INVESTIGATING FAIRNESS OF MALAWI NURSES' LICENSURE EXAMINATIONS THROUGH TEST SCORE EQUATING: THE CASE OF SELECTED NURSES' TRAINING COLLEGES UNDER THE CHRISTIAN HEALTH ASSOCIATION OF MALAWI (CHAM)

M.Ed (TESTING MEASUREMENT AND EVALUATION) THESIS

ISAAC MATIAS NYIRONGO

UNIVERSITY OF MALAWI

JANUARY , 2025

INVESTIGATING FAIRNESS OF MALAWI NURSES' LICENSURE EXAMINATIONS THROUGH TEST SCORE EQUATING: THE CASE OF SELECTED NURSES' TRAINING COLLEGES UNDER THE CHRISTIAN HEALTH ASSOCIATION OF MALAWI (CHAM)

M.Ed (TESTING MEASUREMENT AND EVALUATION) THESIS

By

Isaac Matias Nyirongo

B.Ed. (Sciences) –University of Malawi

Submitted to the Department of Education Foundations, Faculty of Education, in Partial fulfillment of the requirements of a Master of Education (Testing, Measurement, and Evaluation)

UNIVERSITY OF MALAWI

JANUARY 2025

DECLARATION

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

Isaac Matias Nyirongo

_____

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

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

Signature:_____Date:_____

Gerson Mutala Phiri, PhD (Research Fellow)

Main Supervisor

Signature:_____Date:_____

Yohane Chakasika, M.Ed (Senior Lecture)

Postgraduate Coordinator (Education Foundation)

DEDICATION

I dedicate this Study to the loving memory of my dear mother.

ACKNOWLEDGEMENTS

I extend my sincere gratitude to my supervisor, Dr. Gerson Mutala Phiri for his invaluable guidance and unwavering encouragement throughout the research period. His expertise and support were instrumental in the success of this study, and I am truly thankful for his mentorship.

I would also like to express my appreciation to the 2021 TME students who generously shared their insights and experiences, contributing significantly to the progress of this research. Their assist with various inquiries greatly enriched the study, and their contributions are sincerely acknowledged.

Special thanks are extended to Mr. Joseph Malamba and all my dedicated workmates for their continuous encouragement and support. Their belief in my abilities and steadfast encouragement helped me navigate through challenges and remain focused on the study's objectives.

Lastly, I am deeply grateful to God for His boundless love and guidance throughout this journey. His blessings and divine intervention have been evident, and I am humbled by His grace.

ABSTRACT

This study was carried out to investigate the fairness of Nurses' Licensure Examinations (NLE) in Malawi by examining potential biases due to differences in test forms, specifically through score equating. Nurses' Licensure Examinations (NLEs), administered biannually, involve different test forms each year, which may vary in difficulty and psychometric properties. This discrepancy in test forms can lead to unfair comparisons between cohorts, as the difficulty of the test impacts pass rates and classification into grade categories. The study aimed to assess the unidimensionality of the exams, the significance of mean score differences across test forms, and the impact of equating on student classifications. Using the 2020 and 2021 Nurses' and Midwives Technicians (NMT) Examination Paper 1 forms, the study employed Statistical Package for Social Sciences (SPSS) for factor analysis and T-tests, and the R-EQUATE package with log-linear smoothing for equipercentile equating. Results revealed that both test forms were not unidimensional, with no dominant factor identified. No significant difference in difficulty was found between the two test forms. However, before equating, the pass rate for the 2021 test form was 76.34%, compared to 91.40% for the 2020 form, reflecting a 15.06% difference. After equating, the 2021 pass rate improved to 83.87%, reducing the gap to 7.53%. These findings underscore the importance of equating test scores to ensure fairness and validity in comparisons as unadjusted results can disadvantage students who take a more difficult version of the test.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ACRONYMS AND ABBREVIATIONS

CB:	Counterbalanced
CHAM:	Christian Health Association of Malawi
CTT:	Classical Test Theory
EE:	Equipercentile Equating
HSSP:	Health Sector Strategic Plan
ICC:	Item Characteristic Curve
IRB:	Institution Review Board
IRT:	Item Response Theory
LE:	Linear Equating
MANEB:	Malawi National Examinations Board
ME:	Mean Equating
MLE:	Medical Licensing Examination
NEAT:	Non – Equivalent Anchor Test
NEB:	Nurses' Examinations Board
NECO:	National Examinations Council
NLE:	Nurses' Licensure Examinations
NMT:	Nurses' and Midwives Technician

NMT:	Nursing and Midwives Technician
OSE:	Observed Score Equating
PCA:	Principle Factor Analysis
PSLCE:	Primary School Leaving Certificate of Education
RG:	Random Group
SG:	Single Group
SPSS:	Statistical Package for Social Sciences
TFF:	Test Fairness Framework
TIMSS:	Trends in International Mathematics and Science Study
UNIMAREC:	University of Malawi Research and Ethics Committee
WAEC:	West Africa Examinations Council

CHAPTER 1

INTRODUCTION

1.0 Chapter Overview

This Chapter presents the background of the study. The Chapter then presents the statement of the problem, the purpose of the study and the research questions. Finally, the significance of the study, its limitations and the definition of the operational terms are presented.

1.1 Background of the study

Comparison of test scores obtained from different test forms has been a center of attention in psychometrics. Examination agencies, policymakers, media houses, and the public at large make high-stakes decisions (e.g, admissions, placement, certification, diagnosis) based on test scores. Sanzivieri et al (2017) found out that examination agencies administer new editions of tests over a specified period mainly for security purposes. They cannot use the same test form on different administrations. However, Dorans, et al (2010) observed that even if different test editions may be built to a common blueprint and be designed to measure the same constructs, they always differ in their psychometric properties.

While some test forms consist of easy items, others may have difficult items that can cause examinees' scores to differ. Unfortunately, Chulu and Sirec (2011) noted that in some cases educational tests are not statistically equated to account for test score differences over time,

leading to wrong interpretations of students' performance. Without any doubt, some examinees who might have passed can unfairly be failed based on being compared to other examinees who have written a test whose statistical properties are different.

Although the health sector administers equally important high–stakes examinations, the use of test score equating to guarantee fairness has been overlooked. Langer and Swanson (2010) claim that one of the most important psychometric requirements of progress testing in medical education assessment has typically been neglected in past applications is that scores across the time and test forms are not commonly placed on the same scale. As such, there is limited information regarding the fairness administered in the health sector.

Since the Malawi Government adopted the Health Sector Strategic Plan (HSSP I, 2011 - 2016) whose goal was to improve the quality of life of all the people of Malawi by reducing the risk of ill health and the occurrence of premature deaths, thereby contributing to the social and economic development of the country, the Ministry of Health is challenged to cast its nets wider to achieve this goal (HSSP I, 2011 – 2016).

The Nurses and Midwives Council of Malawi is the sole regulatory body of nursing and midwifery education, training, practice and professional conduct of nursing and midwifery personnel in Malawi established in 1966 under an Act of Parliament and Laws of Malawi and Cap 36:02 (Nurses and Midwives Act, 2014). This aims to confirm that the nurses are competent and safe to practice nursing and midwifery. Success in the nursing and midwifery licensure examination is the only legal prerequisite to practice as a nurse and midwife in Malawi (Nurses & Midwives Council of Malawi, 2012)

1.2 Statement of the problem

Nurses' Licensure Examinations (NLE) are high-stakes tests that determine the eligibility of nurses and midwives to practice in the healthcare sector in Malawi and globally (Price et al., 2018). These criterion-referenced exams award a pass based on candidates' ability to demonstrate a required level of knowledge and skill (Yim & Huh, 2006). In Malawi, the NLE is administered twice a year by the Nurses' and Midwives' Council, and the results are often publicly scrutinized. Comparisons between cohorts based on pass rates are common, but such comparisons fail to account for the differences in test form difficulty, which can lead to unfair conclusions about candidates' abilities.

The NLE uses varying test forms in each cohort to mitigate item exposure and accommodate changing examinee populations. However, these different test forms may not be psychometrically equivalent, complicating comparisons of cohort performance. For example, Sanagala (2017) highlighted that discrepancies in pass rates between cohorts, like the 15% pass rate for the 2015 cohort versus the 25% rate for the 2016 cohort, could be due to test difficulty rather than candidate ability. Therefore, it is critical to consider the impact of these differences on the fairness of pass rate comparisons.

In Malawi, there is a significant gap in research on the use of score equating for professional exams like the NLE. Chakwera, Khembo, and Sireci (2004) noted that while score equating is common in Europe and North America, it is underutilized in many African countries, despite the high stakes of these exams. Similarly, Holmes (1986) pointed out that the use of test score equating in professional licensure has received limited

attention. This gap in research necessitates the application of score equating to address fairness concerns in Malawi.

Few empirical studies on test score equating have been conducted in Malawi and Sub-Saharan Africa, particularly in the context of professional licensure exams. A study by Chakwera et al. (2004) in Malawi showed that fairness in large-scale assessments remains a challenge, and Mkandawire et al. (2015) in Zambia highlighted the need for similar research in the health sector.

This study aims to fill this gap by investigating the fairness of NLEs through test score equating. It hypothesizes that some nurses and midwives may unfairly fail the NLE due to comparisons with candidates who sat for different test forms with distinct psychometric properties. Using data from the Christian Health Association of Malawi (CHAM) licensure examinations, administered by the Nurses' Examinations Board (NEB), this study seeks to provide a comprehensive analysis of score equating, offering valuable insights into the fairness of the licensure process.

1.3 Objectives of the Study

1.3.1 The main Research objective

To investigate the fairness of Nurses' Licensure Examinations (NLE) in Malawi through test score equating, using Nurses and Midwives Technician (NMT) 2020 and 2021 paper 1 test forms. Specifically, the study will examine whether differences in test forms across cohorts may impact the fairness of the results, particularly in terms of psychometric properties, and how these factors could affect candidate performance

1.3.1 Specific Research objectives

To achieve this, the following specific Research objectives were investigated:

i. To measure the unidimensionality of the Nurses' Licensure Examinations

ii. To compare the relative difficulty of the different test forms.

iii. To examine the presence of bias in the classification of students into grade categories before and after equating.

iv. To determine the interchangeability of scores from the two test forms.

v. To measure the correlation between the test forms to assess their equivalence

1.4 Research questions:

1.4.1 Main Research question

To what extent does test score equating using parallel test forms ensure fairness in Malawi's Nurses' Licensure Examinations? Specifically, how does it address biases related to differences in psychometric properties across test editions?

1.4.2 Specific Research questions

The following were the specific Research questions for the Study

i. To what extent do items on the test forms manifest unidimensionality?

ii. How do indices of item difficulty across test forms differ?

iii. How does the classifications of students into grade categories across forms before and after equating cause inequalities?

iv. How do the mean scores obtained from each test form compare?

v. What is the degree of correlation between the test forms in order to assess their interchangeability?

1.5 Significance of the study

Test fairness studies are important for several reasons, and this particular study is significant in the following ways:

Firstly, the Study provides research-based information on the fairness of licensure examinations in Malawi. This information will be helpful to other researchers who would want to conduct related studies in the future. As commended by Fraenkel and Wallen (2000) that before planning the details of a study, researchers usually dig into the literature to find out what has already been written on the topic to be investigated. The information would not only help researchers gather the concepts of others in particular research, but also allow them to learn about the results of other similar studies.

Secondly, results from this study informs authorities in the health sector especially the credentialing board members and other individuals who hold major responsibility for preparing, administering, and scoring credentialing examinations as well as other stakeholders in credentialing health professionals on best practices when administering nurses' licensure examinations.

In this regard, the results of the current study provides a research-based framework to authorities in the health sector through the Nurse's Examination Board on how best to improve the administration of fair licensure examinations across cohorts in Malawi. Such information is critical in pursuit of raising the standards of health systems in Malawi and beyond.

Thirdly, it is anticipated that by equating NLE scores for each test administration, the fairness of the licensure process will be improved, ensuring that nurses are not unfairly judged based on differences in test difficulty between cohorts. This approach could help address the issue of nurses being incorrectly assessed due to cohort-specific variations, ultimately contributing to a more accurate reflection of their qualifications and helping to alleviate the shortage of nurses in Malawi hospitals.

A study by Perera et al (2015) revealed that the doctor-to-population ratio in Malawi is 0.2:10,000, and the nurse-to-population is 3.4:10,000. Perera (2015) further argues that this nursing ratio is one-third of the World Health Organization's (WHO) recommended ratio of 10:10,000 people. It can be noted that the situation is a serious one that needs various interventions to mitigate the challenge.

Last, but not least, the study contributes to the efforts in reducing suicide cases in the health sector in Malawi. High suicide rates have been reported over the past three years. A study by Mwale and Mafuta (2017) revealed that 9 out of 100,000 people commit suicide in Malawi, compared to the global rate of 11 out of 100,000. It has been estimated that up to 90% of suicide attempters had been depressed before (Brendel et al., 2010; Mann, 2003).

Again, Atemafac (2014) pointed out that the consequences of failing the Licensure Examinations are terrible and extensive for the student emotionally and financially. On the other hand, McCumpsey (2011) found out that the emotional impact of failing Licensure Examinations is usually devastating to the candidate's employment potential and financial situation. Licensure examinations being one of the high–stakes examinations have the potential to frustrate students when failed to the extent of some of them committing suicide.

1.6 Limitations of the study

This study was based on the Observed score equating methods, adopting the Classical Test Theory (CTT) equating models, hence used observed scores to draw its conclusions. As such the results from this study would not be generalized to the true scores of examinees that need the Item Response Theory (IRT) equating methods. Similar studies have to be done to establish the trends of such scores.

Secondly, the study used data drawn from the health sector, hence, other studies have to be done to understand the situation in other sectors that administer licensure examinations such as transportation and communication. Different sectors may have unique challenges, standards, and procedures in their licensure processes, which could affect how fairness and test validity are perceived and measured. As a result, the conclusions drawn from this study may not fully reflect the circumstances or practices in other sectors that administer licensure exams

Thirdly, even though equating scores do not require validity to be a prerequisite, the study assumed that the nurse's examinations are valid, hence issues to do with validity was not covered in the study. Further studies have to be undertaken to understand the extent of the validity of nurses' examinations.

In addition, the study assumed that the subjects were sitting for Licensure Examinations for the first time. Hence, the study did not take into account the effect of subjects who were sitting for the Licensure Examination for the second time or more during data analysis. The effect of repeaters on the performance of Licensure Examinations is left for future researchers to explore.

Finally, much that comparing equated scores in a testing procedure might provide an in-depth picture of the fairness of Licensure Examinations, this study only focused on establishing the presence of bias in Licensure Examinations. Studies aiming at comparing equated Licensure Examination scores are left for future researchers to explore

1.7 Definitions of operational terms

For a better understanding of this study, the following terms were defined in the context of this research:

Examination Blueprint. Refers to a template used to define the content of an examination (Sales, Sturrock et al. 2010). This can take the form of a table in which the axes are labeled content area and competency area.

Nurses' Licensure Examinations. These are examinations taken by nurses and medical doctors close to the point of graduation from medical school (Price et al, 2018). The examinations are used to determine whether an applicant is qualified for licensure by an occupational board.

Parallel Test Forms: These are different subsets of the same universe of items, which capture the same attribute with the same accuracy (Hilger & Beauducel, 2017). They are interchangeable versions of a test in terms of construct and content and equivalence in test performance of test takers with similar abilities across test administrations.

Score Equating. This is a statistical process used for adjusting scores obtained from test forms so that the scores can be used interchangeably (Kolen & Brennan, 2004). After equating, the scores can be used as if they came from the same test.

Smoothing. This is a process that is used to produce a new observed-score distribution by eliminating irregularities without changing the distribution's range, shape, or location (Livingston, 2004). The main aim of conducting smoothing is to minimize the sampling errors.

Test Fairness. The impartial treatment of all test takers during the testing process, absence of measurement bias, equitable access to the measured constructs, and justifiable test score interpretation validity for the intended purpose(s) (AREA, APA, & NCME, 2014).

1.8 Chapter summary

This Chapter introduced the study's focus on fairness in Nurses' Licensure Examinations (NLE) in Malawi. The Chapter then outlined the statement of the problem and the purpose of the study. Furthermore, the Chapter detailed the objectives such as examining test unidimensionality, comparing test difficulties, assessing grading bias, and evaluating score interchangeability using data from parallel test forms before presenting the research questions. It emphasized the study's significance for health sector policies while acknowledging limitations related to methodology and scope concerning nursing licensure examinations. The Chapter concluded by addressing the study's limitations and defining key operational terms

CHAPTER 2

LITERATURE REVIEW

2.0 Chapter Overview

This Chapter discusses related literature to the concept of Test score equating. The Chapter then discusses the reasons for multiple equating, horizontal and vertical equating. The Chapter further discusses the conditions for equating, observed score equating methods (OBE), conditions, and their assumptions. The Chapter then presents the equating designs and smoothing methods. Finally, the score-equating studies are also briefly reviewed

2.1 Concept of test score equating

Different scholars have defined test score equating in various ways. Kolen and Brennan (2004) define test score equating as the statistical process used for adjusting scores obtained from test forms so that these scores can be used interchangeably. Similarly, Crocker and Algina (1986) have defined test equating as a process that establishes equivalent scores from two different measurement instruments.

The definitions mean that test score equating is a process that establishes equivalent scores from two different measurement instruments (Crocker & Algina, 1986). They argue that when the percentiles corresponding with the X and Y scores obtained from different tests that have equal reliability and measure the same construct are equal, the tests that these X and Y scores were obtained from are equal. This is a statistical process that is applied to

confirm that scores on different test forms are comparable. Through equating, scores on one test are statistically adjusted for the difficulty to the level of scores on another test (Livingston, 2004; van Davier et al., 2004).

2.2 Parallel test forms

Parallel forms of a test are different subsets of the same universe of items, which capture the same attribute with the same accuracy (Hilger & Beauducel, 2017). Parallel tests need to be comparable or equivalent because the incomparability of parallel tests is likely to cause fairness issues and thus erode the value of tests. As Wendler and Walker (2015) indicated, parallel tests must be equivalent to ensure the interchangeability of test scores. However, there is a lack of evidence for the comparability and equivalence of parallel tests and testing agencies are usually criticized for failure to provide such vital evidence (Bachman et al., 1995; Chalhoub-Deville & Turner, 2000; Weir & Wu, 2006).

Evidence for the comparability of parallel tests usually includes both test construct and content comparability and test score equivalence (Bae & Lee, 2010; Wendler & Walker, 2015). Test score equivalence can be examined by examining test takers' performance on different parallel tests (Bae & Lee, 2010; Weir & Wu, 2006). Although the use of parallel test forms seems to be a reasonable way to ensure fairness (Kan, 2010) and exam security, the issue of the comparability of the scores obtained from these different tests is a source of concern.

2.3 Reasons for multiple testing

Sanzivieri et al (2017) identified at least three reasons for agencies that administer high – stakes examinations to have multiple forms of a test (and consequently equating). The first

is security. The same test form cannot be administered on two different testing dates. Testing programs administer high-stakes examinations in which performance has an important impact on the examinee and the public: conferring a license or certificate to practice a profession, permitting admittance to a college or other training program, or granting credit for an educational experience.

A second reason is a current movement to open testing. Braun (1982) argued that many programs find it necessary and desirable to release test items to the public. When this occurs, the use of the released items on future test forms will provide some examinees an unfair advantage.

