

**ASSESSING MEASUREMENT INVARIANCE WITH
DICHOTOMOUS ITEMS: THE CASE OF EGMA FROM THE
ZAMBIAN SAMPLE**

MASTER'S THESIS

BRIAN MUMBA

**MERSIN UNIVERSITY
INSTITUTE OF EDUCATIONAL SCIENCES**

**FACULTY OF EDUCATION
DEPARTMENT OF MEASUREMENT AND EVALUATION IN
EDUCATION**

Supervisor

Prof. Devrim Alici

Co-Supervisor

Assoc. Prof. N. Bilge Uzun

**MERSIN
JUNE - 2020**

ETHIC DECLARATION

This thesis is prepared in accordance with the rules specified in Mersin University Graduate Education Regulation, and I declare to comply with the following conditions:

- I have obtained all the information and the documents of the thesis per the academic rules.
- I presented all the visual and written information and study results per scientific ethics.
- I refer under the norms of scientific works about the case of exploitation of others' works.
- I used all of the referred works as references.
- I did not do any tampering with the data.
- I did not present any part of this thesis as another thesis at Mersin University or another university.
- I transfer all copyrights of this thesis to the Mersin University.

ETİK BEYAN

Mersin Üniversitesi Lisansüstü Eğitim-Öğretim Yönetmeliğinde belirtilen kurallara uygun olarak hazırladığım bu tez çalışmada,

- Tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlâk kurallarına uygun olarak sunduğumu,
- Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- Atıfta bulunduğum eserlerin tümünü kaynak olarak kullandığımı,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Bu tezin herhangi bir bölümünü Mersin Üniversitesi veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı,
- Tezin tüm telif haklarını Mersin Üniversitesi'ne devrettiğimi beyan ederim.

June 30, 2020

İmza / Signature
Öğrenci Adı ve Soyadı / Brian Mumba

ABSTRACT

ASSESSING MEASUREMENT INVARIANCE WITH DICHOTOMOUS ITEMS: THE CASE OF EGMA FROM THE ZAMBIAN SAMPLE

Assessment of measurement invariance is an essential component of construct validity in psychological measurement. However, the procedure for evaluations is dependent on the type of data. The assessment of measurement invariance with dichotomous items partially differs from those of invariance testing with continuous items. With dichotomous items, assessing invariance requires that factor loadings and thresholds are the same since both parameters follow the same item distribution curves. However, many studies have focused on invariance testing with continuous items or ordered polytomous items. Hence, the purpose of this study was to assess measurement invariance with dichotomous items.

This study demonstrates the assessment of measurement invariance using Confirmatory Factor Analysis with dichotomous items. We analysed the study with data from primary school students who participated in the Early Grade Mathematics Assessment Survey (EGMA) from different provinces of Zambia in 2018. The study determined measurement invariance for three subtests of EGMA according to gender. The total sample consisted of 4698 individuals (49% male, $n = 2305$; and 51% female, $n = 2393$). The method used was a confirmatory factor analysis with three highly correlated factors: number identification, number discrimination and recognition of number patterns.

The findings revealed measurement invariance of the EGMA model as defined by three subtests (number identification, number discrimination and recognition of number patterns), across gender subgroups. However, individual analysis with each subtest revealed that the test scores have the same unit of measurement for both male and female on two factors, excluding those of number identification that lack measurement invariance. Finally, these results contribute to evidence of validity for EGMA scores in terms of their internal structure from the perspective of equality of measurement.

Lastly, our study results, analysis and discussions contribute significantly to the modelling decisions of applied researchers in measurement invariance studies. More importantly, it has provided a strong implication of the validity of some items to researchers using EGMA in psychological assessment.

Keywords: Measurement Invariance with Dichotomous Items; Early Grade Mathematics Assessment; Multiple Group Confirmatory Factor Analysis; Validity

Supervisor: Prof. Devrim Alici, Mersin University, Department of Educational Measurement and Evaluation, Mersin. Co-supervisor: Assoc. Prof. N. Bilge Uzun, Mersin University, Department of Educational Measurement and Evaluation, Mersin.

