MANTEL-HAENSZEL ANALYSIS OF GENDER DIFFERENTIAL ITEM FUNCTIONING (DIF) IN MALAWI SCHOOL CERTIFICATE OF EDUCATION (MSCE) COMPUTER STUDIES EXAMINATION ITEMS

MASTER OF EDUCATION (TESTING, MEASUREMENT AND EVALUATION) THESIS

BETTY TAFWAKUWANASI ZIMBA

UNIVERSITY OF MALAWI
CHANCELLOR COLLEGE

JUNE, 2018

MANTEL-HAENSZEL ANALYSIS OF GENDER DIFFERENTIAL ITEM FUNCTIONING (DIF) IN MALAWI SCHOOL CERTIFICATE OF EDUCATION (MSCE) COMPUTER STUDIES EXAMINATION ITEMS

MASTER OF EDUCATION (TESTING, MEASUREMENT AND EVALUATION) THESIS

By

BETTY TAFWAKUWANASI ZIMBA

Bachelor of Science (Biology & Chemistry) –University of Malawi

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

University of Malawi

Chancellor College

June, 2018

DECLARATION

I, the undersigned, hereby declare that this thesis is my own original work which has not been submitted to any other institution for a similar purpose. Acknowledgements and reference citations have been made where other people's work has been used.

BETTY TAFWAKUWANASI ZIMBA

Full legal names

Signature

Date

CERTIFICATE OF APPROVAL

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

Signature: ________________________________Date:_________________________

Bob W Chulu, PhD (Lecturer)

Supervisor

DEDICATION

This thesis is dedicated to Chalo and Mabowoza families.

ACKNOWLEDGEMENTS

I wish to thank Dr. Bob Wajizigha Chulu, my thesis supervisor, for guidance throughout the entire process of writing this thesis. His advice, suggestions, criticisms and encouragement coupled with his expertise were all of great assistance.

To all lecturers in the Department of Education Foundations, University of Malawi, Chancellor College, for their constructive criticisms and guidance during the research proposal presentation.

To the Executive Director, Malawi National Examinations Board for assisting in the provision of data for the research.

To Angel M. Fidalgo for sending me a free GMHDIF program which I used in the analysis of data; my heartfelt thanks.

To Messers I. Chang, A. Muyaya, K. Kumwenda, B.K. Pindani, M. Bandawe and C. Kadango for their much needed support in debugging the computer program and all other technical assistance pertaining to this research.

To all Ministers and Christians of the New Apostolic Church; for their earnest prayers which guided me through the whole process of writing this research. Without their prayers, nothing would have materialized.

Last, but by no means least, I would like to thank my husband Charles Godwin Zimba; my children, Wanangwa, Godwin, Josephine, Richard and Margaret; my son in-laws, Hardy and Stanley; my grandchildren, Victoria, Tiwonge and Lughindiko for their patience, support and love.

ABSTRACT

The purpose of this study was to re-examine gender Differential Item Functioning (DIF) in Malawi School Certificate of Education (MSCE) Computer Studies examination items using a Classical Test Theory (CTT)-based Mantel-Haenszel (MH) DIF detection procedure. MSCE Computer Studies examinations are CTT-based. Analysis of DIF is there to enhance the validity of examinations. The study was a partial replication of a research study by Jailosi (2012) who used Item Response Theory-Likelihood Ratio (IRT-LR) DIF detection procedure on a CTT- based examination. IRT procedures have very stringent assumptions that are practically very difficult to meet. One of the assumptions is that there must be model data fit where even a slight model misfit to the data can be mistaken for DIF (Teresi, 2004). The current study was undertaken to find out the extent to which MH DIF detection procedure could detect DIF in MSCE Computer Studies examination items; determine the group which the detected DIF disadvantaged and establish the type of DIF that was detected. Thirty-four items were analysed for a random sample of 500 examinees (300 males and 200 females). Data was analysed using Generalized Mantel-Haenszel DIF (GMHDIF) Program by Fidalgo (2011). Results have shown that category ‘B’ and ‘C’ DIF magnitude was detected in 3 out of 34 items representing 8.82%. Out of the detected items, 66.67% had uniform DIF whereas 33.33% had non-uniform DIF. Both items with uniform DIF disadvantaged females. Results were discussed concerning the appropriateness of the DIF detection method based on the nature of the data. It was recommended that the examining board consider adopting gender DIF analysis to enhance examination quality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FORMULAE

LIST OF APPENDICES

LIST OF ABBREVIATIONS AND ACRONYMS

CTT	Classical Test Theory
DIF	Differential Item Functioning
ETS	Educational Testing Service
GMHDIF	Generalised Mantel-Haenszel DIF
IRT	Item Response Theory
IRT-LR	Item Response Theory-Likelihood Ratio
MANEB	Malawi National Examinations Board
MH	Mantel-Haenszel
MSCE	Malawi School Certificate of Education
PSLCE	Primary School Leaving Certificate of Education
SIBTEST	Simultaneous Item Bias Test

CHAPTER 1

BACKGROUND

1.1 Chapter Overview

This chapter provides the background information to the research topic in question, statement of the problem, purpose of the study, research objectives the study was trying to fulfill and significance of the study.

1.2 Background to the study

Tests, in most cases, are used for making important decisions such as selection, training and promotion (Bryant & Forde, 2002) which concern the future of individuals who have taken that test. Test items are designed to provide information about the examinee. Difficult items are supposed to be more demanding and easy items are supposed to be less demanding. However, sometimes test items carry with them demands other than those intended by the test developer (Abedalaziz, 2010). As such, it is imperative that the tests used for making such decisions be both reliable and valid (Maccines, 2009). One of the threats to test validity is bias; a situation which results when performance on a test is not the same for individuals from different subgroups of the population, but with the same level of trait measured by the test. Concerns about bias at the item level emerged from within the framework of test bias since a test is comprised of items (Maccines, 2009).

This concern for validity in testing led to a surge in the development of statistical methodology for item bias detection in the 1960s and 1970s. Incipient statistical investigations into item bias were predicated on the identification of items that displayed unusually large differences in the mean performance between demographic groups relative to that observed for the other items on the test (Penfield, 2001). Although these methods enabled the identification of items that were differentially difficult for one group of examinees, they lacked a rigorous method for determining whether differences in group performance on an item were caused by some form of unfairness in the item or simply differing levels of proficiency in the groups being compared. By using the framework of DIF, confounding effects of differing levels of group proficiency (item impact) are controlled because subjects are matched on ability.

A test item is said to function differentially (is a DIF item) when subjects from different groups, but with the same ability, have different probabilities of answering the item correctly. As such, DIF items can lead to biased measurement of ability because the measurement is affected by the so-called nuisance factors (Magis, Sebastien, Tuerlinckx, & De Boek, 2010). It is, therefore, important to identify items in a test that function differently between groups of examinees with same ability because such items are a threat to the validity of interpretation and use of the test scores. Issues of construct-related evidence of validity and issues of DIF are interrelated in the sense that if a test lacks construct-related evidence of validity, it means that the test contains items that are measuring constructs other than those that are intended to be measured, indicating that there is a potential for bias against or for a certain group of examinees (Atar & Kamata, 2011).The presence of DIF, therefore, jeopardizes the ideal of a correct measurement procedure (Magis, Sebastien, Tuerlinckx, & De Boek, 2010).

Claims that tests are biased against certain minority groups have led to numerous lawsuits or bans on certain types of tests. As such, tests and testing practices have come under close scrutiny to ensure that tests are free of bias against minorities (Hambleton, Swaminathan, & Rogers, 1991) because item functioning is intended to be invariant with respect to irrelevant aspects of test takers such as gender, ethnicity and socio-economic status (Badia, Prieto, & Linacre, 2002). Therefore, given the serious consequences of DIF for the test-takers, test-users and test- developers, the identification and use of procedures that can effectively detect DIF is very important (Lopez, 2012).

In recent years, two methods; judgmental and statistical, have been employed to detect item bias. Judgmental methods consist of the employment of one or more expert judges to assess items in a test in order to determine if there is a likelihood that a given group would find those items more difficult than another group because of something other than the trait being measured. However, the problem with judgmental methods is that they are subjective because expert judges are human. Regardless of their subjectivity, judgmental methods are still widely employed in screening items for item bias with the complement of statistical or empirical methods which are more objective (Hadley, 1994).

There are numerous statistical or empirical methods of detecting DIF. Some are parametric such as those based on Item Response Theory (IRT) and others are non-parametric such as those based on Classical Test Theory (CTT).

In Malawi, only two DIF studies had been done on the Malawi National Examinations Board examination items, i.e. Primary School Leaving Certificate of Education (PSLCE) Mathematics and Malawi School Certificate of Education (MSCE) Computer Studies examinations' items. Both studies used IRT-LR DIF detection method and the results (of both studies) indicated that some items exhibited DIF (Gondwe, 2012) and (Jailosi, 2012). However, according to the history of item construction by MANEB, all MANEB examinations have been constructed and analyzed based on Classical Test Theory (CTT) (Mankhomwa, 2013). As such, item analyses on MANEB examinations which are based on a different theory may not be expected to yield results of appropriate validity. Consequently, results of earlier DIF studies which were done on MANEB examination items using IRT-LR DIF detection method may be of questionable validity because IRT-based DIF detection methods, apart from being more labor-intensive and very complicated to carry out, are based on very stringent assumptions which are, practically, very difficult to meet. One of the assumptions that IRT-based methods are based on is the assumption of model-data fit where even a slight model misfit to the data can be mistaken for DIF (Teresi, 2004).

Another common problem of IRT approaches to DIF is that item parameters must be estimated in both groups; and for proper estimation, a large number of examinees with a large ability range is needed. However, in typical DIF analyses, the number of examinees in the minority group is usually small and, furthermore, the group may have a restricted ability range. In such a situation, item parameters are estimated poorly and the obtained DIF statistics may lead to erroneous decisions about the presence of DIF (Hambleton, Swaminathan, & Rogers, 1991).

The present study used CCT-based MH DIF detection procedure to re-examine DIF in MSCE Computer Studies examination items.

MH DIF detection procedure has been reported widely in literature that it has reasonable power for effective DIF detection in small samples as compared to IRT-based DIF detection methods which require extremely large samples for effective DIF detection (Clauser & Mazor, 1998; Bryant & Forde, 2002; Hambleton, Swaminathan & Rogers, 1991; Lopez, 2012; Maccines, 2009 & Teresi, 2004). So far, for MSCE Computer Studies examination results, the number of examinees at each score level is usually less than 100 which is a very small population that provides a very small sample for DIF analysis. As such, MH DIF detection procedure is expected to be more effective in detecting DIF in MSCE Computer Studies examination items compared to IRT-LR DIF detection procedure that was used in Jailosi's (2012) research.

1.3 Statement of the problem

MSCE Computer Studies examination is one of high-stakes national examinations which is developed and offered by the Malawi National Examinations Board (MANEB). The examinees' scores from this examination have a significant influence on decisions concerning selection to higher levels of education, i.e. tertiary education and placement into various programs at institutions of higher learning. The examinees' scores also influence decisions regarding certification and placement into different sectors of employment. They are also an indicator of the quality of education in the country. As such, proper analysis of items should be carried out before examinees take such high-stakes examinations to ensure validity and reliability of the scores that emanate from them.

According to literature, one of the things that enhance validity of examinations is analysis of possible DIF in items using proper DIF detection procedures which are appropriate to the data being analyzed. The identification of procedures for effective detection of DIF is very important given the serious consequences of DIF for the test-takers, users and developers (Lopez, 2012). In Malawi, only two DIF analysis research studies, so far, had been conducted on items of MANEB examinations (PSLCE Mathematics and MSCE Computer Studies). Both studies used IRT-LR DIF detection method, a parametric method based on Item Response Theory; a theory with assumptions which are, practically, very difficult to meet (Teresi, 2004).

Several research studies have been conducted on DIF in an attempt to identify the most appropriate method for the detection of DIF in various situations. For example, Hadley (1994) conducted research on performance of Mantel-Haenszel (MH) and Logistic Regression (LR) DIF detection procedures across sample size and effect size where it was found that the two methods were comparable at detecting uniform DIF with LR performing slightly better than MH. Hadley also reported that in their 1993 study, Swaminathan and Rogers had results which indicated that MH performed marginally better than LR at DIF detection (Hadley, 1994). Lopez (2012) found out that neither IRT-Likelihood Ratio Test (IRT-LRT) nor Crossing-SIBTEST and LR were effective at identifying the type of DIF that was simulated. Hambleton and Rogers (1989) compared IRT Area and MH methods in detecting potentially biased test items where they found that the results were very similar. To resolve this conflict on the identification of the most appropriate DIF detection method in various situations, there is need for more research to be conducted as the current researcher has done.

1.4 Purpose of the study

The purpose of this study was to re-examine DIF in MSCE Computer Studies examination items using a CTT-based Mantel-Haenszel DIF detection method.

The current study was a partial replication of a research study conducted by Jailosi in 2012 on the basis of methodology. Jailosi (2012) detected DIF in MSCE Computer Studies examination items using an IRT-based IRT-LR DIF detection method (Jailosi, 2012).

1.5 Research objectives

The above purpose was accomplished by fulfilling the following objectives:

1.5.1 Overall objective

To analyse gender differential item functioning (DIF) in MSCE Computer Studies examination items using Mantel-Haenszel DIF detection procedure.

1.5.2 Specific objectives

The overall objective of the study was met through accomplishment of the following specific objectives:

1. To find out the extent to which MH DIF detection procedure can detect DIF in MSCE Computer Studies examination items.
2. To determine the group (focal or reference) which the DIF in MSCE Computer Studies examination items disadvantages.
3. To establish the type of DIF (uniform or non-uniform) that MH DIF detection procedure is able to detect in MSCE Computer Studies examination items.

1.6 Significance of the study

Examination results are used for making very important decisions as regards certification, selection and placement. As such, it is important that assessment instruments should be unbiased in order to enhance equity and improve assessment. Analysis of DIF in items of an assessment instrument is one of the most important assessment practices that ensure non-biasness. Finch and French (2007 in Awuor, 2008) stated that it is evident that DIF continues to receive attention both in applied and methodological studies (Awuor, 2008).

In Malawi, only two DIF research studies had been conducted on items of MANEB examinations (PSLCE Mathematics and MSCE Computer Studies). Both studies used IRT-LR DIF detection method which is a parametric method based on Item Response Theory. This means that there is very little in literature concerning DIF studies on MANEB examinations and in Malawi as a nation. The current research will, therefore, add to the existing literature concerning DIF studies globally and in Malawi in particular.

As mentioned earlier on, the current research study is a partial replication of Jailosi's 2012 research study on detection of DIF in MSCE Computer Studies examination items. Jailosi used IRT-LR DIF detection method and the current researcher used MH DIF detection procedure. According to literature, one of the considerations in the choice of a DIF detection method is the type of data that is being examined (Teresi, 2004 & Hambleton, 2006). If MANEB may find it necessary to implement results of

DIF research, results of the current research will be easier and faster to implement because both MSCE Computer Studies examination data and MH DIF detection procedure are based on the same theory (CTT). Major manipulations in the system, which may be both economically and practically difficult, may not be necessary.

According to Teresi (2004) non-parametric DIF detection procedures are the easiest and MH is one of the easiest of the non-parametric DIF detection procedures. Teresi (2004) further stated that ease of use, although not the only consideration, does play a role in the decision regarding a method. Results of this research will, therefore, also enable MANEB to have a chance of using one of the easiest DIF detection procedures so that information about items that are identified as having DIF is presented for review by experts in time. This will enable quick informed decisions about inclusion or exclusion of such items in the examination or item bank.

Furthermore, the Mantel-Haenszel procedure provides both a statistical test of significance and an effect size estimate for DIF (magnitude of DIF) (Maccines, 2009). There are schemes for interpreting this effect size of differential performance which can inform the decision as to whether to retain those items in the common item pool or to discard them (Michaelides, 2008). The interpretation of effect size of differential performance that will be determined in results of this research will, therefore, enable MANEB to make such informed decisions.

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter overview

This chapter provides literature review of related studies. It includes subgroups in DIF analysis, types of DIF, DIF detection, MH DIF detection procedure, contingency table in general, odds and odds ratio, converting data from the general contingency table to proportions, logarithm of odds ratio, magnitude of DIF, Educational Testing Service (ETS) classification of magnitude of DIF, MSCE Computer Studies examination, comparison of 2007, 2008 and 2012 MSCE Computer Studies Paper 1 examination items, software packages used in DIF analysis, statistics implemented by GMHDIF program and theoretical framework.

2.2 Literature review of related studies

Differential Item Functioning (DIF) is present when examinees from different groups have a different probability or likelihood of answering the item correctly, after controlling for overall ability (Gierl, Khaliq, & Boughton, 1999). Once identified, DIF may be attributed to item bias or item impact. Item bias is the invariability or systematic error in how a test item measures a construct for the members of a particular group. It is systematic because it constantly distorts performance for members of the group. When a test item unfairly favors one group of examinees compared to another, the item is biased. On the other hand, group disparity in item performance that reflects actual knowledge and experience difference on the construct of interest is called item impact. Impact is also constant for members of a particular group, but these effects reflect performance differences that the test is intended to measure (Gierl, Khaliq, & Boughton, 1999).

The link between DIF, bias and impact is largely methodological. Statistical methods are used to identify items with DIF. Judgmental analyses are used to determine if DIF is attributable to bias or to impact for members of a specific group. Therefore, both procedures are needed for reviews intended to evaluate the fairness of an item (Gierl, Khaliq, & Boughton, 1999).