A third reason for administering different test forms is that of test content, and therefore test items, by necessity change gradually over time. However, the use of different test forms on different dates raises concerns over whether the difficulty level of these forms differs (Kolen & Brennan, 2014). If no adjustment for difficulty differences is made, it is impossible to fairly compare test-takers who have taken different test forms and interpretations from the scores will be unfair to one group

2.4 Horizontal and vertical score equating

Cook and Eignor (1991) categorize equating into horizontal and vertical. Horizontal equating is appropriate when several forms of tests are needed for the security of the tests. These forms are not the same, but they are expected to be similar in their content and difficulty. When the difficulty, reliability, and content of tests are so different from one form to another, only a few equating methods can properly work (Cook & Eignor, 1991). Furthermore, it is expected that examinees must have equal ability levels. When the ability

levels are very diverse, linear equating and equipercentile equating cannot be used (Kolen & Brennan, 2004).

On the other hand, vertical equating is concerned with equating test scores of two tests that are deliberately set with different difficulty levels and yet measuring the same broad realm of knowledge. Furthermore, this equating procedure differs from horizontal equating because the distribution of abilities between examinees is different from one level to another. Vertical equating is out of the scope of the current study as it limits its discussion to horizontal equating.

2.5 Observed score equating methods

There are several methods of equating scores, some of which are within classical Test Theory (CTT) while others are within Item Response Theory (IRT). Those within the CTT are usually called the methods of observed score equating (OSE) which have been discussed extensively by Kolen and Brennan (2004) and Livingston (2004). In their discussion they presented three methods of observed score equating, namely: (1) mean equating (ME); (2) linear equating (LE), and (3) equipercentile equating (EE). These methods differ in the way each one defines relative positions (Chulu and Sireci, 2011). They are further discussed below.

2.5.1 Mean equating

Mean equating defines relative position in terms of the number of points above or below the mean in the target population of examinees (Livingston, 2004). Therefore, in mean equating, equivalent scores are obtained by setting equal scores on the two equal test forms (assigned) distance away from their respective means. One test form is considered to differ

in difficulty from another test form by a constant amount along the score scale (Kolen & Brennan, 2004).

$$y(x) = x + (\mu_y - \mu_x) \quad (1)$$

Where *y(x)* is a function that transforms scores of Form X to the scores on Form Y, *x* and *y* are the raw scores μ_x and μ_y are the means of Form X and Form Y respectively. In mean equating, the difference between the means of the two populations who take the two forms of a test is computed. This difference is then added to the scores of all the examinees who have taken the harder test form or subtracted from the score of those who have taken the easier version (Kolen & Brennan, 1995). Mean equating assumes that differences in difficulty between the two forms are constant throughout the entire score range (Barnard, 1996). Thus, this method considers that Form X is differentiated by Form Y in difficulty by a constant amount over the score scale.

2.5.2 Linear equating

In contrast to mean equating, linear equating defines relative position in terms of both the mean and standard deviation. Angoff (1984, p. 564) defined linear equating as scores being equivalent when the scores on two test forms correspond to the same standard – score deviations. It allows for the test forms to be differentially difficult along the score scale (Kolen & Brennan, 2004).

Equivalent scores are obtained by transforming scores on the new form to scores on the old form that are the same number of standard deviations above or below the mean of the group (Livingston, 2004) – setting the standardized deviation scores (z-scores) on the two tests

to be equal. Concerning Form X (new form) and Y reference (old form), μ_X and μ_Y gives means of the forms and S_X and S_Y gives standard deviations of the forms. On this basis, Kolen and Brennan (2004) define linear equating using the following equation:

$$y(x) = \frac{S_Y}{S_X}x + \left[\mu_x - \frac{S_Y}{S_X}\mu_x\right] \quad (2)$$

2.5.3 Equipercentile equating

In equipercentile equating, percentile ranks for each form are first calculated. Scores that have the same percentile rank are taken to be equivalent (Kolen, 1988; Kolen & Brennan, 2004; Livingston, 2004). This procedure uses percentile rankings to scale scores from test form X to the scale of test form Y. A score of *x* on form X equates to a score of *y* on form Y provided they have the same percentile rank (Petersen, Cook, & Stocking, 1983).

Equipercentile equating involves four main stages. Determine the percentile ranks for the score distributions on each of the two instruments. Percentile ranks are then plotted against the raw scores for each of the two instruments A percentile rank-raw score curve is then drawn for each instrument. Equivalent scores can then be obtained from the graph.

This study adopted the equipercentile procedure provided by Hanson et al (1994). However, for an original version of the procedure, refer to Holland and Thayer (1989). According to Hanson et al (1994), in the random groups equating design the new and old forms are each administered to a random sample from a common population. Let the random variables X and Y represent the test scores on the new and old forms of the test, respectively, for a random examinee from the population of interest.

The test score X is to be equated to the test score Y. The equipercentile equating function is determined by the cumulative distribution functions of X and Y. If the random variables X and Y were continuous, then, Hanson et al (1994) defined the equipercentile equating function:

$$\mathrm{F_Y^{-1}[F_X(x)]} \tag{3}$$

Where $F_Y(y) = P_r(Y < y)$ and $F_X(x) = P_r(X < x)$

Because X and Y are discrete random variables the equipercentile equating function is not defined. To define an equipercentile equating function based on X and Y the common practice is to use the equipercentile equating function based on continuous approximations of X and Y. The most widely used continuous approximation is based on a uniform kernel being applied to X and Y to produce approximating continuous distributions (Holland & Thayer, 1989).

The uniform kernel spreads the density at each score point uniformly in a unit interval one-half point above and below the score point. This results in a continuous distribution on the interval (-1/2, K +1/2), where K is the number of items on the test. Based on the continuous distribution given by the uniform kernel, Hanson et al (1994) defined the equipercentile equivalent of raw score on the new form as follows:

$$e_Y(X) = \frac{p^*(i) - P_r(Y < u^*(i))}{P_r(Y = u^*(i))} + u^*(i) - 0.5$$

Where $u^*(i)$ is the smallest integer such that $p^*(i) < P_r(Y \leq u^*(i))$

In this study equipercentile equating had been chosen for equating scores on the tests. As Chulu and Sirec (2011) noted that equipercentile method has three advantages over linear

equating in that: (1) it is based on a better definition of the "relative position" of a particular score in the distribution of scores than the linear and mean equating; (2) it takes into account the possibility that the target population's score distributions on the new form and on the old form may have different shapes; and (3) it minimizes the problem of out-of-range adjusted scores (Chulu & Sirec, 2011)

2.6 Conditions for observed score equating

For the scores from two test forms to be equated based on Classical Test Theory (CTT), they have to satisfy four conditions of equating. The conditions are equal construct, equal reliability, symmetry and equivalent difficulty levels (Angoff, 1984; Dorans & Holland, 2000; Kolen and Whitney, 1982).

A test is assumed to be unidimensional only when the individual items in the two tests measure the same trait (Hambleton et al., 1991). That is, individual test items in the test form should measure one thing. This means that the test developer oftentimes works with a test blueprint to assure that each of the forms meet set specifications and contain the same format of items measuring the same construct. This includes making the tests the same length, approximately the same difficulty, and designed for the same audience.

Equal reliability needs that the test must have the same level of reliability. Lord (1980) pointed out that scores X and Y on two tests cannot be equated unless either (1) both scores are perfectly reliable or (2) the two tests are strictly parallel. According to the property of the same specifications, the test forms to be equated are required to have the same content and statistical properties. The scores obtained from an equation that ignores these statistical properties cannot be used interchangeably (Kolen & Brennan, 2004).

The third requirement as argued by Dorans and Holland (2000) and Kolen and Brennan (2004) is that of symmetry which assumes that the equating function for *X* to *Y* must be the inverse of the function for *Y* to X. To elucidate further, if a score on form *X* equates to a particular score on form Y, then the score on *Y* should equate back to the original score on X. For instance, if a score of 150 on form *X* equates to a 180 on form Y, then a score of 180 on *Y* should reverse-equate to a 150 on form X.

The property of symmetry differentiates equating from prediction, and statistics such as regression since these are not necessarily symmetric in nature. To check for this property, an equating of Form X to Form Y and an equating of Form Y to Form X could be conducted. If these equating relationships are plotted, then the symmetry property requires that these plots be indistinguishable.

Again, equality pursues a lack of difference resulting from taking form X or Y of individuals. In order for equity to be achieved, the tests must be measures of the same construct or characteristic (Dorans, 1990).

Although the same construct is a prerequisite for equity, it does not ensure equity. Tests forms measuring the same construct may differ in terms of difficulty and other psychometric characteristics. For instance, test X may be easier than test Y. So, if test X and test Y measure the same construct, examinees would decide to take the easier test X because they would get higher scores on it.

In this study unidimensionality of each form was checked by conducting the Factor Analysis (FA) using Statistical Package for Social Sciences (SPSS). The assumptions of

equal reliability equivalent difficulty levels were checked by examining the Cronbach alpha (α) and the z – transformation statistic.

2.7 Limitations of equipercentile equating

One limitation of equipercentile equating according to Livingston (2014) is that the equating relationship cannot be determined for the parts of the score range above the highest score observed and below the lowest score observed. But, it is not usually a problem for very low scores, because test users rarely need to discriminate at score levels below the lowest score observed. However, it can be a problem at high score levels on a difficult test, because some future examinees may get a raw score higher than the highest score in the data used for the equating.

This problem was solved by smoothing scores because many smoothing methods produce a smoothed score distribution with nonzero probabilities (possibly very small, but not zero) at the highest and lowest score levels, even if no test takers actually attained those scores (Livingston, 2014). However, the equating relationship computed from the smoothed distributions at those very high and very low score levels will be based on scores that were not actually observed

2.8 Equating errors

Equating being a statistical procedure, one important aspect to take into account is statistical errors. Equating errors according to Kolen and Brennan (2004) are divided into two sources, random error and systematic error. A random equating error (sampling error) occurs when the parameters of a sample that are drawn from the whole population, such as the mean, standard deviation, and percentile rank, are estimated (Kolen, 1988).

Random errors may also be defined as the difference between the estimated equating relationship for the samples and for the whole population (Aşiret & Sünbül, 2016). Fortunately, Kolen (1988), pointed out that random sample errors can be minimized by increasing the sample size and selecting an appropriate equating design. Consequently, Kolen and Brennan (2004) pointed out that when the whole population was available during equating, no random errors would be present.

Systematic errors occur when there are violations of the statistical assumptions or conditions of the equating methods (Aşiret & Sünbül, 2016). For instance, in the single-group design, an examinee failing the exam because of fatigue or getting a high score due to practicing results in systematic errors. Again in a random group design, if the spiraling process is unable to group comparably, systematic errors arise. As a result, Kolen (1988) pointed out that if Form X and Form Y differ in difficulty, content, and reliability, systematic errors can be concluded to appear.

2.9 Equating designs

Selecting an equating design is one of the most important steps of test form equating. A variety of designs are used for collecting data for equating, and the choice of a design involves considering both practical and statistical issues (Kolen, 1988). An equating design refers to the basic structure of an equating study, just as a research design refers to the structure of a research study (Albano, 2010).

There are four basic designs discussed in the literature: single group (SG), counterbalanced (CB), random group (RG), and non – equivalent group with anchor test (NEAT) (Crocker & Algina, 1986; Kolen & Brennan, 2004). For purpose of the research presented here, this

chapter will only focus on random group design, but for the sake of comprehensiveness the other three designs, SG, CB, and NEAT are briefly discussed.

2.9.1 Single – group design

In a single-group design, the same individuals are given both test forms X and Y. The order of test form administration remains uniform for all examinees. When the order of administration is alternated, then, the design is termed counterbalanced. Although, the design is simple and there are no errors arising due to the ability levels of individuals since forms are answered by the same individuals, fatigue and familiarity with the test are challenges that cannot be overlooked (Kolen & Brennan, 2004). Fatigue would make the second test more difficult and familiarity would make the second test form easier than it would be. For this reason, this equating design is rarely used in practice. Table 2.1 shows the SG design: one examinee group takes both test forms, X and Y (Godfrey, 2007, p. 11)

TABLE 2. 1: Single Group Design

	X	*Y*
Examinee Group	✓	✓

Source: Godfrey (2007)

2.9.2 Counterbalanced design

The counterbalanced design is very similar to the SG design. However, while the SG design does not take order effects into account in the administration of test forms, the CB does. Half of the examinee group is given test form X first and then test form Y second. The

other half of the examinee group takes the same two tests in reverse order. This design is illustrated in Table 2.2. (Godfrey, 2007, p. 12)

TABLE 2. 2: Counterbalanced design

	X_1	Y_2	Y_1	X_2
Examinee Group *P*	✓	✓		
Examinee Group *Q*			✓	✓

Source: Godfrey (2007)

2.9.3 Random groups design

In a random group design, examinees are randomly assigned the form to be administered. A spiraling process is one procedure that can be used to randomly assign forms using this design. In this method, Form X and Form Y are alternated when the test booklets are packaged. When the booklets are handed out, the first examinee receives Form X, the second examinee Form Y, the third examinee Form X, and so on. This spiraling process typically leads to comparable, randomly equivalent groups taking Form X and Form Y.

2.9.4 Non – equivalent groups design

A NEAT Design is often used when more than one form per test date cannot be administered because of test security or other practical concerns. In this design, Form X and Form Y have a set of items in common, and different examinee groups P and Q are administered the two forms. A group tested one year might be administered Form X and a group tested another year might be administered Form Y. In this regard, the target population, T, is the combination of P and, and is defined by (Godfrey, 2007)

$$T = wP + (1 - w)Q \quad (4)$$

If P and Q are equal in size, w is equal to 0.5 (Godfrey, 2007). The NEAT design is not applicable especially for research purposes because of the time factor. Again, to avoid the credibility issues for fatigue when using the single group designs the current study will adopt the random group design. **Table 2.3** demonstrates this design. (Godfrey, 2007, p. 12)

TABLE 2. 3: Non-equivalent Groups with an Anchor Test Design

		X	*Y*	*A*
T	Population *P*	✓		✓
	Population *Q*		✓	✓

Source: Godfrey (2007)

2.10 Smoothing

Samples for examinees who will take the test forms for equipercentile equating can be drawn from one or more populations. For this reason, some irregularities can appear as a result of sampling errors when the raw score distribution is graphed (Kolen & Brennan, 2004). These sampling errors can be minimized by increasing the sample size (Aşiret & Sünbül, 2016). Kolen and Brennan (1995) suggested a sample size of 400 per form for linear equating and a sample size of 1,500 per form is sufficient for the equipercentile-equating method. However, it may not always be possible to attain this sample size. To minimize these sampling errors, smoothing methods are used (Cui & Kolen, 2009; Donlon, 1984) such as the log-linear smoothing developed by Livingston (1993) and the collateral information method by Wingersky (1993).

The process that produces a new observed-score distribution by eliminating irregularities without changing the distribution's range, shape, or location is called smoothing (Livingston, 2004). The study used equipercentile equating, however, the researcher would not be able to attain a sample size of 1,500, due to financial and time limitations, as such, smoothing methods was used to minimize the sampling errors.

One type of smoothing that has numerous applications is Log-linear smoothing which is used mostly in educational assessment as a preliminary step in the equating of scores on different forms of a test. The log-linear smoothing is first applied and then the smoothed results can be used with nonlinear equating procedures such as the traditional equipercentile procedure or the kernel procedure (von Davier, Holland, & Thayer, 2004; Hanson, 1996; Holland & Thayer, 1989; Rosenbaum & Thayer, 1987). According to Holland and Thayer (1987, 2000), the polynomial log-linear method fits a model of the following form to the distribution:

$$\log[N_X f(x)] = \omega_0 + \omega_1 x + \omega_2 x^2 + \cdots + \omega_C x^C \quad (5)$$

When the mean and standard deviation of the distribution is preserved, then the model reduces this quadratic as follows

$$\log[N_X f(x)] = \omega_0 + \omega_1 x + \omega_2 x^2 \quad (6)$$

The ω parameters in the model can be estimated by the method of maximum likelihood. Holland and Thayer (1987) described the algorithms for maximum likelihood estimation with this method. Furthermore, the choice of C is an important consideration when using this method. The fitted distribution can be compared, subjectively, to the empirical distribution. One such method is to use goodness-of-fit statistical significance testing

methods. These procedures were articulated and investigated by Moses and Holland (2009a)

2.11 Previous studies on score equating

To understand the scope of score equating research, this section reviews key studies, categorized by study objectives and research questions, focusing on Classical Test Theory (CTT) and Item Response Theory (IRT). The ongoing debate between the superiority of CTT and IRT equating methods is reflected in various empirical studies that provide mixed results, with some studies favoring IRT (e.g., Peterson, Cook, & Stocking, 1983), others favoring CTT (e.g., Clemans, 1993; Kolen, 1981; Skaggs & Lissitz, 1986a), and some finding both methods produce comparable results (Skaggs & Lissitz, 1988). Researchers have suggested that no single method can be universally superior across all test types (Skaggs & Lissitz, 1986b).

2.11.1 Unidimensionality of Test Items

The unidimensionality of test items is a key assumption in both Classical Test Theory (CTT) and Item Response Theory (IRT), as it underpins the accuracy of the measurement process. Several studies have explored the unidimensionality of items across various test forms, with results supporting the consistency and comparability of both theories. For instance, Fan (1998) conducted a study in the United States using data from the Texas Assessment of Academic Skills (TAAS), a state-mandated criterion-referenced test. His study revealed very high correlations among both person parameters (all higher than 0.96) and item difficulties (all higher than 0.90), suggesting that both CTT and IRT approaches yield similar results when assessing unidimensionality. This finding indicated that the

assumption of unidimensionality was supported across the measurement frameworks, reinforcing the validity of both approaches.

Courville (2005) replicated Fan's study with a larger sample of 80,000 examinees who took the ACT Assessment. His findings were consistent with those of Fan (1998), further corroborating the notion that both CTT and IRT produced highly comparable estimates for person ability and item difficulty. This consistency across different test populations and sample sizes reinforced the robustness of the unidimensionality assumption in both frameworks. Building upon these findings, Tate and Baird (2014) conducted a study that examined item dimensionality in standardized testing in the context of higher education performance assessments. Their study found that both CTT and IRT methods revealed high consistency in factor loadings when testing for unidimensionality, supporting the assumption of unidimensionality for the majority of item sets. Specifically, the study reported that factor loadings for both CTT and IRT approaches were consistently above 0.90, suggesting a strong one-dimensional structure in the item sets used for standardized assessments. This study further supported the conclusion that both CTT and IRT can be relied upon to maintain unidimensionality in large-scale educational assessments, aligning with previous findings in K-12 testing contexts.

2.11.2 Item Difficulty

In terms of item difficulty, several studies have compared the performance of Classical Test Theory (CTT) and Item Response Theory (IRT) in assessing test difficulty. MacDonald and Paunonen (2002) conducted a Monte Carlo study at the University of Ontario, where they controlled the spread of item difficulty and item discrimination. Their

study found that both CTT and IRT produced highly correlated difficulty indices when the item spread was controlled, suggesting that both frameworks yield comparable results in terms of assessing test difficulty. Specifically, the correlation between the difficulty estimates from CTT and IRT was found to be greater than 0.95 when the spread of item difficulties was controlled. However, IRT's discrimination indices performed better when the spread of item difficulty values was small. On the other hand, the CTT discrimination estimates were less accurate in conditions where the spread of item difficulty values was large, suggesting that IRT might offer a more precise measurement in those contexts.