ÖZET

İKİLİ PUANLANAN MADDELERLE ÖLÇME DEĞİŞMEZLİĞİNİN İNCELENMESİ: EGMA ZAMBİYA ÖRNEĞİ

Ölçme değişmezliğinin incelenmesi, psikolojik ölçmede yapı geçerliliğinin önemli bir bileşenidir. Ancak, inceleme süreci veri türüne bağlıdır. İkili maddelerle ölçme değişmezliğinin incelenmesi, sürekli maddelerle değişmezlik testinden kısmen farklıdır. İkili maddelerle, değişmezliğin incelenmesi, her iki parametre de aynı madde dağılım eğrilerini izlediğinden, faktör yüklerinin ve eşiklerinin aynı olmasını gerektirir. Bununla birlikte, birçok çalışma sürekli maddeler veya sıralı politom maddelerle değişmezlik testine odaklanmıştır. Bu nedenle, bu çalışmanın amacı, ölçme değişmezliğini ikili puanlanan incelemektir.

Bu çalışma, ikili puanlanan Doğrulamalı Faktör Analizi kullanılarak ölçme değişmezliğinin incelenmesini göstermektedir. Çalışmayı 2018 yılında Zambiya'nın farklı illerinden Early Grade Mathematics Assessment (EGMA) katılan ilkokul öğrencilerinin verileri ile analiz ettik. Çalışmada, EGMA'nın üç alt testi için ölçme değişmezliğini cinsiyete göre belirlenmiştir. Toplam örneklem 4698 kişiden oluşmaktadır (% 49 erkek, $n = 2305$; ve % 51 kadın, $n = 2393$). Kullanılan yöntem yüksek korelasyonlu üç faktör ile doğrulamalı bir faktör analizidir: sayı tanımlama, sayı ayrımcılığı ve sayı örüntülerinin tanınması.

Bulgular, cinsiyet alt gruplarında üç alt testle (sayı tanımlama, sayı ayrımcılığı ve sayı modellerinin tanınması) tanımlanan EGMA modelinin ölçme değişmezliğini ortaya koymuştur. Bununla birlikte, her bir alt test ile yapılan bireysel analizler, test puanlarının, ölçme değişmezliği olmayan sayı tanımlaması hariç olmak üzere, hem erkek hem de kadın için aynı ölçme birimine iki faktör üzerinde sahip olduğunu göstermiştir.

Son olarak, bu sonuçlar EGMA puanlarının ölçme değişmezliğini açısından yapı geçerliliğine ilişkin kanıtlara katkıda bulunmaktadır. Son olarak, çalışma bulguları, analiz ve tartışmalar ölçme değişmezliği çalışmalarında uygulamalı araştırmacıların modelleme kararlarına önemli ölçüde katkıda bulunmaktadır. Ayrıca, psikolojik değerlendirmede EGMA kullanan araştırmacılara bazı maddelerin geçerliliğinin güçlü bir şekilde yansımaları sağlamıştır.

Anahtar Kelimeler: İkili Puanlanan Maddelerle Ölçme Değişmezliği; Early Grade Mathematic Değerlendirmesi; Çoklu Grup Doğrulamalı Faktör Analizi; Geçerlilik

Danışman: Prof. Dr. Devrim Alıcı, Mersin Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Mersin. Eş danışman: Doç. Dr. N. Bilge Uzun, Mersin Üniversitesi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Mersin.

Acknowledgements

I am grateful for the grace given to me by my Heavenly Father, who has given me the desire to pursue studies in Educational Measurement and Evaluation. Indeed, all things are possible with Him. He has provided all strength to persevere to continue this dream. To my family, I say thank you for your constant support and patience. I appreciate the Department of Measurement and Evaluation in Education for the opportunity to learn with the incredibly knowledgeable and caring professors at Mersin University. The staff and administrators have always been caring and supportive, and their encouragement has been crucial to my progress.

I especially thank Prof. Devrim Alıcı and Assoc. Prof. N. Bilge Uzun for continually working to help direct me forward with their never-give-up attitude! The knowledge, direction, and positive approach they provide give me hope to thrive and improve in my academics. I am very much blessed to learn from such gifted teachers. This opportunity would not have been possible without the support from the Turkish Government through Turkish Scholarship, to them I cannot say enough about their love towards humanity and support they provide to many bright students, thank you. Last but not least, I wish to acknowledge that I did not use any services of the editor in constructing this thesis.

