discrimination distribution range	Average question frequency
Main Category	Middle classification				
Language	communication skills	30	−0.91 ~ 3.05	0.13 ~ 2.22	–

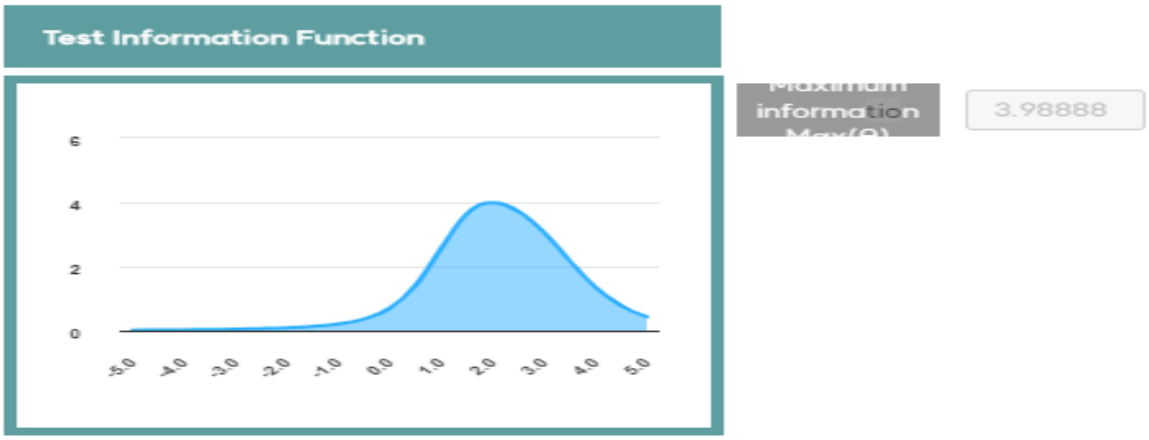


Figure 16: Item bank information.

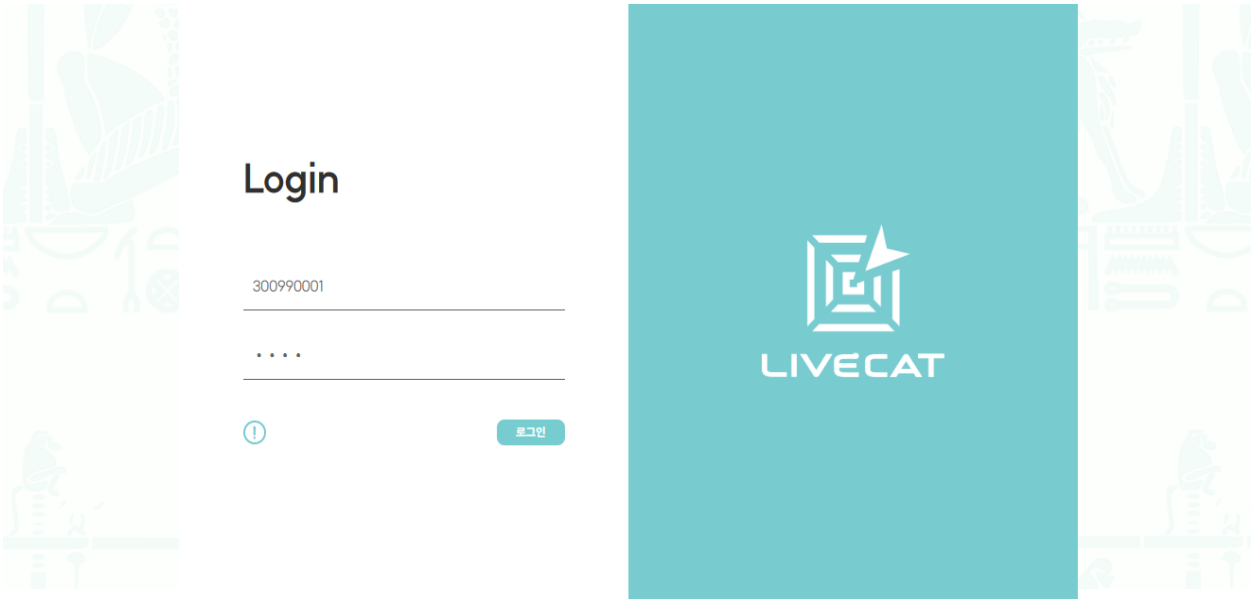


Figure 17: CATKOREA Login page.

The CAT platform is in Korean for item registration, item bank registration, and test registration but when using Chrome browser there is an option for translation to English. Nevertheless, the delivery of the test supports English. Examinees were told to click on correct response among the options 1 to 4 and when they are certain they should click the blue enter button to submit their response. Thereafter they were required to wait for some few seconds as the algorithm would be selecting the next question. It was observed that examinees were confident in using desktop computers, Laptops, and smartphones. However, the internet network was at times, an issue. Another observation during the Live CAT administration was that there were at times 30 seconds delay to provide the next question where the platform was estimating the ability of candidates to select the next appropriate item. These logistical problems did not affect the results as the candidates were told how the platform works prior to taking the examination. Most examinees were excited to understand that using a calibrated item bank, examinees can write different items but the measurement of their ability is comparable, as well as the immediate score reports after writing the examinations.

4.6 Comparability Analysis of PBT and CAT

PBT and CAT were compared in terms of test time, frequency of items administered, mean values of theta, and standard error of measurement. The methodological approaches of PBT and CAT in terms of test time are different. PBT provides a fixed time for a test whilst CAT can may or may not prescribe time. In practice, many CATs include a function of time per item whilst in literature time is of no essence. The PBT administrations provided examinees 30 minutes to complete the 30 test items, averaging

one minute per item and when the time elapsed the candidates were told to stop writing. In contrast, the CAT administration does not end the test when time elapses. Candidates were free to respond at their own pace since precision of measurement is the utmost goal in CAT other than speed in responding. However, it was found that on average examinees answered 20 items in 14 minutes. Thus, on average an examinee used 42 seconds to respond to a question. The time aspect is informed by Cloe et al (2017) who expressed that despite common operationalization in CAT, measurement efficiency of computerized adaptive testing should not only be assessed in terms of the number of items administered but also the time it takes to complete the test. To this end, their study introduced a novel item selection criterion that maximizes Fisher information per unit of expected response time (RT), which was shown to effectively reduce the average completion time for a fixed-length test with minimal decrease in the accuracy of ability estimation. The results affirm that computerized adaptive testing (CAT) offers the potential for significantly reduced test times, making it an attractive option for large-scale testing programs. This efficiency is achieved through the use of item response theory (IRT) which targets item difficulty to examinee proficiency, maximizing information in the estimation of proficiency

PBT requires candidates to attempt all questions whilst CAT provides the candidates with varying test lengths and item administration patterns. The correct response in PBT is used for calculation of total score whilst the correct response for CAT is used to determine the next question to be administered. These two contrasting methodologies necessitate that we compare the item administration frequency (Exposure rate) and correct response

frequency for the test items. In terms of item exposure rate in PBT, all examinees were exposed to all the questions representing a 100% exposure rate of the items whereas in CAT the exposure rate ranged from 28% for question 1 to 94% for question 10 with a mean of 64%. This explains that some items in the test were not necessary for some examinees as they provided the same information about the candidates. The item exposure rate is a security feature and from the results of this study, CAT provides security for test items by not exposing about 36% of the items to all examinees.

This study also compared the percentage of providing correct responses for the items. Table 6 summarises the percentage of correct responses per item for both CAT and PBT, going on to provide the number of items in the same range for both CAT and PBT.

Table 6: Percentage of correct responses per item.

% range of correct responses per item	CAT	PBT	Items in the same range for both CAT & PBT
0 - 25	**8 items**(Q23, Q28, Q15, Q16, Q27, Q17, Q19, Q29)	**13 items** (Q23, Q29, Q4, Q9, Q27, Q15, Q22, Q21, Q30, Q11, Q8, Q24, Q7)	4
26 -50	**13 items** (Q30, Q21, Q18, Q4, Q25, Q22, Q26, Q24, Q3, Q10, Q1, Q2, Q13)	**10 items** (Q2, Q5, Q10, Q26, Q20, Q25, Q14, Q12, Q18, Q17)	5
51 -75	**7 items** (Q7, Q14, Q5, Q20, Q12, Q6, Q8)	**7 items** (Q6, Q3, Q1, Q19, Q16, Q28, Q13)	1
76 - 100	**2 items** (Q12, Q9)	**No item**	0

The results show that candidates provide more correct responses in CAT than in PBT. This is because the program selects appropriate questions for every candidate based on their performance in the previous item. If a candidate provides a correct response the algorithm will select a slightly more challenging question. If a candidate provides an incorrect response, the algorithm will select an easier question. The results also show that the algorithm worked appropriately in this study and that the results found could be relied upon. Having questions tailored to a candidate's ability relieves stress and increases the motivation to write examinations (Kalender & Berberoglu, 2017). Likewise, the instances of no correct response in the test are minimal for CAT since the item bank has items with a wide range of abilities. However, the number of correct responses does not signify competence as the ability of candidates is based on the difficulty level of the item correctly responded to. This then supports that precision of measurement ability can be achieved with CAT.

Lastly, the study compared the values of theta and standard error of measurement for PBT and CAT to determine if the theta values of the two test delivery methods are comparable, such that they can be used interchangeably or if there is a conversion from PBT to CAT we could be certain that the estimates will be equivalent. PBT provided theta values ranging from -7.00 to +2.874, with an average of -0.327. CAT provided theta values ranging from -5.00 to +2.97, with an average of 0.047. The PBT provided an extremely low-ability estimate. A theta of -7.00 is way below the usual range of -3 to +3. Hence that value is regarded as an outlier and not counted for. Thus, from these values, the latent trait seems to be within the same range between the two testing methods whilst the means seem to be far from each other. Nevertheless, an independent t-test was

conducted to compare the theta means for examinees who sat for PBT and those for CAT versions of the test. The null hypothesis for this test was that there is no significant difference between the mean values of theta for the two test delivery modes. Theta values for PBT were grouped as 1 and theta values for CAT were grouped as 2. Table 7 summarises the results of an independent sample t-test.

Table 7: Results of independent sample T-Test.