Gonzalez and Lemos (2016) further explored the comparison between CTT and IRT methods for assessing item difficulty in high-stakes testing. Their study found that both CTT and IRT produced similar difficulty indices across test forms, with a correlation of 0.97 between the two methods for item difficulty. However, in cases of extreme item difficulty, IRT provided a more nuanced interpretation of item parameters. Specifically, IRT was able to better differentiate between items with extreme difficulty levels, where the difficulty index from CTT might have been less discriminating. This finding reinforced the idea that while both methods can yield comparable difficulty measures, IRT may offer more detailed insights in certain testing scenarios.

2.11.3 Bias in Classification and Test Form Comparison

The impact of test form differences on student classification has been the subject of several studies, particularly in the context of equating methods. Ozdemir (2017) compared TIMSS mathematics subtest scores from two different years (2011 and 2007) using different nonlinear observed score equating methods under a Non-Equivalent Anchor Test (NEAT)

design. The study found that when using equipercentile equating, the equivalent scores for the TIMSS 2011 mathematics subtest ranged from -0.38 to 23.08. However, when raw scores were presmoothed before equating, the equivalent scores ranged from 0 to 23.08, indicating that equipercentile equating with presmoothing yielded more accurate results. Furthermore, when employing circle-arc equating methods, the equivalent scores for TIMSS 2011 ranged from 0 to 23 for both raw and presmoothed scores. Interestingly, the results showed that all raw scores for the TIMSS 2011 mathematics subtest were smaller than the equivalent scores for the TIMSS 2007 mathematics subtest based on circle-arc equating, suggesting that the TIMSS 2007 mathematics tests were easier than those from TIMSS 2011.

Similarly, Temitope (2021) compared the NECO and WAEC Chemistry tests in Nigeria and found that NECO test forms were easier than WAEC. The CTT-equated scores for examinees in NECO were higher, ranging from 5 to 56, compared to WAEC, where the scores ranged from 4 to 49. Likewise, the IRT-equated scores for examinees in NECO ranged from 13 to 53, while those in WAEC ranged from 11 to 48, demonstrating no statistical equivalence between the WAEC and NECO Chemistry examinations in terms of difficulty. These findings highlight the significance of using appropriate equating methods to ensure fairness in comparing test forms.

Wang and Li (2019) investigated differential item functioning (DIF) in parallel test forms and its implications for classification decisions. Their study showed that classification biases arising from test form differences could be corrected using both CTT and IRT methods, particularly through adjustments to the equating process. Specifically, their study

revealed that when DIF adjustments were made, the classification error rate decreased by 5-7% across both methods, with IRT showing slightly better correction of biases than CTT. This highlights the importance of using robust equating methods to ensure fairness and accuracy in student classification across different test forms.

2.11.4 Interchangeability of Scores across Test Forms

The question of interchangeability of scores between test forms is crucial for ensuring fairness in high-stakes testing. Research has consistently demonstrated that both Classical Test Theory (CTT) and Item Response Theory (IRT) yield highly interchangeable results across different test forms, though differences in the underlying statistical properties may arise. For instance, Progar et al. (2008**)** found that the person parameter estimates derived from both CTT and IRT were highly correlated, with correlations of 0.984 for the Math item pool and 0.990 for the Science item pool. Despite these high correlations, the study also highlighted that the distributions of item difficulties and discriminations differed between the two methods. This suggests that while the overall person parameter estimates are similar, the way in which item characteristics are interpreted might vary between CTT and IRT. Nevertheless, these results point to a high degree of interchangeability between the two approaches when it comes to estimating the ability of test-takers.

Similarly, Ndalichako and Rogers (1997) found an almost perfect correlation of 0.988 between the ability estimates obtained from CTT and IRT, reinforcing the notion of high interchangeability between the two methods. This study, conducted using data from a school-leaving reading comprehension exam in Canada, concluded that the ability

estimates produced by both frameworks were sufficiently comparable to allow for meaningful interchangeability of scores across different test forms.

Additionally, Bowers, A. J., & Pearson, M. (2015) conducted a longitudinal study examining the effects of score equating on student performance classification. The study found that both CTT and IRT methods led to highly interchangeable scores across different test forms. However, the authors noted that while both methods produced similar results overall, the choice of equating method could impact the interpretation of marginal scores, particularly in longitudinal studies where the focus is on tracking changes over time. Specifically, the study reported that the correlation between test forms was 0.986 for IRT and 0.983 for CTT, suggesting that, although both methods produce comparable scores, IRT's ability to capture more granular variations in item characteristics may offer a slight advantage in interpreting marginal scores. This underscores the importance of selecting the appropriate equating method depending on the context and the specific goals of the testing program.

2.11.5 Psychometric Properties and Fairness of Test Forms

Several studies have underscored the importance of equating procedures in ensuring fairness across different test forms. In their study, Chulu and Sirec (2011) applied equating procedures to the Primary School Leaving Certificate of Education (PSLCE) mathematics exams in Malawi. Prior to equating, they found a substantial difference in pass rates between the 2003 and 2004 test forms. Specifically, the pass rate for the 2004 test was **69.96%,** while that for the 2003 test was **81.41%,** representing a difference of **11.45%.** However, after equating the scores, it was revealed that 52 students who had taken the 2004

form would have failed if the difficulty level of the test had not been adjusted to match that of the 2003 test form. This significant discrepancy highlights the crucial role of equating in ensuring fairness, as without it, score interpretations and, consequently, the decision-making process could be deemed unfair.

Similarly, Baghaei (2010) found that the lack of equating in a reading comprehension exam led to inconsistent pass/fail decisions, further emphasizing the necessity of equating to ensure the fairness of score classifications. These studies underscore the critical importance of equating in maintaining the fairness and consistency of educational assessment processes.

Finally, Zhu and Liu (2020) conducted a study on medical licensure exams in China, examining the impact of test form differences on fairness. They discovered that without equating, the fairness of medical licensure exams was compromised due to significant disparities between the test forms. Specifically, candidates who took the easier test form scored significantly higher, with an average score difference of 15.7 points compared to those who took the more difficult form. This disparity resulted in substantial inconsistencies in candidate classification, with **22%** of candidates being incorrectly classified as either passing or failing due to the difficulty imbalance. Zhu and Liu (2020) concluded that these discrepancies could be rectified through robust equating procedures, which would help align the passing rates across test forms and, ultimately, ensure fairness in candidate evaluation. This study further emphasizes that without equating, inequities in the assessment process are inevitable, leading to unfair outcomes for candidates.

2.12 Gaps in the Literature

Although there is a growing body of literature on test score equating, particularly in the context of large-scale examinations, there are several gaps and weaknesses that necessitate this study on Nurses' Licensure Examinations (NLE) in Malawi. These weaknesses primarily stem from the limited application of test equating in the health sector, particularly in African countries like Malawi.

One major gap in the existing literature is the lack of equating studies specifically focused on health-sector licensure examinations in Malawi. While there has been substantial research on equating procedures in educational assessments (e.g., Chulu & Sireci, 2011; Chakwera et al., 2004), similar studies targeting professional licensure examinations in the medical field are scarce. This creates an information void regarding the fairness and validity of NLE results, which is crucial given the high stakes involved in such examinations for career advancement and healthcare quality.

Additionally, studies in other contexts, such as those by Ozdemir (2017) and Temitope (2021), highlight the widespread issue of unfair score interpretations caused by differences in test difficulty across forms, but there is limited application of these findings to the NLE context in Malawi. While the importance of equating procedures in examinations administered in the education sector is well-documented, especially in international studies, the application to medical licensure exams, which have different psychometric properties and consequences, has not been adequately explored.

Furthermore, the literature on score equating methodologies reveals a lack of consensus regarding the best approach to use (CTT vs. IRT), with some studies indicating comparable results between both methods (e.g., Fan, 1998; Ndalichako & Rogers, 1997). However, studies examining the specific characteristics of licensure examinations, particularly in African contexts, often fail to account for the unique challenges faced in such assessments, such as cohort differences, limited item pools, and the variability in test forms.

The absence of research on the fairness of NLEs in Malawi is compounded by the reliance on pass rate comparisons across cohorts, which can be misleading due to differences in test difficulty, as highlighted by Sanagala (2017). The absence of rigorous statistical tools like test score equating in Malawi's NLEs further exacerbates the problem, leaving the practice vulnerable to biased interpretations of examinee performance.

Given these gaps, this study seeks to fill the void by applying test score equating methods to assess the fairness of NLE results in Malawi, using real data from the Christian Health Association of Malawi (CHAM) licensure examination. By investigating whether the psychometric properties of different test forms lead to unfair pass/fail classifications, the study will contribute to ensuring that NLEs are equitable, reliable, and valid measures of a candidate's competence, thereby enhancing the integrity of Malawi's healthcare sector.

2.13 Theoretical frameworks

The study was based on two frameworks namely: the Classical Test Theory (CTT) and the Test Fairness Framework (TFF). The CTT is one of the earliest frameworks that conceptualized the nature of associations between measured values and target properties via mathematical models (Novick, 1966). CTT is based on the equation (X = T + E),

where *X* is the measured value, *T* is the true score that is the expected value of the target construct and *E* is the error that represents the discrepancy between the true score and the measured values (Hayes and Embretson, 2012; Novick, 1966). Under CTT, *E* is assumed to be random and independent from (uncorrelated with) *T* and has the expected value of zero. *X* for each single target *T* score also has the expected value of this target *T* score and variance equal to the variance of *E*.

Test fairness Framework (TFF) has been defined as the impartial treatment of all test takers during the testing process, absence of measurement bias, equitable access to the constructs being measured, and justifiable validity of test score interpretation for the intended purpose(s) (AREA, APA, & NCME, 2014). This ethic–inspired theory was proposed by Kunnan (2000) with a set of principles and sub – principles. Kunnan's (2000) framework was originally motivated by three test qualities namely: validity, absence of bias, and social consequences then, the improved TFF added qualities of access and administration as other qualities of test fairness.

Further scrutiny of TF indicates that Kunnan (2004) regarded fairness as the whole system of a testing exercise, not just the test itself (Moghadam & Nasirzadeh, 2020). The author argues that test fairness is affected by the various facets of fairness that include multiple uses, multiple stakeholders in the testing procedure (test takers, test users, teachers, and employers), and numerous steps in the test development process (test design, development, administration, and use) and established universal ideologies of fairness and beneficence and sub-principles that are basic to the TFF (Kunnan, 2004).

The first, according to Kunnan (2004) of such ideologies, is that of justice which tries to ensure that a test must be fair to all examinees. Kunnan (2004) argued that this aspect

includes two critical intertwined sub-principle which state that any test ought to have comparable construct validity in terms of its test-score interpretation for all examinees and secondly, that the test ought not to be biased against any group of examinees.

The second principle in Kunnan (2004) is that of beneficence which states that a test ought to bring good to society. This means that it should not be harmful to society, rather, ought to promote good in society by providing test score information and social impacts that are beneficial to society and it ought not to cause harm by providing test score information or social impacts that are inaccurate and misleading (Kunnan, 2000).

The present study conceptualizes test score equating as another tool that can be used to bring fairness among examinees perceived to be on the same level who have taken the parallel test. The equating procedure can be used to eliminate biases by setting scores from two groups of test takers on the same scale so that they can be used interchangeably.

2.14 Empirical Dimensions of the Study

The empirical dimensions of this study focused on the psychometric properties of the Nurses' Licensure Examinations (NLE) in Malawi, with particular attention to the fairness of the examination process. To achieve a comprehensive understanding of test fairness, the study investigated various empirical aspects such as test difficulty, unidimensionality, score equating, and bias in classification. These dimensions are crucial for ensuring that the licensure examinations are both valid and equitable, as they directly impact the professional certification of nurses and midwives in Malawi.

2.14.1 Test Difficulty and Item Analysis

One of the first empirical dimensions explored is the comparative difficulty of the two NLE test forms (2020 and 2021). According to Chulu and Sirec (2011), differences in test difficulty can lead to inaccurate conclusions about candidate performance, especially if the difficulty level of the two forms is not accounted for. The study applies item difficulty analysis to examine how the distribution of correct responses varies between the two test forms, ensuring that test difficulty is comparable and that neither cohort is unfairly advantaged or disadvantaged. This dimension is essential for making valid comparisons across different cohorts that may have sat for different versions of the test.

2.14.2 Unidimensionality of the Test Forms

The concept of unidimensionality refers to the extent to which a test measures a single underlying construct. A test that lacks unidimensionality may be measuring multiple factors, which could introduce bias and reduce the reliability of the test scores (Novick, 1966). This study investigated the unidimensionality of the 2020 and 2021 NLE test forms using factor analysis. If the test forms are not unidimensional, it may imply that the test is assessing unrelated skills or knowledge areas, which could affect fairness in scoring. Studies in other African countries, such as those by Chakwera et al. (2004), have noted that addressing multidimensionality in tests is crucial for ensuring that the tests accurately measure what they intend to.

2.14.3 Score Equating

The application of score equating is a central empirical dimension of the study. As noted by Dorans et al. (2010), even when test forms are designed to measure the same construct,

differences in their psychometric properties—such as item difficulty—can lead to score discrepancies. This study employed equipercentile equating using the R-EQUATE package to adjust scores between the two test forms. The goal was to ensure that test takers from different cohorts are evaluated on the same scale, which mitigates the impact of test form differences on performance comparison. Previous studies in the region, such as those by Sanagala (2017), have demonstrated that score equating can significantly reduce disparities in pass rates between cohorts, thereby ensuring fairness in the licensure process.

2.14.4 Bias in Classification and Grade Categorization

Another critical dimension is the bias in classification, which refers to how differences in test difficulty can affect candidates' classification into different grade categories. Before equating, the study examined whether candidates who took the more difficult 2021 test form were unfairly classified into lower grade categories compared to those who took the 2020 test form. After score equating, the study reassessed the classification to determine if the disparity in grades was reduced. Kunnan (2000) emphasizes that test fairness must account for the equitable classification of all test takers, regardless of which test form they have taken. The study explored this empirical dimension to ensure that candidates are classified fairly based on their true abilities, rather than on the test form they were assigned.

2.14.5 Equivalence of Test Forms

Lastly, the study examined the equivalence of the two test forms through correlation analysis. High correlation between the test forms indicates that they are measuring the same construct with similar psychometric properties. If the two test forms are found to be equivalent, it would support the use of scores from both forms interchangeably, ensuring

fairness in comparing the performance of candidates across cohorts. Studies in the Malawian context, such as those by Chakwera et al. (2004), have shown that equating and establishing test form equivalence are essential for ensuring the credibility of licensure examinations.

2.15 Chapter summary

The Chapter delved into equating methodologies, focusing on Classical Test Theory (CTT) and Item Response Theory (IRT) in educational and licensure examinations. It discussed empirical studies comparing CTT and IRT, noting varied perspectives on their effectiveness. Emphasizing the role of equating in promoting fairness and accurate score interpretation, particularly in healthcare licensure examinations, the Chapter also discussed ongoing debates and the need for further research. It then highlighted studies from diverse countries, such as Turkey, Nigeria, and Malawi, underscoring the importance of equating procedures in different contexts before presenting the gaps and weaknesses in the reviewed literature, theoretical framework and the empirical dimensions of the study.

CHAPTER 3

METHODOLOGY

3.0 Chapter Overview

This Chapter is dedicated to presenting the research paradigm, the study design for score equating and the setting for the study. The study population, sampling procedure and data collection tools and procedure are also presented in the Chapter. Finally, the Chapter presents the inclusion and exclusion criteria, data collection procedures, data analysis methods and the ethical considerations for the study.

3.1 Research paradigm

A paradigm is a set of beliefs or philosophical assumptions that guide a researcher when conducting a study (Creswell, 1998). Guba (1970) refers to these worldviews as a basic set of beliefs that guide action, as such, says Kuhn (1962), that a paradigm directs research. Therefore, without a paradigm, a research study lacks direction. This study used a positivism research paradigm using a quantitative approach. The decision to use a **quantitative research approach** in this study is grounded in its ability to provide objective, data-driven insights into the fairness of Nurses' Licensure Examinations (NLE) in Malawi. By applying statistical techniques like **Classical Test Theory (CTT)**, the study

can systematically assess the impact of test form differences on candidate scores. These methods allow for a clear comparison of test forms with varying difficulty levels, helping to determine if differences in pass rates are due to test properties or the abilities of the examinees (Kolen & Brennan, 2004). Quantitative research ensures that findings are based on measurable data rather than subjective interpretations, promoting fairness in the analysis.

The quantitative approach is also essential for examining key **psychometric properties** such as **unidimensionality** and **item difficulty** across different test forms. Techniques like **factor analysis** and **item difficulty indices** enable a rigorous assessment of whether the test forms measure the same construct and whether they differ in difficulty. This objective approach ensures that the results are grounded in empirical data, offering an accurate evaluation of the NLE's fairness (Sireci, Thissen, & Wainer, 1991). Quantifying these properties helps in understanding whether any observed discrepancies in performance can be attributed to differences in test properties rather than candidate abilities.

3.2 Study design

Neuman (2013) emphasized the importance of descriptive non-experimental designs in providing insights into real-world contexts without introducing artificial conditions, thereby contributing to a deeper understanding of natural phenomena. These designs allow researchers to systematically observe and document variables and conditions without manipulating them, which is crucial for understanding the natural state of complex systems (Leedy et al, 2023). They provide a detailed exploration of relationships and characteristics

within real-world settings, offering valuable insights that inform theory development and practical applications in various fields of study.

The present Study followed a descriptive non – experimental design since the aim was to unearth the characteristics of an already existing system. The study used data from a high-stakes Nurses Examinations Board (NEB) in Malawi, nurses' licensure examinations (NLE), called the Nurses' and Midwives Technician (NMT) paper 1 administered to CHAM students. This allowed for a more controlled examination of factors affecting exam outcomes, as CHAM students typically undergo similar educational training and face comparable challenges in their academic and professional paths. This homogeneity within the study population reduces variability from other student groups, making it easier to isolate the effects of the test forms and equating process.

3.2.1 Study setting

The study was conducted in the selected nurses training Colleges of Health sciences under CHAM in the Southern and South Eastern regions of Malawi. The study was conducted between December 2022 and December 2023.

3.2.2 Study population, sample and sampling procedure

To investigate the fairness of nurses' licensure examinations, the population of the study comprised of 358 final year nurses who sat for nurses' licensure examinations in November 2023 in all the training Colleges in the Southern and South Eastern regions of Malawi. Miaoulis and Michener (1976) recommends that in addition to the purpose of the study and population size, three criteria need to be specified to determine the appropriate sample size: the level of precision, level of confidence, and the degree of variability in the attributes being measured. In this study, a stratified sampling was used to draw a sample of 186

nurses (70% Females and 30% Males) using a 95% confidence level, a population proportion (or standard deviation) of 0.5, and a confidence interval (margin of error) of ± 5%.