DIF is a common way of evaluating item bias. DIF refers to a difference in item performance between subgroups of the population that have been conditioned or matched on the level of the targeted trait or ability (Maccines, 2009). Conditioning on the level of targeted trait or ability is a critical step of DIF analysis because it ensures that examinees are matched on a common measure before they are compared (Gierl, Khaliq, & Boughton, 1999). This is what distinguishes the detection of DIF and item impact which is the existence of true group differences on item performance (Maccines, 2009).

In their essence, all DIF techniques match test takers from different groups according to their total test scores and then investigate how the different groups performed on individual test items to determine whether the test items are creating problems for a particular group (Zumbo, 1999). In the current study, the total test score was the employed criterion for ability where examinees were matched to ensure that performance differences that were found were not confounded by differences in trait or ability levels.

2.2.1 Subgroups in DIF analysis

There are two subgroups of the population that are compared in a DIF analysis; the reference group (R) and the focal group (F). The reference group is the comparison group and is chosen to provide standard performance on the item of interest e.g. male examinees whereas the focal group is the group whose differential performance, if any, is to be detected and measured e.g. female examinees (Hadley, 1994). These are the groups which are matched on the level of the intended trait or ability. Therefore, any differences between the focal and the reference groups are not confounded by differences in trait or ability levels (Maccines, 2009). In the current study, the reference group comprised of male examinees and the focal group comprised of female examinees.

2.2.2 Types of DIF

Swaminathan and Rogers (1990 in Awuor, 2008) and Rogers and Swaminathan (1993 in Awuor, 2008) had classified two types of DIF; uniform and non-uniform DIF. According to their classification, uniform DIF exists when there is no interaction between ability level and group membership and non-uniform DIF exists when there is interaction between ability level and group membership (Awuor, 2008). Uniform DIF, therefore, refers to differences in performance between the focal and reference groups that are the same in direction across all levels of the ability and indicates that one group has an advantage on the item across the continuum of ability. Non-uniform DIF refers to differences in performance between the focal and reference groups that are not the same in direction across all levels of ability indicating that the disadvantaged group changes depending on the ability level (Maccines, 2009 & Padilla, Hidalgo, Benitez, & Gomez-Benito, 2012). The current study determined whether MH DIF detection procedure was able to detect both uniform and non-uniform DIF in MSCE Computer Studies examination items in an attempt to accomplish objective number 3 of the study.

2.2.3 DIF detection

Tan and Gierl (2005 in Awuor, 2008) noted that in the process of establishing the validity of specific score inferences, one important step is to assess the fairness of the test through the analysis of DIF. DIF analysis is a procedure used to determine if test items are fair and appropriate for assessing the knowledge in a specific subject area across similar groups of examinees (Awuor, 2008). DIF statistical techniques are, therefore, based on the principle that if different groups of test-takers, e.g. males and females, have roughly the same level of something, e.g. knowledge, then they should perform similarly on individual test items regardless of group membership (Zumbo, 1999). The presence of DIF, therefore, signals that factors related to group membership affect the probability of correct response and threaten fair assessment. As such, procedures have been developed for DIF detection in an attempt to improve equity among examinees with same ability on the latent trait that the item is intended to measure (Awuor, 2008).

As stated by Zumbo (1999) there are two approaches to examining potential measurement bias. These approaches are Judgmental and statistical. Judgmental approaches rely solely on one or more expert judges' opinions to select potentially biased items. However, the problem with judgmental methods is that they are subjective because expert judges are human. Nevertheless, regardless of their subjectivity, judgmental methods are still widely employed in screening items for item bias with the complement of statistical or empirical methods which are more objective. However, Zumbo (1999) recommended that in a high-stakes context faced by human resource organizations, one should rely on statistical techniques for investigating potential bias because of this method's defensibility.

There are numerous statistical or empirical methods of detecting DIF. Some are parametric such as those based on IRT and others are non-parametric such as those based on CTT. Parametric methods assume a specific item response model while non-parametric methods do not assume a specific item response model (Awuor, 2008). In general, parametric procedures require larger data sets and have the risk of model misspecification (Maccines, 2009).

Finch (2008 in Awuor, 2008) observed that, for detecting DIF in dichotomous items with the non-parametric approach, the Mantel Haenszel and Logistic Regression procedures were the two most popular ones (Awuor, 2008).

In the current research, Mantel-Haenszel (MH) DIF detection procedure, which is a non-parametric statistical technique, was used.

2.2.4 MH DIF detection procedure

MH DIF detection procedure is CTT- based and belongs to contingency table procedures (Wiberg, 2007). It was developed by Mantel and Haenszel in 1959 for use in epidemiological research as a measure of association and was later applied to DIF by Holland and Thayer in 1988 (Penfield, 2001).The Mantel-Haenszel procedure provides both a statistical test of significance and an effect size estimate for DIF (magnitude of DIF) using an estimated constant odds ratio (Maccines, 2009 & Bao, 2007). Applying the MH procedure to DIF detection begins by establishing a criterion for ability. The total test score is the employed criterion whereby test scores are organized into score intervals or ability groups for both the reference and the focal groups (Hadley, 1994). MH detects DIF by comparing the odds ratios of item endorsement frequencies across reference and focal groups after matching on the trait measured by the test. The reference and focal groups are split into *k* subgroups, representing different levels of the total test score, and at each score level, a 2*2 contingency table is constructed showing group membership as a function of item response frequency. In the current study, the reference and focal groups were split into 21 subgroups representing different levels of the total score. Male and female examinees who got the same score formed one subgroup.

CTT is the theory on which item construction for MANEB examinations (MSCE Computer Studies examination inclusive) is based so far (Mankhomwa, 2013). Bolt (2002 in Teresi, 2004) reported that the author of a simulation study comparing different models and methods concluded by saying that practitioners should begin to move away from selection of a DIF method based on personal preference, but should select methods that are appropriate to the data that is being analysed (Teresi, 2004). Awuor (2008) also indicated that, for organizations that construct assessment instruments, false flagging of items for DIF is not a simple matter because test preparation is an expensive and time-consuming exercise. Therefore, discarding perfectly good items from a test because of inaccurate functioning of a DIF detection procedure has serious educational and economic implications (Awuor, 2008). Again, removal of perfectly good items from a test lowers both test score reliability and validity (Hambleton, 2006). As such, the fact that both MH DIF detection procedure and construction of MSCE Computer Studies examination items are based on the same theory (CTT) means that the MH DIF detection procedure that was selected in the current study was appropriate to the data that was analysed. As such, effective DIF detection was expected.

Furthermore, MH DIF detection procedure has been reported widely in literature that it has reasonable power for effective DIF detection in small samples as compared to IRT-based DIF detection methods which require extremely large samples for effective DIF detection (Clauser & Mazor, 1998; Bryant & Forde, 2002; Hambleton, Swaminathan & Rogers, 1991; Lopez, 2012; Maccines, 2009 & Teresi, 2004). So far, for MSCE Computer Studies examination results, the number of examinees at each score level is usually less than 100 which is a very small population that provides a very small sample for DIF analysis. As such, MH DIF detection procedure was expected to be more effective in detecting DIF in MSCE Computer Studies examination items compared to IRT-LR DIF detection procedure that was used in Jailosi's (2012) research.

Although MH procedure is popular, straightforward, adaptable, has an effect size measure and test of significance and works well with small samples, it has been extensively reported that it has one very strong limitation of being insensitive to non-uniform DIF (Hambleton, Swaminathan & Rogers, 1991; Hadley, 1994; Lopez, 2012; Swaminathan & Rogers, 1990; Wiberg, 2007 & Maccines, 2009). However, it has been reported that non-uniform DIF is very rare in items that undergo extensive qualitative analysis as is the case with items constructed at Educational Testing Services (ETS) where MH procedure is considered as a screen that can identify first level DIF (Teresi, 2004). Likewise, before being compiled into an examination paper, items of MSCE Computer Studies and all other MANEB examinations go through an extensive process of item construction to ensure that only those items that are deemed suitable are the ones that are used to compile the examination paper. As such, non-uniform DIF was expected to be very rare in items of MSCE Computer Studies examination. Consequently, MH DIF detection procedure was expected to detect first level DIF in the current research.

2.2.5 Contingency table in general

Table 1 shows an example of a contingency table in general as used in MH DIF detection procedure (Chulu, 2012).

Table 1: CONTINGENCY TABLE IN GENERAL

		SCORE ON A GIVEN ITEM		**TOTAL**
		1	0	
GROUP	R	A_j	B_j	nR_j
	F	C_j	D_j	nF_j
	TOTAL	$M1_j$	$M0_j$	T_j

In the table;

R = Reference Group

F = Focal Group

A_j = the number of subjects or examinees in the reference group, at score level "j", which answered the concerned particular item correctly.

B_j = the number of subjects or examinees in the reference group, at score level "j", which answered the concerned particular item incorrectly.

C_j = the number of subjects or examinees in the focal group, at score level "j", which answered the concerned particular item correctly.

D_j = the number of subjects or examinees in the focal group, at score level "j", which answered the concerned particular item incorrectly.

nR_j = the total number of subjects or examinees in the reference group at score level "j".

nF_j = the total number of subjects or examinees in the focal group at score level "j".

T_j = the total number of examinees (both reference and focal group examinees) at score level "j".

2.2.6 Odds and odds ratio

Macinnes (2009) and Mellenbergh (1982) indicated that the first step in DIF analysis is to calculate the common odds ratio, (α_{MH}), which is the ratio of the odds that an individual from the reference group will answer an item correctly to the odds that an individual from the focal group, of the same ability level, will answer the same item correctly. The estimate of α_{MH} has a range of zero to positive infinity. If the estimate for α_{MH} is equal to 1, then there is no difference in performance between the reference and focal groups. Values of α_{MH} between 0 and 1 indicate that the item favors the focal group while values greater than 1 indicate that the item favors the reference group α =0.05 (Abedalaziz, 2010 & Chulu, 2012).

2.2.7 Converting data from the general contingency table (Table 1) to proportions to calculate the odds ratio

A_J, B_J and $n_{Rj}$are divided by n_{Rj}; then label $P_{Rj}= A_j / n_{Rj}$ and $Q_{Rj} = B_j / n_{Rj}$ so that $P_{Rj}+Q_{Rj} = 1$. Likewise, $P_{Fj} = C_j / n_{Fj}$ and $Q_{Fj} = D_j / n_{Fj}$ so that $P_{Fj} + Q_{Fj} = 1$(Chulu, 2012).

Therefore, the data for the proportions will be as indicated in Table 2.

Table 2: **CONVERTED DATA FROM THE GENERAL CONTINGENCY TABLE TO PROPORTIONS**

		SCORE ON A GIVEN ITEM		
		1	**0**	**TOTAL**
GROUP	R	P_{Rj}	Q_{Rj}	1
	F	P_{Fj}	Q_{Fj}	1

Since the two groups have been conditioned on the same total score, it is expected that the proportions in the two groups would be the same or equal, i.e.

$\Delta_{Pj} = P_{Rj} - P_{Fj} = A_j / n_{Rj} - C_j / n_{Fj}$ ………………. 2(i)

If the proportions are not the same, then the item is functioning differentially in the two groups.

From Table 2, the odds that a reference group examinee will answer an item correctly is defined as:

$\Omega_{Rj}= P_{Rj} /Q_{Rj}$ …………………………… 2 (ii)

Whereas the odds of a correct response in the focal group are defined as:

$\Omega_{Fj} = P_{Fj} / Q_{Fj}$ …………………………….. 2(iii)

The ratio of the above two proportions is what is regarded as the odds ratio and is indicated as follows:

$\alpha_j = \dfrac{P_{Rj} /Q_{Rj}}{P_{Fj} / Q_{Fj}} = \dfrac{\Omega_{Rj}}{\Omega_{Fj}}$ ……………………….. 2 (iv)

Formula 2(iv) was the one which was used by the current researcher in calculating the odds ratio. The odds ratio was the one which was used to determine the magnitude of the DIF that MH DIF detection procedure was able to detect in MSCE Computer Studies examination items.

2.2.8 Logarithm of odds ratio

The odds ratio can be anywhere in the interval zero to infinity; and it cannot be symmetrical. As such, a simple natural logarithm transformation (Log or ln) on the odds ratio is recommended compared to the common odds ratio because the resulting statistic is symmetrical about zero and zero has the interpretation of equal odds indicating that the concerned item is not functioning differentially between the two groups. Positive values indicate that the item is easier for the reference group and negative values indicate that the item is easier for the focal group (Chulu, 2012 & Lopez, 2012).

In the current research, the logarithm of the odds ratio for the concerned data was calculated. The values that were obtained assisted in determining which item was easier for which group and vice-versa.

2.2.9 Magnitude of DIF

Magnitude of DIF refers to the degree of DIF, and can be measured by examining parameters or statistics associated with the method, for example, the odds ratio, beta-coefficient or increment with R-Square associated with the DIF term for the studied item (Teresi, Ramirez, Lai, & Silver, 2008). Kirk (1996 in Zumbo, 1999) reported that as per usual, in statistical hypothesis testing, the test statistic should be accompanied by some measure of magnitude of the effect because, for example, small sample sizes can hide interesting statistical effects whereas large samples (like the ones found in typical psychometric studies) can point to statistically significant findings where the effect is quite small and meaningless (Zumbo, 1999 & Li & Zumbo, 2009). Scott, et al. (2010) also said that many authors have proposed DIF classifications that incorporate both statistical significance and the magnitude of DIF because statistical significance does not necessarily imply practical significance. Without an examination of effect size, trivial effects could be statistically significant when the DIF test is based on a large sample size due to too much statistical power (Zumbo, 1999).

In the current research, the magnitude of DIF was calculated using the Odds Ratio statistic to determine both the statistical and practical significance of the DIF that was detected.

2.2.10 The Educational Testing Service (ETS) classification of DIF magnitude

As reported by Abedalaziz (2010), the common odds ratio, $\alpha_{MH,}$ which is the same as α_j in Formula 2(vi), is often transformed onto the ETS "delta metric" (D Statistic) which represents the differences in item difficulty for the reference and focal groups after the total score has been taken into account. This ETS "delta metric" (D Statistic) is represented by the following formula (Abedalaziz, 2010):

$\Delta_{MH} = -2.35\ln(\alpha_{MH})$ ………………………………………………………….. 2 (v)

A "D" value of 0.0 indicates no DIF

A positive "D" value indicates DIF favoring the focal group

A negative "D" value indicates DIF favoring the reference group

The present researcher also transformed the calculated common odds ratio to the scale of differences in item difficulty used by the ETS using the same above formula as shown in Tables 4.1 and 4.3 in the results' section.

The Educational Testing Service had proposed values of Δ_{MH} for classifying the magnitude of DIF as negligible, moderate or large using the following guidelines as indicated by Abedalaziz (2010); Maccines (2009) & Zwick & Thayer (1999).

Type A items – negligible DIF and $|\Delta_{MH}| < 1$

Type B items – Intermediate or moderate DIF and $1.0 \leq |\Delta_{MH}| < 1.5$

Type C items – large DIF and $|\Delta_{MH}| \geq 1.5$

The above type of classification was the one which was used in the current research for classification of DIF magnitude.

2.2.11 MSCE Computer Studies examination

MSCE Computer Studies examination was only introduced recently in 2005 by MANEB. This was in response to the introduction of a secondary school Computer Studies curriculum in 2004. MSCE or secondary school Computer Studies was introduced by the Ministry of Education, Science and Technology to help learners identify different educational and job opportunities available in computing and other related fields (Education, 2001). Information from this examination is also utilized by university colleges and other institutions of higher learning for selection and placement of students into various programs. Therefore, it is a high-stakes examination whose fairness amongst learners should be guaranteed through, among other things, proper identification of gender DIF.

MSCE Computer Studies examination is comprised of two papers; Paper I (Theory) and Paper II (Practical). Before being compiled into an examination paper, items of MSCE Computer Studies and other MANEB examinations go through an extensive process of item construction to ensure that those items that are deemed suitable are the ones that are used to compile the examination paper. The process of test construction begins with the training of item writers who then construct items in accordance with Bloom's Taxonomy and Table of Test Specification. Thereafter, the items undergo further review by subject specialists called editors or moderators who discard some of the items that may be deemed unsuitable. The accepted items are then compiled into examination papers which undergo pretesting which is followed by item analysis (item difficulty and item discrimination indices) basing on Classical Test Theory. After item analysis, the final examination papers are compiled, proof-read and printed by security printers (Mankhomwa, 2013). Furthermore, the processes of administration and scoring of all MANEB examinations are standardized and highly secured to ensure that fair treatment is accorded to all candidates. This is in line with what is reported by Buzick and Stone (2011) that focal and reference groups that are being compared should be tested under the same conditions. However, regardless of this great effort that goes into test construction, printing, administration and scoring of MANEB examinations, MSCE Computer Studies examination pass rates have persistently revealed that boys perform better than girls since its inception in 2005 (Jailosi, 2012). This scenario prompted Jailosi (2012) to investigate if some items in this examination were differentially functioning against girls. The results of Jailosi's DIF investigation, using IRT-LR DIF detection procedure, indicated that, really, some items in MSCE Computer Studies examination were functioning differentially against girls (Jailosi, 2012). 14 out of 20 (70%) items were identified as having significant DIF. Furthermore, the DIF in 70% of the significantly identified DIF items was against girls (Jailosi, 2012).