Independent Samples T-Test												
				Levene's Test		t-test for Equality of Means						
	Mean		SD	F	Sig.	T	df	Sig.	Mean Diff.	SE Diff.	95% CI	
											Lower	Upper
Theta	PBT	-0.327	1.72	17.18	.000	-4.488	1049	.000	-.3701	.0825	-.5319	-.2083
	CAT	0.047	1.18			-4.993	1467	.000	-.3701	.0741	-.5155	-.2247

The results show that there was a significant difference (t (1049) = -4.488, p=.000) in the theta value with mean theta values of PBT (M = -0.327, SD=1.72) lower than mean values of CAT (M=0.047, SD=1.18). The magnitude of the differences in the mean (mean difference = -0.3701 at 95% CI: -1.5085 to -0.4813) was significant. This explains that PBT and CAT are not comparable in terms of mean values of theta. It also shows that the values obtained for CAT are higher than those obtained in PBT. This indicates that CAT precisely estimates the examinee's ability within the test's maximum information range as compared to PBT. The Standard Error for PBT was found to be 0.755 whilst CAT used a fixed standard error of 0.35. SE is directly related to the reliability of a test; that is, the larger the SE, the lower the reliability of the test and the less precision there is

in the measures taken and scores obtained. This study thus shows that in PBT the precision of measurement was low as compared to the CAT delivery mode. However, even though the SE of 0.35 is commonly used in practice the test did not terminate for any examinee using standard error as the termination rule. All candidates attempted 20 items in CAT. The results in this study are similar to what Cikrikci et al (2018) found in their study on the development of a computerized adaptive version of the Turkish driving license exam. They found theta range of -3.00 to +2.96 with a mean of 0.38 and a standard error of measurement of 0.35 to 0.56. This was explained as conforming with the theoretical knowledge suggested and accepted in the literature, even though the CAT application with a fixed number of questions was used (Embretson & Reise, 2000; Linacre, 2006). Standard error of measurement is related to the reliability of a test. When standard error is minimal the reliability is achieved. Hence with a standard error of 0.35 CAT could be seen as an alternative to PBT.

4.7 Correlation Analysis of PBT and CAT

Correlation is an important aspect when decisions to transition from paper-based test to computer adaptive test are being made or where the two testing methods are to be used interchangeably. The study checked the relationship of the theta values using Pearson moment correlation. A Pearson moment correlation statistic is a measure of the strength of a linear relationship between paired data. The value of the correlation coefficient is from 0 to 1. Evans (1996) classified the correlation values in the ranges as 0.0 to 0.19 = Very weak, 0.20 to 0.39 = weak, 0.40 to 0.59 = moderate, 0.60 to 0.79 = strong, and 0.80 to 1.0 = very strong.

This analysis first grouped the examinees to come up with the frequency distribution table with 22 points from theta -4 to 4 with a point taking a range of .4 (like 0.0 to 0.4) (Appendix 6). This was done to come up with bivariate linearly related data since the sample size of the data sets were different and we could not perform a correlation analysis with the raw theta values. Thereafter, a Pearson moment correlation was computed to understand the linear relationship in theta values for the Paper-based mode to those in computer adaptive mode. Table 8 summarises the correlation statistic and the significance of the relationship.

Table 8: Results of Pearson Moment Correlation

		Frequency of examinees (CAT)	Frequency of examinees (PBT)
Frequency of examinees (CAT)	Pearson Correlation	1	0.717
	Sig. (2-tailed)		.000
	N	22	22
Frequency of examinees (PBT)	Pearson Correlation	0.717	1
	Sig. (2-tailed)	.000	
	N	22	22

A Pearson moment correlation statistic of 0.717 explains that there is a strong positive relationship in the ability estimations of candidates in PBT and CAT. This confirms that if we change the test delivery mode from a Paper-based test to a computer-based test or use the two test delivery modes interchangeably we will certainly get reliable estimates in 71.7% of the cases. The P-value of .000 for a 2-tailed analysis provides evidence that the relationship is significant. The results are consistent with the findings of Kalender and Berberoglu (2017) where they found the correlations of CAT's ability estimations with

the mathematics subtest scores of the full PBT versions to be 0.83, 0.68, and 0.77 for public, Anatolian, and private high schools. They discussed such results as supporting the use of CAT in the admission system since it seems to serve a similar function as the mathematics test on the PBT version. Similarly, it is in tandem with the findings of Oz and Ozturan (2018) in their study on whether the test administration mode influences the reliability and validity of achievement tests, computer-based or paper-based testing. It was deduced that at a correlation of 0.84 there was a statistically significant relation between the test administration modes. This study thus adds to the theoretical knowledge that CAT and PBT scores in college admission examinations are strongly correlated.

4.8 Chapter Summary

This chapter has extensively highlighted the results found after numerous analyses, provided interpretations, and decisions that could be made. The data analysis has responded clearly to the research questions. For research question 1, what is the quality of the communication skills entrance examination paper? Analysis of item parameters and test information function has shown that the 2022 DCE communication skills paper which was used in this study was of moderate quality. This implies that the test was best suited for candidates with moderate ability levels. This test could fail to precisely estimate ability estimates of candidates with low ability and those with high ability. Likewise, Only 6 to 8 items which could be selected based on the information they provide about the candidate could be used instead of administering all 30 items. For research question 2, how comparable is the frequency of correct responses to items in CAT and PBT? Data analysis shows that candidates provide more correct responses in CAT than in PBT. This is inherent in the methodology of CAT as it selects appropriate

questions for every candidate based on their performance in the previous item. This will enable optimal measurement of candidates' ability rather than ability items which are not of their ability level. CAT takes the test to the comfort of a candidate's ability. For research question 3, how comparable are the candidate's ability measurements in CAT to PBT? Data analysis explains that PBT and CAT are not comparable in terms of mean values of theta. It also shows that the values obtained for CAT are higher than those obtained in PBT. This indicates that CAT precisely estimates the examinee's ability within the test's maximum information range as compared to PBT. For research question 4, what is the relationship between candidates' scores in CAT and PBT? Data analysis shows that scores obtained from CAT and PBT are strongly correlated. This is an indication that the two modes can be used interchangeably. The biggest practice should be pre-testing of items so that the items administered should have known parameters. The next chapter makes the conclusions based on the findings and provides recommendations for practice and further research.

CHAPTER 5

CONCLUSION, IMPLICATIONS AND RECOMMENDATIONS

5.1 Chapter Overview

After a thorough study, this chapter presents the research conclusion. The implications of the findings are also highlighted, providing strong insights into the reliability of paper-based and computer-based test delivery methods. The conclusion is based on the test's quality as measured by item parameters, comparability, and reliability analysis. Following the conclusion and implications, recommendations, study addition to knowledge, and potential areas for further research and practice are discussed.

5.2 Conclusion

The research aimed to understand the reliability in the estimation of examinees' ability using computer adaptive testing as an alternative to paper-based tests in communication skills entrance examinations in Malawi. It has been argued throughout this study that comparability and reliability of examinees' latent traits cannot be assumed between the test delivery methods (Wang & Kolen, 2001). Similarly, technological improvements need the modernization of all systems, including assessments and entrance examinations. As a result, this study has addressed methodological concerns in order to assist in deciding on whether to transition from paper-based testing to computer-adaptive testing

or to use the two approaches concurrently. This study has made three conclusions as follows:

The test items for the 2022 DCE communication skills entrance examination paper were of moderate quality. The test comprised 63% items with appropriate difficulty levels and 66.6% items with appropriate discrimination levels with a 22.7% possibility of correct response through guessing. Further to this, the test information shows that the selection of students into college could be achieved with only 6 to 8 items as opposed to the 30 items. This reveals that a crucial stage in the assessment process is the pre-testing of items to determine their parameters. For the test's purpose to be effectively fulfilled, it must have items with acceptable levels of difficulty and discrimination. Likewise, the test assembly should include items that offer sufficient information, with the test information covering a broad spectrum of the latent trait continuum. Against this disclosure, the selection of test items is currently done solely by the judgement of subject matter experts. Hence, this study questions the quality of tests that follow this methodology.

The study found significant variance in mean examinee ability estimations between paper-based tests and computer-adaptive test delivery methods, with the latter having higher mean ability levels. This clarifies that computer adaptive testing assesses examinees' abilities more precisely than paper-based assessments. Similarly, it has been demonstrated that CAT decreases test time, improves test security by restricting item exposure to examinees, and enriches the test-taking experience by allowing examinees to provide more accurate responses because the test adapts to their potential. This study

agrees with Wang et al. (2008) that PBT and CAT provide measurements of ability that are minimally equivalent.

The results show that there is a strong positive relationship between the PBT and CAT theta estimates. Considering a Pearson moment correlation statistic of 0.717, Tertiary institutions can make an informed conclusion that using computer adaptive testing will not jeopardize the selection of college students. This implies that only outstanding students will be accepted into the college, with the assumption that they will thrive academically. The college can make use of CAT's benefits without sacrificing its purpose of admitting outstanding students.

To sum up, college entrance examinations should be composed of test items that are of high quality, which have been pre-tested, and with known parameters. Tertiary institutions can reliably adopt Computer-adaptive testing as an alternative to paper-based tests in the selection of students. This will enable them to leap the benefits of CAT without compromising on the purpose.

5.3 Implications for policy and practice

The use of paper-based testing in the selection of students into college may have the following implications. Firstly, the selection process might have left out deserving students because the communication skills paper was composed of high-difficulty test items. Likewise, moderate discrimination could have resulted in poor distinction between outstanding students and those with low ability. In addition, the knowledge from this study would ensure boldness in decisions made by institutions when evaluating the

selection process of candidates, and deciding whether to transition from conventional assessment methods to modern assessment methods leveraging the technological advancements. Furthermore, In Practice, test items should be pre-tested to know how they function. Test developers will thus select only the appropriate item for administration.

5.4 Recommendations

5.4.1 Study Contribution to Knowledge

This study confirms that CAT and PBT estimates strongly correlate. A transition from a Paper-based test to a computer-adaptive test could ensure the reciprocation of the candidate's scores. Hence, stipulates that the computer adaptive test is a reliable alternative method in selecting students for college and impressively has practical benefits in assessment.