Stratified sampling is a type of probability sampling technique in which the population is divided into distinct subgroups, known as strata that share similar characteristics. The purpose is to ensure that every subgroup is represented proportionally in the sample, making it more likely to reflect the diversity of the entire population (Cochran, 1977). Stratified sampling is a highly appropriate method for this research on the fairness of the Nurses' Licensure Examinations (NLE) in Malawi, as it ensures the representation of key subgroups within the population. Given that the study involves both male and female nurses, stratified sampling guarantees proportional representation of these gender groups, addressing potential biases that could arise from societal roles or educational differences (Miaoulis & Michener, 1976)

The sample size was determined using the formula proposed by Krejci and Morgan (1970). To them the sample size (n) can be calculated using the formula below:

$$\mathrm{n} = \frac{\chi^2 \times \mathrm{N} \times \mathrm{P}(1-\mathrm{P})}{\left(\mathrm{ME}^2(\mathrm{N}-1) + (\chi^2 \times \mathrm{P}(1-\mathrm{P})\right)} \tag{7}$$

Where:

n = Sample size

χ^2 = Chi-square for the specified confidence level at 1 degree of freedom ($\chi^2 = 3.841$)

N = Population size

P = Population proportion (P = 0.5 since this would provide maximum variability)

ME = Desired marginal of error (expressed as a proportion ($\mp 5\%$ / 0.05)

In Classical Test Theory (CTT)-based equating methods, such as linear equating, sample sizes of at least 200 to 300 examinees per group (for two test forms) are typically recommended to ensure reasonable accuracy (Kolen & Brennan, 2004). However, in practical scenarios, smaller sample sizes (fewer than 100 examinees per group) are sometimes unavoidable due to contextual and logistical factors. These factors may include specialized testing populations, pilot testing, or high-stakes exams with limited candidates. The use of small sample sizes can lead to instability in the equating results, making it essential to apply statistical techniques, such as smoothing, to address these issues and improve the reliability of the equating process (Kolen & Brennan, 2004). Additionally, more advanced methods, such as item response theory-based equating, can help stabilize results when sample sizes are small (Zheng & van der Linden, 2010). In the current study, the sample size was less than 100 per group due to logistical constraints and limitations in the number of candidates available. Consequently, smoothing techniques were employed to ensure the stability of the equating results.

3.3 Inclusion and exclusion criteria

Researchers use inclusion and exclusion criteria to determine the characteristics of the subjects or elements in a study. Inclusion and exclusion criteria define who can be included or excluded from the study sample (Garg, 2016). Establishing these criteria for subjects is an important step in designing high-quality research (Connelly, 2020). The study targeted

Nurses who were in their final year of their nursing studies in training Colleges under CHAM and were to sit for the Nurses' Licensure Examinations in November 2023. The Nurses who were in their final year but were not qualified to sit for the Licensure Examinations in November did participate in the current study. This was done to ensure that the sample reflects the reality on the ground.

Nurses who were not sitting for the November form would not be serious about the test or might have been missing classes which would interfere with the findings of the study. Again, nurses who were in the final year of their studies but were enrolled in a college that was not in the Southern or South-Eastern regions of Malawi did not participate in the study because it would be hard for the investigator to reach them.

3.4 Data collection tools and procedure

The study used data from the administration of the Nursing and Midwives Technician (NMT) Paper 1 administered in 2020 and 2021. The tests forms were administered to a total of 186 nurses (70% Females and 30% Males) in Southern and South Eastern Region of Malawi. 93 nurses (50%) sat for the 2020 (Old form) form and 93 nurses (50%) sat for the 2021 form (New form). The 2020 test form (Old form) was named Form X and the 2021 test form (New form) was named Form Y.

NMT paper 1 is one of the four papers administered to Nurses at the end of their Nursing study. The paper comprises of seven (7) sections namely: Medical Nursing (18%), Surgical Nursing (22%), Gynecology Nursing (5%), Pediatrics/Child Health Nursing (15%), Community Health Nursing (25%), Mental Health and Psychiatric Nursing (9%) and Leadership and management (6%). The NMT paper 1 was formulated by the Nurse's

Examination Board (NEB). Both question papers were written in three (3) hours. All the questions in this paper were multiple choice and the pass mark for each paper was 50%.

A spiraling process was used to create two random groups of examinees (Group A and Group B) who sat for the two test forms. When an examinee from Group A receives the Form X script (Old Form), an examinee from Group B receives the Form Y script (New Form), an examinee from Group A receives the Form X script, then, another Group B examinee receives the Form Y script, and so on. In this way, Group A examinees sat for Form X while Group B examinees sat for Form Y at the same time of administration.

The researcher with the assistance from the research assistant facilitated data collection procedure. In accordance with section 45 of Nurses and Midwives Act 16 of 1995, the research assistant was responsible for the mass production of the tools to ensure that any damaged tools during production was returned at the Nurses Examinations Board. On the day of data collection, the researcher and the research assistant together distributed the scripts to the subjects and supervised as the subjects were writing the examinations until completion.

The researcher and the research assistant collected the scripts. The researcher and the research assistant together scored the scripts and entered the scores in the researcher's computer. The job of the research assistant ended here and data analysis was done by the investigator.

3.5 Validity and Reliability of the data collection tools

The data collection tools in this study consisted of the Nursing and Midwives Technician (NMT) Paper 1 administered in 2020 and 2021. Validity was ensured through the use of

standardized test forms developed by the Nurse's Examination Board (NEB) which aligns the test items with the content areas of the nursing curriculum through the use of test blueprints.

Reliability was ensured by consistently following procedures for administering and scoring the tests. Random assignment of examinees to different test forms helped reduce bias, thereby improving the reliability of comparisons between the forms. Additionally, both the researcher and research assistant were actively involved in administering, scoring, and entering data, ensuring consistency and accuracy, which further enhanced the reliability of the data collected. Adherence to regulatory guidelines, such as section 45 of the Nurses and Midwives Act, also ensured the integrity of the data collection process. Overall, the rigorous procedures employed in data collection support the validity and reliability of the study's findings

3.6 Data analysis

To determine the extent of the unidimensionality of Nurses' Licensure Examinations, the Principle Component Analysis (PCA) was conducted using a Statistical Package for Social Sciences (SPSS). A test is assumed to be unidimensional only when there is just one dominant factor or ability being measured by items (Hambleton et al., 1991). The explained variance ratio at and above 30% is regarded as adequate (Büyüköztürk, 2007).

To compare the test difficulties of the test forms, an independent samples T–test was used to check the differences in the means of the two forms on Statistical Package for Social Sciences (SPSS) using a level of significance of 0.05, the null hypothesis would be rejected if the p-value for the t-test were less than 0.05.

Equal reliability was checked using the Chronbanch alpha (α) coefficients. The reliability correlation coefficients and Fischer's Z transformation was used to check if there was a difference between the two reliability coefficients (Akhun, 1984).

To assess bias when classifying students into grade categories before and after equating, scores from the new form (Form X) were equated to the scores of the old form (Form Y). Equipercentile equating was conducted using R-EQUATE on R software with log-linear smoothing. The average score and pass rates before and after equating of the new form were compared.

To examine whether scores from the two test forms could be used interchangeably, the average scores for each test form were computed and subsequently compared by using an F – test statistic. The null hypothesis (H_0) for this test was that the variances of proportions between the Group that sat for 2020 test form and the Group that sat for the 2021 test form were not significantly different. The alternative hypothesis (H_A) was that the variances of proportions between the two groups were significantly different. While various methods, such as visually examining histograms or box plots of the scores for each test form, can be employed, this study opted for this numerical approach. Specifically, the difference between the mean scores was calculated. The F – test helped to evaluate and to ascertain if there were any noteworthy disparities between the variances derived from each test form.

3.7 Ethical considerations

Swain (2016) defines the term ‘ethics’ as the moral principles or rules of conduct held by a group or profession that guide the conduct of the research. Ethical considerations in research are a set of principles that guide a research design and practices (Bhandari, 2022).

Bhandari (2022) states that these principles include voluntary participation, informed consent, anonymity, confidentiality, potential harm, and results communication. To ensure that the current study comply with the research ethics, the following ethical issues were considered.

Firstly, the researcher submitted the proposal to the University of Malawi Research Ethics Committee (UNIMAREC), the Institution Review Board (IRB), for approval before data collection. The approval letter from UNIMAREC and a letter of introduction from the University of Malawi to conduct the study at the selected colleges were presented to the colleges from where the subjects were drawn and to all stakeholders involved in the data collection exercise.

Secondly, to ensure autonomy, the subjects were briefed on the research regarding what it intends to achieve and how the data would be used. Therefore, all the subjects had to volunteer for themselves and no one was forced or coerced to participate in the study. Prior to data collection, all the subjects signed an informed consent form.

Finally, the research guaranteed the anonymity of participants by not collecting any personally identifying information, for example, names, phone numbers, email addresses, physical characteristics, photos, and videos. To ensure this ethical issue, the researcher used a serial number for each subject instead of names and other identifying information.

3.8 Chapter summary

The Chapter began by discussing the importance of research paradigms in guiding researchers' beliefs and actions. The Chapter also highlighted the interdependence of the

health system with national development. Factors affecting health system efficiency, including funding and cultural beliefs, are also identified.

The Chapter also outlined, utilizing a descriptive non-experimental design, detailing sampling techniques, inclusion/exclusion criteria and data collection procedures. Key statistical analyses conducted include Principle Component Analysis (PCA) for assessing test unidimensionality, independent samples T-tests to compare test difficulties, and reliability checks using Chronbach's alpha coefficients and Fischer's Z transformation.

The Chapter concluded by discussing the ethical considerations, including participant consent and confidentiality, are thoroughly addressed.

CHAPTER 4

RESULTS AND DISCUSSION

4.0 Chapter Overview

This Chapter gives the results of the study. The results are presented under four subheadings namely unidimensionality of test forms, difficulties across test forms, inequalities caused due to the classifications of students into grade categories across forms before and after equating and determining whether scores from the two test forms can be used interchangeably.

4.1 Unidimensionality of the test forms

The study aimed to assess the unidimensionality of two test forms using factor analysis. According to Hambleton et al. (1991), a test is considered unidimensional when it primarily measures one dominant factor. Factor analysis was conducted using SPSS, and the results were presented in **Tables 4.1A** and **4.1B**, which depict the outcomes of factor analyses for each test form.

TABLE 4. 1A: Principle Component Analysis for Form X

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.696	10.696	10.696	10.696	10.696	10.696
2	5.787	5.787	16.483	5.787	5.787	16.483
3	4.722	4.722	21.206	4.722	4.722	21.206
4	4.096	4.096	25.302	4.096	4.096	25.302
5	3.638	3.638	28.940	3.638	3.638	28.940
6	3.560	3.560	32.500	3.560	3.560	32.500
7	3.290	3.290	35.791	3.290	3.290	35.791
8	3.050	3.050	38.841	3.050	3.050	38.841
9	2.907	2.907	41.747	2.907	2.907	41.747
10	2.765	2.765	44.512	2.765	2.765	44.512

TABLE 4. 1B: Principle Component Analysis for Form Y

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11.471	11.471	11.471	11.471	11.471	11.471
2	7.585	7.585	19.057	7.585	7.585	19.057
3	4.694	4.694	23.750	4.694	4.694	23.750
4	4.357	4.357	28.108	4.357	4.357	28.108
5	3.658	3.658	31.766	3.658	3.658	31.766
6	3.510	3.510	35.276	3.510	3.510	35.276
7	3.321	3.321	38.597	3.321	3.321	38.597
8	3.070	3.070	41.667	3.070	3.070	41.667
9	2.716	2.716	44.383	2.716	2.716	44.383
10	2.604	2.604	46.986	2.604	2.604	46.986

Table 4.1A shows eleven (11) of the thirty-three (33) components that have eigenvalues greater than 1 for form X. The first factor has an initial eigenvalue of, **10.696** greater than the second factor of **5.787** (See Initial Eigenvalues in **Table 4.1**).

Similarly, **Table 4.1B** shows the first eleven (11) of the thirty-three (33) components whose eigenvalues are greater than 1. The first factor has an initial eigenvalue of **11.471**, greater than the second factor of **7.585** (See Initial Eigenvalues in **Table 4.1B**). All the remaining factors are less important because they are less than 1 since the percentage of total variance explained by the first principal component on Form X and Y is **10.696%** and **11.471%** respectively. These values are less than 30%. Hence, according to Büyüköztürk (2007) who argues that for a test to be unidimensional, the first factor has to explain 30% or more of the variance. In this regards, both the 2020 and 2021 NMT test forms were not unidimensional since in each case the first factor explains less than 30%.

The unidimensionality of test items is a central assumption in both Classical Test Theory (CTT) and Item Response Theory (IRT), as it ensures the accuracy and validity of the measurement process (Fan, 1998; Courville, 2005). Several studies, including those by Fan (1998) and Courville (2005), have demonstrated that when the assumption of unidimensionality holds, both CTT and IRT provide similar and reliable results regarding person ability and item difficulty. For example, Fan (1998) found that in the Texas Assessment of Academic Skills (TAAS), the correlation between person parameters and item difficulties was extremely high (above 0.96 and 0.90, respectively), supporting the assumption of unidimensionality and reinforcing the validity of both measurement frameworks. This is further corroborated by Courville's (2005) replication study with a

larger sample, which also found highly comparable results, indicating that both CTT and IRT were consistent in measuring a dominant factor.

The results of the current study, which aimed to assess the unidimensionality of two test forms using factor analysis, echo these findings but also reveal some important nuances. According to Hambleton et al. (1991), a test is considered unidimensional when it primarily measures one dominant factor, a concept that aligns with the findings from Fan (1998) and Courville (2005) regarding the consistency across measurement frameworks. The factor analysis results for both Form X and Form Y, however, do not support the assumption of unidimensionality.

As shown in Tables 4.1A and 4.1B, the first factor for both test forms accounts for less than 30% of the total variance (10.696% for Form X and 11.471% for Form Y), which contradicts the criterion proposed by Büyüköztürk (2007) for unidimensionality, i.e., that the first factor should explain at least 30% of the variance. This finding is consistent with previous research suggesting that tests with low variance explained by the first factor may not exhibit unidimensionality (Tate & Baird, 2014). The relatively low percentage of variance accounted for by the first factor in this study suggests the presence of multiple underlying dimensions, leading to the conclusion that both forms of the test were multidimensional.

Having used eigenvalues to describe the unidimensionality of the NMT multiple-choice 2021 and 2020 test items, scree plots were used to affirm the number of factors retained. In the scree plot, the point of interest is where the curve starts flattened as shown in **Figures** 4.1A and **Figures** 4.1B.

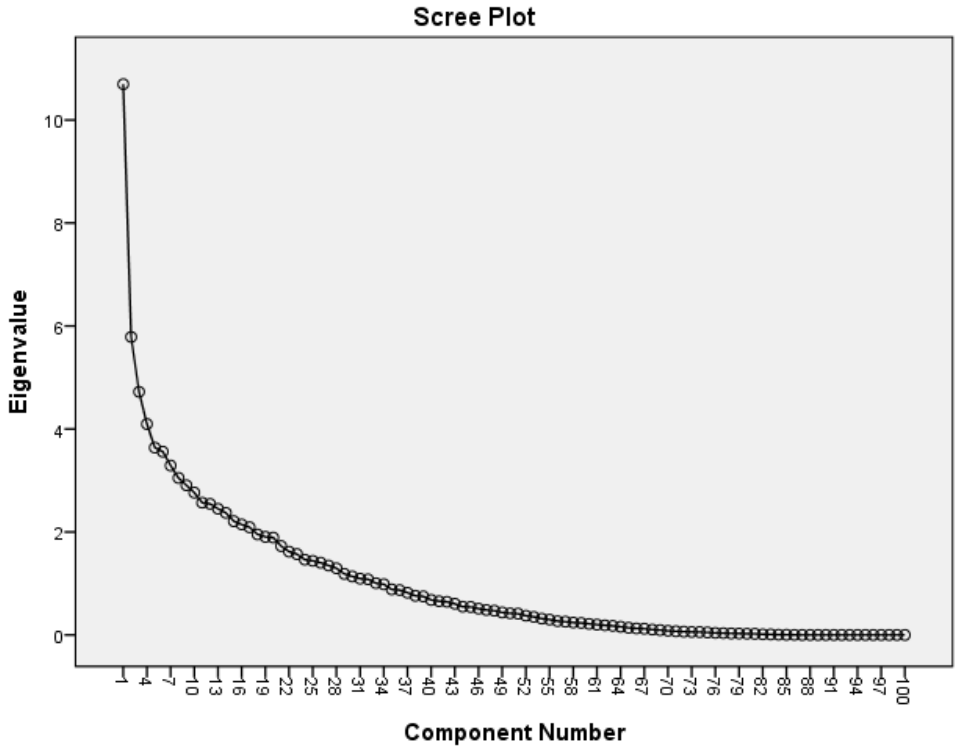


FIGURE 4. 1A: The scree plot for Form X

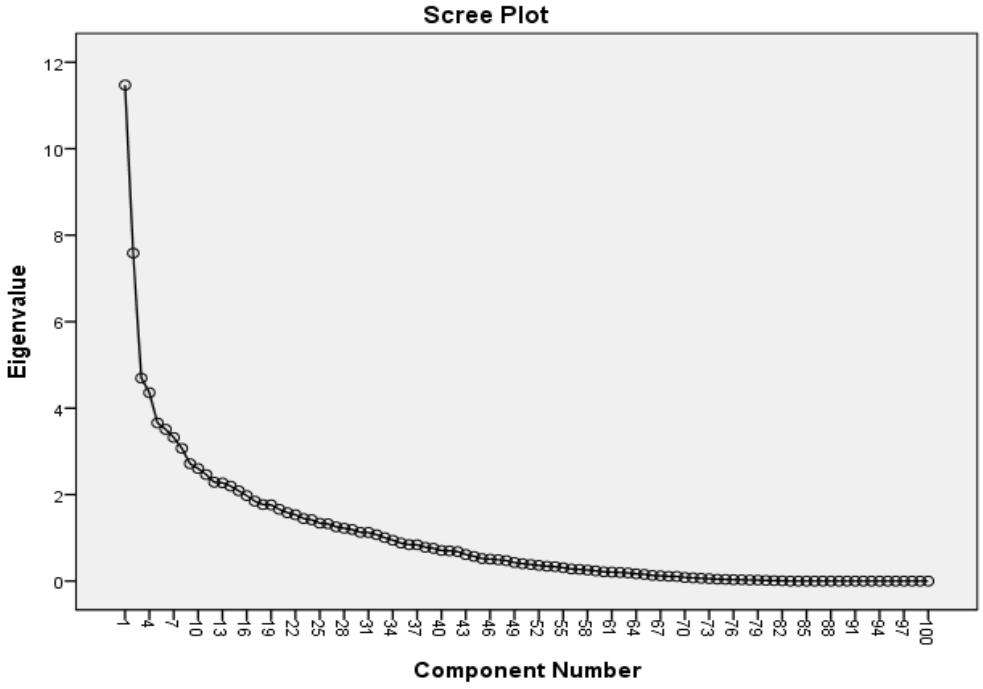


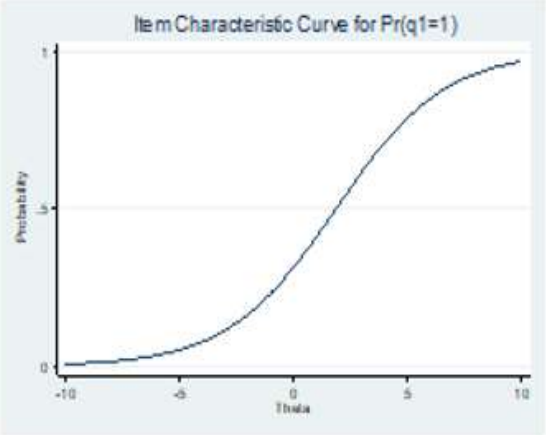
FIGURE 4. 1B: The scree plot for Form Y

Figure 4.1A and **Figure** 4.1B show the scree plots of NMT multiple-choice items constructed in the 2021 and 2020 in that order. From **Figure** 4.**1A**, the Scree plot indicates the total variance associated with each factor. The steep slope indicates the large factors associated with the loading greater than the eigenvalue of 1. The first thirty three factors show a slope but with a steep between the first and fourth factors. The rest of the factors from 34 are lower than an eigenvalue of 1.