However, since MANEB, currently, formulates and analyses its examinations using CTT, to some extent, the results of the earlier research which used IRT-LR DIF detection procedure may have questionable validity and reliability considering the fact that IRT-LR DIF detection procedure, though, theoretically, considered to be one of the best DIF detection procedures, is based on IRT which, apart from being a different theory altogether, is based on very stringent assumptions which are very difficult to meet if the baseline model that is chosen does not fit the data. Lack of model fit can be mistaken for DIF (Teresi, 2004). This scenario can, to some extent, give false indicators about the presence of DIF. The present researcher replicated Jailosi's research on the basis of methodology. CTT-based MH DIF detection procedure was used to re-investigate DIF in MSCE Computer Studies examination items. Jailosi used IRT-LR DIF detection method and analyzed DIF in 2007 and 2008 MSCE Computer Studies examination Paper 1 items. The current researcher used MH DIF detection procedure to analyze DIF in 2012 MSCE Computer Studies examination Paper 1 items (Appendix A).

As stated earlier on, DIF had been identified in items of MANEB's PSLCE Mathematics (Gondwe, 2012) and MSCE Computer Studies (Jailosi, 2012) examinations. The present researcher preferred to replicate Jailosi's research because MSCE is a higher grade than PSLCE and according to Gierl et al, 1999, test developers should anticipate more DIF in Mathematics and Science in the higher grades. Gierl et al. (1999) had further indicated that test developers should also expect more DIF, overall, in Science compared to Mathematics.

2.2.12 Comparison of 2007, 2008 and 2012 MSCE Computer Studies Paper 1 examination items

As mentioned earlier on, Jailosi (2012) used 2007 and 2008 MSCE Computer Studies Paper 1 examination results. The current researcher used 2012 MSCE Computer Studies Paper 1 examination results. All these results were sourced from MANEB and their respective examinations or tests are regarded as parallel tests. As such, the three results are comparable because they were assessing the same content areas which are spelled out in the Malawi Secondary School Computer Studies Syllabus (Education, 2001). All the test forms were administered and marked under similar conditions.

2.2.13 Software packages used in DIF analysis

A number of statistical software packages that permit calculation of MH statistics to a greater or lesser degree, exist. The most complete package is the SAS software (2002, SAS Institute, Version 9.0). Unfortunately, the program is not very easy to use for DIF detection (Fidalgo, 2011). Amongst programs that are easier to use than SAS or that are specifically designed for DIF detection are SPSS (2007, SPSS for windows, Rel. 16.0.0), MHDIF and DIFAS. Unfortunately these other easier-to-use programs calculate only a small number of MH statistics (Fidalgo, 2011). For instance, SPSS only computes MH statistic with continuity correction and the MH test of linear association. MHDIF calculates only the MH statistic with the continuity correction and that can be used for calculating non-uniform DIF. DIFAS calculates the MH statistics with continuity correction and the Mantel test (Fidalgo, 2011).

Apart from the above-mentioned software packages, there is a new approach for DIF detection. This approach uses Generalized Mantel-Haenszel DIF (GMHDIF) program (Fidalgo, 2011). The program constitutes the most complete and single option for assessing DIF even in multiple groups using MH statistics. The program is advantageous over SAS because it has been developed to provide an easy-to-use program for conducting DIF analyses and runs in any windows operating system. It is also free of charge (Fidalgo, 2011).

Furthermore, results of a research study by Randall generally proved that Generalized Mantel-Haenszel (GMH) statistic was the most appropriate of the three MH procedures (the MH chi-square statistic with no adjustment to the alpha level, the MH chi-square statistic with a Bonferroni adjusted alpha level and the GMH Statistic) because its type 1 error rate remained at the nominal level of 0.05 and its power was consistently among the highest (1st Penfield, 2001).

In view of the above, data for the current research was analyzed using GMHDIF program.

2.2.14 Statistics implemented by GMHDIF program

The GMHDIF program implements the following Generalized Mantel-Haenszel (GMH) statistics:

1. QGMH (1) or the Generalized Nominal MH statistic.
2. QGMH (2) or the Generalized Ordinal MH statistic.

The following equivalences between the above-cited generalized MH statistics and the MH statistics usually employed in the DIF literature exist (Fidalgo, 2011):

1. For the special case of two factor levels (2 groups), QGMH (1) is identical to the Generalized Mantel-Haenszel (GMH) test proposed by Mantel and Haenszel (1959).
2. For the special case of two factor levels (2 groups), QGMH (2) is identical to the extended MH test (Mantel test) proposed by Mantel (1963).
3. For a dichotomously scored item and 2 groups, QGMH (1) = QGMH (2) = χ^2_{MH} (Mantel-Haenszel Chi-square).

According to Fidalgo (2011), in the case of dichotomously scored items, it would be the same to use QGMH (1) or QGMH (2) since they are equivalent. QGMH (1) is further recommended if a single method is used given that QGMH (1) is capable of detecting more complex patterns of association than QGMH (2) (Fidalgo, 2011).

In this research, QGMH (1) statistic was used to analyze data since a single (Mantel-Haenszel) method was used to detect DIF.

2.2.15 Theoretical framework

MH DIF detection method is based on Classical Test Theory (CTT). Classical Test Theory, also known as true score theory, is a traditional quantitative approach to testing the reliability and validity of a scale based on its items (Cappelleri, Lundy, & Hays, 2014). In CTT, the true score is defined on the examinee’s observed test score (Hambleton & Jones, 1993). CTT assumes that each observed score (X) on an instrument is a combination of an underlying True Score (T) on the concept of interest and unsystematic (random) error (E), i.e. X = T + E. It is assumed that the true score, T, would be obtained if there were no errors in the measurement. A person’s true score is, therefore, defined as the expected score over an infinite number of independent administrations of the scale. As such scale users never observe a person’s true score, T, but they observe score X (Cappelleri, Lundy, & Hays, 2014). It is further assumed that the mean of the distribution of random errors over a hypothetical infinite number of independent administrations of the scale on the same subject is taken to be zero (0). In addition, random errors are assumed to be uncorrelated with the true score. There is no systematic relationship between a person's true score and whether that person's observed score has positive or negative errors (Cappelleri, Lundy, & Hays, 2014). Because of these assumptions, in CTT a person's observed score is assumed to be that person's true score.

This theory (CTT) was, therefore, considered to be the theoretical framework on which to base the current research study because, currently, construction and analysis of items in MANEB examinations is based on CTT (Mankhomwa, 2013); as such, the expectation that assumptions which underline CTT could be met, was very high. Again, according to literature, in CTT, analyses can be performed with smaller representative samples of examinees (Bichi & al, 2015). For MSCE Computer Studies examination results, the number of examinees at each score level is usually less than 100 which is a small population that provides a very small sample for DIF analysis. As such, a DIF detection procedure based on CTT was expected to perform much better with such small samples.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Chapter overview

This chapter describes the methodology that was used in carrying out this research. It includes research design, data and data source, data collection instrument, research population and participants, sample and sampling, data analysis, ethical issues considered in the current research and limitations of the study.

3.2 Research design

Mbambo (2009) states that Polit and Hungler describes research design as a blueprint or outline for conducting the study in such a way that maximum control will be exercised over factors that could interfere with the validity of the research results. It is the researcher's overall plan for obtaining answers to the research questions or objectives guiding the study. This study used secondary quantitative descriptive design to identify and analyze DIF in MSCE Computer Studies examination items. The design was secondary quantitative descriptive design because the study utilised secondary quantitative data in form of total scores from marked scripts of 2012 MSCE Computer Studies national examination. The observed and analyzed DIF then provided information which described the quality of each individual item as having DIF or not.

3.3 Data and data Source

Mbambo (2009) defines data as information obtained during the course of an investigation or study. This study utilized secondary quantitative data from marked scripts of 2012 MSCE Computer studies Paper 1 national examination which was developed, administered and scored by the Malawi National Examinations Board (MANEB) to assess students' proficiency in Computer at the end of the secondary cycle. These data were archival records for all individual examinees who sat for the 2012 MSCE Computer Studies Paper 1 national examination. Literature has it that

examination of archival records is one of the data collection techniques for making empirical observations and measurements (Shaughnessy & Zechmeister, 1990).
2012 MSCE Computer Studies Paper 1 examination was administered to 3001 examinees. 1741 of these examinees were male and 1260 were female. Results for this examination were regarded as being free of bias, valid and reliable because they were results of a national examination whose production, administration and all other processes were standardized.

2012 MSCE Computer Studies Paper 1 is a free response examination paper. In this research, the items were treated as binary items and were scored as "1" or "0" depending on whether the examinee got the question correct or not. This is in accordance to Nagelkerke who treat the binary variable as nominal to be in line with the conventional strategy in DIF analysis (Zumbo, 1999). This is a traditional scoring rubric for binary items in aptitude, achievement, knowledge or personality/social measures. Responses scored as "1" are indicators for more knowledge, aptitude or personality construct (depending on the setting) than a score of "0". For example, in a knowledge test, an item score of "1" is indicative that the respondent knows more than an individual who gets a score of "0" (Zumbo, 1999).

3.4 Data collection instrument

Data collection is the process by which the researcher collects the information needed to answer the research problem and data collection instrument is a device that is used to collect data (Nalzaro, 2012). These instruments include questionnaires, tests and interviews (Mbambo, 2009 & Nalzaro, 2012). In the current research study, the instrument that was used was the 2012 MSCE Computer Studies Paper 1 national examination (Appendix A). The 2012 paper was chosen because it was the only paper available at MANEB at the time of commencement of this study which had corresponding examinees' scores. These examinees' scores were the ones which were used to obtain data relevant to the study's objectives.

2012 MSCE Computer Studies Paper 1 national examination consisted of thirty-four (34) items. This length of the examination was in accordance with what is recommended in literature that the minimum number of items on the entire test should be 20 to allow for greater variance in the score distribution which results in development of meaningful ability levels (Zumbo, 1999). All the 34 items in the 2012 MSCE Computer Studies Paper 1 national examination were used in this study because, although no standard exists as to how many items should be used for DIF detection as this changes from study to study, literature recommends that all test items should be tested for DIF. This will provide information about all items in the test; shedding light on problematic items as well as those that function similarly for both the reference and focal groups.

3.5 Research population and participants

Mbambo (2009) defines a research population as the totality of all subjects that conform to a set of specifications, comprising the entire group of persons that is of interest to the researcher and to whom the research results can be generalized. The research population for this study comprised of a total of 3,001 examinees who sat for the 2012 MSCE Computer Studies national examination; 1,741 were male and 1,260 were female examinees. This included all examinees who sat for 2012 MSCE Computer Studies national examination.

Research participants are living individuals or groups of living individuals about whom a researcher conducting research obtains data through intervention or interaction with the person or identifiable private information (Committee, 2013). In this study, research participants were examinees of 2012 MSCE Computer Studies national examination.

3.6 Sample and sampling

A sample is described as a portion or subset of the research population selected to participate in a study, representing the research population (Mbambo, 2009). Participants used in this study were selected using random sampling aided by a Table N of random numbers (Vernoy & Vernoy, 1992) to ensure generalizability of results (external validity) from the sample to the population because, according to literature, in random sampling, each member of the population of interest has an equal chance of being selected as a respondent (Gravetter & Forzano, 2003).

In this research, a sample of 500 examinees was selected; 300 were male and 200 were female. Literature widely recommends a sample of at least 500 members with at

least 200 members from the focal group and at least 300 members from the reference group (Zwick, 2012; Hadley, 1994 & Wiberg, 2007). Twenty-one (21) total score (ability) levels were identified from the sample of 500 examinees that was selected for this study. Examinees were then grouped according to the total test scores that they obtained. For each group, the number of male and female examinees was determined.
All examinees at each score level were included in the study because the total number of male and female examinees at each of the 21 score levels was less than 100. Literature has it that if the number of subjects in a population is less than 100, the recommendation is to analyze the entire population (Simon & Goes, 2012). Thereafter, a 2*2 Contingency Table was established for each of the 34 items at each of the 21 score/ability levels. This data provided the first set of statistical results.

According to literature, it is advisable to have two or more sets of statistical results so that items which consistently show up as DIF items can be identified. Hambleton (2006) indicated that two or more sets of statistical results can be achieved by using different procedures for the same data or using split samples of the same data and replicate the process. Items that consistently show up as DIF items across different procedures or different samples can be chosen to be the focus of explaining the DIF that has been detected because replication produces stable results (Hambleton, 2006 & Lei, Moses, Puhan and Dorans, 2008). As such, the second set of statistical results was provided by five (5) groups of examinees randomly drawn from a particular score level designated “Q”aided by a Table N of random numbers (Vernoy & Vernoy, 1992). The sample sizes were 12, 18, 24, 30 and 48 with respective sample size combinations of 6F:6M, 6F:12M, 12F:12M, 12F:18M and 12F:36M; where "F" stands for female and "M" stands for male.

Score level “Q”, from which the five groups of examinees were drawn, was purposively sampled from amongst the 21 identified score/ability levels. Purposive sampling is recommended in literature because, in purposive sampling, the investigator handpicks the elements to be included in the sample on the basis of expert judgment. The individuals selected may be either those judged to have certain special characteristics or those who are likely to provide the most useful information for the purposes for which the study is being done (Shaughnessy & Zechmeister, 1990). Literature also recommends that the largest possible number of examinees should be used to obtain stable DIF estimates and to ensure sufficient power for DIF detection (Rezaee & Shabani, 2010). Score Level “Q” provided sample size combinations which were able to accommodate the actual maximum number of examinees in the focal and reference groups (12F:36M) at that particular score level. Score level “Q” was, therefore, selected because it was likely to provide the most useful information for the purpose of accomplishing research objectives of this study. Another 2*2 contingency table was established for each of the 34 items, at each of the five (5) sample size/sample size combination.

The first set of data for the twenty-one (21) score levels provided twenty-one (21) replications of the analysis for each of the 34 items and the second set of data for the sample sizes provided five (5) replications of the analysis for each of the 34 items. This, therefore, gave a total of 26 replications of the analysis for each of the 34 items.

3.7 Data analysis

Data analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data (Mbambo, 2009). In DIF analysis, there are a number of statistical software packages that permit calculation of MH statistics to a greater or lesser degree. These software packages include SAS, SPSS, MHDIF, DIFAS and GMHDIF (Fidalgo, 2011). Analysis of data in the current research was done using Generalised Mantel-Haenszel DIF (GMHDIF) program described in Chapter 2. GMHDIF program is advantageous over the other programs because it has been developed to provide an easy-to-use program for conducting DIF analyses and runs in any windows operating system. It is also free of charge (Fidalgo, 2011).This program utilized the notepad form of data that formed a 2*2 contingency table for each item for both data sets (score level data and sample size/sample size combination data). Of the two Generalised Mantel-Haenszel (GMH) statistics that the GMHDIF program implements, i.e. Generalised Nominal MH statistic{QGMH (1)} and Generalised Ordinal MH statistic {QGMH (2)}, QGMH (1) was used to analyse the data in this research because a single method (MH) was used. Literature recommends the use of QGMH (1) if a single method is used because of its greater capability of detecting more complex patterns of association than QGMH (2) (Fidalgo, 2011).With GMHDIF program, data begins on the first line of the file which is a delimited text (ASCII) file. Variable 1 is the grouping variable. The other variables start from variable 2 (Fidalgo, 2011). As such, the 34 items in the current study ranged from variable 2 to variable 35. The matching variable (total score) was divided into intervals of amplitude equal to 1.

A 1-stage analysis (i.e. no purification) was used in the current study. It was assumed that 2-stage analysis (purification) was not necessary because the largest sample sizes (all examinees at each score level) were used in the analysis. Literature has it that 2-stage analysis is costly and time consuming and the procedure becomes unnecessary when large samples are used (Gierl, Jodoin, & Ackerman, 2000).

3.8 Ethical issues considered in the current study

Consideration of ethical issues is integral to the research process. Researchers have two basic categories of ethical responsibility, i.e. responsibility to the individuals both human and non-human; and responsibility to the discipline of science; to be accurate and honest in the reporting of their research (Gravetter & Forzano, 2003).

Considering that the current study dealt with examinees' national examination results, ethical issues concerning examinees' scores were observed because this is very confidential information. A letter introducing the researcher to various institutions and organizations was written by the University of Malawi, Chancellor College to assist the researcher with data collection (Appendix B). Permission was also granted by the Malawi National Examinations Board (MANEB) for the researcher to use 2012 MSCE Computer Studies examination results (Appendices C & D). The data collected was treated with ultimate confidentiality.

Ethical issues concerning the discipline of science were observed through the use of real data from a national examining body (MANEB). The data was analyzed by a reputable program (GMHDIF program) which was specifically e-mailed to the researcher by the owner, Angel M. Fidalgo, University of Oviedo, Spain; who is an Associate Professor and contributor (in the area of Psychometrics) to a book on Research Studies, Literature Reviews and Perspectives in Psychological Science (Patelis, 2012). Furthermore, reference citations have been made throughout the research report as evidenced by the reference section. Acknowledgements have also been made to all those who assisted in one way or another which led to the success of this research study.

3.9 Limitations of the study

Camilli and Shepard (1994 in Lopez, 2012) repeatedly emphasized the importance of using several methods for testing and measuring DIF so that they can complement each other because different methods have different characteristics (Lopez, 2012).

As mentioned earlier on, Jailosi (2012) used 2007 and 2008 MSCE Computer Studies examination items. The current researcher used 2012 MSCE Computer Studies examination items. Though items in the three examinations, i.e. 2007, 2008 and 2012 are regarded as similar in nature because they are parallel tests, there is still a reasonable probability that the results of DIF analyses on these three examinations may not be identical because parallel tests, though similar, are not identical. Again, individual experimental errors, performed by different researchers at different times could affect the final results of the analysis thereby leading to a compromise in terms of the validity and reliability of the conclusions drawn. As such, the best way of comparing the results would have been done through simultaneous comparison of the two concerned DIF methods on the same data using e.g. difR package which contains several traditional methods (IRT-based or not) to detect DIF in dichotomously scored items (Magis, Beland, & Raiche, 2013). In such a way, the two methods would have been set with a similar structure and would have been compared in one run and detection of items that are never, sometimes or always flagged as DIF by the two methods would have been simultaneously and easily detected. In so doing, probable individual experimental errors would be minimized. Otherwise there is chance that these probable errors could have negatively affected the validity and reliability of results of the two concerned methods.