The study suggests that issues such as computer inexperience or test/computer anxiety seem not to constitute a problem. One of the criticisms about using computers as the testing medium is that individuals may have computer anxiety or have varying degrees of computer experience (Abele & Spurk, 2009). However, the findings show that neither computer-based testing nor paper-based testing affected the success of the test-takers, showing that the candidates were not anxious or stressed about using computers, so it can be deduced that the computer-based can be used alternatively to the paper-based version.

The study shows that the IRT framework enhances the reliability of entrance examinations. By pre-testing the items to get the parameters it is possible to predict how a student with a certain ability can fair. This enables test developers to select items with

parameters that will serve the purpose of the test, at the same time rephrase or completely discard items that are not functioning in a desired way.

5.4.2 Proposed Areas for Further Research Studies

Further research studies to complement this study can be in the following related areas:

Understanding the perception of students with the transition from paper-based tests to computer adaptive tests. This will uncover the level of technology acceptability, as this study found that there are still gaps in the testing environment that do not encourage computer tests, despite various forces driving institutions to adopt modern methodologies.

A study on consequences of using a misfit model in Computer adaptive tests. In this study, the model data fit did not perfectly fit. However, the IRT framework assumes using the model that fits the data. The fit is not quantified and the consequences of a misfit or non-perfect model-data fit are not known. Such a study will give an understanding of the repercussions in terms of precision of measurement when a misfit model is used.

A study on minimum items for calibrating an item bank. This study used 30 items for item bank calibration. There were no issues because a maximum number of 20 items was specified. However, for a large group testing such an item bank will expose the items and affect how they function. This study will give direction for item bank calibration.

5.5 Chapter Summary

This chapter drew three major conclusions from the study's findings. The implications of the findings for the selection process, as well as the study's contribution to knowledge, were also discussed. Finally, recommendations were made to higher learning institutions that administer entrance examinations, and possible study areas to complement the research were provided

REFERENCE

Abele, A. E. & Spurk, D. (2009). The longitudinal impact of self-efficacy and career goals on objective and subjective career success. *Journal of Vocational Behaviour,* 74(1), 53-62.

Adedoyin, O., & Mokobi, T. (2013). Using IRT Psychometric Analysis in Examining the Quality of Junior Certificate Mathematics Multiple Choice Examination Test Items. *International Journal of Asian Social Science, 3*(4), 992–1011.

American Education Research Association (2014). *The Standards for Educational and Psychological Testing.* Author

Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable length CATs are not biased. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing* (pp. 1–21). https://publicdocs.iacat.org/cat2010/cat09babcock.pdf

Babcock, B., & Weiss, D. J. (2012). Termination Criteria in Computerized Adaptive Tests: Do Variable-Length CATs Provide Efficient and Effective Measurement? *Journal of Computerized Adaptive Testing*, 1(1), 1-18. http://dx.doi.org/10.7333/1212-0101001

Baker F. B. (2001). *The Basics of Item Response Theory* (2nd ed.). ERIC. http://ericae.net/irt/baker.

Bartlett, M. S. (1954). A Note on the Multiplying Factors for Various Chi Square Approximation. *Journal of Royal Statistical Society,* 16(Series B), 296-8.

Bennett, S., Maton, K. & Kervin, L. (2008). The ‘digital natives’ debate: a critical review of the evidence. *British Journal of Educational Technology*, 39(5),775–786.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 384)*, 295-317

Cacciattolo, M. (2015). Ethical Considerations in Research. In M. Vicars, S. Steinberg, T. McKenna, M. Cacciattolo, (Eds.), *The Praxis of English Language Teaching and Learning (PELT). Critical New Literacies* (pp. 123-37). Sense Publishers, Rotterdam. https://doi.org/10.1007/978-94-6300-112-0_4

CAT KOREA. (2023). LIVECAT [Computer software]. https://www.thecatkorea.com

Chen C. T., Chen Y. L., Lin Y. C., Hsieh C. L., Tzeng J. Y., & Chen K. L. (2018). Item saving assessment of self-care performance in children with developmental disabilities: A prospective caregiver-report computerized adaptive test. *PLoS ONE*, 13(3), e0193936. https://doi.org/10.1371/journal.pone.0193936

Chulu, B. (2013). *Institutionalisation of assessment capacity in developing nations: The case of Malawi*. https://doi.org/10.1080/0969594X.2013.843505.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33*(5), 593–602. https://doi.org/10.1111/1467-8535.00294

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando.

Cronbach, L. J. (1947). Test "reliability": its meaning and determination. *Psychometrika, 12(1)*, 1–16. https://doi.org/10.1007/BF02289289

Culligan, B. (2008). *Estimating word difficulty using Yes/No tests in an IRT framework and its application for pedagogic objectives* (Masters dissertation). Temple University Japan.

Cyrinus, B., Essen, B. C., Idaka E., & Metibemu A. M. (2017). Item level diagnostics and model-data fit in item response theory (IRT) using bilog-mg v3.0 and irtpro v3.0 programmes. *Global Journal of Educational Research*, 16(1), 87-94. http://dx.doi.org/10.4314/gjedr.v16i2.2

Davey, T., & Pitoniak, M. J. (2006). *Designing computerized adaptive tests. Handbook of test development*. Lawrence Erlbaum Associates.

Eggen, T. & Veldkamp, B. (2012). *Computerized Adaptive Testing Item Selection in Computerized Adaptive Learning Systems.* BMC Medical Informatics and Decision Making.

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists* (1st ed.). Psychology Press. https://doi.org/10.4324/9781410605269

Franke, G.R. (2010). Product Moment Correlation. In J. Sheth and N. Malhotra (Eds.), *Wiley International Encyclopedia of Marketing.* https://doi.org/10.1002/9781444316568.wiem020688

González, M., Minuesa, C., & Del Puerto, I. M. (2016). Maximum likelihood estimation and Expectation-Maximization algorithm for controlled branching processes. Computational Statistics & Data Analysis. http://dx.doi.org/10.1016/j.csda.2015.01.015

Ministry of Education, Science and Technology (2016). *National Education Policy*. Author

Grant, L. & Villalobos, G. (2008). *Designing educational technologies for social justice*. Futurelab.

Guyer, R., & Thompson, N.A. (2014). *User's Manual for Xcalibre item response theory calibration software, version 4.2.2 and later*. Assessment Systems Corporation.

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues & Practice*, 12(3), 38-47.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: principles and applications.* Springer.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Sage.

Han, K. T. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement*, 36(1), 64-66. http://dx.doi.org/10.1177/0146621611414407

He L., & Min S. (2017). Development and validation of a computer adaptive EFL test. *Language Assessment Quarterly*, 14(2),160–176. https://doi.org/10.1080/15434303.2016.1162793

Hsu, S. F., & Liou, S. (2021). *Artificial Intelligence Impact on Digital Content Marketing Research*. https://doi.org/10.1109/ICOT54518.2021.9680666

International Test Commission (2005). *International Guidelines on Computer-Based and Internet Delivered Testing*. Author

Jiang, Q., & Xuyang, G. (2020). *Research on the Reform of Chinese College Entrance Examination System*. https://doi.org/10.2991/assehr.k.200205.024

Johnson, R., Onwuegbuzie, A., & Turner, L. (2007). Toward a Definition of Mixed Methods Research. *Journal of Mixed Methods Research,* 1(1), 112-133. http://dx.doi.org/10.1177/1558689806298224

Junker, B. W. (1993). *Conditional association, essential independence and monotone unidimensional item response models*. The Annals of Statistics.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement,* 34(1),111–117.

Kalender, I. & Berberoglu, G. (2017). *Can computerized adaptive testing work in students' admission to higher education programs in Turkey?* http://doi.org/10.12738/estp.2017.2.0280

Kaya, E. (2021). *A comparability and classification analysis of computerized adaptive and conventional paper-based versions of an English language proficiency reading subtest.* Ankara

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and Reliability of Measurement Instruments Used in Research. *American Journal of Health-System Pharmacy*, 65(1), 2276-2284. https://doi.org/10.2146/ajhp070364

Kingsbury, G. G. & Houser, R. (2008). *ICAT: An Adaptive Testing Procedure for the Identification of Idiosyncratic Knowledge Patterns*. http://dx.doi.org/10.1027/0044-3409.216.1.40

Kitchin, H. A. (2007). *Research ethics and the internet: Negotiating Canada's Tri-Council policy statement*. Halifax & Winnipeg.

Li, M., Shavelson, J. R., Yue, Y., & Wiley, E. (2015) Generalizability Theory. http://dx.doi.org/10.1002/9781118625392.wbecp352

Livingston, S. A. (2018). *Test reliability Basic concepts* (Research Memorandum No. RM-18-01). Educational Testing Service. Princeton.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. https://doi.org/10.4324/9780203056615

Lynch, S. (2022). Adapting Paper-Based Tests for Computer Administration: Lessons Learned from 30 Years of Mode Effects Studies in Education. *Practical Assessment, Research, and Evaluation*, 27(Article 22). https://doi.org/10.7275/pare.1317

Mead, A. D. (2005). *Psychometric Reliability: Definition, Estimation, and Application*. https://doi.org/10.1002/0470013192.bsa672

Ministry of Education (2020). *National Education Sector Investment Plan 2020-2030*. Author

Mislevy, R.J (1995). *Evidence and inference in educational assessment: Educational testing service*. Princeton.

Mocorro, R. (2017). The First Step in Research. International Journal of Science and Research (IJSR). http://dx.doi.org/10.21275/ART20178777

Morizot J., Ainsworth A. T., & Reise S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In Robins R. W., Fraley R. C., Krueger R. F. (Eds.), *Handbook of research methods in personality* (pp. 407–423). Guilford.

National Planning Commission (2020). Malawi 2063.

Oz, H., & Ozturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? Journal of Language and Linguistic Studies, 14(1), 67-85.

Piaw, C. Y. (2012). Effects of computer-based testing on test performance and testing motivation. *Computers in Human Behavior, 28*(5), 1580 1586. https://doi.org/10.1016/j.chb.2012.03.020

Ree, M. J., & Jensen, H. E. (1983) 7 - Effects of Sample Size on Linear Equating of Item Characteristic Curve Parameters in Chikoko. T. A Small sample estimation of item parameters in item response theory models using operational data. Unpublished thesis, University of Malawi

Rudner, L. M., & Guo, F. (2011). Computer Adaptive Testing for Small Scale Programs and Instructional Systems. *Journal of Applied Testing Technology*.