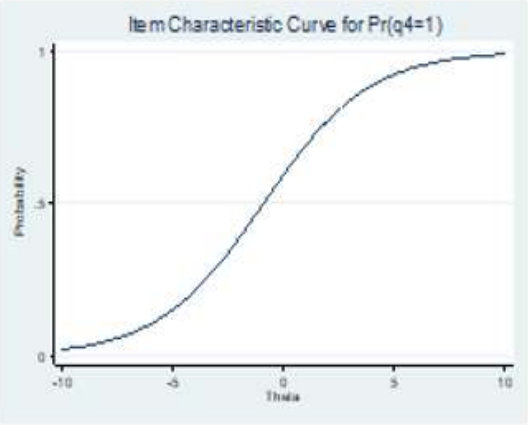
Similarly, For Form Y, **Figure** 4.1B the first thirty three factors show a slope but with a steep between the first and third factors. The rest of the factors from 34 are lower than an eigenvalue of 1. This, therefore, shows that the NMT multiple-choice items constructed by the Nurses Council in the 2021 and 2020 session were not unidimensional.

Furthermore, the scree plots in Figures 4.1A and 4.1B provide additional support for the lack of unidimensionality of the two test forms. In line with the research of Tate and Baird (2014), who found high consistency in factor loadings in standardized assessments, the scree plots for both test forms indicated a steep slope between the first and a few subsequent factors, followed by a flattening of the curve, indicating that multiple factors contribute to the total variance. This pattern, as shown in Figures 4.1A and 4.1B, reinforces the lack of unidimensionality in the test forms, confirming that these tests measure more than one dominant factor.

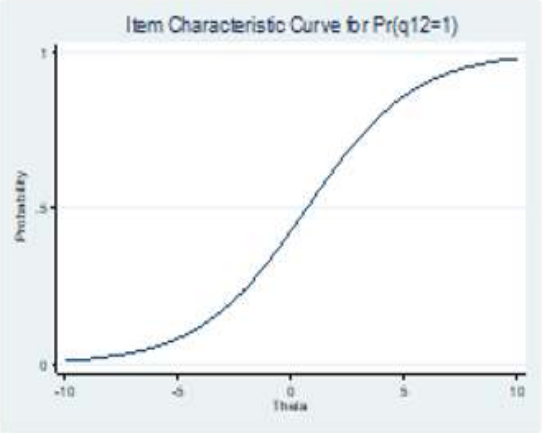
Having carried out the Principal Component, the researcher computed Item Characteristic Curve (ICC) using STATA 14 software as reported in figures **4.1C (a)** to 4.1C **(f)**. The researcher identified the number of items that favour unidimensionality by examining these ICCs.



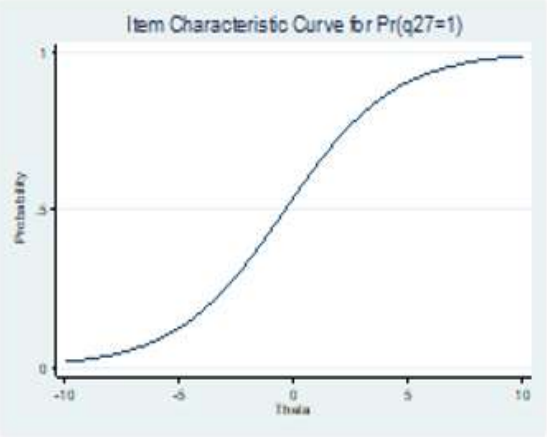
(a). Item Characteristic Curve for Item 1



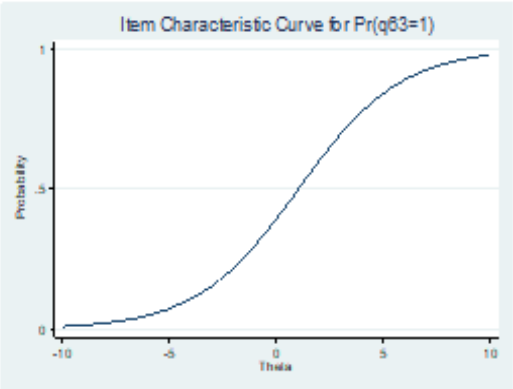
(b). Item Characteristic Curve for Item 4



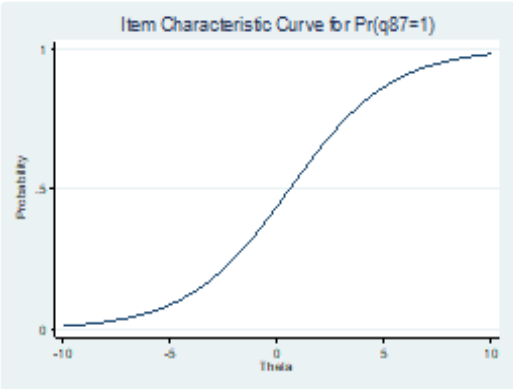
(c). Item Characteristic Curve for Item 12



(d). Item Characteristic Curve for Item 27



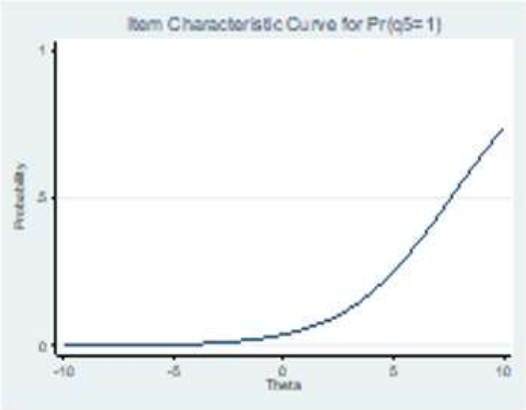
(e). Item Characteristic Curve for Item 63



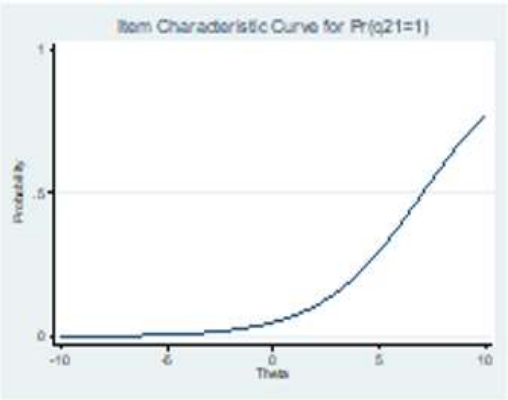
(f). Item Characteristic Curve for Item 87

FIGURE 4. 1C: Item Characteristic Curves of the items that assumed an "S" shape for Form X

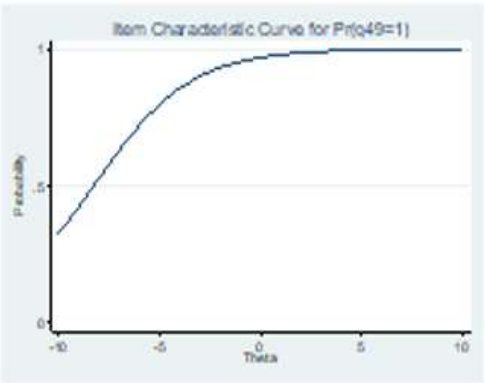
Figures 4.1C (a) to **4.1C (f)** were randomly selected from the Item Characteristic Curve of items 1 to 100 for Forms X. Examination of the 100 figures reveals that **73 (73%)** of the items assumed an "S" shape. This "S" shape suggests that the item is discriminating well between individuals with lower and higher ability levels. It implies that the test item is able to effectively differentiate between candidates who are less knowledgeable or skilled and those who are more proficient. Among them are violated the assumption of "S" shape. In the 73 items, as the ability of the test-takers increases, the probability of getting the answer correct increases. Some of the twenty-seven (27) items that violated local item independence are shown in **Figure 4.1D (a)** to **4.1D (d).** The presence of some items that violate the "S" shape suggests that a portion of the items might not be as effective or may suffer from issues such as **local item dependence** or other forms of bias, potentially impacting the fairness and accuracy of the test.



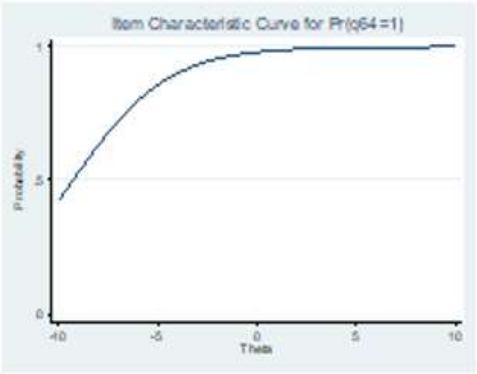
(a). Item Characteristic Curve for Item 5



(b). Item Characteristic Curve for Item 21



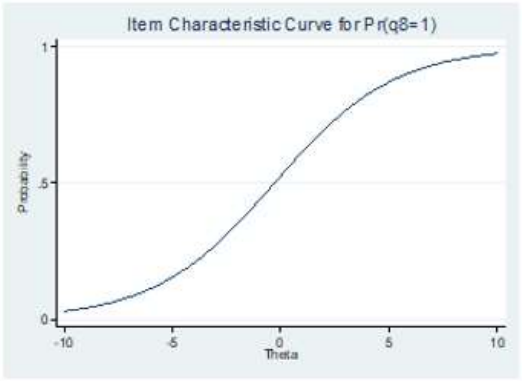
(c). Item Characteristic Curve for Item 49



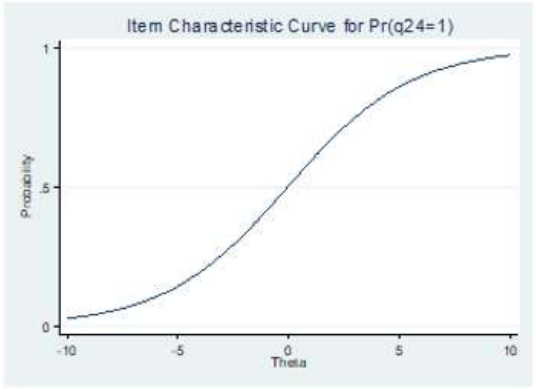
(d). Item Characteristic Curve for Item 64

FIGURE 4. 1D: Item Characteristic Curves of the items that assumed an "S" shape for Form X

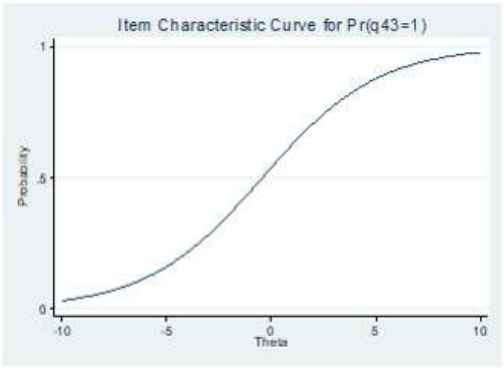
For Form Y, examination of the 100 figures reveals that **79 (79%)** of the items assumed an "S" shape. Among them are: **Figures 4.1E (a)** to **4.1E (f)** show randomly selected Item Characteristic Curve of items for Forms Y. The rest twenty one **(21%)** did not assume an "S" shape. Likewise, in the **79** items, as the ability of the test-takers increase, the probability of getting the answer correct increases.



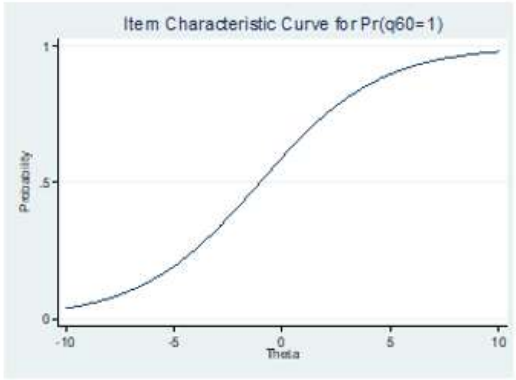
(a). Item Characteristic Curve for Item 8



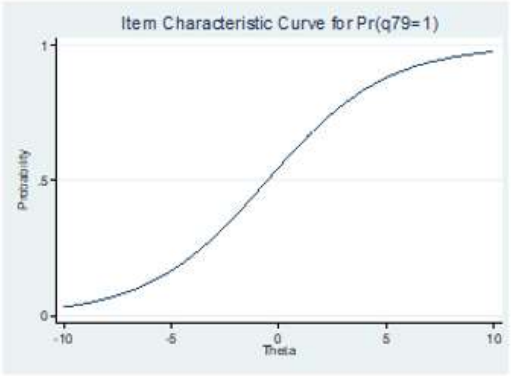
(b). Item Characteristic Curve for Item 24



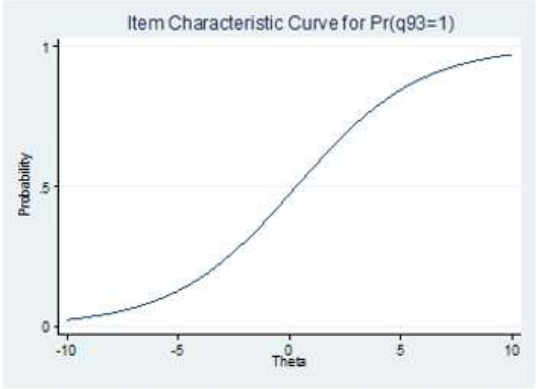
(c). Item Characteristic Curve for Item 43



(d). Item Characteristic Curve for Item 60



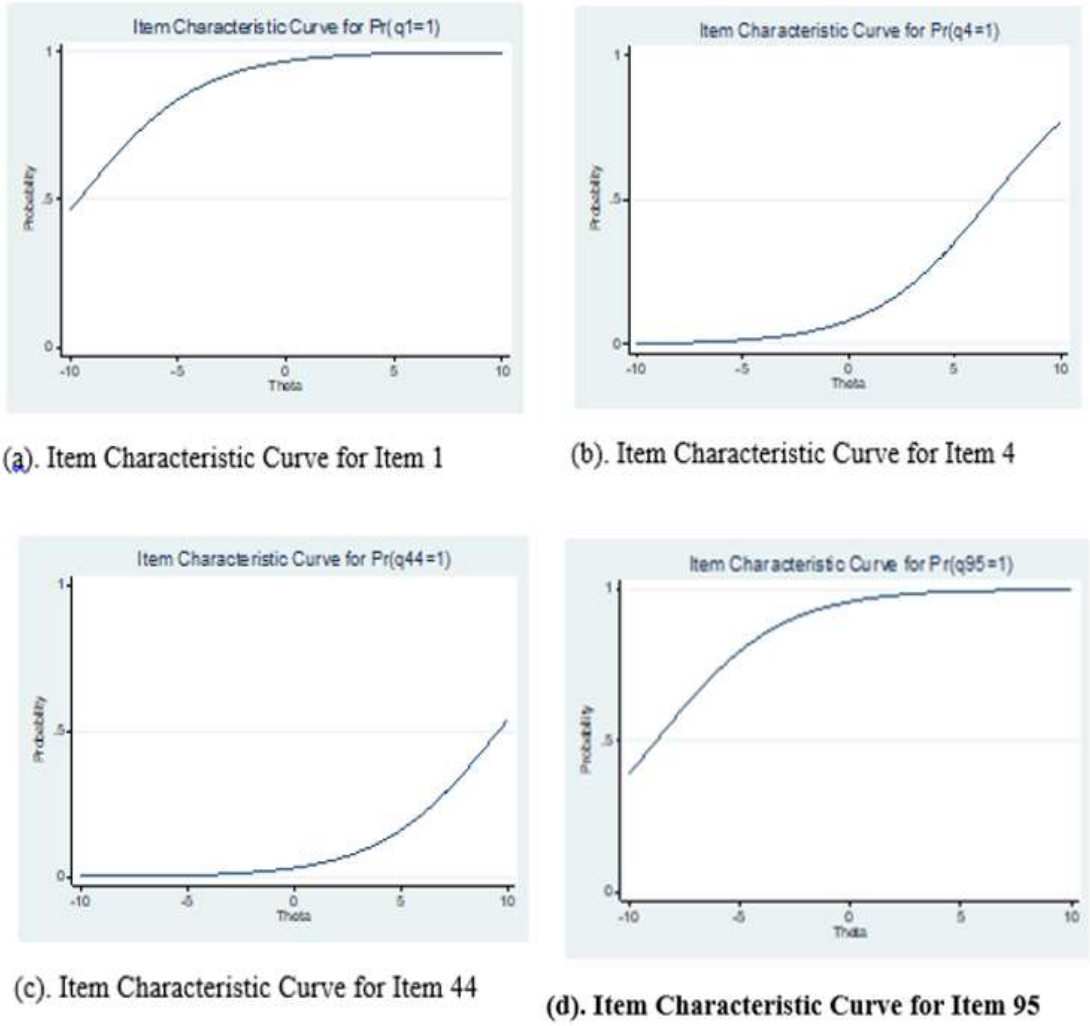
(e). Item Characteristic Curve for Item 79



(f). Item Characteristic Curve for Item 93

FIGURE 4. 1E: Item Characteristic Curve of Items with an "S" shape for Form Y

A sample of the twenty one items that did not have an "S" shape are shown in **Figure 4.1F (a)** to **4.1F(d)**.



(a). Item Characteristic Curve for Item 1

(b). Item Characteristic Curve for Item 4

(c). Item Characteristic Curve for Item 44

(d). Item Characteristic Curve for Item 95

FIGURE 4. 1F: Item Characteristic Curves of Items that did not have an "S" shape for Form Y

Both the eigenvalues and the scree plots confirms that the two test forms were not unidimensional since their first components explains less than 30%. Büyüköztürk (2007) argues that for a test to be unidimensional, the first factor has to explain 30% or more of the variance. In this regards, the two test forms were not unidimensional due to lack of one

principle factor since in both case the first factor explains less than 30%. For forms X. Examination of the 100 figures reveals that 73 (73%) of the items assumed an “S” shape and 27% violated the local independence. For Form Y, examination of the 100 figures reveals that 79 (79%) of the items assumed an “S” while twenty one (21%) did not assume an “S” shape.

To determine the extent of unidimensionality of the test forms, factor analysis was conducted on each test form using SPSS. The results were presented in Tables 4.1A and 4.1B. These tables displayed the initial eigenvalues for each factor extracted, with a focus on the first factor. It was found that the first factor's eigenvalue for both forms was less than 30%, indicating that the tests were not unidimensional according to the criterion proposed by Büyüköztürk (2007).