TABLE OF CONTENTS

JÜRİ KARAR.....	
ETHIC DECLARATION.....	
ETİK BEYAN.....	
ABSTRACT	i
ÖZET.....	ii
Acknowledgements	iii
ABBREVIATIONS.....	vi
LIST OF TABLES	vii
LIST OF FIGURES	viii
1. INTRODUCTION.....	1
1.1. Problem Statement.....	4
1.1.1. Theoretical Framework of Measurement Invariance.....	8
1.1.2. Measurement Invariance Models.....	9
1.1.3. Measurement Invariance with Confirmatory Factor Analysis.....	10
1.1.4. Support for Using CFA with Dichotomous Items.....	12
1.1.5. Measurement Invariance with Dichotomous Items.....	14
1.1.6. Procedures for Measurement Invariance and Fit Indices.....	16
1.2. Significance of the research	17
1.3. Purpose of the research.....	18
1.4. Research Objectives	18
1.5. Research Questions	18
1.6. Assumptions.....	19
1.7. Delimitations.....	19
1.8. Chapter Summary	19
2. REVIEW OF LITERATURE	20
2.1. Introduction	20
2.2. Measurement Invariance Related Studies	20
2.2.1. PISA Measurement Invariance Related Studies.....	20
2.2.2. TIMSS Measurement Invariance Related Studies	21
2.2.3. Measurement Invariance Studies with Dichotomous Items.....	22
2.1. Causes of Measurement Invariance in Dichotomous Items.....	23
2.2. Gender Differences in Mathematics Performance.....	24
2.3. Chapter Summary	26
3. METHODOLOGY	27
3.1. Introduction	27
3.2. Research Design.....	27
3.2. Participants.....	27
3.3. Instrument	27

3.4.	Data Analysis.....	28
4.	RESULT.....	34
4.1.	Introduction	34
4.2.	Assessment of Measurement Invariance with Dichotomous Items	34
4.3.	Measurement Invariance for the EGMA Model.....	34
4.4.	Measurement Invariance by each Factor of the EGMA Model.....	35
4.3.1.	Measurement Invariance for Number Identification.....	35
4.3.2.	Measurement Invariance for Number Discrimination.....	36
4.3.3.	Measurement Invariance for Recognition of Number Patterns	37
4.4.	Implications of using MGCFA in Measurement Invariance with Dichotomous Items ...	38
4.5.	Summary of the Findings	38
5.1.	Introduction	40
5.2.	Study Discussion.....	40
5.3.	Strength and Limitations of the study	43
5.4.	Recommendations	44
5.5.	Conclusion.....	44
	REFERENCES.....	45
	APPENDIXES.....	50
	Appendix 1: Timetable	50
	Appendix 2. Request Letter for Dataset.....	51
	Appendix 3: R Syntax Measurement Invariance Analysis.....	52
	CURRICULUM VITAE.....	57

ABBREVIATIONS

χ^2 – Chi-Square Test

ANOVA – Analysis of Variance

CFA - Confirmatory factor analysis

CFI - Comparative Fit Index

DIF – Differential Item Functioning

ECZ – Examination Council of Zambia

EGMA - Early Grade Mathematics Assessment Survey

EGRA - Early Grade Reading Assessment

IRT - Item Response Theory

MESVTEE - Ministry of Education, Science, Vocational Training, and Early Education

MGCFA – Multiple Group Confirmatory factor analysis

NFI - Normed Fit Index

PISA - Programme for International Student Assessment

RMSEA - Root Mean Square Error of Approximation

RTI International - Research Triangle Institute

SEM - Structural Equation Modelling

SSME - Snapshot of School Management Effectiveness

STME - Science, Technology and Mathematics Education

TIMSS - Trends in International Mathematics and Science Study

USAID - United States Agency for International Development

WLSMV - Weighted Least Squares Mean and Variance

LIST OF TABLES

Table 1. summary description of subtests	28
Table 2. CFA model fit by gender	29
Table 3. Summary of model fit with estimates and standardised values	30
Table 4. Measurement invariance for the three-structure model	35
Table 5. Measurement invariance for number identification	36
Table 6. Measurement invariance for number discrimination.....	36
Table 7. Measurement Invariance for recognition of number patterns	37
Table 8. The study's summary table for measurement invariance analysis	37