CHAPTER 4

RESULTS AND DISCUSSION OF RESULTS OF THE STUDY

4.1 Chapter overview

This Chapter presents results of the study and discusses the findings with reference to research objectives presented in Chapter 1 and in comparison to research findings by Jailosi (2012) who conducted a similar study, but using IRT-based IRT-LR DIF detection procedure.

4.2 Results of the study

This research study was conducted with the aim of re-examining DIF in MSCE Computer Studies examination items using CTT-based MH DIF detection procedure.

Appendix E shows results of DIF analysis (GMH Statistics) for each of the 34 items at each of the 21 score levels. Appendix F shows results of DIF analysis (GMH Statistics) for each of the 34 items at each of the five sample sizes and sample size combinations. In some cases, GMHDIF statistic could be computed for an item at a particular score level and unable to be computed at another score level (indicated N/P) regardless of sample size. For example, at score level No. 1 of sample size 20, GMH Statistic was not able to be computed for variables 2 and 8, but was able to be computed at score level No.2 with a larger sample size of 34. However, for variable No. 4, a GMH Statistic was able to be computed at score level No. 1 of sample size 20 and it was not able to be computed at score level No. 3 with a larger sample size of 30. This supports what is recorded in literature that DIF statistics are, sometimes, surprisingly, unstable even with large samples (Hambleton, 2006).

An asterisk (*) shows statistically significant DIF for the concerned item at $\alpha = 0.05$.

Table 3: RESULTS OF DIF MAGNITUDE FOR STATISTICALLY SIGNIFICANT DIF IDENTIFIED IN ITEMS THAT DISPLAYED SIGNIFICANT DIF AT EACH OF THE 21 SCORE LEVELS

ITEM No.	SCORE LEVEL AT WHICH THE ITEM WAS IDENTIFIED AS A DIF ITEM	ODDS RATIO (α_{MH})	LOGARITHM OF ODDS RATIO {ln (α_{MH})}	DIF MAGNITUDE (Δ_{MH})= -2.35ln (α_{MH})	DIF AGAINST	CATEGORY
1	15	0.22	-0.66	1.55	Male	C
	20	2.45	0.39	-0.92	Female	A
2	11	0.7	-0.15	0.35	Male	A
	6	2	0.3	-0.71	Female	A
	11	1.64	0.21	-0.49	Female	A
3	16	1.82	0.26	-0.61	Female	A
	21	2.03	0.31	-0.73	Female	A
	1	1.37	0.14	-0.33	Female	A
	2	1.1	0.04	-0.09	Female	A
	3	1.67	0.22	-0.52	Female	A
	4	1.69	0.23	-0.54	Female	A
	6	0.39	-0.41	0.96	Male	A
	7	2.7	0.43	-1.01	Female	B
5	8	1.04	0.02	-0.05	Female	A
	9	0.59	-0.23	0.54	Male	A
	11	1.23	0.09	-0.21	Female	A
	12	1.83	0.26	-0.61	Female	A

	13	1.88	0.27	-0.63	Female	A
	14	1.68	0.23	-0.54	Female	A
	15	3.02	0.48	-1.13	Female	B
	16	29.58	1.47	-3.45	Female	C
	17	8.69	0.94	-2.21	Female	C
	18	8	0.9	-2.13	Female	C
	19	0.78	-0.11	0.26	Male	A
	20	0.42	-0.38	0.89	Male	A
	21	3	0.48	-1.13	Female	B
6	4	0.25	-0.6	1.41	Male	B
	11	2.55	0.41	-0.96	Female	A
7	6	0.17	-0.77	1.81	Male	C
	16	0.03	-1.52	3.57	Male	C
	1	2.28	0.36	-0.85	Female	A
8	6	1.69	0.23	-0.54	Female	A
	11	1.83	0.26	-0.61	Female	A
	2	0.66	-0.18	0.42	Male	A
	3	1.07	0.22	-0.07	Female	A
	7	0.78	-0.11	0.26	Male	A
9	8	0.93	-0.03	0.07	Male	A
	12	3.13	0.5	-1.18	Female	B
	14	1.48	0.17	-0.43	Female	A
	6	1	0	0	NO DIF	NODIF
	8	0.67	-0.17	0.4	Male	A
	9	1.09	0.04	-0.09	Female	A
	11	0.6	-0.22	0.52	Male	A
	12	0.73	-0.14	0.33	Male	A

	13	1.83	0.26	-0.61	Female	A
10	14	0.91	-0.04	0.09	Male	A
	15	1.56	0.19	-0.45	Female	A
	16	0.2	-0.7	1.65	Male	C
	18	0.61	-0.21	0.49	Male	A
	19	1.79	0.25	-0.59	Female	A
	14	0.12	-0.92	2.16	Male	C
11	19	1.12	0.05	-0.12	Female	A
	20	2.45	0.39	-0.92	Female	A
12	11	0.48	-0.32	0.75	Male	A
	6	1.19	0.08	-0.19	Female	A
13	11	0.51	-0.29	0.68	Male	A
	16	0.87	-0.06	0.14	Male	A
	1	0.89	-0.05	0.12	Male	A
	3	1.24	0.09	-0.21	Female	A
	4	0.18	-0.74	1.74	Male	C
	6	0.79	-0.1	0.24	Male	A
	7	0.7	-0.15	0.35	Male	A
15	8	1.06	0.03	-0.07	Female	A
	9	0.47	-0.33	0.78	Male	A
	11	0.52	-0.28	0.66	Male	A
	12	7.32	0.86	-2.02	Female	C
	14	2.67	0.43	-1.01	Female	B
	16	1.94	0.29	-0.68	Female	A
	19	1.29	0.11	-0.26	Female	A
16	11	1.4	0.15	-O.35	Female	A
	19	0.67	-0.17	0.4	Male	A

17	6	2	0.3	-0.71	Female	A
	6	0.33	-0.48	1.13	Male	B
18	11	0.6	-0.22	0.52	Male	A
	16	1.27	0.1	-0.24	Female	A
	2	2.5	0.4	-0.94	Female	A
	4	0.45	-0.35	0.82	Male	A
	7	0.37	-0.43	1.01	Male	B
	8	2.55	0.41	-0.96	Female	A
19	12	0.38	-0.42	0.99	Male	A
	13	0.67	-0.17	0.4	Male	A
	14	1.86	0.27	-0.63	Female	A
	17	0.31	-0.51	1.2	Male	B
	18	1.43	0.16	-0.38	Female	A
	4	1.33	0.12	-0.28	Female	A
	7	0.67	-0.17	0.4	Male	A
	8	1.11	0.05	-0.12	Female	A
20	9	1.15	0.06	-0.14	Female	A
	11	1.1	0.04	-0.09	Female	A
	12	1.43	0.16	-0.38	Female	A
	13	1	0	0	NO DIF	NO DIF
	14	5.13	0.71	-1.67	Female	C
	16	1	0	0	NO DIF	NO DIF
	17	4.73	0.67	-1.57	Female	C
	18	2.32	0.37	-0.87	Female	A
	19	0.42	-0.38	0.89	Male	A
	21	2.03	0.31	-0.73	Female	A
21	11	1.14	0.06	-0.14	Female	A

22	11	0.64	0.19	-0.45	Female	A
	1	1.37	0.14	-0.33	Female	A
23	6	0.21	-0.68	1.6	Male	C
	11	1	0	0	NO DIF	NO DIF
	16	7.72	0.89	-2.09	Female	C
	2	3.6	0.56	-1.32	Female	B
24	7	0.79	-0.1	0.24	Male	A
	12	3.13	0.5	-1.18	Female	B
	4	2.47	0.39	-0.92	Female	A
	6	0.67	-0.17	0.4	Male	A
	7	1.6	0.2	-0.47	Female	A
	8	0.79	-0.01	0.24	Male	A
	9	2.72	0.43	-1.01	Female	B
	11	1.03	0.01	-0.02	Female	A
25	12	1.92	0.28	-0.66	Female	A
	13	1.03	0.01	-0.02	Female	A
	14	0.36	-0.44	1.03	Male	B
	16	1.27	0.1	-0.24	Female	A
	17	0.49	-0.31	0.73	Male	A
	18	0.49	-0.31	0.73	Male	A
	19	0.33	-0.48	1.13	Male	B
	20	0.33	-0.48	1.13	Male	B
26	11	2.67	0.43	-1.01	Female	B
27	6	0.29	-0.54	1.27	Male	B
	16	1.31	0.12	-0.28	Female	A
28	11	0.65	-0.19	0.45	Male	A
	16	0.37	-0.43	1.01	Male	B

29	1	4.56	0.66	-1.55	Female	C
	2	2.92	0.47	-1.1	Female	B
	3	0.23	-0.64	1.5	Male	C
	7	0.78	-0.11	0.26	Male	A
	8	1.45	0.16	-0.38	Female	A
	13	3.14	0.5	-1.18	Female	B
	6	0.33	0.52	-1.22	Female	B
	7	3	0.48	-1.13	Female	B
	8	1.8	0.26	-0.61	Female	A
	9	3.18	0.5	-1.18	Female	B
30	12	1.4	0.15	-0.35	Female	A
	13	1.24	0.09	-0.21	Female	A
	16	0.47	-0.33	0.76	Male	A
	17	1.1	0.04	-0.09	Female	A
	18	2.39	0.38	-0.89	Female	A
	20	0.33	-0.48	1.13	Male	B
31	9	0.75	-0.12	0.28	Male	A
32	1	1.78	0.25	-0.59	Female	A
	11	1.38	0.14	-0.33	Female	A
	1	4.56	0.66	-1.55	Female	C
33	6	0.82	-0.09	0.21	Male	A
	11	0.96	-0.02	0.05	Male	A
	16	4.41	0.64	-1.5	Female	C
34	12	0.17	-0.77	1.81	Male	C

Table 4: RESULTS OF CATEGORY "B" AND "C" DIF MAGNITUDE FOR THE STATISTICALLY SIGNIFICANT DIF IDENTIFIED IN ITEMS THAT DISPLAYED SIGNIFICANT DIF AT EACH OF THE 21 SCORE LEVELS

ITEM No	SCORE LEVEL AT WHICH THE ITEM WAS IDENTIFIED AS A DIF ITEM	ODDS RATIO (α_{MH})	LOGARITHM OF ODDS RATIO {ln (α_{MH})}	DIF MAGNITUDE (Δ_{MH}) = -2.35ln (α_{MH})	DIF AGAINST	CATEGORY	TYPE OF DIF
1	15	0.22	-0.66	1.55	Male	C	Uniform
5	7	2.7	0.43	-1.01	Female	B	Uniform
	15	3.02	0.48	-1.13	Female	B	
	16	29.58	1.47	-3.45	Female	C	
	17	8.69	0.94	-2.21	Female	C	
	18	8	0.9	-2.13	Female	C	
	21	3	0.48	-1.13	Female	B	
6	4	0.25	-0.6	1.41	Male	B	Uniform
7	6	0.17	-0.77	1.81	Male	C	Uniform
	16	0.03	-1.52	3.57	Male	C	
9	12	3.13	0.5	-1.18	Female	B	Uniform
10	16	0.2	-0.7	1.65	Male	C	Uniform
11	14	0.12	-0.92	2.16	Male	C	Uniform
15	4	0.18	-0.74	1.74	Male	C	Non-uniform
	12	7.32	0.86	-2.02	Female	C	
	14	2.67	0.43	-1.01	Female	B	
18	6	0.33	-0.48	1.13	Male	B	Uniform
19	7	0.37	-0.43	1.01	Male	B	Uniform
	17	0.31	-0.51	1.2	Male	B	
20	14	5.13	0.71	-1.67	Female	C	Uniform
	17	4.73	0.67	-1.57	Female	C	

23	6	0.21	-0.68	1.6	Male	C	Non-uniform
	16	7.72	0.89	-2.09	Female	C	
24	2	3.6	0.56	-1.32	Female	B	Uniform
	12	3.13	0.5	-1.18	Female	B	
	9	2.72	0.43	-1.01	Female	B	
25	14	0.36	-0.44	1.03	Male	B	
	19	0.33	-0.48	1.13	Male	B	Non-uniform
26	20	0.33	-0.48	1.13	Male	B	Uniform
	11	2.67	0.43	-1.01	Female	B	
27	6	0.29	-0.54	1.27	Male	B	Uniform
28	16	0.37	-0.43	1.01	Male	B	Uniform
29	1	4.56	0.66	-1.55	Female	C	
	2	2.92	0.47	-1.1	Female	B	Non-uniform
	3	0.23	-0.64	1.5	Male	C	
	13	3.14	0.5	-1.18	Female	B	
	6	0.33	0.52	-1.22	Female	B	
30	7	3	0.48	-1.13	Female	B	Non-uniform
	9	3.18	0.5	-1.18	Female	B	
	20	0.33	-0.48	1.13	Male	B	
33	1	4.56	0.66	-1.55	Female	C	
	16	4.41	0.64	-1.5	Female	C	Uniform
34	12	0.17	-0.77	1.81	Male	C	Uniform

Table 5: RESULTS OF DIF MAGNITUDES FOR ITEMS THAT DISPLAYED SIGNIFICANT DIF AT EACH OF THE FIVE SAMPLE SIZES AND SAMPLE SIZE COMBINATIONS

ITEM No.	SAMPLE SIZE	SAMPLE SIZE COMBINATION {RATIO: FEMALE (F) : MALE(M)}	ODDS RATIO (α_{MH})	LOGARITHM OF ODDS RATIO {ln (α_{MH})}	DIF MAGNITUDE (Δ_{MH})= -2.35ln (α_{MH})	DIF AGAINST	CATEGORY
3	12	6:6	0.1	-1	2.35	Male	C
	48	12:36	0.45	-0.35	0.82	Male	A
	12	6:6	4.14	0.62	-1.46	Female	B
	18	6:12	4.14	0.62	-1.46	Female	B
5	24	12:12	4.14	0.62	-1.46	Female	B
	30	12:18	7.27	0.86	-2.02	Female	C
	48	12:36	8.69	0.94	-2.21	Female	C
7	48	12:36	0.8	-0.1	0.24	Male	A
8	12	6:6	1	0	0	NO DIF	NO DIF
	48	12:36	0.7	-0.15	0.35	Male	A
11	30	12:18	0.6	-0.22	0.52	Male	A
12	12	6:6	1	0	0	NO DIF	NO DIF
13	48	12:36	0.37	-0.43	1.01	Male	B
16	30	12:18	0.8	-0.1	0.24	Male	A
17	48	12:36	2.36	0.37	-0.87	Female	A
18	12	6:6	10.15	1.01	-2.37	Female	C
	48	12:36	2.31	0.36	-0.85	Female	A
	24	12:12	9.09	0.96	-2.26	Female	C
20	30	12:18	6.15	0.79	-1.86	Female	C

	48	12:36	4.73	0.67	-1.57	Female	C
23	12	6:6	0.1	-1	2.35	Male	C
	48	12:36	3.21	0.51	-1.2	Female	B
24	24	12:12	1	0	0	NO DIF	NO DIF
	12	6:6	1	0	0	NO DIF	NO DIF
	18	6:12	0.42	-0.38	0.89	Male	A
25	24	12:12	1	0	0	NO DIF	NO DIF
	30	12:18	0.63	-0.2	0.47	Female	A
	48	12:36	0.49	-0.31	0.73	Male	A
26	30	12:18	9.96	1	-2.35	Female	C
27	48	12:36	0.23	-0.64	1.5	Male	C
28	48	12:36	0.61	-0.21	0.49	Male	A
	12	6:6	0.41	-0.39	0.92	Male	A
	18	6:12	0.67	-0.17	0.4	Male	A
30	24	12:12	0.46	-0.34	0.8	Male	A
	30	12:18	0.54	-0.27	0.63	Male	A
	48	12:36	1.1	0.04	-0.09	Female	A
31	48	12:36	0.19	-0.72	1.69	Male	C
33	48	12:36	0.72	-0.14	0.33	Male	A

Table 6: CATEGORY "B" AND "C" DIF MAGNITUDE FOR ITEMS THAT DISPLAYED SIGNIFICANT DIF ATFIVE SSAMPLE SIZES AND SAMPLE SIZE COMBINATIONS

ITEM No	SAMPLE SIZE	SAMPLE SIZE COMBINATION {RATIO: FEMALE (F) : MALE(M)}	ODDS RATIO (αMH)	LOGARITHM OF ODDS RATIO {ln (αMH)}	DIF MAGNITUDE (ΔMH)=-2.35ln (αMH)	DIF AGAINST	CATEGORY	TYPE OF DIF
3	12	6:6	0.1	-1	2.35	Male	C	Uniform
5	12	6:6	4.14	0.62	-1.46	Female	B	Uniform
	18	6:12	4.14	0.62	-1.46	Female	B	
	24	12:12	4.14	0.62	-1.46	Female	B	
	30	12:18	7.27	0.86	-2.02	Female	C	
	48	12:36	8.69	0.94	-2.21	Female	C	
13	48	12:36	0.37	-0.43	1.01	Male	B	Uniform
18	12	6:6	10.15	1.01	-2.37	Female	C	Uniform
20	24	12:12	9.09	0.96	-2.26	Female	C	Uniform
	30	12:18	6.15	0.79	-1.86	Female	C	
	48	12:36	4.73	0.67	-1.57	Female	C	
23	12	6:6	0.1	-1	2.35	Male	C	Non-
	48	12:36	3.21	0.51	-1.2	Female	B	uniform
26	30	12:18	9.96	1	-2.35	Female	C	Uniform
27	48	12:36	0.23	-0.64	1.5	Male	C	Uniform
31	48	12:36	0.19	-0.72	1.69	Male	C	Uniform

Tables 3 and 5, respectively, show magnitudes of DIF for the respective statistically significant DIF for each of the 34 items at each of the 21 score levels and five sample sizes/sample size combinations. Negative values of logarithms of odds ratios or positive values of magnitudes of DIF indicate that the DIF was in favor of female examinees and positive values of logarithms of odds ratios or negative values of magnitudes of DIF indicate that the DIF was in favor of male examinees.