Additionally, scree plots were employed to confirm the number of factors retained in the analysis. The scree plots illustrated the variance associated with each factor, with a particular focus on where the curve began to flatten. Both Figures 4.1A and 4.1B showed that after the first few factors, there was a significant decrease in variance, further supporting the conclusion that the tests were not unidimensional.

Furthermore, Item Characteristic Curves (ICCs) were computed using STATA 14 software to examine the shape of item responses and assess local item independence. Figures 4.1C to 4.1F presented randomly selected ICCs for both forms, highlighting the percentage of items exhibiting an "S" shape, which indicates appropriate response patterns, and those violating local item independence.

For Form X, 73% of the items showed an "S" shape, while 27% violated local item independence. Similarly, for Form Y, 79% of the items displayed an "S" shape, while 21% did not. These findings further supported the conclusion that both test forms were not unidimensional.

Moreover, the examination of Item Characteristic Curves (ICCs) provides further insights into the nature of item responses and local item independence. According to Gonzalez and Lemos (2016), the "S" shape of the ICC is a key indicator of unidimensionality, as it shows that the item is discriminating effectively between individuals with higher and lower ability levels. In this study, 73% of the items for Form X and 79% for Form Y exhibited the "S" shape, suggesting that a majority of the items discriminate effectively. However, as noted by Progar, Smith, and Taylor (2008), the presence of a portion of items that violate the "S" shape (27% for Form X and 21% for Form Y) indicates potential issues such as local item dependence, which further challenges the assumption of unidimensionality. This local dependence may be a contributing factor to the lack of a single dominant factor across the test forms, supporting the argument that both forms are multidimensional.

In summary, while previous studies such as those by Fan (1998), Courville (2005), and Tate & Baird (2014) provide evidence that both CTT and IRT can support unidimensionality under certain conditions, the results of this study suggest that the 2020 and 2021 NMT multiple-choice test forms do not meet the unidimensionality criterion. The low percentage of variance explained by the first factor, combined with the findings from the scree plots and ICC analysis, all point to the conclusion that these test forms were not unidimensional. This aligns with the concerns raised by Büyüköztürk (2007) and others

regarding tests with multiple underlying dimensions, which may impact the accuracy and fairness of the measurement process.

4.2 Difficulties across test forms

Descriptive statistics were calculated using the Excel spreadsheet. **Tables 4.2A** and **4.2B** provide the descriptive statistics for Form X and Form Y. From **Table 4.2A**, the mean of Form X was **56.38** and the mean for Form Y was **58.51**. The highest score of Form X was **73%** while the highest score of Form Y was **76%.** The Kurtosis for forms X and Y were **0.65** and **0.34** respectively while the Skewness was 0.30 and 0.48 in that order. This shows that both test forms are positively skewed, however, form Y is more skewed than form X. This meant that the examinees who took Form Y did not perform well as compared to their counterparts who sat for form X.

TABLE 4. 2A: Descriptive statistics I for Form X and Form Y

Form X ***(New Form, 2021)***		***Form Y*** ***(Old Form, 2020)***	
Mean	56.376344	Mean	58.50537634
Median	55	Median	59
Standard Deviation	8.5477101	Standard Deviation	7.99677062
Kurtosis	0.6540156	Kurtosis	0.319479817
Skewness	0.3020401	Skewness	0.479736339
Range	34	Range	36
Minimum	39	Minimum	40
Maximum	73	Maximum	76
Cronbanch alpha reliability	0.76	Cronbanch alpha reliability	0.72

In this study, item difficulty was compared across two test forms—Form X (new, 2021) and Form Y (old, 2020)—using descriptive statistics, with findings indicating that both test forms had similar mean difficulty scores. According to Table 4.2A, the mean difficulty for

Form X was 56.38, while Form Y had a slightly higher mean of 58.51. These means suggest that Form Y, on average, was slightly more difficult than Form X. Both test forms had a similar range of scores, with Form X ranging from 39 to 73, and Form Y ranging from 40 to 76, further suggesting that the tests were comparably difficult in terms of the spread of scores. However, the kurtosis and skewness values for both forms suggest some differences in the distribution of scores. While both forms showed positive skewness, indicating that the majority of examinees performed well, Form Y was more positively skewed than Form X, indicating a greater proportion of examinees performed below the mean. This suggests that test-takers on Form Y were generally less successful compared to those who took Form X.

The reliability of the two test forms was also evaluated using Cronbach's alpha, with Form X yielding an alpha of 0.76 and Form Y yielding a slightly lower alpha of 0.72 (see Table 4.2B). According to Gonzalez and Lemos (2016), reliability coefficients above 0.70 are generally considered acceptable, indicating that both test forms were reliable. Furthermore, the difference in reliability between the two forms was not statistically significant, as confirmed by the Z statistics ($z = 0.438$, $p > 0.05$), suggesting that both forms provided equally reliable measures of test performance.

TABLE 4. 2B: Descriptive statistics II for Form X and Form Y

Statistic	Mean difficulty	Cronbanch alpha (α)	Standard deviation
Form X	0.571	0.76	8.55
Form Y	0.584	0.72	8.00

To further explore the differences in difficulty between the two forms, an independent samples t-test was conducted. The results (see Table 4.2C) showed that there was no significant difference in the mean difficulty between the two forms ($t(184) = -1.754$, $p > 0.05$). This suggests that the slight difference in mean difficulty between the two forms was not large enough to be statistically significant, reinforcing the notion that the difficulty levels of the two test forms were comparable.

TABLE 4. 2C: The Two sample t – Test Assuming Unequal Variances

	Form X (New Form, 2021)	***Form Y (Old Form, 2020)***
Mean	56.37634409	58.50537634
Variance	73.06334736	63.94834035
Observations	93	93
Hypothesized Mean Difference	0	
Df	183	
t Stat	-1.75406187	
P(T<=t) two-tail	0.081093844	
t Critical two-tail	1.973011915	

These findings align with the literature on item difficulty. MacDonald and Paunonen (2002) and Gonzalez and Lemos (2016) both highlighted that CTT and IRT can yield similar difficulty indices, particularly when the spread of item difficulty is controlled. In this study, the lack of a significant difference between the two forms in terms of difficulty supports the idea that, even though the test forms were developed at different times (2020 and 2021), they are sufficiently comparable in terms of difficulty, as suggested by the high correlation between the difficulty indices of the two methods (CTT and IRT) discussed by these authors. Thus, the results suggest that both forms of the NMT, despite some

differences in skewness, can be considered to have similar overall difficulty levels, in line with the findings of previous research.

In summary, this study confirms that the 2020 and 2021 test forms demonstrate similar item difficulty, both in terms of mean scores and reliability. The lack of significant differences between the two forms in difficulty, as shown by the t-test and supported by descriptive statistics, aligns with the literature on CTT and IRT assessments of test difficulty, reaffirming that both theories can provide consistent and reliable results in evaluating the performance of test items

4.3 Inequalities caused by the classification of students into grade categories across forms before and after equating the test scores

The purpose of this analysis was to examine whether classifying students into grade categories based on scores from two different test forms (2020 and 2021) led to inequalities. To do so, the researcher employed equipercentile equating, a statistical technique that adjusts scores from the 2021 form to align with those from the 2020 form. This approach was implemented using the R package "equate" (Albano, 2014), which transformed scores from the newer test form (2021) to their equivalent scores on the older form (2020), ensuring that students' performance was compared fairly across the two forms. The equating process involved matching students' percentile ranks on the 2021 form to the corresponding ranks on the 2020 form, as described by Kolen and Brennan (2004).

The results of the equating process, as presented in Table 4.3A, indicated that the 2021 test form was slightly more difficult than the 2020 test form. For instance, a score of 56 on the

2021 test form was equivalent to a score of 58 on the 2020 test form. This suggested that the distribution of scores on the 2021 test was lower than on the 2020 test, which could lead to students on the 2021 form being unfairly classified as failing even if they performed similarly to those on the 2020 form. For example, a student who scored 48% on the 2021 test would have been classified as failing, while the same score would have been classified as passing on the 2020 test. This demonstrates how unadjusted differences in test difficulty could lead to inequitable outcomes, as students taking the more difficult test (2021 form) might unfairly fail despite comparable performance.

TABLE 4. 3A: Conversion Table for Equipercentile Equating

2021 Score	**2020 Equivalent**	**2021 Score**	**2020 Equivalent**	**2021 Score**	**2020 Equivalent**
0	0.0000	34	37.5713	68	69.3798
1	6.6983	35	38.5068	69	70.3154
2	7.6338	36	39.4424	70	71.2509
3	8.5694	37	40.3779	71	72.1865
4	9.5049	38	41.3135	72	73.1220
5	10.4405	39	42.2490	73	74.0576
6	11.3760	40	43.1846	74	74.9931
7	12.3116	41	44.1201	75	75.9287
8	13.2471	42	45.0557	76	76.8642
9	14.1827	43	45.9912	77	77.7997
10	15.1182	44	46.9267	78	78.7353
11	16.0537	45	47.8623	79	79.6708
12	16.9893	46	48.7978	80	80.6064
13	17.9248	47	49.7334	81	81.5419
14	18.8604	48	50.6689	82	82.4775
15	19.7959	49	51.6045	83	83.4130
16	20.7315	50	52.5400	84	84.3486
17	21.6670	51	53.4756	85	85.2841
18	22.6026	52	54.4111	86	86.2197
19	23.5381	53	55.3467	87	87.1552
20	24.4737	54	56.2822	88	88.0907

21	25.4092	55	57.2177	89	89.0263
22	26.3447	56	58.1533	90	89.9618
23	27.2803	57	59.0888	91	90.8974
24	28.2158	58	60.0244	92	91.8329
25	29.1514	59	60.9599	93	92.7685
26	30.0869	60	61.8955	94	93.7040
27	31.0225	61	62.8310	95	94.6396
28	31.9580	62	63.7666	96	95.5751
29	32.8936	63	64.7021	97	96.5106
30	33.8291	64	65.6377	98	97.4462
31	34.7647	65	66.5732	99	98.3817
32	35.7002	66	67.5087	100	99.3173
33	36.6357	67	68.4443		

To better understand the impact of these differences on student classification, Table 4.3B shows the pass rates for the 2020 and 2021 forms before and after the equating process. Before equating, the pass rate for the 2021 form was 76.34%, while the pass rate for the 2020 form was 91.40%, a difference of 15.06%. This disparity highlights the advantage of students who took the 2020 test, as the test was easier, and they were more likely to pass. However, after equating the scores, the difference in pass rates was reduced to 7.53%, with the pass rate for the 2021 form increasing to 83.87%. This adjustment indicates that the equating process helped to account for the differences in test difficulty, making the classification of students into pass and fail categories more equitable.

The reduced gap in pass rates after equating aligns with findings from previous studies, such as those by Chulu and Sires (2011), who emphasized the importance of equating to ensure fairness when comparing scores from different test forms. Their work highlighted how unadjusted differences in test difficulty could result in unjust classification decisions, as observed in the initial pass rate differences between the 2020 and 2021 forms.

TABLE 4.3B: Pass Rates of Candidates before and after Equating

Form	N	Pass rate before equating	Difference	Pass rate after equating	Difference
2021	93	76.34%	15.06%	83.87%	7.53%
2020	93	91.40%		91.40%	

In conclusion, the results from this analysis demonstrate that without equating, test form differences in difficulty can lead to significant inequalities in student classification. By applying equipercentile equating, the researcher was able to mitigate these biases and ensure that students were classified more fairly. The reduction in the pass rate difference after equating underscores the importance of using robust equating methods to ensure that classification decisions are based on valid comparisons of student performance, rather than differences in test difficulty. This approach helps to provide a more accurate and equitable assessment of student outcomes across different test versions.

4.4 Determining whether scores from the two test forms can be used interchangeably

In this study, the interchangeability of scores between the 2020 and 2021 test forms was assessed using an F-test statistic, following the theoretical framework provided by Kolen & Brennan (1987) and Dorans (2004). The F-test helps evaluate whether the variances of proportions between the two test groups (2020 and 2021 test takers) are significantly different, which is a crucial step in determining whether scores from different test forms can be used interchangeably.

The null hypothesis (H0) for the F-test was that the variances of proportions between the two groups were not significantly different, while the alternative hypothesis (HA) posited that there was a significant difference between the variances. The results, as shown in Table 4.3C, revealed that the calculated F-statistic was 0.8136, which is greater than the critical value (F Critical = 0.7084), and the p-value (0.1622) was greater than 0.05. Given that the p-value was above the typical significance threshold of 0.05, the null hypothesis was rejected, and the alternative hypothesis was adopted.

This result indicates that the variances of proportions between the two test groups are significantly different. In other words, the two test forms (2020 and 2021) exhibit different statistical properties that make them less interchangeable. As a result, caution should be exercised when using the two test forms interchangeably, as the differences in their statistical characteristics could lead to inequitable comparisons between test-takers. Therefore, while previous studies have shown high interchangeability between test forms using CTT and IRT, the results of this F-test suggest that the specific test forms in this study (2020 and 2021) may not meet the criteria for being fully interchangeable, emphasizing the importance of validating test form equivalence before making direct comparisons.

TABLE 4.3C: F-Test Two-Sample for Variances

	Form X	*Form Y*
Variance	72.2533	88.8011
Observations	93	93
Df	92	92
F	0.8136	
P(F<=f) one-tail	0.1622	
F Critical one-tail	0.7084	

4.5 Chapter summary

The chapter began by discussing the importance of assessing the unidimensionality of a test, which is crucial for ensuring that the test measures a single, dominant factor. To evaluate the unidimensionality of two test forms (Forms X and Y), factor analysis was conducted. The results from Tables 4.1A and 4.1B showed that both forms failed to meet the unidimensionality criterion, with the first factor explaining less than 30% of the variance, as suggested by Büyüköztürk (2007). Scree plots were used to confirm the number of factors, with Figures 4.1A and 4.1B showing a clear drop in variance after the first factor, supporting the conclusion of multidimensionality.

Additionally, Item Characteristic Curves (ICCs) were analyzed using STATA 14 software to assess item response patterns and local item independence. Figures 4.1C to 4.1F demonstrated that while most items showed an "S" shape (indicating good discrimination), several items violated local item independence, further supporting the conclusion of multidimensionality for both test forms.

Descriptive statistics for Forms X and Y indicated positive skewness and highlighted differences in student performance between the two forms. Despite these differences, reliability analyses showed that both forms had acceptable reliability, with Cronbach's alpha values of 0.76 for Form X and 0.72 for Form Y.

The chapter also examined the impact of equating the test scores between the two forms to ensure fair classification. The equipercentile equating process adjusted the scores from the 2021 form (Form X) to align with the 2020 form (Form Y), significantly reducing the

disparity in pass rates. Before equating, the pass rate for Form X was 76.34%, while for Form Y it was 91.40%, but after equating, the gap was reduced to 7.53%, with Form X increasing to 83.87%. This demonstrated the importance of equating in ensuring fair comparisons and mitigating the effects of test difficulty.

Finally, an F-test was conducted to assess the interchangeability of the two test forms. The results indicated that the two forms had significantly different variances, suggesting that they were not fully interchangeable. Therefore, the study emphasized the need for careful validation before using the two test forms interchangeably in order to avoid inequitable comparisons

CHAPTER 5

CONCLUSION, RECOMMENDATIONS

5.0 Research journey and Chapter Overview

This chapter summarizes the key findings of the research and offers conclusions based on the analysis of the fairness of the Malawi Nurses' Licensure Examination (NLE). The study aimed to explore the impact of test form differences between the 2020 (Form X) and 2021 (Form Y) test editions on the fairness and equivalence of the exam scores for nurses in Malawi. The research journey involved identifying a critical issue in the examination process—the potential psychometric differences between test forms over time and their effects on candidate performance. This issue was central to the study, as it addressed the validity and fairness of high-stakes decisions made based on exam scores, such as licensure and employment opportunities.

Following an extensive review of existing literature, which underscored the importance of test score equating and fairness in high-stakes examinations, the research established a clear framework for the study. The objectives were set to assess the **unidimensionality** of the exams, compare the **relative difficulty** of the test forms, detect any **bias** in grading, and evaluate the **interchangeability** of scores between the two forms.

To achieve these objectives, a **quantitative approach** was adopted, utilizing **stratified sampling** to ensure appropriate representation of both male and female candidates, as well as those from different regions of Malawi. A total of 186 nurses participated, with 93 sitting for the 2020 exam (Form X) and 93 sitting for the 2021 exam (Form Y). The exams were administered using a spiraling process, ensuring that both test forms were given simultaneously to groups of participants. The collected data were subsequently entered into a secure database for analysis.

To achieve these objectives, a **quantitative methodology** was employed, involving the use of **stratified sampling** to ensure representative participation from male and female candidates, as well as those from diverse regions of Malawi. The data collection process followed a well-structured procedure, with the researcher and research assistant ensuring standardized test administration and securing the integrity of the exam scripts. Once the data were collected, the researcher used various **statistical techniques**, including **linear equating**, **equipercentile equating**, **factor analysis**, and **analysis of variance (ANOVA)**, to analyze the differences between the two test forms and assess their fairness.

The findings revealed important insights into the psychometric properties of the two test forms, their difficulty levels, and any potential biases present in the classification of candidates. The analysis showed whether the two test forms were **interchangeable** and whether equating methods could help adjust for any observed differences in test difficulty, ensuring fairness in the licensure process.

In this chapter, the **conclusions** drawn from these findings will be presented, followed by a set of **recommendations** aimed at improving the fairness and transparency of future

licensure examinations. These recommendations are based on the study's outcomes and the identified gaps in the current testing process, with the goal of ensuring that all candidates are evaluated equitably, regardless of the test form they take. The chapter concludes with a reflection on the implications of the study for policymakers, examination bodies, and the broader health sector, highlighting the importance of continued efforts to enhance fairness in high-stakes testing

5.1 Conclusion

The Study focused on investigating fairness of Nurses' Licensure examinations through score equating. The data was collected using the 2020 and 2021 Nurses' and Midwives Technicians (NMT) paper 1. In general, the observations and the findings made by the researcher showed that NMT paper 1 constructed and administered by the Nurses' Council in 2020 and 2021 sessions did not fully comply with unidimensionality parameters, lacking one principle factor. These findings are in support of Plake (1995) who argues that many licensure examinations consist of subcategories or sub-disciplines that may not be strongly unidimensional as a set. The test forms used for this study were no exceptional.

Moreover, results indicated no significant difference in difficulty across the **2020** and **2021** forms, indicating their suitability for equating since equating is done on test form whose difficulty levels are not totally different. This is an indication that the Nurses' Council used the test blueprint adequately to ensure that each of the forms meet set specifications and contain the same format of items. However, certain questions were not formulated to measure the same construct.

Further, it has been established that before equating, the pass rate on the **2021** form was **76.34%** while the pass rate on the **2020** form was **91.40%** representing a difference of **15.06%.** After equating, the pass rate of the **2021** form improved to **83.87%** reducing the difference to **7.53%.** These differences caused inequalities in decision making. An examinee who has taken Form X (the harder form), for instance, failed because s/he had scored **48%**. Another examinee who took Form Y test (the easier Form) passed because s/he had scored **50%**. However, **Table 4.6** shows that after equating, that is, after scores were adjusted for form difficulty, an examinee who got **48%** turned out to be more proficient than an examinee who got a score of **50%** on Form Y.