Sahin, A., & Weiss, D. (2015). Effects of Calibration Sample Size and Item Bank Size on Ability Estimation in Computerized Adaptive Testing. Educational Sciences: Theory & Practice. 15. 1585-1595. http://dx.doi.org/10.12738/estp.2015.6.0102

Sahın, A., & Ozbası, D. O. (2017). Effects of Content Balancing and Item Selection Method on Ability Estimation in Computerized Adaptive Tests. Eurasian Journal of Educational Research, 17(69), 21-36.

Sandars, J. & Dearnley, C. (2008). Twelve tips for the use of mobile technologies for work based assessment. Medical teacher. 31. 18-21. http://dx.doi.org/10.1080/01421590802227966

Schaeffer, G., Bridgeman, B., Golub-Smith, M., Lewis, C., Potenza, M. & Steffen, M. (1998). Comparability of Paper-and-Pencil and Computer Adaptive Test Scores on the GRE" General Test. ETS Research Report Series. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01787.x

Seo. G. D., & Choi. J. (2020) Introduction to the LIVECAT web-based computerized adaptive testing platform. Journal of Educational Evaluation for Health Professions.

Singini. G (2014, December 30) Council wants ’varsity entrance exams scrapped. *Nation Online*. https://mwnation.com/council-wants-varsity-entrance-exams-scrapped/

Stepanek, L., & Martinkova, P. (2020) Feasibility of computerized adaptive testing evaluated by Monte-Carlo and post-hoc simulations. The 15th Conference on *Computer Science and Information Systems (FedCSIS)*. https://doi.org/10.15439/2020F197

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Research Report 94–5). https://doi.org/10.1002/j.2333-8504.1994.tb01578.x

TDR-IR Toolkit (2023). *TDR Implementation Research Toolkit* (2nd ed). Access and Delivery Partnership

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In *Computerized adaptive testing* (pp. 101-133). Routledge.

Thompson, N. (2021, February 2). “*IRT test information function”: Assessment systems.* https://assess.com/irt-test-information-function/

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report 44). National Center on Educational Outcomes.

Townsend, L. & Wallace, C. (2016). *Social Media Research: A Guide to Ethics*. University of Aberdeen.

University of Malawi (2015). *University entrance examinations*.Author

Van der Linden, W. J., & Pashley, P. J. (2000). Item selection and theta estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp.1-25). Kluwer Academic.

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement*, 60(3), 371–384. https://doi.org/10.1177/00131640021970600

Wang, H., & Shin, C. D. (2010). Comparability of computerized adaptive and paper-pencil tests. *Test, Measurement and Research Service Bulletin*, 13(1), 1-7

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, 68(1), 5-24.

Wang, T. & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement, 38*(1), 19 -49. https://doi.org/10.1111/j.1745-3984.2001.tb01115.x

Way, W. D., & Robin, F. (2016). The history of computer-based testing. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 185–207). The Guilford Press.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive theory. *Applied Psychological Measurement*, 6(4), 473–492. https://doi.org/10.1177/014662168200600408

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 1(4), 361-75.

Wilson, L.T. (2018). *Pearson Product-Moment Correlation*. Semantic Scholar.

Wise, S. L. (2018). Computer-based testing. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 341–344). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Yao, D. (2019). *A Comparative Study of Test Takers' Performance on Computer-Based Test and Paper-Based test across different CEFR levels*. http://dx.doi.org/10.5539/elt.v13n1p124

Yongbo, L. (2020). *Improvement of College Entrance Examination Enrolment System Based on Fairness, Efficiency and Autonomy.* http://dx.doi.org/10.2991/assehr.k.201202.147

Zanon, C., Hutz, C.S., & Yoo, H. (2016). An application of item response theory to psychological test development. *Psicol. Refl. Crít.,* **29(1)**, 18-27 https://doi.org/10.1186/s41155-016-0040-x

Zucker, S. (2003). *Fundamentals of Standardized Testing: Assessment report*. Harcourt.

APPENDICES

Appendix 1: DCE and NCE 2022 communication skills aptitude test paper.

Examination number .....

MINISTRY OF EDUCATION

DOMASI AND NALIKULE COLLEGES OF EDUCATION

2022 ENTRANCE EXAMINATIONS

PAPER I: COMMUNICATION SKILLS

Date: 18 May 2022 **Time :** 8:00 – 9:00 am

Time Allowed: 1 Hour

INSTRUCTIONS

1. This paper contains **9 pages**; please check.
2. Write your full examination number in the space provided on each page.
3. There are **50** questions in three sections: **A, B** and **C**. Answer **all** questions in each section.
4. In sections **A** and **C**, show your answer by circling the letter (**A, B, C** or **D**) that represents the best option for each question.
5. In **Section B**, write your answers in the spaces provided in the passage.
6. If you decide to erase your answer, rub it off completely and mark your final answer clearly.
7. If you need the assistance of an invigilator during the test, raise your hand.
8. You have exactly **1 hour** in which to answer all the questions in this paper.

Page **1** of **9**

Examination number .....

In questions 1 – 25, choose the option (A, B, C or D) that best completes the given sentence.

1. It's high time Gomezgani _____ another job.
   A. found
   B. should find
   C. finds
   D. must find

2. As I write this test, the sun _____ setting.
   A. will have been
   B. was
   C. is
   D. has been

3. Hard work and dedication ____ all that a student requires in order to succeed in college.
   A. is
   B. was
   C. were
   D. are

4. The way Tamara is living, she _____ all her pocket money by the end of next month.
   A. will spend
   B. will have been spending
   C. will have spent
   D. will be spending

5. If our Member of Parliament _____ consistently, he would never have been re-elected.
   A. did not campaign
   B. had not campaigned
   C. had been campaigning
   D. wouldn't campaign

6. Although he tried hard to _____ his pen, there was no trace of it in the house.
   A. look at
   B. look out for
   C. look for
   D. look to

7. I must _____ a doctor, he will find out what is wrong with my son.
   A. send to
   B. send off
   C. send for
   D. send to

8. You must all _____ your assignments by four o'clock, today.
   A. hand in
   B. hand on
   C. hand up
   D. hand over

9. Neither Grace nor her sons wanted to expose _____ to danger.
   A. herself
   B. himself
   C. himself or herself
   D. themselves

10. Some of the students who visited the boy paid ____ tuition fees by working part time.
   A. his
   B. their
   C. his or her
   D. her

11. If all of us _____ actively, we can finish this assignment in thirty minutes.
   A. participates
   B. participate
   C. participated
   D. have participated

12. Public servants _____ discharge their duties with integrity.
   A. could
   B. would
   C. should
   D. may

13. If I _____ run three kilometers in one minute, I would be famous.
   A. can
   B. could
   C. would
   D. should

14. Thieves stole a lot of money _____ my grandmother's house.
    A. from
    B. in
    C. to
    D. into

15. He devoted a lot of time, money and effort ____ his studies.
    A. for
    B. on
    C. in
    D. to

16. The two men fought _____ the money that they found along the road.
    A. for
    B. on
    C. over
    D. about

17. What she is doing will soon land her _____ trouble.
    A. on
    B. in
    C. for
    D. onto

18. The professor's speech was _____; we learnt a lot from her.
    A. an eye-opener
    B. mind-numbing
    C. out of order
    D. a perfect storm

19. When I raised a good point during the discussion, the teacher said, "_____."
    A. You can do that again
    B. You can say that again
    C. Look on the bright side
    D. You should call a spade a spade

20. I have been extremely busy today. _____ since morning.
    A. It's been one thing after another
    B. I have been on the loose
    C. I've been on the run
    D. I have been on the go

21. He was _____ for his hard work.
    A. rewarded
    B. awarded
    C. repelled
    D. compensated

22. The _____ for submitting applications is 23rd November, 2022.
    A. dateline
    B. timeline
    C. deadline
    D. maximum period

23. Any score below 40 per cent doesn't _____ to get you into college
    A. Supplement
    B. suffice
    C. sufficient
    D. complement

24. My progress in school was _____ by poor study habits and poor attitude.
    A. aided
    B. prohibited
    C. reinforced
    D. hindered

25. The _____ sentenced the thief to three years in prison.
    A. police
    B. Attorney General
    C. magistrate
    D. lawyer

In questions 26 and 27, choose the option (A, B, C or D) which represents a sentence that is correctly punctuated.

26. A. Looking across the street, I saw a person who was running faster than a cheetah.
    B. Looking across the street; I saw a person who was running faster than a cheetah.
    C. Looking across the street, I saw a person, who was running faster than a cheetah.
    D. Looking across the street I saw a person who was running, faster than a cheetah.

Examination number .....

27. A. "No," Said Sam to Andrew. "I will not give you anything."
    B. "No," said Sam to Andrew. "I will not give you anything."
    C. "No," said Sam to Andrew, "I will not give you anything".
    D. "No," said Sam to Andrew: "I will not give you anything."

In questions 28 – 30, choose the option (A, B, C, or D) that correctly answers the question.

28. Isn't there any bank near by? _____.
    A. Yes, there isn't
    B. No, there is
    C. No, there isn't
    D. Yes, it is

29. You don't have a problem, do you? _____.
    A. No, I do
    B. Yes, I don't
    C. Yes, I do
    D. Yes, I haven't

30. You do realize that this will not benefit you, don't you? _____.
    A. No, I don't
    B. No, I do
    C. No, it will
    D. Yes, it won't

SECTION B: CLOZE PASSAGE

Complete the following passage by filling in each blank space (31 - 40) with one most suitable word.

Stress is an inevitable aspect of all our _____(31). Learning to deal _____(32) stress in a positive, intelligent way is _____ (33) to good health. One way to _____ (34) stress is to work it off in physical activities. For example, jogging around the neighbourhood or exercising on the dance floor can _____(35) stress and give you more energy to cope with life. Stress can also be controlled by changing your mental attitude. Learn to accept things; fighting against something that is inevitable or _____ (36) is useless. Learn to take one thing at a _____ (37). Rather than trying to do _____(38) at once, deal with more important problems first, and leave the rest to another day. Finally, talking about stress is important. When events in your life seem overwhelming, talk about your _____ (39). This can be done by opening up to your _____(40) and friends.