4.2.1 Accomplishment of Objective No. 1 (To find out the extent to which MH DIF detection procedure can detect DIF in MSCE Computer Studies Examination items)

Results in Tables 3 and 5 show that all categories of DIF (A, B and C) were detected in MSCE Computer Studies examination items. Disregarding items in the "A" category, results in Table 4 show that 11 items had consistently been detected as DIF items at more than 1 score level representing a consistent detection rate of 32.35% across score levels. These items are 5, 7, 15, 19, 20, 23, 24, 25, 29, 30 and 33.

Results in Table 6 show that 3 items, i.e. 5, 20 and 23 were consistently detected as DIF items at more than one sample size/sample size combination; representing a consistent detection rate of 8.82% across sample sizes/sample size combinations.
Results in Tables 4 and 6 indicate that three items, i.e. 5, 20 and 23 were consistently detected as DIF items both at score level data and sample size/sample size combination data; representing a consistent detection rate of 8.82% across all score levels and all sample sizes/sample size combinations.

Therefore, results in Tables 3, 4, 5 and 6 accomplish objective No. 1 of this research. The results indicate that MH DIF detection procedure was able to detect all categories of DIF magnitude in MSCE Computer Studies examination items with an overall consistent detection rate of 8.82%.

4.2.2 Accomplishment of Objective No.2 {To determine the group (focal or reference) which the DIF in MSCE Computer Studies examination items disadvantages}

Results in Table 4 show that items displayed DIF in favor of males at more score levels compared to that displayed in favor of females. For example, in item 15, DIF was in favor of females at score level No. 4 whereas it was in favor of males at score levels 12 and 14. In item20, DIF was in favor of only males at score levels 14 and 17. In item 23, DIF was favoring females at score level No. 6 and it was favoring males at score level No. 16. In item29, DIF was favoring females at score level No. 3 and was favoring males at score levels 1, 2 and 13. In item 30, DIF was favoring females at score level No. 20 and was favoring males at score levels 6, 7 and 9.

All in all, results in Table 4 show that DIF in favor of males had been detected in the concerned items 24 times at different score levels and DIF in favor of females had been detected in the concerned items 20 times at different score levels indicating that more DIF disadvantaging females was detected in the items.

Similarly, results in Table 6 also show that more DIF that was detected in the concerned items at different sample sizes and sample size combinations, disadvantaged females. For instance, at sample size combination of 6F:6M, DIF disadvantaging males was detected two (2) times in items 3 and 23 while that disadvantaging females was also detected two (2) times in items 5 and 18; at sample size combination of 6F:12M, DIF disadvantaging females was detected once in item 5 and no DIF disadvantaging males was detected; at sample size combination of 12F:12M, DIF disadvantaging females was detected two (2) times in items 5 and 20 and no DIF disadvantaging males was detected; at sample size combination of 12F:18M, DIF disadvantaging females was detected three (3) times in items 5, 20 and 26 and no DIF disadvantaging males was detected; at sample size combination of 12F:36M, DIF disadvantaging females was detected three (3) times in items 5,20 and 23 and DIF disadvantaging males was also detected three (3) times in items 13, 27 and 31. All in all, DIF disadvantaging females was detected in the concerned items 11 times at different sample sizes and sample size combinations and DIF in favor of females was detected in the concerned items only 5 times at different sample sizes and sample size combinations.

Therefore, results in Tables 4 and 6 accomplished research objective No. 2. The DIF in MSCE Computer Studies examination items disadvantaged females.

4.2.3 Accomplishment of Objective No. 3 {To establish the type of DIF (Uniform or Non-uniform) that MH DIF detection procedure is able to detect in MSCE Computer Studies examination items}

Results in Tables 4 and 6 show that three items, i.e. 5, 20 and 23, were consistently detected as DIF items both at score level data and sample size data analyses. Items 5 and 20 were consistently detected as having uniform DIF while item 23 was consistently detected as having non-uniform DIF.

These results accomplished objective No. 3. MH DIF detection procedure was able to detect both uniform and non-uniform DIF in MSCE Computer Studies examination items.

4.3 Discussion of Results of the Study

The purpose of this study was to re-examine DIF in MSCE Computer Studies examination items using CTT-based Mantel-Haenszel DIF detection procedure. To accomplish this purpose, three objectives were considered; namely: (1) To find out the extent to which MH DIF detection procedure can detect DIF in MSCE Computer Studies examination items (2) To determine the group (focal or reference) which the DIF in MSCE Computer Studies examination items disadvantages (3) To establish the type of DIF (uniform or non-uniform) that MH DIF detection procedure is able to detect in MSCE Computer Studies examination items.

In the current study, DIF analysis was conducted on two sample conditions, namely; score level data and sample size/sample size combination data. At score level data, the analysis for each item was done at each of the twenty-one score levels which means that there were twenty-one replications for each of the 34 items. Likewise, at sample size/sample size combination data, the analysis for each item was done at each of the five sample sizes/sample size combinations which means that there were five replications for each of the 34 items for this sample condition making a total of twenty-six (26) replications for each of the 34 items.

Results in Tables 3, 4, 5 and 6 indicate that items 5, 20 and 23 were the ones that consistently showed up as DIF items both at score level data and sample size/sample size combination data after several replications. Items 5, 20 and 23 were, therefore, the ones that were deemed to have been effectively detected as DIF items by MH DIF detection procedure in 2012 MSCE Computer Studies examination items; representing 8.82% detection rate. These findings are in line with what is widely reported in literature that MH DIF detection procedure is more conservative compared to most of the DIF detection procedures including IRT-based DIF detection procedures (Hadley, 1994; Teresi, 2004; Wiberg, 2007 & Lopez, 2012). This also confirms other research findings which indicate that, compared with other methods, in science, MH DIF detection procedure consistently flagged the smallest number of items ranging from 8.4% (Gierl, Khaliq, & Boughton, 1999). Results in Tables 4 and 6 further show that the DIF in 2 out of the 3 effectively detected DIF items (5, 20 and 23) was against female examinees both at score level data and sample size/sample size combination data representing 66.67% of the effectively detected DIF items. In Jailosi's (2012) study, 7 out of 10 significantly detected DIF items (representing 70%) disadvantaged females (Jailosi, 2012).

All the three objectives of this study have consistently been fulfilled by DIF results in Tables 3, 4, 5, and 6 with the three items (5, 20 and 23) which were deemed to have been effectively detected as DIF items; all the three categories of DIF magnitude (A, B and C) were detected; both uniform and non-uniform DIF was detected and more DIF disadvantaged females. For instance, at score level data, results in Table 4 for the three concerned items, show that DIF in favor of males was detected nine times whilst that in favor of females was detected only once. Likewise, at sample size/sample size combination data, results in Table 6 show that DIF in favor of males was detected nine times while that in favor of females was detected only once.

Results in Tables 4 and 6 further show that, of the effectively detected DIF items (5, 20 and 23), one item (item 23) was consistently detected as a non-uniform DIF item representing 33.33% detection rate of non-uniform DIF. Two items (5 and 20) were consistently detected as having uniform DIF representing 66.67% detection rate of uniform DIF. Both items with uniform DIF consistently disadvantaged females. As shown, uniform DIF was detected to a greater extent than non-uniform DIF. This is in support of what is widely reported in literature that, in general, MH DIF detection procedure is good at detection of uniform DIF and poor at detection of non-uniform DIF (Hadley, 1994; Teresi, 2004; Wiberg, 2007 & Lopez, 2012). These results are also in support of what is again reported in literature that non-uniform DIF is very rare in items that undergo extensive qualitative analysis as is the case with items constructed at ETS where MH DIF detection procedure is considered as a screen that can identify first level DIF (Teresi, 2004). Likewise, all items constructed by MANEB (MSCE Computer Studies examination items inclusive) go through an extensive process of item construction to ensure that only those items that are deemed suitable are the ones that are used to compile examination papers (Mankhomwa, 2013). As such, very rare non-uniform DIF was expected in MSCE Computer Studies examination items as portrayed by results of the current research. Furthermore, it is also reported in literature that non-uniform DIF is quite rare in practice (Gierl, Jodoin, & Ackerman, 2000).

To some extent, results of DIF study by Jailosi and those of the current research agree. Both studies have indicated that there is DIF in MSCE Computer Studies examination items and that this DIF is disadvantaging females more than males. However, it has been observed that in Jailosi's (2012) study, 14 out of 20 items, representing 70%, were detected as DIF items. In the current study, only 3 out of 34 items were effectively detected as DIF items both at score level data and sample size/sample size combination data; representing 8.82% detection rate.

4.4 Comparison of results of the current research and those of Jailosi's (2012) study

DIF results of the current research have confirmed Jailosi's (2012) DIF results that DIF is present in MSCE Computer Studies examination items and that the DIF is in favor of male examinees.

However, it has been observed that IRT-LR DIF detection method that was used by Jailosi (2012) detected a bigger percentage (70%) of DIF items in MSCE Computer Studies examination than that detected in the current research (8.82%) by MH DIF detection procedure. The high DIF detection rate by IRT-LR DIF detection procedure in the earlier research could have been contributed by:

1. Poor model-data fit of the underlying model

There is a very reasonable probability that, in Jailosi's (2012) DIF research, there could have been no good model fit to the MSCE Computer Studies examination items' data because MSCE Computer Studies examination items' data is based on CTT and the method that was used for the detection of DIF (IRT-LR DIF detection method) is based on IRT. Therefore, it can be said that, in a good number of items, this model-data misfit could have been mistaken for DIF; hence the high DIF detection rate in the earlier research.

2. Poor estimation of item parameters

It could be that item parameters were poorly estimated in the earlier research by Jailosi due to the usually small sample of examinees at a given ability level in MSCE Computer Studies examination. IRT-LR DIF detection methods require large samples for proper estimation of item parameters. Poorly estimated item parameters could have contributed to erroneous decisions about the presence of DIF in some items; resulting in the high percentage of the detected DIF.

3. Lack of determination of magnitude of DIF

Literature recommends that DIF classifications should incorporate both statistical significance and the magnitude of DIF because statistical significance does not necessarily imply practical significance (Scott, et al., 2010). In Jailosi's (2012) research, the magnitude of the detected DIF was not determined. As such, the high DIF detection rate in the earlier research could again be attributed to the lack of determination of the magnitude of the statistically significant DIF that was detected. It could be that some of the statistically significant DIF that was detected was in the **"A"** category and, therefore, negligible. Consequently, such items with negligible DIF could have contributed to the high percentage of DIF detection rate by IRT-LR DIF detection method in the earlier research.

4. Lack of replications

Lack of replications of the analysis in the DIF research by Jailosi (2012), as per the recommendation in literature, could also have contributed to the high DIF detection rate in the earlier research because through replications, it is not uncommon to see approximately 50% of the DIF in one direction and 50% in another making the overall impact of DIF very minimal (Hambleton, 2006).

On the other hand, MH DIF detection procedure, which is CTT-based, might have fitted well to the MSCE Computer Studies examination items' data which is also CTT-based; as such, there was nothing or very little, if any, that could have been mistaken for DIF other than the actual DIF. Again, incorporation of both statistical significance and the magnitude of DIF in the current research helped in determining only those items with both statistically and practically significant DIF; hence, the lower rate of DIF detection. Several replications of the analysis in the current research could also have made the possibility of cancellation of some DIF which was operating in the opposite direction; hence, the overall minimal impact of DIF and the consequent smaller percentage of DIF that was detected. The effectiveness of MH DIF detection procedure in small samples could also have assisted the MH DIF detection procedure employed in this research to effectively detect the actual DIF that was present in the usually small sample sizes of examinees at various ability levels in MSCE Computer Studies examination items.

CHAPTER 5

CONCLUSION

5.1 Chapter Overview

This chapter provides conclusions, implications of the study, recommendations to MANEB and recommendations for future research.

5.2 Conclusions

Results of this research study have positively fulfilled all the research objectives presented in Chapter 1. In general, MH DIF detection procedure employed in the current research has proved to have performed according to the expectation as outlined in the statement of the problem and in accordance with the literature; in the detection of gender DIF in MSCE Computer Studies examination items. Apart from the fact that a greater percentage of DIF was detected in Jailosi's (2012) research, current research results have confirmed results of the earlier research that there is DIF in MSCE Computer Studies examination items and that the DIF is in favor of male examinees.

It can, therefore, be, categorically, concluded that:

1. MH DIF detection procedure was able to, satisfactorily; detect both statistically and practically significant gender DIF in MSCE Computer Studies examination items.
2. MH DIF detection procedure was able to detect both uniform and non-uniform DIF in MSCE Computer Studies examination items.
3. As is the case with what was reported in Jailosi's study, the current research also found that more DIF in MSCE Computer Studies examination items was in favor of male examinees.
4. According to the expectation as outlined in the statement of the problem and in accordance with literature, MH DIF detection method was more conservative in the detection of gender DIF in MSCE Computer Studies

examination items compared to IRT-LR DIF detection method which was used in Jailosi's research in 2012.

5. MH DIF detection procedure seemed to have worked very effectively in the current research with small sizes of samples of both reference and focal groups at each particular score level; supporting what is reported in literature that MH DIF detection procedure works very effectively with small samples.

5.2.1 Implications of the study

Results from this study have implications on test quality. Test scores from MANEB examinations (MSCE Computer Studies examination inclusive) are used for making important decisions regarding certification, selection to institutions of higher learning and placement into different programs at institutions of higher learning and into various sectors of employment. They are also an indicator of the quality of education in Malawi. For this reason, a complete evaluation of test quality is paramount and must include a proper evaluation of each item to ensure that items are assessing only knowledge or skills that are identified as part of the domain being tested and should avoid irrelevant factors such as gender. Examination items should be fair among examinees from all possible subgroups of the population of examinees. Otherwise, the results can show an incomplete and wrong picture of the quality of education in the country vis-a-vis certificates for different groups and this can lead to the resources for education being distributed in an unfair manner. Similarly, it is necessary for item writers to develop test items and subject them to pilot studies which should include proper analysis of DIF to ensure that items that are selected are free from differential functioning.

5.2.2 Recommendations to Malawi National Examinations Board (MANEB)

MANEB is a national examining body that develops MSCE Computer Studies examination items and other national examination items.

Hambleton (2006) has recommended that, after formal judgmental review of a measuring instrument and DIF analyses, a feedback of the findings should be given to the developers of the instrument for future use. As such, the following recommendations can be considered by MANEB:

1. It is recommended that MANEB should seriously consider conducting gender DIF analyses in MSCE Computer Studies examination items more especially now that two DIF research studies (the current study and Jailosi, 2012), which used different DIF detection procedures, have confirmed the presence of gender DIF in MSCE Computer Studies examination items.

2. Should consideration to conduct gender DIF analyses in MSCE Computer Studies examination items be made by MANEB, it is recommended that MH DIF detection procedure should be preferred to IRT-LR DIF detection procedure because, apart from the numerous advantages that MH DIF detection procedures have over IRT based DIF detection procedures, the current research found that MH DIF detection procedure that was employed in this study seemed to have fitted well to the MSCE Computer Studies examination items' data as manifested by its considerable consistency in the DIF detection and results which are comparable to those reported in literature. As mentioned earlier on, both MSCE Computer Studies examination items' data and MH DIF detection procedure are based on CTT; as such, major manipulations in the system, which may be both economically and practically difficult, will not be necessary if implementation of results of the current research will be considered.

3. MANEB should also consider looking at its other examinations' items for other subjects so that, should they show some elements of bias, DIF analysis could also be performed on them to, again, ensure validity of the test scores that would emanate from them. Of course, some have questioned the utility of examining DIF in "professionally developed measures" in education; arguing that tests of cognitive ability and educational achievement have been shown not to be test biased, producing equivalent performance outcomes, given equivalent test performance. However, it can always be assumed that DIF will always exist and that sources are unpredictable because too many factors are at work and too many cultural background variables exist (Teresi & Fleishman, 2007). Gierl, Jodoin & Ackerman (2000), also indicated that, because all tests are imperfect, they, in fact, may contain some items that do have DIF.

Furthermore, retaining DIF items in a test may have an effect on other psychometrics' procedures such as linking and equating (Li & Zumbo, 2009).

5.2.3 Recommendation for future research

For future research, consideration should be made in the simultaneous comparison of different DIF methods in the analysis of DIF using the same data. This will enable more meaningful interpretation of results because, by implication, individual experimental errors will be minimized. Rezaee and Shabani (2010) also indicated that, generally, an ideal DIF study is the one which uses a combination of DIF techniques.

REFERENCES

1st Penfield, R. D. (2001). Assessing Differential Item Functioning Among Multiple Groups: A Comparison of Three Mantel-Haenszel Procedures. *Applied Measurement in Education*, 27(3), 235-239.