The statistical analysis revealed crucial insights into the interchangeability of the **2020** and **2021** test forms. With a t-statistic of t(184) = -1.754 and a p-value exceeding 0.05, there is insufficient evidence to reject the null hypothesis, indicating that the means of the two test forms do not significantly differ. Moreover, the correlation coefficient between the forms, calculated at -0.289, suggests a modest but noticeable negative relationship. While this correlation indicates some degree of association between the Test forms, its magnitude is relatively low, implying that the forms are not perfectly interchangeable. Therefore, while the results suggest some degree of similarity between the forms, caution should be exercised in assuming complete interchangeability of the **2020** and **2021** test forms, as there may be differences impacting test performance.

5.2 Recommendations

Based on the findings of this Study, the following recommendations are being made for policymakers and stakeholders to enhance the reliability and equity of the Nurses' Licensure Examination in Malawi:

i. **Enhancement of Test Construction**. The Nurses' Council may consider revisiting the construction process of the NMT paper 1 to ensure alignment with unidimensionality parameters. Attention should be given to construct items that reflect a single principle factor, thereby improving the overall fairness and validity of the examination. A test is assumed to be unidimensional only when the individual items in the two tests measure the same trait (Hambleton et al., 1991).

ii. **Equating Procedures:** To maintain fairness in decision-making, Nurses' Council may consider equating scores across different test forms, especially due to observed pass-rate disparities between 2020 and 2021 versions. Dorans, et al (2010) emphasize that despite similar blueprints, test editions vary in psychometric properties, necessitating accurate equating to adjust for form difficulty discrepancies. This process ensures fairness and prevents inequalities stemming from divergent difficulty levels.

iii. **Consistency in Test Specifications:** The Nurses' Council may consider using test specification tables across different test forms to ensure consistency in difficult levels of test items.

iv. **Educational Implications:** There is need for educators and administrators of nursing education to be aware of the implications of equating scores and its impact on decision-making processes. Understanding that scores on different forms may not directly reflect proficiency levels can aid in making informed decisions regarding examinees' competence.

v. **Continuous Monitoring and Improvement:** Continuous monitoring of examination processes and outcomes is essential to identify and address any

inequalities. The Nurses' Council may consider implementing mechanisms for ongoing evaluation and improvement of examination practices to uphold fairness and integrity in licensure assessments.

5.3 Suggestions for further study

Further studies could explore the following areas:

i. The present Study utilized Observed score equating methods based on Classical Test Theory (CTT) equating models. However, it acknowledged that these results might not be generalized to true scores of examinees, which require Item Response Theory (IRT) equating methods. Future studies should employ IRT equating methods to establish trends in true scores with respect to examinees abilities.

ii. Future researchers should explore the effect of repeaters on examination performance to better understand potential biases and factors influencing outcomes.

iii. Future research should aim to compare equated scores across different testing procedures. This comparative analysis can provide a more comprehensive understanding of fairness and equity in licensure examinations.

5.4 Contributions of the study

This study makes several key contributions to the field of educational measurement, particularly in the areas of test form comparison, equating, and validation.

1. **Unidimensionality and Multidimensionality in Test Forms**: One of the primary contributions of this study is the assessment of the unidimensionality assumption in two different test forms (2020 and 2021). The factor analysis results revealed

that neither test form was unidimensional, as both forms failed to meet the 30% variance criterion for the first factor, as suggested by Büyüköztürk (2007). This finding challenges the assumption that tests are often unidimensional, and highlights the need for multidimensional models to accurately represent the complexity of student performance (Büyüköztürk, 2007). Furthermore, the violation of local item independence, as revealed by the Item Characteristic Curves (ICCs), suggests that multidimensionality is a factor that can affect both test forms, providing valuable insight into the limitations of using a unidimensional framework in such assessments (Embretson & Reise, 2013). This analysis contributes to the growing body of research emphasizing the importance of considering multidimensionality in test design and analysis.

2. **Test Equating and Fairness in Student Classification**: Another significant contribution is the application of **equipercentile equating** to adjust for differences in test difficulty between the two forms. Before equating, the study found a substantial difference in pass rates, with the 2021 form having a lower pass rate than the 2020 form. The equating process helped reduce this disparity, ensuring a more equitable classification of students into passing and failing categories. This finding supports previous research (Chulu & Sires, 2011) that emphasizes the importance of equating in ensuring fairness across test forms. By equating the scores, the study demonstrated how statistical methods can be used to correct for inherent test differences, making comparisons of student performance more valid and equitable. This contributes to the literature on test fairness and highlights the

need for rigorous equating processes to mitigate the effects of test difficulty (Kolen & Brennan, 2004).

3. **Interchangeability of Test Forms**: The study also contributes to the understanding of **test form interchangeability** by employing an F-test to compare the variances between the 2020 and 2021 test forms. The results indicated that the two forms exhibited significantly different statistical properties, suggesting that the forms were not fully interchangeable. This finding underscores the importance of validating test equivalence before using different forms interchangeably. Previous studies have shown that while Classical Test Theory (CTT) and Item Response Theory (IRT) often demonstrate high interchangeability between test forms, this study suggests that test forms with different statistical characteristics require careful evaluation before being used interchangeably (Dorans, 2004; Kolen & Brennan, 1987). This contributes to the ongoing discussion on the need for validation and careful assessment when applying test forms in educational contexts.

REFERENCES

Albano, A. D. (2010). *Equate Statistical methods for test equating [Computer software manual]*. http://CRAN.R-project.org/package=equate (R package)

Albano, A. (2014). *The equate package: Equipercentile equating in R*. https://cran.r-project.org/web/packages/equate/equate.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H. (1987). Technical and practical issues in equating: A discussion of four papers. *International Journal of Testing*, *1*(1), 38-52. http://doi.org/10.1080/15305058.2010.528096

Angoff, W. H. (1984). Test score equating. In H. W. Reese & P. N. Chase (Eds.), *Handbook of research in educational measurement* (pp. 467-501). American Council on Education.

AREA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aşiret, S., & Sünbül, S. Ö. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory & Practice*, *16*(2), 123-38.

Atemafac, J. (2014). *Consequences for nursing graduates of failing the national council Licensure examination (NCLEX)* (Doctoral dissertation). Walden University.

Atsua, T. G., uzoeshi, I. V., Oludi, P., & wagbara, E. S. (2018). Equating 2015 and 2016 Basic Education Certificate Examination on Civic Education using Classical Test

Theory and Item Response Theory in Oyo State, NIGERIA. *Journal of Pristine, 14*(1), 59-67.

Akhun, R. (1984). Reliability coefficients and Fischer's Z transformation. *Journal of Educational Measurement*, 21(2), 95-108.

Baghaei, P. (2010). Test score equating and fairness in language assessment. *JELS,* 1(3), 113-128

Braun, H. I. (1982). Educational testing: A review of some recent developments. *Review of Educational Research,* 52(4), 447-479. http://doi.org/10.3102/00346543052004447

Barnard, A. (1996). Test equating: A critical review of methodologies and applications. *Measurement and Evaluation in Counseling and Development*, 29(4), 193-206. http://doi.org/10.1080/07481756.1996.12068981

Bhandari, P. (2022). *Ethical Considerations in Research. Types & Examples.* https://www.scribbr.com/methodology/research-ethics/.

Büyüköztürk, Ş. (2007). *Factor analysis: A practice book for research*. Pegem A Yayıncılık.

Bowers, A. J., & Pearson, M. (2015). Longitudinal study on the effects of score equating on student performance. *Educational Assessment, 20*(3), 238-254.

Chakwera, E., Khembo, D., & Sireci, S. (2004). High-stakes testing in the warm heart of Africa: The challenges and successes of the Malawi National Examinations Board. *Analysis Archives,* 12(29), 269-80.

Chulu, B. W. & Sireci, S.G (2011) Importance of Equating High-Stakes Educational Measurements. *International Journal of Testing*,11(1), 38-52. http://doi.org/10.1080/15305058.2010.528096

Clemans, W. V. (1993). Item response theory, vertical scaling, and something's awry in the state of test mark. *Educational Assessment, 1*(4), 329-347

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley

Connelly, M. (2020). Inclusion and exclusion criteria. *Medsurg Nursing: Pitman*, 29(2), 125-116.

Cook, L. L., &Eignor, D.R. (1991). An NCMF instructional module on IRT equating methods. *Educational Measurement: Issues and Practice,10(1)*,37-45.

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics*. Texas A & M University

Creswell, J.W. (1998). *Qualitative inquiry and research design: Choosing among five traditions*. Sage

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

Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic B-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement, 46*(2), 135-158.

DBC (2022). *Principle Component Analysis (PCA) – Easily explained*. https://databasecamp.de/en/statistics/principal-component-analysis-en .

Donlon, T. F. (1984). *The College Board technical handbook for the scholastic aptitude test and achievement tests*. College Board.

Dorans, N. J., & Walker, M. E. (2004). Methods and procedures for the verification of equivalence of examinations. *Applied Psychological Measurement, 28*(1), 3-26.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series, 2010*(2), i-41.

Dorans, N. J. (1990). Equity in test scores: Psychometric perspectives. *Journal of Educational Measurement,* 27(4), 377-387. http://doi.org/10.1111/j.2044-2980.1990.tb00675.x

Dorans, N. J., & Holland, P. W. (2000). The role of statistical equating in test interpretation. *Journal of Educational Measurement*, 37(2), 267-282. http://doi.org/10.1111/j.2044-2980.2000.tb01244

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement, 58(3), 357–381.* http://doi.org/10.1177/0013164498058003001

Fraenkel, J.R. & Wallen, N.E. (2000). *How to design and evaluate research in education.* (4th ed.). Library of Congress

Garg, R. (2016). Methodology for research I. *Indian J Anaesth*., 60(9), 640-645. http://doi.org/10.4103/0019-5049.190619.

Godfrey, K. E. (2007). *A comparison of kernel equating and IRT true score equating methods*. The University of North Carolina at Greensboro.

Gonzalez, M., & Lemos, J. M. (2016). Comparison of IRT and CTT for item difficulty estimation in high-stakes tests. *Journal of Educational Measurement, 53*(4), 301-319.

Guba, E. G., & Gephart, W. J. (1970). *Training Materials for Research, Development and Diffusion Training Programs.* Final Report. Office of Education, USA

Hanson, B. A. (1996). Testing for differences in test score distributions using log-linear models. *Applied Measurement in Education, 9(1),* 305–321

Hanson, B. A., Zeng, L., & Colton, D. A. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (No. 94). American College Testing Program

Hambleton, R. K. (1991). *Fundamentals of item response theory*. SAGE Publications

Hayes, H., & Embretson, S. E. (2012). Psychological measurement: Scaling and analysis. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed., pp. 169–188). American Psychological Association. https://doi.org/10.1037/0000318-009

Holland, P. W., & Thayer, D. T. (1989). The kernel method of equating score distributions. *ETS Research Report Series*, *1989*(1), i-45.

Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS RR-89-07). ETS

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

Holmes, J. P. (1986). Equating and score comparability: A historical overview. *Educational Researcher, 15*(3), 6-11. http://doi.org/10.3102/00346543150003006

Jodoin, M. G., Keller, L. A., &Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education*, *71*(3), 229-250.

Keboola (2022). *A guide to Principle Component Analysis (PCA) for Machine Learning*. https://www.keboola.com/blog/pca-machine-learninn,

Kolen MJ, Brennan RL (2004). *Test Equating, Scaling, and Linking*. Springer

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, 18(1), 1-11

Kolen, M. J. (1988). Traditional equating methodology. Educational Measurement: Issues and Practice. *Psychometrika*, 71(1), 207-209

Kolen, M. J., Brennan, R. L., Kolen, M. J., & Brennan, R. L. (2014). Score scales. *Test Equating, Scaling, and Linking: Methods and Practices*. http://doi.org/ 10.1007/978-1-4939-0317-7

Kolen, M. J., & Brennan, R. L. (1987). *Test equating: Methods and practices*. Springer-Verlag

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of general educational development. *Journal of Educational measurement*, 45(2), 279-293.

Kolen, M. J., Brennan, R. L., Kolen, M. J., & Brennan, R. L. (1995). Introduction and concepts. *Test equating: methods and practices*, 12(6), 1-27.

Kolen, M.J., & Brennan, R.L. (2004). *Test Equating, scaling, and linking*. Spring.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and psychological measurement*, *30*(3), 607-610.

Kuhn, H. W., & Quandt, R. E. (1962). An experimental study of the simplex method. *Proceedings of Symposia in Applied Mathematics*, 15 (1962), 107-124

Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (Vol. 9). Cambridge University Press.

Langer, M. M., & Swanson, D. B. (2010). Practical considerations in equating progress tests. *Medical teacher*, *32*(6), 509-512.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum

Leedy, P. D., & Ormrod, J. E. (2023). *Practical Research: Planning and Design*. Pearson

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement*, *30*(1), 23-39.

Livingston, S. A. (2014). *Equating test scores (without IRT),* (2nd. ed.). Educational testing service.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person parameters based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62(1),* 921–943.

Marco, G.L., Petersen, N.S. & Steward, E.E.A. (1979). *A test of the adequacy curvilinear score equating models*. Paper presented at the 1979 Computer Adaptive Testing Conference, Minneapolis.

McCumpsey, K. (2011). Assisting nursing graduates who have failed the National Council Licensure Examination (NCLEX®). *ASBN Update*, 1(2), 24-25.

Mgawadere, F., Unkels, R., Kazembe, A., & van den Broek, N.(2017). Factors associated with maternal mortality in Malawi: Application of the three delays model. *BMC Pregnancy Childbirth*, *17*(1). http://doi.org/10.1186/s12884-017-1406-5.

Miaoulis, G., & Michener, R. D. (1976). *An introduction to sampling*. Kendall.

Miaoulis, G., & Michener, W. (1976). *Sampling: A guide for social scientists*. Macmillan.

Mislevy, R. J., Sheehan, K. M., &Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*(1), 55-78.

Moghadam, M., &Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a reading comprehension test. *Language Testing in Asia*, *10*(1), 7-12.

Moses, T., & Holland, P. W. (2009). Selection strategies for univariate log-linear smoothing models and their effect on equating function accuracy. *Journal of Educational Measurement, 46*(2), 159-176. http://doi.org/10.1111/j.1745-3984.2009.00075

Mwale, C.M, & Mafuta, C. (2019). *Annual Prevalence of Suicide in Malawi.* https://sjog.uk/pdf/Research/Suicide-Prevalence-in-Malawi.pdf.

Ndalichako, J. L., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice. *Review of Educational Research,* 57(4), 697-744. https://doi.org/10.3102/0034654319862495

Neuman, W. L. (2013). *Understanding Research: Pearson New International Edition*. Pearson Higher Ed.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of mathematical psychology*, *3*(1), 1-18.

Nworgu, B. G. &Agah, J. J.(2012). Application of three parameter logistic model in the calibration of a Mathematics achievement test. *Journal of Educational Assessment in Africa,* 7 (1), 162-172.

Ozdemir, B. (2014). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia - Social and Behavioral Sciences, 174* (2015), 2075 – 2083

Perera, M., Schmiedeknecht, K., Schell, E., Jere, J., Geoffroy, E., & Rankina, S. (2015). *Predictors of Workforce Retention Among Malawian Nurse Graduates of a Scholarship Program: A Mixed-Methods Study. Glob Health Sci Pract.,* 3(1),85-96. http://doi.org/10.9745/GHSP-D-14-00170

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*(2), 137-156.

Plake, B. S. (1995). *9. Differential Item Functioning In Licensure Tests*. University of Nebraska-Lincoln

Price, T., Lynn, N., Coombes, L., Roberts, M., Gale, T., Bere, S., & Archer, J. (2018). The International Landscape of Medical Licensing Examinations: A typology Derived From a Systematic Review. *International Journal of Health Policy Management*, 7(9), 782–790.

Progar, R., Smith, J., & Taylor, C. (2008). Person parameter estimates: A comparison of CTT and IRT. *Educational and Psychological Measurement, 68*(5), 687-704.

Progar, Š.,Sočan, G., &Peč, M. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology*, *17*(3), 5-24.

Rosenbaum, P. R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology*, *40(1),* 43–49

Sangala, T. (2017, March 2). *Seventy-five percent fail Nurse Midwifery Technicians examinations.* https://archive.times.mw/index.php/2017/03/02/75-fail-nurse-midwifery-technicians-examinations/

Sanzivieri, V., Wiberg, M. & Matteucci, M. (2017). A review of test equating methods with a special focus on irt-based approaches. *Statistica,* 77(4), 329-352

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement*, *12*(1), 69-82

Skaggs, G., & Lissitz, R. (1986b). The relationship between test equating methods and score comparability. *Applied Psychological Measurement,* 10(2), 195-205.

Swain, J. (2016). *Ethical considerations in research and education*. Sage Publications.

Tate, G. R., & Baird, L. L. (2014**).** Item dimensionality in higher education performance assessments. *Educational Psychology*, *36*(6), 764-781.

Temitope, B. (2021). Determination of the equivalence of WAEC and NECO SSCE chemistry items using linear equating approaches of classical test theory and item response theory. *Bulgarian Journal of Science & Education Policy*, *15*(1), 187-207.

The Association of Chartered Certified Accountants (2013) *Key Health Challenges for Zambia*. Author

Thorndike, R. L. (1982). *Educational measurement: Theory and practice*. American Council on Education

Von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer.

Wang, T., & Li, W. (2019). Differential item functioning and bias correction in test form comparisons. *Educational Assessment, 24*(3), 207-225.

Williamson, Karp, Dalphin& Gray (1982). The research craft: *An Introduction to Social Research Methods.* Little, Brown, and Company

Yim, M. K., & Huh, S. (2006). Test equating of the medical licensing examination in 2003 and 2004 based on the item response theory. *Journal of Educational Evaluation for Health Professions*, *3*(2). http:// doi.org/10.3352/jeehp.2006.3.2

Zheng, L., & van der Linden, W. J. (2010). Equating with small sample sizes: A comparison of methods. *Applied Psychological Measurement, 34*(5), 357-375. http://doi.org/ 10.1177/0146621609343510

Zhu, X., & Liu, Y. (2020). Medical licensure exams in China: The impact of test form differences on fairness. *International Journal of Testing, 20*(2), 147-163.

APPENDICES

APPENDIX 1: Research Ethics and Regulatory approval and permit

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

Our Ref: P.06/23/273

Your Ref.:

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

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

29th September 2023

Mr Isaac Nyirongo,
Department of Education,
University of Malawi.
P.O. Box 280.
Zomba.

Dear Mr Nyirongo,

RESEARCH ETHICS AND REGULATORY APPROVAL AND PERMIT FOR PROTOCOL NO. P.06/23/273. "AN INVESTIGATING FAIRNESS OF MALAWI NURSES EXAMINATIONS THROUGH TEST SCORE EQUATING: THE CASE OF SELECTED NURSES' TRAINING COLLEGES UNDER THE CHRISTIAN HEALTH ASSOCIATION OF MALAWI (CHAM).

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.

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.