LIST OF FIGURES

Figure 1. CFA Model with Model Estimates and Standardised Parameters by All Gender Group	31
Figure 2. CFA Model with Model Estimates and Standardised Parameters by Male Group.....	32
Figure 3. CFA Model with Model Estimates and Standardised Parameters by Female Group	33



1. INTRODUCTION

Measurement invariance is a vital element in psychological testing that allows for a meaningful comparison of test scores across the subgroups of gender, age or ethnicity or different times. Today, it is a common understanding among researchers that there is a need to determine the extent to which the measurement models can be generalised between groups or against time. According to Koziol (2010), merely administering the same test to different groups or at multiple times does not ensure the measurement of the same underlying construct. For instance, a maths test designed for secondary school students will likely measure mathematics ability when administered to a group of secondary school students. However, applying the same test to a group of primary school students will probably measure random guessing. Hence, comparing maths ability of primary school students with that of secondary school students based on their scores from this test would not be meaningful, considering that we did not capture the presumed construct of mathematics ability with primary school students.

Thus, based on the example above, evidence of measurement invariance necessarily takes precedence of any meaningful comparisons of the construct across groups or time points (Brown, 2006). Measurement invariance, then, is based on the notion that a measuring device should function in the same across different groups on the attributes we are assessing them on irrespective of group membership, and that any observed differences should be based on the real differences among the individuals regardless of their membership status (Ergül, 2016). For instance, a maths test should capture varied abilities in mathematics across test takers whose ability levels in mathematics differ (Millsap, 2011).

According to Millsap and Meredith (2007), historically testing for measurement invariance has been enlivened as a necessitous for the meaningful comparison of test scores. Elosua highlighted some of the most common applications of measurement invariance; (1) it allows to check for evidence of measurement bias in scales administered groups or to different culture of a population, (2) allows us to examine the stability of a measurement model over time, and (3) can allow us to check for differences in latent means (2011). The distinction between measurement bias and invariance is essential. According to van de Vijver and Poortinga, bias results from the presence of a nuisance factor that produces an undesirable source of measurement variance (2005). Measurement invariance, on the other hand, is the comparability of scores between groups and examines if items are being conceived equally across the groups on the latent variables (Meredith, 1993). Bias in measurement is equivalent to systematic inaccuracies in measurement.

These inaccuracies are replicable even in repeated measurement. However, bias is different from random errors in the sense that the latter determines the reliability of a test. Millsap (2011), notes that a reliable test may or may not be biased. The unreliable test may still

give unbiased measurement across repeated tests. On the other hand, a biased test maybe highly reliable, but still give consistently inaccurate measurement even after many repeated measurements (Larry, 2019). The critical question, therefore, to ask when defining measurement bias is whether the knowledge of the examinees' group membership can tell us anything new about the examinees score on the measured variable. To the extent that the examinee's group membership adds something new apart from the examinee's level of ability on the latent variable, we will say that the measured variable is biased to the population and the underlying construct under study (Millsap, 2011). Examining measurement invariance also include checking for biased, and this can be checked by testing for scalar invariance. Thus, making bias part of the measurement invariance analysis.

According to Edwards and Wirth (2009), investigating measurement invariance provides valuable sources of information on construct validity that is critical in test construction. For instance, an instrument intended to measure reading ability across a group of primary school students in the same grade (e.g. third graders) should produce test scores that are invariant across various groups (e.g., gender sub-groups males versus females). On the other hand, if measurement invariance fails to hold, it gives us enough evidence to question the validity of such scores in measuring the underlying construct of reading ability (Koziol and Bovaird, 2009).