Abedalaziz, N. (2010). Detecting Gender Related Differential Item Functioning Using Logistic Regression and Mantel-Haenszel Approaches. *Procedia - Social and Behavioural Sciences*, 7(C), 406-413.

Acar, T. (2011). Sample Size in Differential Item Functioning: An Application of Hierarchical Linear Modeling. *Educational Sciences Theory & Practice*, 11(1), 284-288.

Acar, T., & Kelecioglu, H. (2010). Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR. *Educational Sciences: Theory & Practices*, 10(2), 639-649.

Atar, B., & Kamata, A. (2011). Comparison of IRT Likelihod Ratio Test and Logistic Regression DIF Detection Procedures. *H.U. Journal of Education*, 41(1), 36-47.

Awuor, R. A. (2008). *Effect of Unequal Sample Sizes on the Power of DIF Detection: An IRT-Based Monte Carlo Study with SIBTEST and Mantel-Haensel Procedures.* Blacksburg: Virginia Polytechnique Institute and State University.

Badia, X., Prieto, L., & Linacre, J. (2002). *Differential Item Functioning and Differential Test Functionig (DIF & DTF)*. Retrieved 12 September 2016 from https://www.rasch.org/rmt/rmt163g.htm

Bao, H. (2007). *Investigating Differential Item Functionig Amplification and Cancellation in Application of Item Response Testlet Models.* Retrieved March 20, 2016, from drum.lib.umd.edu/hanle/1903/7160

Bao, H., Dayton, M. C., & Hendrickson, A. B. (2009). Differential Item Functioning Amplication and Cancellation in a Reading Test. *Practical Assessment, Research & Evaluation, 14(19), 2*

Bryant, D. U., & Forde, D. S. (2002). *Detecting and Interpreting Uniform and Non-Uniform Differential Item Functioning in a Multidimensional Test*. Florida: University of Central Florida.

Buzick, H., & Stone, E. (2011). *Recommendations for Conducting Differential Item Functioning (DIF) Analyses for Students With Disabilities ased on Previous DIF Studies.* Princetown: ETS

CheckMarket, (2016). *How to Determine Population and Survey Sample size.* Retrieved July 19, 2016, from https://www.checkmarket.com/blog/how-to-estimate-your-population-and-survey-sample-size

Chulu, B. W. (2012). *Detecting Differentially Functioning Items*. Psychometrics Lecture notes. University of Malawi, Zomba.

Clauser, B. E., & Mazor, K. M. (1998). *An NCME Instructional Module on Using Statistical Procedures to identify Differentially Functioning Test Items.* Retrieved June 24, 2013, from www. Education.gsu.edu/coshima/eprs8410/dif.pdf

Committee, H. R. (2013, January). *Definition of Health Research and Human Participants Standard Operating Procedure.* Retrieved February 3, 2016, from www.health.uct.ac.za/.../Definitions%20of%20Health%20Res....

Education, M. I. (2001). *Computer Studies Syllabus for Senior Secondary*. Domasi: Ministry of Education, Science and Technology.

Fidalgo, A. M. (2011). A New Approach for Differential Item Functioning Detection Using Mantel-Haenszel Methods: The GMHDIF Program. *The Spanish Journal of Psychology*, 14(2), 1018-1022.

Fun, X. (1998). Item Response Theory and Classical Test THeory: An Empirical Comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357.

Gierl, M. J., Jodoin, M. G., & Ackerman, T. A. (2000). *Performance of Mantel-Haenszel, Simultaneous Item Bias Test and Logistic Regression When the Proportion of DIF Items is Large.* New Orleans: American Educational Research Association (AERA).

Gierl, M., Khaliq, S. N., & Boughton, K. (1999, June 7-11). Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications. Paper Presented at the Symposium entitled "*Improving Large-Scale Assessment in Education*" at the Annual Meeting of the Canadian Society for the Study of Education, Sherbrooke, Québec, CANADA

Gondwe, M. M. (2012). *Differential Measurement of Latent Trait in Primary Mathematics Examination: A case of multiple choice test items* (Master's Thesis). University of Malawi, Zomba.

Gravetter, F. J., & Forzano, L. A. B. (2003). *Research Methods for the Behavioural Sciences*. Belmont: Wadsworth.

Hadley, P. (1994). *The Performance of the Mantel-Haenszel and Logistic Regression DIF Detection Procedures Across Sample Size and Effect Size: A Monte Carlo Study* (Master's Thesis). University of Ottawa, Ottawa, Canada.

Hambleton, R. K. (2006). Good Practices for Identifying Differential Item Functioning. *Medical Care*, 44(11), 182-188.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Test Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12(3), 38-42.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education*, 2(4), 313-334.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. London: SAGE Publications Ltd.

Hinkle, D. E., Wiersman, W., & Jurs, S. G. (1998). *Applied Statistics for The Behavioral Sciences*. New York: Houghton Mifflin Company.

Jailosi, J. F. (2012). *Finding Differential Item Functioning and its sources in Computer Studies test items using Item Response Theory Likelihood Ratio Test* (Master's Thesis). University of Malawi, Zomba.

Kilmen, S. (2016). Effect of DIF Magnitudes, Focal Group Sample SIze and DIF Ratio on the Performance of SIBTEST. *International Journal of Social Science & Education*, 6(1), 91-97.

Lei, Y., Moses, T. P., Puhan, G., & Dorans, N. J. (2008). *DIF Detection With Small Samples: Applying Smothing Techniques to Frequency Distributions in the Mantel-Haenszel Procedure*. Princetown: Educational Testing Service.

Li, Z., & Zumbo, B. D. (2009). Impact of Differential Item Functioning on Subsequent Statistical Conclusions Based on Observed Test Score Data. *Psicologica*, 30(2), 343-370.

Lopez, G. E. (2012). *Detection and Classiffication of DIF Types Using Parametric and Nonparametric Methods: A Comparison of the IRT-Likelihood Ratio Test, Crossing-SIBTEST and Logistic Regression Procedures*. Retrieved June 24, 2013, from http://scholarcommons.usf.edu/etd

Maccines, J. M. (2009). *The Mantel-Haenszel Method for Detecting DIF in Dichotomously Scored Items: A Multilevel Approach.* Retrieved July 24, 2015, from htpp://ufdc.ufl.edu/ufe0041142/00001

Magis, D., Beland, S., & Raiche, G. (2013). *CRAN-Package difR.* Retrieved April 24, 2015, from https://cran.r-project.org/web/packages/difR/difR.pdf

Magis, D., Sebastien, B., Tuerlinckx, F., & De Boek, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavioral Research Methods*, 42(3), 847-862.

Mankhomwa, W. (2013). *Item Analysis in Test Construction: Experience of The Malawi National Examinations Board.* Lilongwe, Malawi: The Southern African Association for Educational Assessment (SAAEA)

Masemola, P. (2013, June 4-6). Summary report for IRT workshop at Sogecoa Golden Peacock Hotel in Lilongwe Malawi. *The Southern African Association for Educational Assessment (SAAEA) Item Response Theory (IRT) Workshop.* Lilongwe, Malawi.

Mellenbergh, G. J. (1982). Contingency Table Models for Assessing Bias. *Journal of Educational and Behavioural Statistics*, 7(1), 105-118.

Michaelides, M. P. (2008). An Illustration of a Mantel-Haenszel Procedure to Flag Misbehaving Common Items in Test Equating. *Practical Assessment, Research and Evaluation*, 13(7), 25.

Narayaman, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and Simultaneous Item Bias Procedures for Detecting Differential Item Functioning. *Applied Psychological Measurement*, 18(4), 315-328.

Padilla, J. L., Hidalgo, D. M., Benitez, I., & Gomez-Benito, J. (2012). Comparison of Three Software Programs for Evaluating DIF by Means of the Mantel-Haenszel Procedures: EASY-DIF, DIFAS AND EZDIF. *Psicologica: International Journal of Methodology and Experimental Psychology*, 33(1), 135-156.

Patelis, T. (2012). *Research Studies, Literature Reviews and Perspectives in Psychological Science.* Retrieved October 14, 2016, from www.atiner.gr/docs/2012PSY-INTRO.pdf

Penfield, R. D. (2001). Assessing Differential Item Functionig Among Multiple Groups: A Comparison of Three Mantel-Haenszel Procedures. *Applied Measurement in Education*, 14(3), 235-259.

Penfield, R., Giacobbi, P. J., & Myers, N. (2007). Using the cumulative common log-odds ratio to identify differential item functioning of rating scale items in the exercise and sport sciences. *PubMed Res Q Exerc Sport*, 78(5), 451-464.

Progar, S., & Socan, G. (2008). An Empirical Comparison of Item Response Theory and Classical Response Theory. *Horizons of Psychology*, 17(3), 5-24.

Rezaee, A. A., & Shabani, E. (2010). Gender Differential Item Functioning Analysis of the University of Tehran English Proficiency Test. *Archive of SID, Pazhuhesh-e zabanha-ye (English)*, 56(Special Issue), 89-108.

Riley, B. (2011). *Considering Large Group Differences in Ability in DIF Analysis.* Retrieved February 3, 2016, from www.rasch.org/rmt/contents.htm: www.rasch.org/rmt/rmt252d.htm

Salubayba, T. M. (2013). Differential Item Functioning Detection in Reading Comprehension Test Using Mantel-Haenszel, Item Response Theory and Logical Data Analysis. *The International Journal of Social Sciences*, 14(1),76-82.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Gundy, M.G. (2010). *Differential Item Functioning (DIF) analyses of health-related quality of life instruments using logistic regression.* Retrieved 12 January, 2015, from http:/www.ncbi.nlm.nih.gov/pmc/articles/PMC2924271/

Shaughnessy, J. J., & Zechmeister, E. B. (1990). *Research Methods in Psychology*. New York: McGraw-Hill, Inc.

Simon, M. & Goes, J. (2012). *Sample Size Matters.* Retrieved July 19, 2016, from www.dissertationrecipes.com/wp-content/uploads/2011/04/sample size matters.pdf

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning using Logistic Regression Procedures. *Journal of Educational Measurement*, 27(4), 361-370.

Teresi, J. A., & Fleishman, J. A. (2007). Differential Item Functioning and Health Assessment. *Qual Life Res*, 16(1), 33-42.

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and health. *Psychology Science Quarterly*, 50(4), 538-612.

Teresi, J. (2004). *Differential Item Functioning and Health Assessment.* New York: Columbia University

Vernoy, M. W., & Vernoy, J. A. (1992). *Behavioral Statistics in Action.* Belmont: Wadsworth, Inc.

Wiberg, M. (2007). *Measuring and Detecting Differential Item Functioning in Criterion-Referenced Licensing Test: A Theoretical Comparison of Methods.* Retrieved 5 December 2016 from http://www.edusci.umu.se/digitalAssets/59/59534_em-no-60.pdf

Zenisky, A. L., & Hambleton, R. K. (2003). *DIF Detection and Interpretation in Large- Scale Science Assessment: Informing Item Writing Practices.* Massachusetts: Centre for Educational Assessment.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a Unitary Framework for Binaryand Likert-Type (Ordinal) Item Scores.* Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zwick, R. (2012). *A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements and Criterion Refinement.* Princeton: ETS

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An Empirical Approach to Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, 36(1), 1-28.

APPENDICES

Appendix A: 2012 Malawi School Certificate of Education Examination Question Paper for Computer Studies Paper 1

EXAMINATION NO.:__________________

THE MALAWI NATIONAL EXAMINATIONS BOARD

2012 MALAWI SCHOOL CERTIFICATE OF EDUCATION EXAMINATION

COMPUTER STUDIES

Subject Number: M039/I

Tuesday, 17 July

Time Allowed: 2 hours
8:30 – 10:30 am

PAPER I
(100 marks)

Question Number	Tick if answered	Do not write in these columns	
1			
2			
3			
4			
5			
6			
7			
8			
9			

Instructions

1. **This paper contains 9 pages. Please check.**
2. There are **two** sections in this paper, **A** and **B**.
3. Answer **all** questions in **Sections A** and **B** using the spaces provided on this question paper.
4. The maximum number of marks for each answer is indicated against each question.
5. Follow instructions for each section.
6. Write your **Examination Number** at the top of each page of the question paper.
7. Use of computers is **strictly prohibited** in this paper.
8. Use of calculators is **not** allowed.
9. In the table provided on this page, **tick** against the number of the question you have answered.

Turn over

Section A (55 marks)

There are **six** questions in this section. Answer **all** the questions in the spaces provided.

1. a. (i) Name the type of interface.

C : > dir

______________________________ **(1 mark)**

(ii) Explain any **one** disadvantage of the interface named in **1a(i).**

____

______________________________ **(2 marks)**

b. What is the importance of a "word wrap" in a word processor?

____

______________________________ **(1 mark)**

c. Mention any **two** ways in which data may be lost in a computer.

____

____

______________________________ **(2 marks)**

d. Explain **two** ways in which computer personnel can practice safe use and management of a computer laboratory.

____

____

______________________________ **(4 marks)**

2. a. Define a "Node".

____

______________________________ **(1 mark)**

Continued/...

2. (Continued)

b. Draw a well labelled diagram of a star topology.

(4 marks)

c. Discuss any **two** ways in which DOS organizes information.

____

____

____

____

(4 marks)

d. Explain any **two** methods of sorting data in database.

____

____

____

____

(4 marks)

Continued/…

3. **Figure 1**is a diagram of an excel window. Use it to answer the questions that follow.

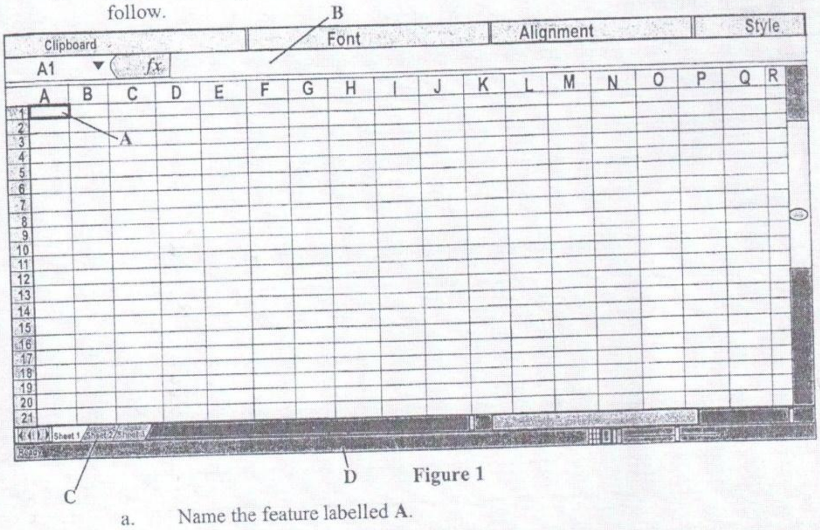

Figure 1

a. Name the feature labelled **A**.

____________________ **(1 mark)**

b. What are the functions of **B** and **D**?

B ____________________

____

D ____________________

____________________ **(4 marks)**

4. a. Define the following computer terms:

(i) "World Wide Web"

____

____________________ **(1 mark)**

(ii) "Internet"

____

____________________ **(1 mark)**

Continued/…

4. (Continued)

b. How would the use of computers lead to loss of jobs in data processing organization?

____

____

____

(2 marks)

c. Explain any **two** advantages of using Local Area Network.

____

____

____

____

____

(4 marks)

5. **Figure 2** shows a diagram of a keyboard button.

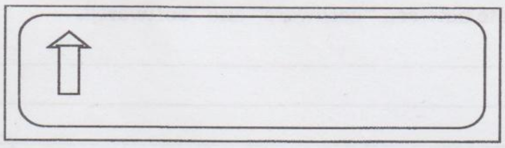

Figure 2

a. Name the button in **Figure 2**.

____

(1 mark)

b. Explain any **two** uses of the button mentioned in **5(a)**.

____

____

(4 marks)

Continued/…

5. **(Continued)**

c. How does a Dot-matrix printer work?

____

____

____

(4 marks)

d. Explain the function of a Recycle bin.

____

____

(2 marks)

6. a. List any **two** types of charts in a spreadsheet.

(i) ______________________________

(ii) ______________________________

(2 marks)

b. Differentiate between "hardware" and "software".

____

____

____

(2 marks)

c. Explain any **two** tasks carried out in an electronic database design.

____

____

____

____

(4 marks)

Continued/...

Section B (45 marks)

There are **three** questions in this section. Answer **all** questions in the spaces provided.

7. a. (i) Define "network".

____

____

____

(1mark)

(ii) Why is it important to scan an attachment before downloading in a computer?

____

____

____

(2 marks)

b. Explain any **three** factors that affect the speed of access to internet.

____

____

____

____

____

(6 marks)

c. How could a keyboard be used when

(i) selecting the entire document?

____

____

(2 marks)

(ii) pasting a block of text in Word Document?

____

____

(2 marks)

Continued/…

7. **(Continued)**

d. Give **two** advantages of standard QWERTY Keyboard.

____

____

____

____

(2marks)

8. a. Mention any **three** job opportunities in the computing field.

____

____

____

(3 marks)

b. Explain any **two** ways in which career talks by computer professionals are important to the students.

____

____

____

____

(4 marks)

c. Describe any **two** types of operating system.

____

____

____

____

(4 marks)

d. Explain any **two** advantages of using spreadsheet as opposed to using a calculator in working out the cost of items.

____

____

____

____

(4 marks)

Continued/…

9. a. Explain any **two** advantages of using mail merge.

____

____

____

(4 marks)

b. What do the following abbreviations stand for?

(i) CU: ______________________ **(1 mark)**

(ii) UPS: ______________________ **(1 mark)**

c. Name the field which uses the computer to prepare payroll.

____

(1 mark)

d. Explain any **two** characteristics of mainframe computers.