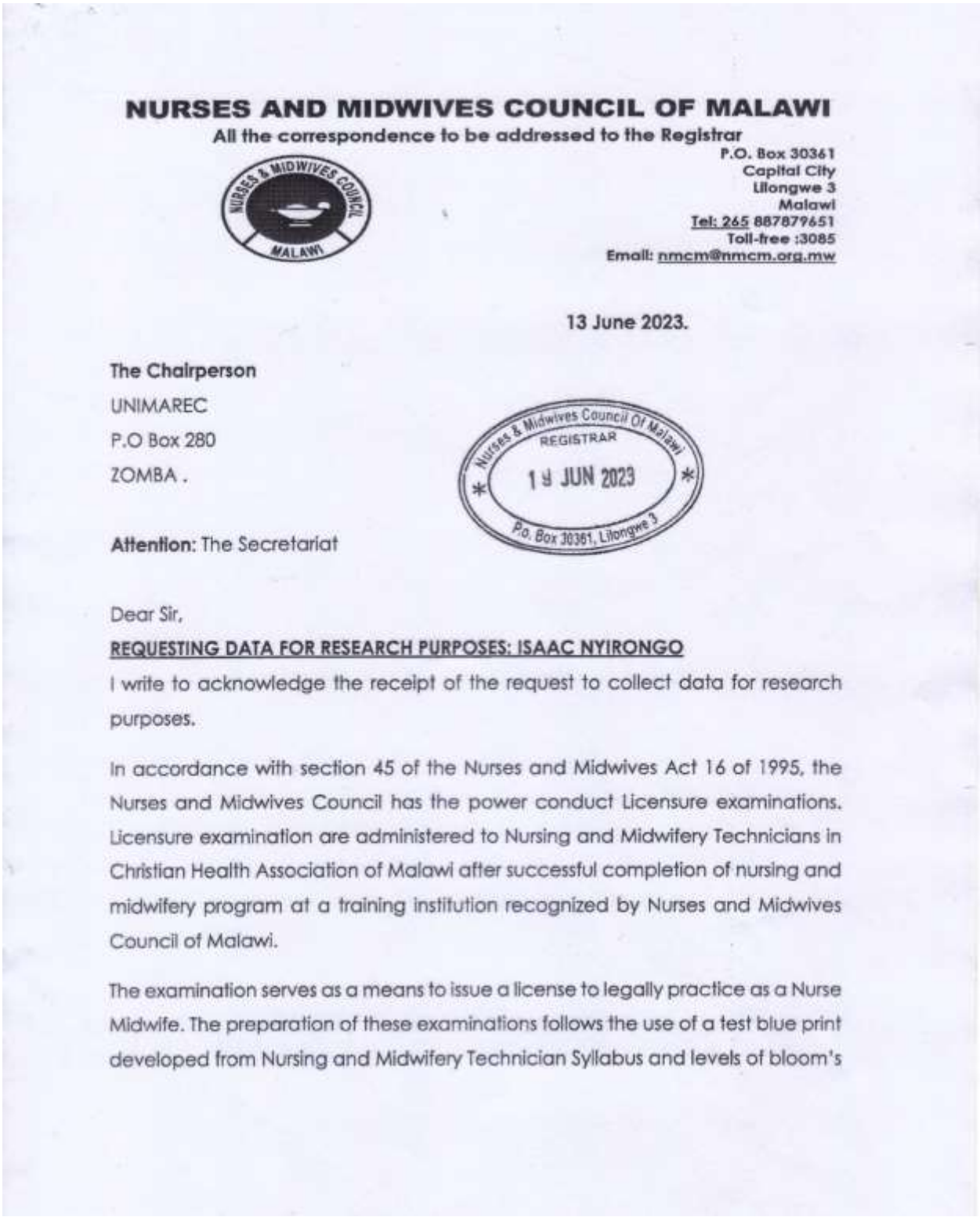
UNIVERSITY OF MALAWI
RESEARCH ETHICS COMMITTEE
29 SEP 2023
APPROVED
P.O. BOX 280, ZOMBA

Yours Sincerely,

Dr Victoria Ndolo
CHAIRPERSON OF UNIMAREC

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

APPENDIX 2: Acceptance to use the Nurse's Council Examinations for data collection



NURSES AND MIDWIVES COUNCIL OF MALAWI

All the correspondence to be addressed to the Registrar

P.O. Box 30361
Capital City
Lilongwe 3
Malawi
Tel: 265 887879651
Toll-free :3085
Email: nmcm@nmcm.org.mw

13 June 2023.

The Chairperson
UNIMAREC
P.O Box 280
ZOMBA .

Nurses & Midwives Council Of Malawi
REGISTRAR
19 JUN 2023
P.O. Box 30361, Lilongwe 3

Attention: The Secretariat

Dear Sir,

REQUESTING DATA FOR RESEARCH PURPOSES: ISAAC NYIRONGO

I write to acknowledge the receipt of the request to collect data for research purposes.

In accordance with section 45 of the Nurses and Midwives Act 16 of 1995, the Nurses and Midwives Council has the power conduct Licensure examinations. Licensure examination are administered to Nursing and Midwifery Technicians in Christian Health Association of Malawi after successful completion of nursing and midwifery program at a training institution recognized by Nurses and Midwives Council of Malawi.

The examination serves as a means to issue a license to legally practice as a Nurse Midwife. The preparation of these examinations follows the use of a test blue print developed from Nursing and Midwifery Technician Syllabus and levels of bloom's

taxonomy. The levels measured in the bloom's taxonomy for this cadre (Technician level) are application and analysis.

The candidates write four papers; 2 general Nursing and 2 Midwifery papers. The candidates are subjected to multiple choice questions and each paper has 100 marks administered in three hours. The pass mark is 50%. The examination is highly protected

We therefore, write to inform you that the request by Mr. Isaac Nyirongo to collect data for research purposes has been granted under the following conditions that:

- Two general Nursing Papers each 100 marks for 2020 and 2021 shall be made available and kept at Nurses and Midwives Council of Malawi up until needed.
- The examinations shall be administered under escort of an Officer from Nurses and Midwives Council of Malawi for security purposes.
- After administration of the examinations, the scripts shall be marked at Nurses and Midwives Council of Malawi Offices and allow the Researcher get the required data.
- The Researcher will cater for the allowances for the Officer who will escort to the administration of examinations.

Nurses and Midwives Council of Malawi wishes you all the best in your studies

Yours Faithfully,

Mrs. Judith Chirembo

REGISTRAR/CHIEF EXECUTIVE OFFICER

Nurses & Midwives Council Of Malawi
REGISTRAR
19 JUN 2023
P.o. Box 30361, Lilongwe 3

APPENDIX 3: Consent form

INFORMED CONSENT FORM (ICF)

Study title:

INVESTIGATING FAIRNESS OF MALAWI NURSES EXAMINATIONS THROUGH TEST SCORE EQUATING: THE CASE OF SELECTED NURSES' TRAINING COLLEGES UNDER THE CHRISTIAN HEALTH ASSOCIATION OF MALAWI (CHAM)

Principal investigator; ISAAC MATIAS NYIRONGO, School of Education, University of Malawi

Introduction

Examining agencies administer new editions of tests over a specified period of time. But, for security reasons, they cannot use the same test form on different administrations. They have to come up with multiple edition of the same test to be administered on different dates. However, research has shown that even if different test editions may be built to a common blueprint and be designed to measure the same constructs, they always differ in their psychometric properties. While some test forms consist of easy items, others may have difficult items that can cause examinees' scores to differ. However, in some cases educational tests are not statistically equated to account for test score differences over time, leading to wrong interpretations of students' performance

Purpose of study

This Study intends to investigate the fairness of Nurses' Licensure examinations through test score equating. It therefore requires a research methodology that interacts with nurses who are in their final year of the nursing studies and will sit for the Licensure Examinations at the end of their training.

You are being invited to take part in this study because you are one of the nurses who will sit for Licensure Examinations. You will have to sit for the either form X of form Y as guided by the main investigator and the research assistant. I, Isaac Matias Nyirongo, the main investigator, with the assistance of my colleague you see besides me (Officer from Nurses' Examinations Board) would like to administer a Licensure Examination which will take 3 hours. You have choice to agree or not. We will follow all COVID-19 preventative measures.

Risks and discomforts of the research study:

There are no major risks involved in this study. An inconvenience may be the time and effort you will lender to be a participant. You may also lose your writing materials and this will be replaced upon establishing evidence through witnesses. You do not have to take part if you feel uncomfortable.

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

Potential benefit of the research study:

Although there are no monetary benefits to you for participating in the study however your participation will help to improve the administration of Nurses' Licensure Examinations (NLEs) in Malawi. The results from this study will inform authorities in the health sector especially the credentialing board members and other individuals who hold major responsibility for preparing, administering, and scoring credentialing examinations as well as other stakeholders in credentialing health professionals on best practices when administering nurses' licensure examinations. Besides, the study will also help future researchers who would wish to conduct similar studies to check fairness of Licensure Examinations in other sectors. The information from this research would not only help researchers gather the concepts of others in particular research, but also allow them to learn about the results of other similar studies.

Alternative procedures:

There are no alternative procedures.

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.

Provisions for confidentiality

I assure your confidentiality during and after the examination. 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. The only exceptions for disclosing your identity and both of them are rare would be:

1. Personal information may be disclosed if required by law
2. The Human Research Ethics Committee of the University may exceptionally require personal data to respond to a formal complaint, or for a compliance audit

Research related injury:

I assure you that there are no major risks for your participation in this study, however if someone gets hurt in course of participating in this study you will be provided with the appropriate medical care and support. However there will be no compensation for any injury or harm during the study.

Contact for additional information

If you have any questions about this research, you can ask me or the Chairperson of Research at University of Malawi who heads the University of Malawi Research Ethics Committee (UNIMAREC) on the following contact details:

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

Isaac Matias Nyirongo

School of Education

University of Malawi

P.O Box 280

Zomba.

isaac4nyirongo@gmail.com; +265 (0) 881 971 001 / +265 (0) 996 414 799

Dr. Victoria Ndolo
Chairperson of University of Malawi Research Ethics Committee (UNIMAREC)
University of Malawi
P.O. Box 280
Zomba.
+265 995 0427 60

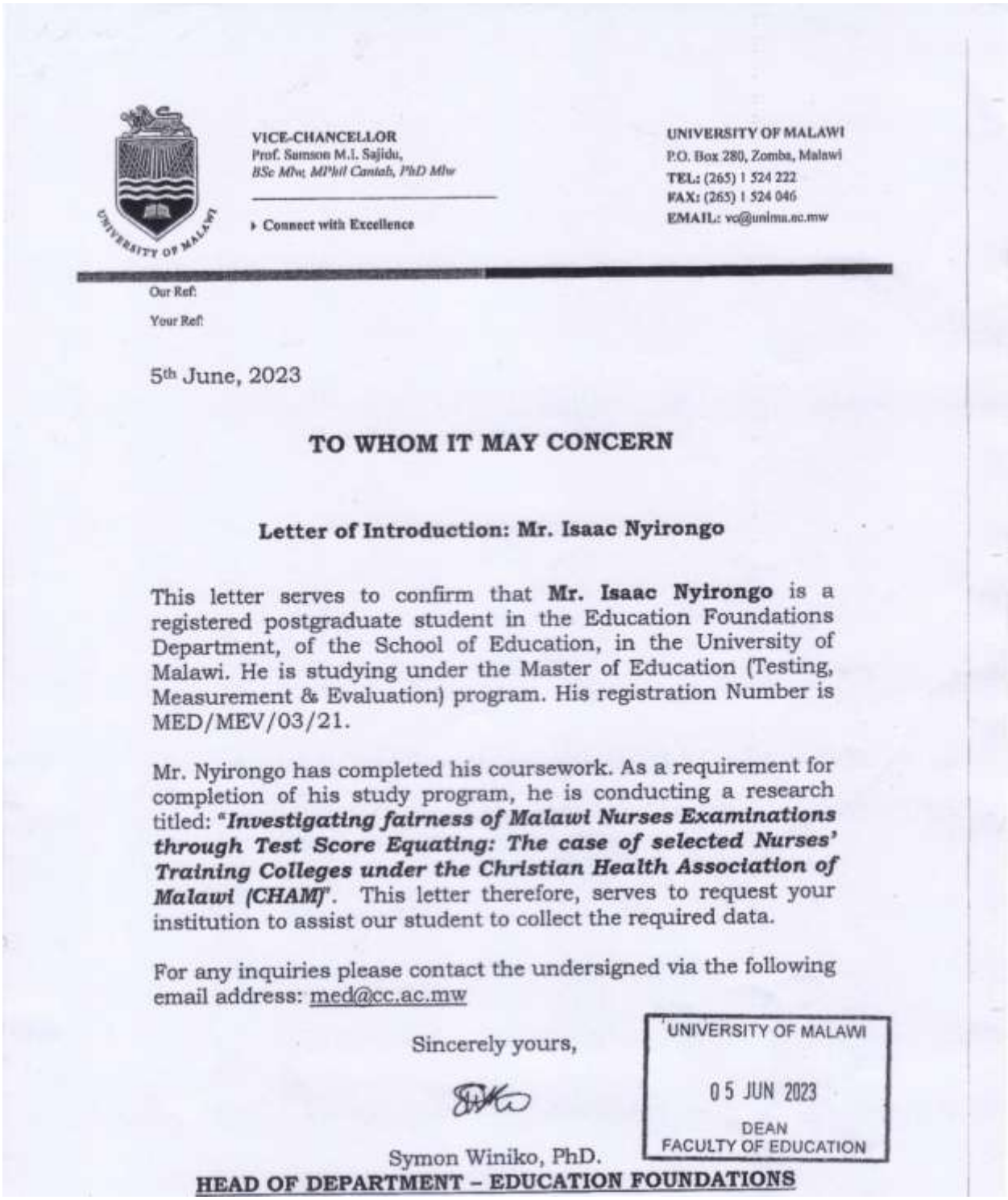
Do you agree to Continue with the study? Yes ✓ No _____

THANK YOU

THE DEAN OF FACULTY
ST. LUKE'S COLLEGE OF
NURSING AND MIDWIFERY
13 NOV 2023
P.O. BOX 21, CHILEMA

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

APPENDIX 4: Introductory letter



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:

5th June, 2023

TO WHOM IT MAY CONCERN

Letter of Introduction: Mr. Isaac Nyirongo

This letter serves to confirm that **Mr. Isaac Nyirongo** 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/03/21.

Mr. Nyirongo has completed his coursework. As a requirement for completion of his study program, he is conducting a research titled: "***Investigating fairness of Malawi Nurses Examinations through Test Score Equating: The case of selected Nurses' Training Colleges under the Christian Health Association of Malawi (CHAM)***". 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

Sincerely yours,

UNIVERSITY OF MALAWI
05 JUN 2023
DEAN
FACULTY OF EDUCATION

Symon Winiko, PhD.
HEAD OF DEPARTMENT – EDUCATION FOUNDATIONS

APPENDIX 5: An output file for equating

Scale	yx	se	se.b
1.00	6.698300	26.937133	3.587058
2.00	7.633800	26.035850	3.515846
3.00	8.569400	25.150738	3.444754
4.00	9.504900	24.281798	3.373787
5.00	10.440500	23.429029	3.302955
6.00	11.376000	22.592432	3.232266
7.00	12.311600	21.772007	3.161729
8.00	13.247100	20.967754	3.091356
9.00	14.182700	20.179673	3.021157
10.00	15.118200	19.407763	2.951145
11.00	16.053700	18.652025	2.881333
12.00	16.989300	17.912459	2.811737
13.00	17.924800	17.189064	2.742373
14.00	18.860400	16.481842	2.673259
15.00	19.795900	15.790791	2.604415
16.00	20.731500	15.115912	2.535863
17.00	21.667000	14.457204	2.467627
18.00	22.602600	13.814669	2.399734
19.00	23.538100	13.188305	2.332214
20.00	24.473700	12.578113	2.265101
21.00	25.409200	11.984092	2.198432
22.00	26.344700	11.406244	2.132248
23.00	27.280300	10.844567	2.066596
24.00	28.215800	10.299062	2.001528
25.00	29.151400	9.769728	1.937104
26.00	30.086900	9.256567	1.873389
27.00	31.022500	8.759577	1.810458
28.00	31.958000	8.278759	1.748397
29.00	32.893600	7.814113	1.687301
30.00	33.829100	7.365638	1.627279
31.00	34.764700	6.933335	1.568454
32.00	35.700200	6.517204	1.510966
33.00	36.635700	6.117245	1.454974
34.00	37.571300	5.733457	1.400656
35.00	38.506800	5.365842	1.348216
36.00	39.442400	5.014397	1.297880
37.00	40.377900	4.679125	1.249904
38.00	41.313500	4.360025	1.204569
39.00	42.249000	4.057096	1.162184
40.00	43.184600	3.770339	1.123083
41.00	44.120100	3.499754	1.087621
42.00	45.055700	3.245340	1.056164

43.00	45.991200	3.007098	1.029079
44.00	46.926700	2.785029	1.006720
45.00	47.862300	2.579130	0.989407
46.00	48.797800	2.389404	0.977408
47.00	49.733400	2.215849	0.970920
48.00	50.668900	2.058466	0.970054
49.00	51.604500	1.917255	0.974824
50.00	52.540000	1.792216	0.985149
51.00	53.475600	1.683348	1.000857
52.00	54.411100	1.590652	1.021700
53.00	55.346700	1.514128	1.047370
54.00	56.282200	1.453775	1.077524
55.00	57.217700	1.409595	1.111796
56.00	58.153300	1.381586	1.149819
57.00	59.088800	1.369749	1.191233
58.00	60.024400	1.374083	1.235697
59.00	60.959900	1.394590	1.282894
60.00	61.895500	1.431268	1.332533
61.00	62.831000	1.484118	1.384353
62.00	63.766600	1.553139	1.438117
63.00	64.702100	1.638333	1.493616
64.00	65.637700	1.739698	1.550663
65.00	66.573200	1.857235	1.609093
66.00	67.508700	1.990943	1.668762
67.00	68.444300	2.140824	1.729540
68.00	69.379800	2.306876	1.791316
69.00	70.315400	2.489100	1.853989
70.00	71.250900	2.687496	1.917471
71.00	72.186500	2.902063	1.981686
72.00	73.122000	3.132802	2.046562
73.00	74.057600	3.379713	2.112041
74.00	74.993100	3.642796	2.178067
75.00	75.928700	3.922050	2.244592
76.00	76.864200	4.217477	2.311573
77.00	77.799700	4.529075	2.378972
78.00	78.735300	4.856844	2.446754
79.00	79.670800	5.200786	2.514887
80.00	80.606400	5.560899	2.583345
81.00	81.541900	5.937184	2.652102
82.00	82.477500	6.329641	2.721135
83.00	83.413000	6.738269	2.790425
84.00	84.348600	7.163070	2.859951
85.00	85.284100	7.604042	2.929698

86.00	86.219700	8.061186	2.999650
87.00	87.155200	8.534501	3.069793
88.00	88.090700	9.023988	3.140114
89.00	89.026300	9.529647	3.210602
90.00	89.961800	10.051478	3.281246
91.00	90.897400	10.589481	3.352035
92.00	91.832900	11.143655	3.422961
93.00	92.768500	11.714001	3.494016
94.00	93.704000	12.300519	3.565192
95.00	94.639600	12.903209	3.636481
96.00	95.575100	13.522070	3.707878
97.00	96.510600	14.157103	3.779375
98.00	97.446200	14.808308	3.850968
99.00	98.381700	15.475685	3.922652
100.00	99.317300	16.159233	3.994421