Appendix 2: UNIMAREC ethical clearance letter.

VICE-CHANCELLOR
Prof. Samson Sajidu, BSc Mlw, MPhil Cantab, PhD Mlw

Our Ref: P.08/23/285

Your Ref.:

CHANCELLOR COLLEGE
P.O. Box 280, Zomba, Malawi

Telephone: (265) 1 526 622
Fax: (265) 1 524 031
E-mail: vc@unima.ac.mw

21st September 2023

Mr Thokozani Chisale
Master's in education, Testing, Measurement
And Evaluation
University of Malawi
P.O. Box 280
Zomba

Dear Mr Chisale

RESEARCH ETHICS AND REGULATORY APPROVAL AND PERMIT FOR PROTOCOL P.08/23/285 ASSESSING THE RELIABILITY OF COMPUTER ADAPTIVE TESTING IN COLLEGE ENTRANCE COMMUNICATION SKILLS EXAMINATIONS IN MALAWI

Having satisfied all the relevant ethical and regulatory requirements, I am pleased to inform you that the above-referred research protocol has officially been approved. You are now permitted to proceed with its implementation. Should there be any amendments to the approved protocol in the course of implementing it, you shall be required to seek approval of such amendments before implementation of the same.

This approval is valid for **one year** from the date of issuance of this approval. If the study goes beyond one year, an annual approval for continuation shall be required to be sought from the University of Malawi Research Ethics Committee (UNIMAREC) in a format that is available at the Secretariat.

1

Once the study is finalized, you are required to furnish the Committee and the Vice Chancellor with a final report of the study. The committee reserves the right to carry out a compliance inspection of this approved protocol at any time as may be deemed by it. As such, you are expected to properly maintain all study documents including consent forms.

UNIMAREC wishes you a successful implementation of your study.

UNIVERSITY OF MALAWI
RESEARCH ETHICS COMMITTEE
21 SEP 2023
APPROVED
P.O. BOX 280, ZOMBA

Yours Sincerely,

Dr Victoria Ndolo
CHAIRPERSON OF UNIMAREC

CC: Vice Chancellor
The Registrar
Director of Finance and Investments
Acting Head of Research
Chairperson UNIMAREC
UNIMAREC Compliance Officer

2

Appendix 3: EDF letter of introduction.

VICE-CHANCELLOR
Prof. Samson M.I. Sajidu,
BSc MSc, MPhil Cantab, PhD Mlw

› Connect with Excellence

UNIVERSITY OF MALAWI
P.O. Box 280, Zomba, Malawi
TEL: (265) 1 524 222
FAX: (265) 1 524 046
EMAIL: vc@unima.ac.mw

Our Ref:

Your Ref:

25th September, 2023

TO WHOM IT MAY CONCERN

LETTER OF INTRODUCTION: THOKOZANI CHISALE

This letter serves to confirm that **Mr. Thokozani Chisale** is a registered postgraduate student in the Education Foundations Department, of the School of Education, in the University of Malawi. He is studying under the Master of Education (Testing, Measurement & Evaluation) program. His registration Number is MED/MEV/01/21.

Mr. Chisale has completed his first year of studies which mainly involves coursework. As a requirement for completion of his study program, he is conducting a research titled: "***Assessing the Reliability of Computer Adaptive Testing in College Entrance Communication Skills Examinations in Malawi***". This letter therefore, serves to request your institution to assist our student to collect the required data.

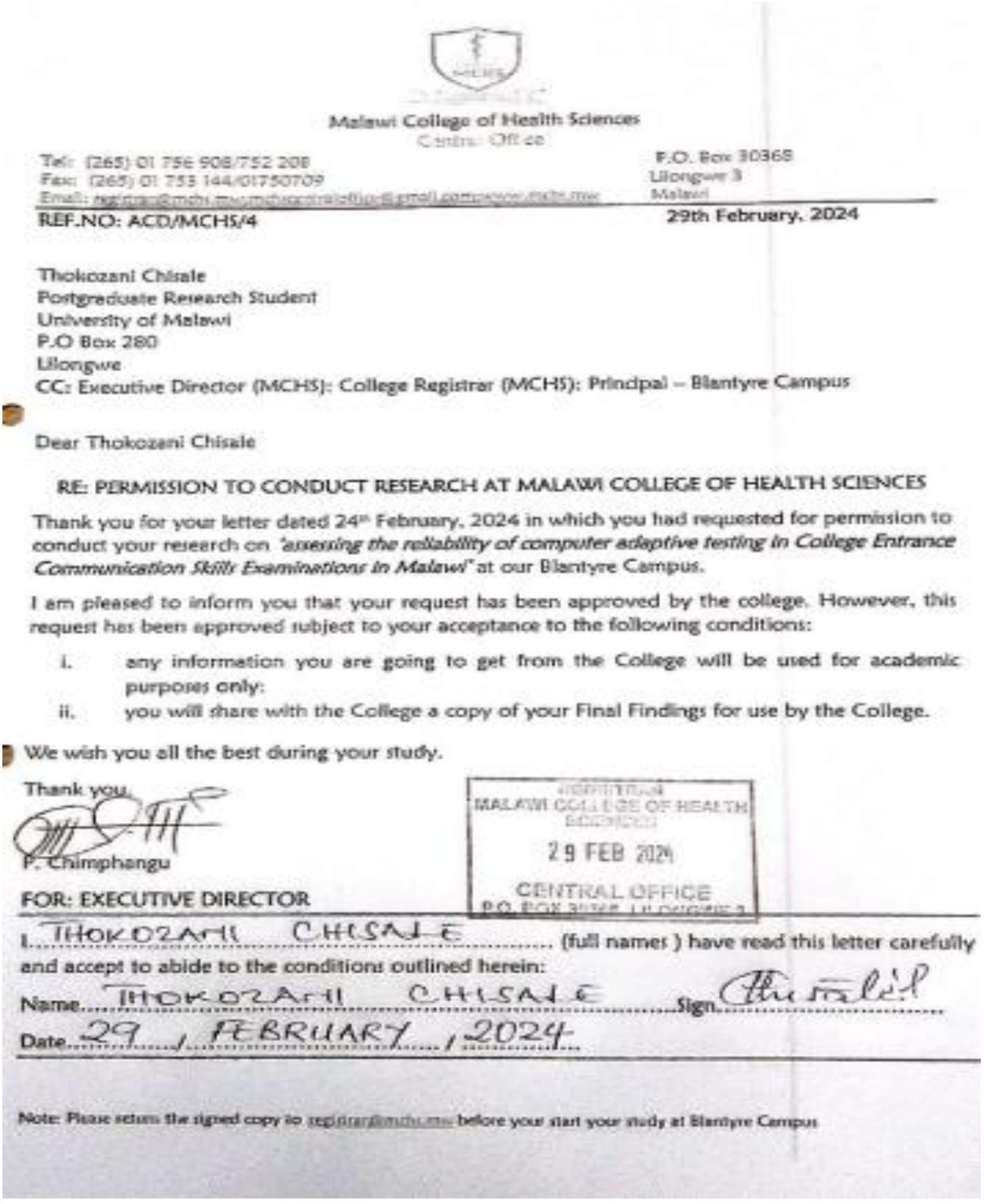
For any inquiries please contact the undersigned via the following email address: med@cc.ac.mw

UNIVERSITY OF MALAWI
25 SEP 2023
SCHOOL OF EDUCATION
P.O. BOX 280, ZOMBA

Sincerely yours,

Symon Winiko, PhD.
HEAD OF DEPARTMENT – EDUCATION FOUNDATIONS

Appendix 4: Permission to conduct research



Malawi College of Health Sciences
Centre Office

Tel: (265) 01 756 908/752 208
Fax: (265) 01 753 144/01750709
Email: registrar@mchs.mw/mchscentraloffice@gmail.com/www.mchs.mw

P.O. Box 30368
Lilongwe 3
Malawi

REF.NO: ACD/MCHS/4

29th February, 2024

Thokozani Chisale
Postgraduate Research Student
University of Malawi
P.O Box 280
Ulongwe

CC: Executive Director (MCHS): College Registrar (MCHS): Prindpal – Blantyre Campus

Dear Thokozani Chisale

RE: PERMISSION TO CONDUCT RESEARCH AT MALAWI COLLEGE OF HEALTH SCIENCES

Thank you for your letter dated 24th February, 2024 in which you had requested for permission to conduct your research on *"assessing the reliability of computer adaptive testing in College Entrance Communication Skills Examinations in Malawi"* at our Blantyre Campus.

I am pleased to inform you that your request has been approved by the college. However, this request has been approved subject to your acceptance to the following conditions:

i. any information you are going to get from the College will be used for academic purposes only;
ii. you will share with the College a copy of your Final Findings for use by the College.

We wish you all the best during your study.

Thank you.

P. Chimphangu

FOR: EXECUTIVE DIRECTOR

MALAWI COLLEGE OF HEALTH SCIENCES
29 FEB 2024
CENTRAL OFFICE
P.O. BOX 30368, LILONGWE 3

I, THOKOZANI CHISALE (full names) have read this letter carefully and accept to abide to the conditions outlined herein:

Name THOKOZANI CHISALE Sign

Date 29 / FEBRUARY / 2024

Note: Please return the signed copy to registrar@mchs.mw before your start your study at Blantyre Campus

Appendix 5: Sample of participants' consent form.

INFORMED CONSENT FORM

TITLE OF STUDY

Assessing the reliability of computer adaptive testing in college entrance communication skills examinations in Malawi.

PRINCIPAL INVESTIGATOR

Thokozani Elvis Chisale
Education Foundations Department
University of Malawi, P.O Box 280, Zomba.
+265 (0) 882 579 775/ +265 (0) 996 673 295
med-mev-01-21@unima.ac.mw

Research Purpose and Procedures

You are being asked to take part in a research study. The purpose of this study is to assess the reliability of computer adaptive testing as an alternative to paper based testing for entrance examinations. Tests can be delivered through oral interviews, paper based and computer based. With the advent of technology is it prominent that institutions will consider to use digital technology in testing. As such this research seek to inquire the reliability of the technology as compared to the convention testing practice. Duration of your participation in this study is approximately **40 minutes**. In this study you will be expected to write examination (multiple choice items) using computer based delivery mode. The computer will adapt to your level of ability based on the responses you will be providing. The test will end once the conditions for its termination have been satisfied. In these examinations, participants will write different test items.