____

____

____

____

(4 marks)

e. Explain any **one** advantage of on-line booking of tickets.

____

____

(4 marks)

END OF QUESTION PAPER

NB: This paper contains 9 pages.

Appendix B: Introductory Letter from the University of Malawi in Respect of Data Collection by the researcher

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

Principal: Christopher Kamlongera,
B.A., Dip TEO., M.A., Ph.D

P. O. Box 280, Zomba, MALAWI
Tel: (265) 01 524 222
Telex: 44742 CHANCOL MI
Fax: (265) 01 524 046

Our Ref.: EDF/6/19
Your Ref.:

26th June 2013

TO WHOM IT MAY CONCERN

INTRODUCTORY LETTER FOR MASTER OF EDUCATION RESEARCH

Ms Betty Zimba is a Master of Education student in Testing, Measurement and Evaluation in the department of Educational Foundations at Chancellor College, University of Malawi.

Ms Zimba is working on her thesis *Mantel Haenszel Analysis of Differentially unctioning Items in MSCE Computer Studies Paper I.* This is meant to be a request to our institution or organization to assist our student in his endeavor to collect data.

iank you

University of Malawi
26 JUN 2013
Dept. of Educ. Foundation
P.O. BOX 280, Zomba

E Kamchedzera
UTY HEAD, DEPARTMENT OF EDUCATIONAL FOUNDATIONS

Appendix C: Permission (1) from MANEB to use 2012 MSCE Computer Studies National Examination Results

(41)

THE MALAWI NATIONAL EXAMINATIONS BOARD

INTER-OFFICE MEMORANDUM

TO: Acting Deputy Executive Director

THR: Director Security

FROM: Mrs B T Zimba 27/1/14

REF: C/

SUBJECT: REQUEST FOR PERMISSION TO USE 2012 & 2013 MSCE COMPUTER STUDIES MARKED SCRIPTS FOR GENDER DIFFERENTIAL ITEM FUNCTIONING (DIF) RESEARCH STUDY

DATE: 27/01/14

Sir,

I am a Master of Education (Testing, Measurement and Evaluation) student registration number MED/MEV/12/12 of Chancellor College, a constituent College of the University of Malawi. I intend to carry out an academic research on Differential Item Functioning in MSCE computer studies. This is in partial fulfilment of the requirements for the degree of Master of Education (Testing, Measurement and Evaluation).

I, therefore, write to humbly ask your good office if I could be allowed to use 2012 and 2013 MSCE Computer Studies Paper I marked scripts.

I promise that the data collected using the requested scripts will be treated with ultimate confidentiality

Thank you,

BTZimba

Betty T. Zimba (Mrs)

II Mrs B T Zimba

Approved,

Kee/ to your word.

05/02/14

Appendix D: Permission (2) from MANEB to Use 2012 MSCE National Examination Results

42

THE MALAWI NATIONAL EXAMINATIONS BOARD

INTER-OFFICE MEMORANDUM

TO: Acting DED

THR: Director (Security) 23/04/14

FROM: Mrs B T Zimba

REF: C/

SUBJECT: **REQUEST FOR PERMISSION TO USE 2012 & 2013 MSCE COMPUTER STUDIES MARKED SCRIPTS AND MSCE COMPUTER STUDIES PASS RATES BY GENDER SINCE 2005 FOR DIFFERENTIAL ITEM FUNCTIONING (DIF) RESEARCH STUDY**

DATE: 21/04/14

Sir,

Reference is made to my memo dated 27/01/14 in which I requested for permission to use 2012 and 2013 MSCE Computer Studies marked scripts.

I would like to thank you for the approval.

I would, however, like to humbly further request your good office to allow me to use MSCE Computer Studies pass rates by gender since 2005.

I further promise that the data collected using the requested pass rates will be treated with ultimate confidentiality.

Attached is a copy of the previous memo.

Done 23/04/2014

Thank you,

BTZimba

Betty T Zimba (Mrs)

II

Dir (CSD)

Pls note the approval by the Ag DED on Folio 41 for the supply of the requested information. Go ahead to supply the requested information, Mrs Zimba however should treat them confidentially.

23/4/14

III Dir (CSD)

Approved

23/04/14

Appendix E: Results of DIF Analysis (GMH Statistics) for each of the 34 Items at each of the Twenty-one score levels

SCORE LEVEL NO. 1
GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	N/P		
3	5.0000	1	0.0253*
4	5.0000	1	0.0253*
5	31.0000	50	0.9840
6	155.0000	22	0.0000*
7	15.0000	33	0.9970
8	N/P		
9	5.0000	1	0.0253*
10	46.5000	51	0.6528
11	155.0000	22	0.0000*
12	15.0000	33	0.9970
13	5.0000	1	0.0253*
14	5.0000	1	0.0253*
15	32.1667	51	0.9818
16	155.0000	22	0.0000*
17	15.0000	33	0.9970
18	N/P		
19	5.0000	1	0.0253*
20	33.0000	49	0.9613
21	155.0000	22	0.0000*
22	N/P		
23	5.0000	1	0.0253*
24	5.0000	1	0.0253*
25	63.0000	51	0.1208
26	155.0000	22	0.0000*
27	15.0000	33	0.9970
28	N/P		
29	5.0000	1	0.0253*
30	86.7500	52	0.0018*
31	155.0000	22	0.0000*
32	15.0000	33	0.9970
33	5.0000	1	0.0253*
34	5.0000	1	0.0253*
35	37.2500	47	0.8449

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 2

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	37.0000	33	0.2894
3	N/P		
4	0.0270	1	0.8694
5	65.6127	50	0.0683
6	4303.0116	22	0.0000*
7	27.5971	33	0.7330
8	0.0270	1	0.8694
9	0.0270	1	0.8694
10	113.0565	51	0.0000*
11	4303.0116	22	0.0000*
12	37.0000	33	0.2894
13	N/P		
14	0.0270	1	0.8694
15	41.1642	51	0.8358
16	3894.5241	22	0.0000*
17	29.1032	33	0.6616
18	0.0270	1	0.8694
19	0.0270	1	0.8694
20	77.7079	49	0.0056*
21	3894.5241	22	0.0000*
22	37.0000	33	0.2894
23	N/P		
24	0.0270	1	0.8694
25	70.8725	51	0.0342*
26	3894.5241	22	0.0000*
27	27.6018	33	0.7328
28	0.0270	1	0.8694
29	0.0270	1	0.8694
30	96.1283	52	0.0002*
31	325.1857	22	0.0000*
32	37.0000	33	0.2894
33	N/P		
34	0.0270	1	0.8694
35	66.5430	47	0.0317*

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 3

GMHDIF Beta Version

GMH statistic = QMH1

Alpha level = 0.05

Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	37.0000	33	0.2894
3	0.0270	1	0.8694
4	N/P		
5	62.4916	50	0.1106
6	4303.0116	22	0.0000*
7	37.0000	33	0.2894
8	0.0270	1	0.8694
9	N/P		
10	123.0959	51	0.0000*
11	4303.0116	22	0.0000*
12	37.0000	33	0.2894
13	0.0270	1	0.8694
14	N/P		
15	52.5303	51	0.4145
16	3894.5241	22	0.0000*
17	35.4800	33	0.3521
18	0.0270	1	0.8694
19	N/P		
20	83.8573	49	0.0014*
21	3894.5241	22	0.0000*
22	37.0000	1	0.2894
23	0.0270	1	0.8694
24	N/P		
25	67.3921	51	0.0617
26	3894.5241	22	0.0000*
27	37.0000	33	0.2894
28	0.0270	1	0.8694
29	N/P		
30	98.2629	52	0.0001*
31	325.1857	22	0.0000*
32	37.0000	33	0.2894
33	0.0270	33	0.8694
34	N/P		
35	47.1951	47	0.4646

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 4
GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	df	p
2	111.4348	33	0.0000*
3	1.0000	1	0.3173
4	N/P		
5	N/P		
6	622.9721	22	0.0000*
7	76.9040	33	0.0000*
8	1.0000	1	0.3173
9	N/P		
10	64.2240	51	0.1011
11	622.9721	22	0.0000*
12	25034514791257200000000000000000000.0000	33	0.0000*
13	1.0000	1	0.3173
14	N/P		
15	7.8234	51	1.0000
16	106.4782	22	0.0000*
17	2095208747277580000000000000000000000000000000000.0000	33	0.0000*
18	1.0000	1	0.3173
19	N/P		
20	71.1558	49	0.0210*
21	106.4782	22	0.0000*
22	N/P		
23	1.0000	1	0.3173
24	N/P		
25	13.1189	51	1.0000
26	106.4782	22	0.0000*
27	64851834634135400.0000	33	0.0000*
28	1.0000	1	0.3173
29	N/P		
30	32.7098	52	0.9833
31	69.0951	22	0.0000*
32	15.9580	33	0.9946
33	1.0000	1	0.3173
34	N/P		
35	17.9576	47	1.0000

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 5

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	15.0000	33	0.9970
3	N/P		
4	5.0000	1	0.0253*
5	33.0000	50	0.9696
6	155.0000	22	0.0000*
7	N/P		
8	5.0000	1	0.0253*
9	5.0000	1	0.0253*
10	59.0000	1	0.2063
11	155.0000	22	0.0000*
12	15.0000	33	0.9970
13	N/P		
14	5.0000	1	0.0253*
15	53.0000	51	0.3969
16	155.0000	22	0.0000*
17	15.0000	33	0.9970
18	5.0000	1	0.0253*
19	5.0000	1	0.0253*
20	42.6667	49	0.7262
21	155.0000	22	0.0000*
22	15.0000	33	0.9970
23	N/P		
24	5.0000	1	0.0253*
25	22.6667	51	0.9998
26	155.0000	22	0.0000*
27	N/P		
28	5.0000	1	0.0253*
29	5.0000	1	0.0253*
30	21.0000	52	1.0000
31	155.0000	22	0.0000*
32	103.0000	33	0.0000*
33	N/P		
34	5.0000	1	0.0253*
35	37.0000	47	0.8520

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 6

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	27.3561	33	0.7439
3	0.0270	1	0.8694
4	0.0270	1	0.8694
5	63.5059	50	0.0950
6	4303.0116	22	0.0000*
7	37.0000	33	0.2894
8	N/P		
9	0.0270	1	0.8694
10	112.5173	51	0.0000*
11	4303.0116	22	0.0000*
12	27.5008	33	0.7374
13	0.0270	1	0.8694
14	0.0270	1	0.8694
15	47.1001	51	0.6293
16	3894.5241	22	0.0000*
17	35.4800	33	0.3521
18	N/P		
19	0.0270	1	0.8694
20	73.0962	49	0.0144*
21	3894.5241	22	0.0000*
22	27.6423	33	0.7309
23	0.0270	1	0.8694
24	0.0270	1	0.8694
25	70.6761	51	0.0354*
26	3894.5241	22	0.0000*
27	37.0000	33	0.2894
28	N/P		
29	0.0270	1	0.8694
30	94.9114	52	0.0003*
31	321.6812	22	0.0000*
32	27.5159	33	0.7367
33	0.0270	1	0.8694
34	0.0270	1	0.8694
35	59.4291		0.1054

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 7

GMHDIF   Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	37.0000	1	0.2894
3	0.0270	1	0.8694
4	N/P		
5	62.4916	50	0.1106
6	4303.0116	22	0.0000*
7	37.0000	33	0.2894
8	0.0270	1	0.8694
9	N/P		
10	123.0959	51	0.0000*
11	4303.0116	22	0.0000*
12	37.0000	33	0.2894
13	0.0270	1	0.8694
14	N/P		
15	52.5303	51	0.4145
16	3894.5241	22	0.0000*
17	35.4800	33	0.3521
18	0.0270	1	0.8694
19	N/P		
20	83.8573	49	0.0014*
21	3894.5241	22	0.0000*
22	37.0000	33	0.2894
23	0.0270	1	0.8694
24	N/P		
25	67.3921	51	0.0617
26	3894.5241	22	0.0000*
27	37.0000	33	0.2894
28	0.0270	1	0.8694
29	N/P		
30	98.2629	52	0.0001*
31	325.1857	22	0.0000*
32	37.0000	33	0.2894
33	0.0270	1	0.8694
34	N/P		
35	47.1951	47	0.4646

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 8
GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	15.9743	33	0.9945
3	0.0645	1	0.7995
4	1.0000	1	0.3173
5	40.7730	50	0.8209
6	579.8254	22	0.0000*
7	15.9429	33	0.9946
8	2.8279	1	0.0926
9	1.0000	1	0.3173
10	N/P		
11	579.8254	22	0.0000*
12	13.5347	33	0.9989
13	0.0645	1	0.7995
14	1.0000	1	0.3173
15	32.9976	51	0.9762
16	579.8254	22	0.0000*
17	15.5679	33	0.9957
18	2.8279	1	0.0926
19	1.0000	1	0.3173
20	18.9305	49	1.0000
21	579.8254	22	0.0000*
22	15.4953	33	0.9959
23	0.0645	1	0.7995
24	1.0000	1	0.3173
25	33.8188	51	0.9695
26	579.8254	22	0.0000*
27	15.8703	33	0.9948
28	2.8279	1	0.0926
29	1.0000	1	0.3173
30	3.4438	52	1.0000
31	579.8254	22	0.0000*
32	190.4552	33	0.0000*
33	0.0645	1	0.7995
34	1.0000	1	0.3173
35	20.2613	47	0.9998

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 9

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	**QMH**	**Df**	**P**
2	103.0000	33	0.0000*
3	5.0000	1	0.0253*
4	5.0000	1	0.0253*
5	20.0000	50	1.0000
6	155.0000	22	0.0000*
7	103.0000	33	0.0000*
8	N/P		
9	5.0000	1	0.0256*
10	12.0000	51	1.0000
11	155.0000	22	0.0000*
12	N/P		
13	5.0000	1	0.0253*
14	5.0000	1	0.0253*
15	27.0000	51	0.9977
16	84.0000	22	0.0000*
17	161.0000	33	0.0000*
18	N/P		
19	5.0000	1	0.0253*
20	61.0000	49	0.1167
21	84.0000	22	0.0000*
22	161.0000	33	0.0000*
23	5.0000	1	0.0253*
24	5.0000	1	0.0253*
25	42.2667	51	0.8031
26	84.0000	22	0.0000*
27	161.0000	33	0.0000*
28	N/P		
29	5.0000	1	0.0253*
30	25.0000	52	0.9994
31	7.0000	22	0.9990
32	N/P		
33	5.0000	1	0.0253*
34	5.0000	1	0.0253*
35	24.6667	47	0.9970

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 10

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	**QMH**	**Df**	**p**
2	37.0000	33	0.2894
3	N/P		
4	0.0270	1	0.8694
5	65.6127	50	0.0683
6	4303.0116	22	0.0000*
7	27.5971	33	0.7330
8	0.0270	1	0.8694
9	0.0270	1	0.8694
10	113.0565	51	0.0000*
11	4303.0116	22	0.0000*
12	37.0000	33	0.2894
13	N/P		
14	0.0270	1	0.8694
15	41.1642	51	0.8358
16	3894.5241	22	0.0000*
17	29.1032	33	0.6616
18	0.0270	1	0.8694
19	0.0270	1	0.8694
20	77.7079	49	0.0056*
21	3894.5241	22	0.0000*
22	37.0000	33	0.2894
23	N/P		
24	0.0270	1	0.8694
25	70.8725	51	0.0342*
26	3894.5241	22	0.0000*
27	27.6018	33	0.7328
28	0.0270	1	0.8694
29	0.0270	1	0.8694
30	96.1283	52	0.0002*
31	325.1857	22	0.0000*
32	37.0000	33	0.2894
33	N/P		
34	0.0270	1	0.8694
35	66.5430	47	0.0317*

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 11

GMHDIF   Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	37.0000	33	0.2894
3	0.0270	1	0.8694
4	N/P		
5	62.4916	50	0.1106
6	4303.0116	22	0.0000*
7	37.0000	33	0.2894
8	0.0270	1	0.8694
9	N/P		
10	123.0959	51	0.0000*
11	4303.0116	22	0.0000*
12	37.0000	33	0.2894
13	0.0270	1	0.8694
14	N/P		
15	52.5303	51	0.4145
16	3894.5241	22	0.0000*
17	35.4800	33	0.3521
18	0.0270	1	0.8694
19	N/P		
20	83.8573	49	0.0014*
21	3894.5241	22	0.0000*
22	37.0000	33	0.2894
23	0.0270	1	0.8694
24	N/P		
25	67.3921	51	0.0617
26	3894.5241	22	0.0000*
27	37.0000	33	0.2894
28	0.0270	1	0.8694
29	N/P		
30	98.2629	52	0.0001*
31	325.1857	22	0.0000*
32	37.0000	22	0.2894
33	0.0270	1	0.8694
34	N/P		
35	47.1951	47	0.4646

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 12

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	df	p
2	711680721041422000000000000000000000000000000000000000000000000000.0000	33	0.0000*
3	3.0567	1	0.0804
4	N/P		
5	6.9962	50	1.0000
6	1155.5913	22	0.0000*
7	135.5200	33	0.0000*
8	0.2149	1	0.6429
9	N/P		
10	123.0965	51	0.0000*
11	1155.5913	22	0.0000*
12	2269449967192730000000000000000000000000000000000000000000000000.0000	33	0.0000*
13	3.0567	1	0.0804
14	N/P		
15	N/P		
16	814.1083	22	0.0000*
17	N/P		
18	0.2149	1	0.6429
19	N/P		
20	96.9791	49	0.0001*
21	814.1083	22	0.0000*
22	N/P		
23	3.0567	1	0.0804
24	N/P		
25	10.4652	51	1.0000
26	814.1083	22	0.0000*
27	N/P		
28	0.2149	1	0.6429
29	N/P		
30	50.5268	52	0.5320
31	26.3977	22	0.2350
32	N/P		
33	3.0567	1	0.0804
34	N/P		
35	24.3621	47	0.9974