Risks and Discomforts of the Research Study

In this study you might feel embarrassed for not being able to operate the computers. However, the research assistants will always be available to assist you. Students who will participate in this study will benefit in terms of exposure to examination delivered through computer mode. Participants will have a chance to interact with the researcher to get a depth knowledge CAT since it provides a novice testing paradigm. The study will inform tertiary institutions of test delivery

modes that provide reliable estimates of ability. This study will provide empirical evidence on how test delivery modes influence the estimation of ability in college entrance examinations in Malawi.

Alternative Procedures

You can as well write the examinations remotely. This will help you to participate in this research at any time and at any place. It will as well be convenient since you will not have to suspend other activities.

Provisions for Confidentiality

Your responses to this study will be anonymous. Please do not write any identifying information. Every effort will be made by the researcher to preserve your confidentiality. Personal data collected will be protected using data encryption and password.

Research Related Injury

There is minimal or no injury you could face associated to your participation in this study. All safety measures will be in place so that your health is sustained and assured.

Voluntariness in Participation and The Right to Discontinue Participation Without Penalty

Your participation in this study is voluntary. It is up to you to decide whether or not to take part in this study. If you decide to take part in this study, you will be asked to sign a consent form. After you sign the consent form, you are still free to withdraw at any time and without giving a reason. If you withdraw from the study before data collection is completed, your data will be returned to you or destroyed.

Contacts for Additional Information

If you have questions at any time about this study, or you experience adverse effects as the result of participating in this study or your rights as a participant have been violated, you may contact the researcher whose contact information is provided on the first page OR the supervisor Dr. Foster Gondwe (+265(0) 888 123 961 or fgondwe@unima.ac.mw) OR UNIMAREC Chairperson Dr. Victoria Ndolo, Chairperson of University of Malawi Research Ethics Committee (UNIMAREC), P.O. Box 280, Zomba. +265 995 0427 60.

Do you agree to continue with the study? ☑ YES ☐ NO

Name of the respondent: CLIFTON KAISA

Age: 18

Male/Female

Signature: [signature]

Date: 22-02-2024

Name of the interviewer: THOKOZANI CHISALE

Signature: [signature]

Date: 22-02-2024

THANK YOU

UNIVERSITY OF MALAWI
RESEARCH ETHICS COMMITTEE
21 SEP 2023
APPROVED
P.O. BOX 280, ZOMBA

FREQUENCIES OF CANDIDATES PER THETA RANGES		
Range	**Number of examinees (CAT)**	**Number of examinees (PBT)**
Below -4	6	47
-4.0 to -3.6	1	6
-3.6 to -3.2	2	3
-3.2 to -2.8	2	8
-2.8 to -2.4	3	12
-2.4 to -2.0	7	18
-2.0 to -1.6	12	34
-1.6 to -1.2	23	52
-1.2 to -0.8	31	56
-0.8 to -0.4	28	98
-0.4 to 0.0	33	132
0.0 to 0.4	46	190
0.4 to 0.8	95	153
0.8 to 1.2	103	110
1.2 to 1.6	72	57
1.6 to 2.0	42	23
2.0 to 2.4	20	3
2.4 to 2.8	15	1
2.8 to 3.2	5	2
3.2 to 3.6	0	0
3.6 to 4.0	0	0
Above +4	0	0

APPENDIX 6

Appendix 7: Description of the software (s) used.

Xcalibre Item Response Theory Calibration Software Version 4.2.2

Guyer, R., and Thompson, N.A. (2014). User's Manual for Xcalibre item response theory calibration software, version 4.2.2 and later. Assessment Systems Corporation.

1. Introduction

Xcalibre™ is a Windows® application designed to perform estimation of item response theory (IRT) item parameters with user-friendly reports. The purpose of these reports is to help testing programs evaluate the quality of test items by examining their psychometric characteristics, as well as provide the IRT parameters necessary for scoring examinees with IRT, especially with computerized adaptive testing (CAT).

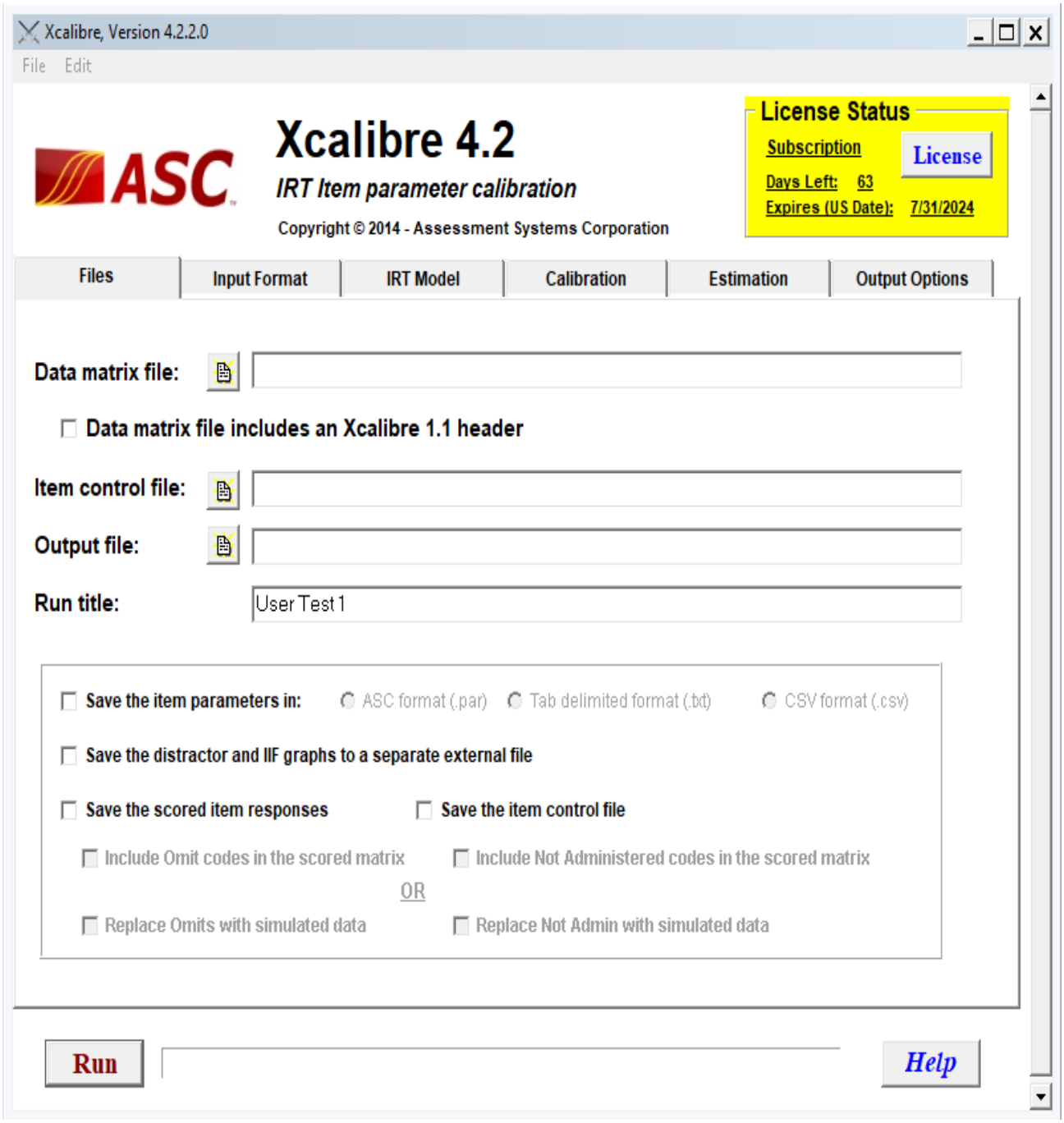
Xcalibre has a friendly graphical user interface (GUI) that makes it easy to run the program, even if you are not an expert on IRT. The GUI is organized into six tabs: *Files, Input Format, IRT Model, Calibration, Estimation, and Output Options*. These are discussed in detail in **Chapter 3: Running the Program**.

Xcalibre 4.2 offers several substantial advantages over *Xcalibre 1.1*:

1. The most important advantage is the addition of graphics. It is now possible to produce a plot of the item response function (IRF) and, optionally, the item information function (IIF) for each item. In addition, you can also include the fit line with the IRF for dichotomous models.
2. *Xcalibre 4.2* now performs IRT item parameter calibration for polytomous IRT models.
3. It is now possible to screen out items with unacceptable classical statistics before performing the IRT item parameter calibration.
4. The item parameter flag ranges can now be specified. In addition, you can easily customize each of the flag labels used by *Xcalibre 4.2*.
5. Instead of simple ASCII text files, the output is now rich text file (RTF) format prepared as a formal report, and also in a comma-separated value (CSV) format that is able to be manipulated (sorted, highlighted, etc.) in spreadsheet software. Xcalibre additionally produces a CSV file of examinee scores.
6. Examinee IRT score (θ) estimates can now be produced for multiple domains as well as the full test.
7. Scores can be classified into two groups at a specified cut score, and the two groups can use user-defined labels.
8. The maximum number of items that can be analyzed has been increased to 1,500.
9. A "batch" type of capability, using a "Multiple Runs File" has been added to allow you to run multiple data sets without having to use the graphic user interface for each run. Multiple Runs Files can be created outside *Xcalibre* in a text editor or interactively within *Xcalibre*.

Your *Xcalibre 4.2* License and Unlocking Your Copy

Unless you have purchased a network or multiple-computer license, your license for *Xcalibre 4.2* is a single-user license. Under this license you may install *Xcalibre 4.2* on two computers (e.g., a desktop and a laptop) so long as there is no possibility that the two copies of the software will be in use simultaneously. If you would like to use *Xcalibre 4.2* on a network or by


Xcalibre, Version 4.2.2.0
File Edit
Xcalibre 4.2
IRT Item parameter calibration
Copyright © 2014 - Assessment Systems Corporation
License Status
Subscription
License
Days Left: 63
Expires (US Date): 7/31/2024
Files
Input Format
IRT Model
Calibration
Estimation
Output Options
Data matrix file:
Data matrix file includes an Xcalibre 1.1 header
Item control file:
Output file:
Run title:
User Test 1
Save the item parameters in:
ASC format (.par)
Tab delimited format (.txt)
CSV format (.csv)
Save the distractor and IIF graphs to a separate external file
Save the scored item responses
Save the item control file
Include Omit codes in the scored matrix
Include Not Administered codes in the scored matrix
OR
Replace Omits with simulated data
Replace Not Admin with simulated data
Run
Help

The CAT KOREA Software Description.