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 13

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	861.0705	33	0.0000*
3	1.0000	1	0.3173
4	1.0000	1	0.3173
5	3.2280	50	1.0000
6	45.4100	22	0.0024*
7	3795.1805	33	0.0000*
8	1.0000	1	0.3173
9	1.0000	1	0.3173
10	1.4059	51	1.0000
11	45.4100	22	0.0024*
12	760.6367	33	0.0000*
13	1.0000	1	0.3173
14	1.0000	1	0.3173
15	6.3340	51	1.0000
16	N/P		
17	20.9263	33	0.9490
18	1.0000	1	0.3173
19	1.0000	1	0.3173
20	1.9165	49	1.0000
21	N/P		
22	0.4587	33	1.0000
23	1.0000	1	0.3173
24	1.0000	1	0.3173
25	3.2766	51	1.0000
26	N/P		
27	0.4587	33	1.0000
28	1.0000	1	0.3173
29	1.0000	1	0.3173
30	1.1714	52	1.0000
31	0.0062	22	1.0000
32	9.0748	33	1.0000
33	1.0000	1	0.3173
34	1.0000	1	0.3173
35	9.0725	47	1.0000

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 14

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	**QMH**	**Df**	**p**
2	12.0000	33	0.9997
3	N/P		
4	4.0000	1	0.0455*
5	18.0000	50	1.0000
6	84.0000	22	0.0000*
7	12.0000	33	0.9997
8	4.0000	1	0.0455*
9	4.0000	1	0.0455*
10	18.0000	51	1.0000
11	84.0000	22	0.0000*
12	12.0000	33	0.9997
13	N/P		
14	4.0000	1	0.0455*
15	23.0000	51	0.9998
16	84.0000	22	0.0000*
17	N/P		
18	4.0000	1	0.0455*
19	4.0000	1	0.0455*
20	6.0000	49	1.0000
21	84.0000	22	0.0000*
22	12.0000	33	0.9997
23	N/P		
24	4.0000	1	0.0455*
25	19.0000	51	1.0000
26	84.0000	22	0.0000*
27	12.0000	33	0.9997
28	4.0000	1	0.0455*
29	4.0000	1	0.0455*
30	18.0000	52	1.0000
31	84.0000	22	0.0000*
32	12.0000	33	0.9997
33	N/P		
34	4.0000	1	0.0455*
35	25.0000	47	0.9965

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 15

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	27.3561	33	0.7439
3	0.0270	1	0.8694
4	0.0270	1	0.8694
5	63.5059	50	0.0950
6	4303.0116	22	0.0000*
7	37.0000	33	0.2894
8	N/P		
9	0.0270	1	0.8694
10	112.5173	51	0.0000*
11	4303.0116	22	0.0000*
12	27.5008	33	0.7374
13	0.0270	1	0.8694
14	0.0270	1	0.8694
15	47.1001	51	0.6293
16	3894.5241	22	0.0000*
17	35.4800	33	0.3521
18	N/P		
19	0.0270	1	0.8694
20	73.0962	49	0.0144*
21	3894.5241	22	0.0000*
22	27.6423	33	0.7309
23	0.0270	1	0.8694
24	0.0270	1	0.8694
25	70.6761	51	0.0354*
26	3894.5241	22	0.0000*
27	37.0000	33	0.2894
28	N/P		
29	0.0270	1	0.8694
30	94.9114	52	0.0003*
31	321.6812	22	0.0000*
32	27.5159	33	0.7367
33	0.0270	1	0.8694
34	0.0270	1	0.8694
35	59.4291	47	0.1054

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 16

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	**QMH**	**Df**	**p**
2	37.0000	33	0.2894
3	0.0270	1	0.8694
4	N/P		
5	62.4916	50	0.1106
6	4303.0116	22	0.0000*
7	37.0000	33	0.2894
8	0.0270	1	0.8694
9	N/P		
10	123.0959	51	0.0000*
11	4303.0116	22	0.0000*
12	37.0000	33	0.2894
13	0.0270	1	0.8694
14	N/P		
15	52.5303	51	0.4145
16	3894.5241	22	0.0000*
17	35.4800	33	0.3521
18	0.0270	1	0.8694
19	N/P		
20	83.8573	49	0.0014*
21	3894.5241	22	0.0000*
22	37.0000	33	0.2894
23	0.0270	1	0.8694
24	N/P		
25	67.3921	51	0.0617
26	3894.5241	22	0.0000*
27	37.0000	33	0.2894
28	0.0270	1	0.8694
29	N/P		
30	98.2629	52	0.0001*
31	325.1857	22	0.0000*
32	37.0000	33	0.2894
33	0.0270	1	0.8694
34	N/P		
35	47.1951	47	0.4646

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 17

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	df	p
2	N/P		
3	0.0862	1	0.7691
4	2.0000	1	0.1573
5	24.5112	50	0.9991
6	509.1170	22	0.0000*
7	N/P		
8	0.8036	1	0.3700
9	2.0000	1	0.1573
10	8.6715	51	1.0000
11	509.1170	22	0.0000*
12	4562697060421470000000000000000000.0000	33	0.0000*
13	0.0862	1	0.7691
14	2.0000	1	0.1573
15	39.1956	51	0.8861
16	436.4858	22	0.0000*
17	18707220957835600000000000000000000000000000000000.0000	33	0.0000*
18	0.8036	1	0.3700
19	2.0000	1	0.1573
20	N/P		
21	436.4858	22	0.0000*
22	N/P		
23	0.0862	1	0.7691
24	2.0000	1	0.1573
25	22.6258	51	0.9998
26	436.4858	22	0.0000*
27	N/P		
28	0.8036	1	0.3700
29	2.0000	1	0.1573
30	3.9407	52	1.0000
31	20.8626	22	0.5292
32	N/P		
33	0.0862	1	0.7691
34	2.0000	1	0.1573
35	35.2460	47	0.8963

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 18

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	861.0705	33	0.0000*
3	1.0000	1	0.3173
4	1.0000	1	0.3173
5	3.2280	50	1.0000
6	45.4100	22	0.0024*
7	3795.1805	33	0.0000*
8	1.0000	1	0.3173
9	1.0000	1	0.3173
10	1.4059	51	1.0000
11	45.4100	22	0.0024*
12	760.6367	33	0.0000*
13	1.0000	1	0.3173
14	1.0000	1	0.3173
15	6.3340	51	1.0000
16	N/P		
17	20.9263	33	0.9490
18	1.0000	1	0.3173
19	1.0000	1	0.3173
20	1.9165	49	1.0000
21	N/P		
22	0.4587	33	1.0000
23	1.0000	1	0.3173
24	1.0000	1	0.3173
25	3.2766	51	1.0000
26	N/P		
27	0.4587	33	1.0000
28	1.0000	1	0.3173
29	1.0000	1	0.3173
30	1.1714	52	1.0000
31	0.0062	22	1.0000
32	9.0748	33	1.0000
33	1.0000	1	0.3173
34	1.0000	1	0.3173
35	9.0725	47	1.0000

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 19

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	N/P		
3	5.0000	1	0.0253*
4	5.0000	1	0.0253*
5	31.0000	50	0.9840
6	155.0000	22	0.0000*
7	15.0000	33	0.9970
8	N/P		
9	5.0000	1	0.0253*
10	46.5000	51	0.6528
11	155.0000	22	0.0000*
12	15.0000	33	0.9970
13	5.0000	1	0.0253*
14	5.0000	1	0.0253*
15	32.1667	51	0.9818
16	155.0000	22	0.0000*
17	15.0000	33	0.9970
18	N/P		
19	5.0000	1	0.0253*
20	33.0000	49	0.9613
21	155.0000	22	0.0000*
22	N/P		
23	5.0000	1	0.0253*
24	5.0000	1	0.0253*
25	63.0000	51	0.1208
26	155.0000	22	0.0000*
27	15.0000	33	0.9970
28	N/P		
29	5.0000	1	0.0253*
30	86.7500	52	0.0018*
31	155.0000	22	0.0000*
32	15.0000	33	0.9970
33	5.0000	1	0.0253*
34	5.0000	1	0.0253*
35	37.2500	47	0.8449

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 20

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	N/P		
3	5.0000	1	0.0253*
4	5.0000	1	0.0253*
5	31.0000	50	0.9840
6	155.0000	22	0.0000*
7	15.0000	33	0.9970
8	N/P		
9	5.0000	1	0.0253*
10	46.5000	51	0.6528
11	155.0000	22	0.0000*
12	15.0000	33	0.9970
13	5.0000	1	0.0253*
14	5.0000	1	0.0253*
15	32.1667	51	0.9818
16	155.0000	22	0.0000*
17	15.0000	33	0.9970
18	N/P		
19	5.0000	1	0.0253*
20	33.0000	49	0.9613
21	155.0000	22	0.0000*
22	N/P		
23	5.0000	1	0.0253*
24	5.0000	1	0.0253*
25	63.0000	51	0.1208
26	155.0000	22	0.0000*
27	15.0000	33	0.9970
28	N/P		
29	5.0000	1	0.0253*
30	86.7500	52	0.0018*
31	155.0000	22	0.0000*
32	15.0000	33	0.9970
33	5.0000	1	0.0253*
34	5.0000	1	0.0253*
35	37.2500	47	0.8449

N/B: N/P means it was not possible to compute the GMH statistic

SCORE LEVEL NO. 21

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	Df	p
2	37.0000	33	0.2894
3	0.0270	1	0.8694
4	N/P		
5	62.4916	50	0.1106
6	4303.0116	22	0.0000*
7	37.0000	33	0.2894
8	0.0270	1	0.8694
9	N/P		
10	123.0959	51	0.0000*
11	4303.0116	22	0.0000*
12	37.0000	33	0.2894
13	0.0270	1	0.8694
14	N/P		
15	52.5303	51	0.4145
16	3894.5241	22	0.0000*
17	35.4800	33	0.3521
18	0.0270	1	0.8694
19	N/P		
21	83.8573	49	0.0014*
22	3894.5241	22	0.0000*
23	37.0000	33	0.2894
24	N/P		
25	67.3921	51	0.0617
26	3894.5241	22	0.0000*
27	37.0000	33	0.2894
28	0.0270	1	0.8694
29	N/P		
30	98.2629	52	0.0001*
31	325.1857	22	0.0000*
32	37.0000	33	0.2894
33	0.0270	1	0.8694
34	N/P		
35	47.1951	47	0.4646

N/B: N/P means it was not possible to compute the GMH statistic

Appendix F: Results of DIF Analysis (GMH Statistics) for each of the 34 Items at each of the Five Sample Sizes and Sample Size Combinations

SAMPLE SIZE / SAMPLE SIZE COMBINATION NO. 1 (6F:6M)
GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	**QMH**	**Df**	**p**
2	N/P		
3	5.0000	1	0.0253*
4	5.0000	1	0.0253*
5	15.8333	21	0.7789
6	155.0000	5	0.0000*
7	15.0000	33	0.9970
8	N/P		
9	5.0000	1	0.0253*
10	14.0000	23	0.9269
11	155.0000	5	0.0000*
12	14.0000	23	0.9269
13	155.0000	5	0.0000*
14	15.0000	33	0.9970
15	5.0000	1	0.0253*
16	155.0000	5	0.0000*
17	15.0000	33	0.9970
18	N/P		
19	5.0000	1	0.0253*
20	10.6667	22	0.9792
21	155.0000	5	0.0000*
22	5.0000	1	0.0253*
23	5.0000	1	0.0253*
24	5.0000	1	0.0253*
25	16.0000	23	0.8553
26	155.0000	5	0.0000*
27	15.0000	33	0.9970
28	N/P		
29	5.0000	1	0.0253*
30	7.0000	23	0.9995
31	155.0000	5	0.0000*
32	15.0000	33	0.9970
33	5.0000	1	0.0253*
34	5.0000	1	0.0253*
35	9.0000	23	0.9960

N/B: N/P means it was not possible to compute the GMH statistic

SAMPLE SIZE / SAMPLE SIZE COMBINATION NO. 2 (6F:12M)

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	df	p
2	37.0000	33	0.2894
3	N/P		
4	0.0270	1	0.8694
5	17.6966	21	0.6681
6	4303.0116	5	0.0000*
7	27.5971	33	0.7330
8	0.0270	1	0.8694
9	0.0270	1	0.8694
10	15.3375	23	0.8821
11	4303.0116	5	0.0000*
12	37.0000	33	0.2894
13	N/P		
14	0.0270	1	0.8694
15	25.1151	23	0.3444
16	3894.5241	5	0.0000*
17	29.1032	33	0.6616
18	0.0270	1	0.8694
19	0.0270	1	0.8694
20	25.6450	22	0.2672
21	3894.5241	5	0.0000*
22	37.0000	33	0.2894
23	N/P		
24	0.0270	1	0.8694
25	16.2151	23	0.8459
26	3894.5241	5	0.0000*
27	27.6018	33	0.7328
28	0.0270	1	0.8694
29	0.0270	1	0.8694
30	24.8950	23	0.3557
31	325.1857	5	0.0000*
32	37.0000	33	0.2894
33	N/P		
34	0.0270	1	0.8694
35	20.5768	23	0.6069

N/B: N/P means it was not possible to compute the GMH statistic

SAMPLE SIZE / SAMPLE SIZE COMBINATION NO. 3 (12F:12M)

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	df	p
2	37.0000	33	0.2894
3	0.0270	1	0.8694
4	N/P		
5	17.6781	21	0.6693
6	4303.0116	5	0.0000*
7	37.0000	33	0.2894
8	0.0270	1	0.8694
9	N/P		
10	11.2737	23	0.9802
11	4303.0116	5	0.0000*
12	37.0000	33	0.2894
13	0.0270	1	0.8694
14	N/P		
15	23.2706	23	0.4451
16	3894.5241	5	0.0000*
17	35.4800	33	0.3521
18	0.0270	1	0.8694
19	N/P		
20	26.7015	22	0.2228
21	3894.5241	5	0.0000*
22	37.0000	33	0.2894
23	0.0270	1	0.8694
24	N/P		
25	56.2545	23	0.0001*
26	3894.5241	5	0.0000*
27	37.0000	33	0.2894
28	0.0270	1	0.8694
29	N/P		
30	26.0150	23	0.3002
31	325.1857	5	0.0000*
32	37.0000	33	0.2894
33	0.0270	1	0.8694
34	N/P		
35	28.3143	23	0.2041

N/B: N/P means it was not possible to compute the GMH statistic

SAMPLE SIZE / SAMPLE SIZE COMBINATION NO. 4 (12F:18M)

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	QMH	df	P
2	N/P		
3	1.1358	1	0.2865
4	0.3215	1	0.5707
5	N/P		
6	1351.5086	5	0.0000*
7	2645643943806080000000000000000000000000000000000.0000	33	0.0000*
8	0.1062	1	0.7445
9	0.3215	1	0.5707
10	16.5520	23	0.8306
11	1351.5086	5	0.0000*
12	6902985412036140000000000000000000000000000000000.0000	33	0.0000*
13	1.1358	1	0.2865
14	0.3215	1	0.5707
15	14.4086	23	0.9146
16	65.2586	5	0.0000*
17	20567253711738600000000000000000.0000	33	0.0000*
18	0.1062	1	0.7445
19	0.3215	1	0.5707
20	17.2983	22	0.7467
21	65.2586	5	0.0000*
22	N/P		
23	1.1358	1	0.2865
24	0.3215	1	0.5707
25	28.4639	23	0.1987
26	65.2586	5	0.0000*
27	5176118793003620000000000000000000000000000000000.0000	33	0.0000*
28	0.1062	1	0.7445
29	0.3215	1	0.5707
30	6.9465	23	0.9995
31	134.1463	5	0.0000*
32	N/P		
33	1.1358	1	0.2865
34	0.3215	1	0.5707
35	24.2760	23	0.3887

N/B: N/P means it was not possible to compute the GMH statistic

SAMPLE SIZE / SAMPLE SIZE COMBINATION NO. 5 (12F:36M)

GMHDIF Beta Version
GMH statistic = QMH1
Alpha level = 0.05
Results statistically significant at the 0.05 alpha level are marked with an asterisk.

VARIABLE	**QMH**	**df**	**p**
2	15.0000	33	0.9970
3	N/P		
4	5.0000	1	0.0253*
5	9.0000	21	0.9892
6	155.0000	5	0.0000*
7	N/P		
8	5.0000	1	0.0253*
9	5.0000	1	0.0253*
10	8.6667	23	0.9970
11	155.0000	5	0.0000*
12	15.0000	33	0.9970
13	N/P		
14	5.0000	1	0.0253*
15	9.3000	23	0.9948
16	155.0000	5	0.0000*
17	15.0000	33	0.9970
18	5.0000	1	0.0253*
19	5.0000	1	0.0253*
20	20.0000	22	0.5830
21	155.0000	5	0.0000*
22	15.0000	33	0.9970
23	N/P		
24	5.0000	1	0.0253*
25	4.0000	23	1.0000
26	155.0000	5	0.0000*
27	N/P		
28	5.0000	1	0.0253*
29	5.0000	1	0.0253*
30	17.5000	23	0.7840
31	155.0000	5	0.0000*
32	103.0000	33	0.0000*
33	N/P		
34	5.0000	1	0.0253*
35	23.0000	23	0.4608

N/B: N/P means it was not possible to compute the GMH statistic