Seo. G. D., and Choi. J. (2020) Introduction to the LIVECAT web-based computerized adaptive testing platform. Journal of Educational Evaluation for Health Professions.

LIVECAT provides examination administrators with an easy and flexible environment for composing and managing examinations. It also stores responses and can track changes in respondents. LIVECAT provides 2 testing forms: computer-based testing and CAT. Examinees can access tests administered using LIVECAT through a variety of internet-connected devices (desktops, laptops, smartphones, and tablet computers).

Development tools

Several tools were used for programming LIVECAT, as follows: (1) operating system, Amazon Linux; (2) web server, nginx 1.18; (3) WAS, Apache Tomcat 8.5; (4) database, Amazon RDMS—Maria DB; and (5) languages, JAVA8, HTML5/CSS, Javascript, and jQuery.

Item response theory models

Several item response theory (IRT) models were implemented in the LIVECAT platform, as follows: (1) Rasch model, (2) 1-parameter logistic model, (3) 2-parameter logistic model, and (4) 3-parameter logistic model. The administrator can choose a specific model for test construction in LIVECAT.

Diagram of the LIVECAT algorithm

Fig. 1 presents the process of constructing a test in the administrator account on LIVECAT. Fig. 2 shows the process of taking a test using an examinee account on LIVECAT. Fig. 3 is a flowchart of the CAT algorithm used to estimate an examinee's ability in LIVECAT [2]. The CAT algorithm is embedded in the "administering test" section, as shown in Fig. 2, and several IRT models and rules of the CAT algorithm can be selected in the "constructing and specifying test" construction, as shown in Fig. 1.

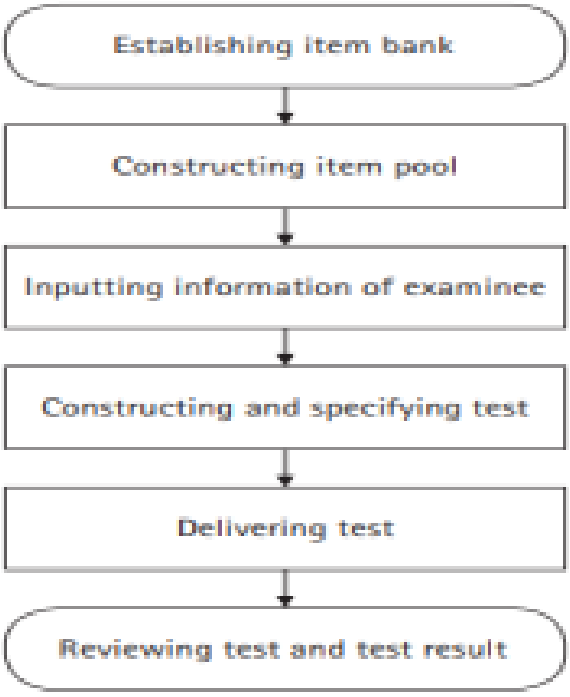


Fig. 1. Flowchart of the process of constructing a test in LIVECAT.

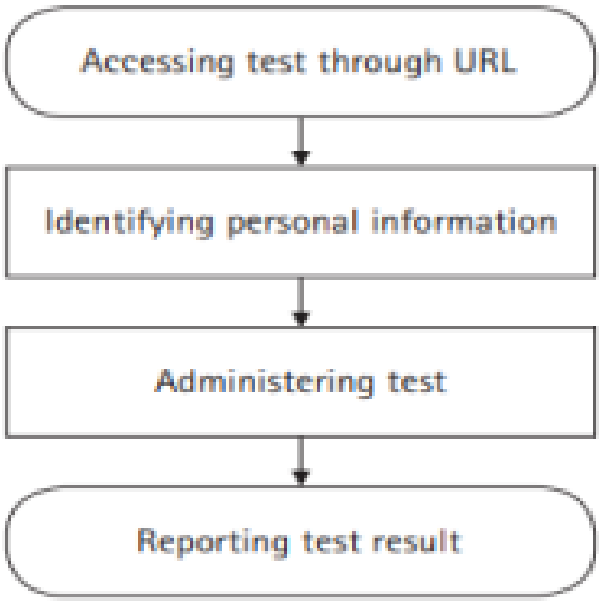


Fig. 2. Flowchart of the process of taking a test in LIVECAT.

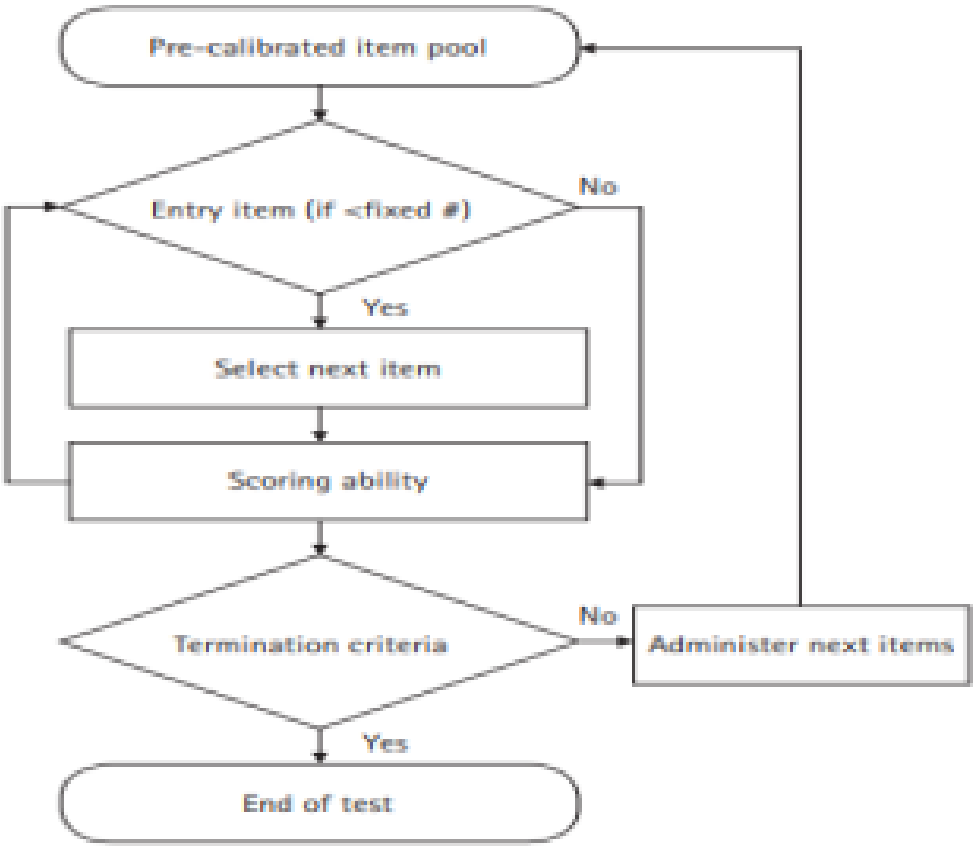
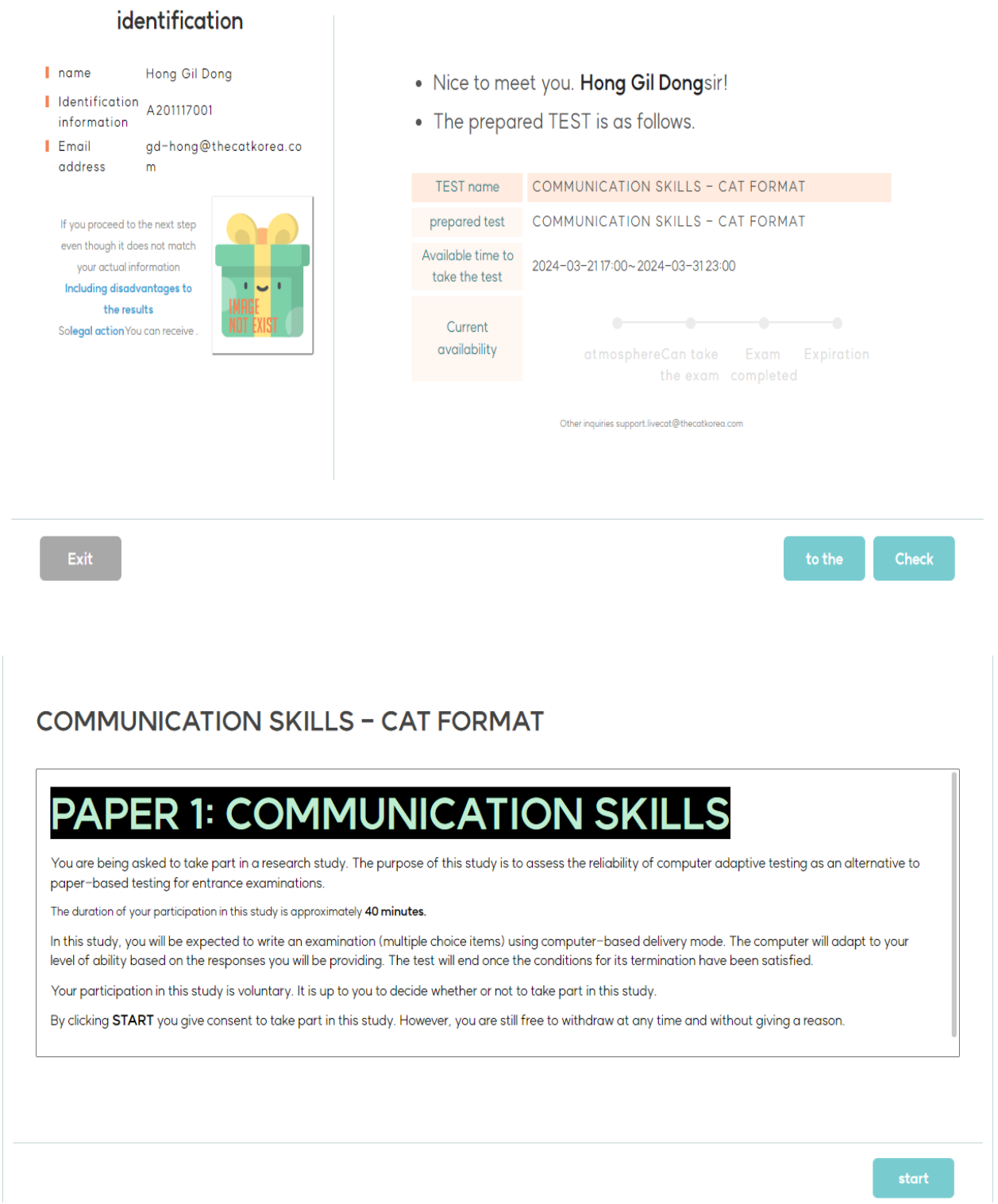


Fig. 3. The computerized adaptive testing algorithm used to estimate an examinee's ability.

Conclusion

The LIVECAT platform has several advantages. First, it saves the expenses of preparing and administering tests (e.g., paper costs, printing costs, time). Since LIVECAT is a web-based platform, there is no restriction on place and time. It is also easy to manage item and test information within the LIVECAT platform, which means that a test administrator does not need to be a psychometrician to apply CAT. However, an important consideration when implementing LIVECAT is that the test server should be stable while a test is being taken. In particular, the test server should not be shut down when a high-stakes examination is being administered. An Amazon Web Services (AWS) cloud server can immediately cope with unexpected external problems and keep a test server stable because AWS has infrastructure distributed over the world. The AWS cloud server was adopted to ensure stability for our LIVECAT platform. The recent version of LIVECAT provides only a dichotomous item response model and basic components of CAT. In the near future, LIVECAT will be updated to include advanced functions, such as polytomous item response models, weighted likelihood estimation method, and content balancing method. Once a polytomous item response model is implemented for psychological testing (e.g., personality, attitude), the LIVECAT platform will be used widely for psychological examinations in addition to certification/licensure examinations.

Appendix 8: Outlook of test items in CATKOREA.



One The _____ for submitting applications is 23rd November, 2022.

One dateline	Choose the option that best completes the sentence.
2 timeline	
3 deadline	
4 maximum period	

response sheet One 2 3 4 to the

2 The two men fought _____ the money that they found along the road.

One for	Choose the option that best completes the sentence.
2 on	
3 over	
4 about	

Previously response sheet One 2 3 4 to the

3 If all of us _____ actively, we can finish this assignment in thirty minutes.

One participates	Choose the option that best completes the sentence.
2 participate	
3 participated	
4 have participated	

Previously response sheet One 2 3 4 to the

4 You must all _____ your assignments by four o'clock, today.

One hand in 2 hand on 3 hand up 4 hand over	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

5 If I _____ run three kilometers in one minute, I would be famous.

One can 2 could 3 would 4 should	Choose the option that best completes the sentence

Previously response sheet One 2 3 4 to the

6 Although he tried hard to _____ his pen, there was no trace of it in the house.

One look at 2 look out for 3 look for 4 look to	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

7 If our Member of Parliament _____ consistently, he would never have been re-elected.

One did not campaign
2 had not campaigned
3 had been campaigning
4 wouldn't campaign

Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

8 Public servants _____ discharge their duties with integrity.

One could
2 would
3 should
4 may

Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

9 The _____ sentenced the thief to three years in prison.

One police
2 Attorney General
3 magistrate
4 lawyer

Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

10 Some of the students who visited the boy paid _____ tuition fees by working part time.

One his 2 their 3 his or her 4 her	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

11 I must _____ a doctor, he will find out what is wrong with my son.

One send to 2 send off 3 send for 4 send by	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

12 He devoted a lot of time, money and effort _____ his studies.

One for 2 on 3 in 4 to	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

13 My progress in school was _____ by poor study habits and poor attitude.

One aided	Choose the option that best completes the sentence.
2 prohibited	
3 Reinforced	
4 hindered	

Previously response sheet One 2 3 4 to the

14 Hard work and dedication _____ all that a student requires in order to succeed in college.

One is	Choose the option that best completes the sentence
2 was	
3 were	
4 are	

Previously response sheet One 2 3 4 to the

15 Choose the option which represents a sentence that is correctly punctuated.

One "No," Said Sam to Andrew. "I will not give you anything."	Choose the option which represents a sentence that is correctly punctuated.
2 "No," said Sam to Andrew. "I will not give you anything."	
3 "No," said Sam to Andrew, "I will not give you anything."	
4 "No," said Sam to Andrew: "I will not give you anything."	

Previously response sheet One 2 3 4 to the

16 It's high time Gomezgani _____ another job.

One found 2 should find 3 finds 4 must find	Choose the option that best completes the sentence

Previously response sheet One 2 3 4 to the

17 Choose the option which represents a sentence that is correctly punctuated.

1 Looking across the street, I saw a person who was running faster than a cheetah. 2 Looking across the street: I saw a person who was running faster than a cheetah. 3 Looking across the street, I saw a person, who was running faster than a cheetah. 4 Looking across the street I saw a person who was running, faster than a cheetah.	Choose the option which represents a sentence that is correctly punctuated.

이전으로 응답지 1 2 3 4 다음으로

18 When I raised a good point during the discussion, the teacher said, "_____."

One You can do that again 2 You can say that again 3 Look on the bright side 4 You should call a spade a spade	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

19 Neither Grace nor her sons wanted to expose _____ to danger.

One herself 2 himself 3 himself or himself 4 ThemSelves	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

20 Isn't there any bank near by? _____.

One Yes, there isn't 2 No, there is 3 No, there isn't 4 Yes, it is	Choose the option that correctly answers the question.

Previously response sheet One 2 3 4 to the

21 The way Tamara is living, she _____ all her pocket money by the end of next month.

One will spend 2 will have been spending 3 will have spent 4 will be spending	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

22 He was _____ for his hard work.

One rewarded	Choose the option that best completes the sentence
2 awarded	
3 repelled	
4 compensated	

Previously response sheet One 2 3 4 to the

23 As I write this test, the sun _____ setting.

One will have been	choose the option that best completes the sentence
2 was	
3 is	
4 has been	

Previously response sheet One 2 3 4 to the

24 I have been extremely busy today. _____ since morning.

One It's been one thing after another	Choose the option that best completes the sentence.
2 I have been on the loose	
3 I've been on the run	
4 I have been on the go	

Previously response sheet One 2 3 4 to the

25 You do realize that this will not benefit you, don't you? _____.

One No, I don't

2 No, I do

3 No, it will

4 Yes, it won't

Choose the option that correctly answers the question.

Previously response sheet One 2 3 4 to the

26 Any score below 40 percent doesn't _____ to get you into college.

One supplement

2 suffice

3 sufficient

4 complement

Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

27 The professor's speech was _____: we learned a lot from her.

One an eye-opener

2 mind-numbing

3 out of order

4 a perfect storm

Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

28 You don't have a problem, do you? _____.

One No, I do 2 Yes, I don't 3 Yes, I do 4 Yes, I haven't	Choose the option that correctly answers the question.

Previously response sheet One 2 3 4 to the

29 Thieves stole a lot of money _____ my grandmother's house.

One from 2 in 3 to 4 into	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 to the

30 What she is doing will soon land her _____ trouble.

One on 2 in 3 for 4 onto	Choose the option that best completes the sentence.

Previously response sheet One 2 3 4 complete

COMMUNICATION SKILLS – CAT FORMAT

THANK YOU FOR TAKING PART IN THIS STUDY.

If you have questions at any time about this study, or you experience adverse effects as the result of participating in this study or your rights as a participant have been violated, you may contact the researcher Mr Thokozani Chisale {+265(0) 882 579 775} OR the supervisor Dr. Foster Gondwe { +265(0) 888 123 961 or fgondwe@unima.ac.mw } OR UNIMAREC Chairperson Dr. Victoria Ndolo, Chairperson of University of Malawi Research Ethics Committee (UNIMAREC), PO Box 280, Zomba. +265 995 0427 60.

check

Appendix 9: Curriculum Vitae for Research Assistant.

CURRICULUM VITAE

PERSONAL INFORMATION

Name:	Humphrey Kunyenge
Date of birth:	10th May 1986
Marital status:	Married
Religion:	Christian
Nationality:	Malawian
Home Address:	Chiradzulu District, T.A Mpama, Malindi 2
Email address:	kunyengeh@gmail.com
Contact number:	0999480138/0999123795
Contact address:	C/o B. Mbirika: Department of Policy and Planning Ministry of Transport and Public Works, Private Bag 322. Lilongwe

PERSONAL STATEMENT

Dependable well trained Water Resources Manager to work in high-stress environments and stay calm under pressure. I do examine sources of water and make effective conservation decisions in maintaining a balance between fragile ecosystems that need clean water and the societal necessity for water resources in many aspects of modern life.

EDUCATION BACKGROUND

2013: Bachelor of Science (Water Resources Management and Development) Mzuzu University

2009: Diploma in Irrigation Technology from Natural Resources College

2005: Malawi School Certificate of Education from Dedza Secondary School

PROFESSIONAL EXPERIENCE

1. Water and Irrigation Technician

Irritech Solution & Water Supply Position

- Irrigation system design and installation
- Borehole drilling and development
- Rehabilitation of irrigation structures
- Rehabilitation of water supply structures

2. Section Supervisor

Illovo Sugar Company Position

- Planning daily operations from weekly resource plan
- Supervision of cane growing operations in the section
- Supervision of all irrigation activities
- Handling staff grievances and disciplinary issues in the section
- Submitting a week and monthly report to the manager of all operations

OTHER COMPETENCEIES

- Quick learning skills
- Good analytical and problem solving
- Self-starter and able to work under minimal supervision.

REFEREES

Mr 0. Jacobs

Irritec Solution and Water Supply Blantyre

Cell: +265993441553

Mr. Jimmy Tamani

P. O. Box 5144

Blantyre

Cell: +265993713285

Mr R. Sopa

Farm Manager

Illovo Sugar Company

Private Bag 50

Blantyre

Cell: +265999221432