The above substantiates the notion that investigating measurement invariance is essential in educational and psychological assessment. However, the procedure for assessing measurement invariance is dependent on the data type. According to Muthén and Asparouhov (2002), the steps for assessing measurement invariance with categorical data differ from those of invariance testing with continuous variables. For instance, with binary items, since both factor loadings and thresholds use the same item characteristic curves, testing invariance involves treating these parameters in the same way (Widaman and Reise, 1997). Hence, we skip the third level of invariance when assessing measurement invariance with dichotomous items (Muthén and Muthén, 2009). However, many studies have focused on invariance testing with continuous items or ordered polytomous items. Hence, the purpose of this study was to assess measurement invariance with dichotomous items. The study is a demonstration of measurement invariance with dichotomous items using multiple-group confirmatory factor analysis while using Early Grade Mathematics Assessment dataset from the Zambian sample as a study case.

EGMA serves as an excellent example for our study because it contains items dichotomously scored. EGMA is one of the three instruments developed by RTI international that is in use, especially in low- and middle-income countries. The other two are the Early Grade Reading Assessment (EGRA) and Snapshot of School Management Effectiveness (SSME). EGRA, EGMA, and SSME assessment tools developed by RTI International and funded by USAID for use by various testing institutions and organisations interested in early literacies in developing

countries (RTI International, 2014). Before administration, the tools are translated and adapted to the local context by assessment experts from educational bodies in respective countries. For instance, in Zambia the tools as mentioned earlier are adapted to local contents based on the grade-appropriate curriculum and textbooks, e.g., reading, writing, and mathematics contents for grade 2, 3 and 4 are analysed and used to inform changes made on the EGRA and EGMA instruments.

Today, EGRA and EGMA are being used by governmental and non-governmental organisations in various countries to measure baseline reading and mathematical competencies of students in low- and middle-income countries whose findings provide a basis for interventions. For instance, in Zambia, the instruments have been in use since 2011 to present at grade 2 and grade 3 level. However, Zambia comprises of approximately 72 ethnic groups. This diversity presents a challenge to instructors, students as well as to those involved in designing tools of instruction for use in preschools and primary schools, together with those tools used in assessing learners. Additionally, the language of instruction from grade 1 to grade 4; Nyanja, Bemba, Lozi, Tonga, Kaonde, Luvale, and Lunda, respective to regions in which the learners come from (USAID, 2016). Although the official language is English, the teaching of initial literacy at early grade is in these seven languages up to grade 4. Hence, the need to assess for item performance across different groups.

According to Alatlı and Bökeoğlu (2018), testing is vital for several reasons. More importantly, through testing we can get information about individual students, as well about schools and teachers, e.g., scores generated from large-scale standardised testing programs like those of PISA and TIMSS provides valuable information critical for decision making. However, scores must valid for them to be appropriate and meaningful (Linn, 2009). That is, for the validity to exist, we must be able to make comparable differences for individuals regardless of their group membership using the scores produced by the testing tool (e.g., gender, race). On the contrary, if the equality of measurement is not assured, the scores resulting from the measurement tool may be carrying different meanings for different individuals, hence compromising their applicability and introducing potential unfairness into decisions in which these scores play a role. Thus, it is worth to find out what EGMA measures in Zambia, and one way to do this is by using various measurement models to screen for the psychometric properties of this instrument.

According to Larry (2019), the development of any assessment instrument is a function of psychological measurement and assessment, where the interest is to make inferences regarding the existence of the underlying construct. While researchers cannot directly observe these properties, researchers focus on the observable variables that are assumed to reflect the underlying construct under investigation (Millsap and Kwok, 2004). Murphy and Davidshofer (2004) stated that users could develop psychological assessment tools for various purposes. Acar

(2008) further adds that regardless of the purpose for which we develop tools for, the problem is determining the psychometric characteristics of the assessment tool. Hence, the screening process is one of the fundamental problems of test development in the assessment field. Moreover, one such problem that is of importance in literature is the assessment of measurement invariance. Because assessing for assessment is based on several models, among them, are Structural Equation Modelling (SEM) and Item Response Theory (IRT).

The most popular model of assessing measurement invariance is the linear factor model or Confirmatory factor analysis (CFA), which models continuous item data to their respective latent variable (Mellenbergh, 1994). However, this model is most appropriate with continuous items and most researchers are using it to assess measurement invariance with continuous items or ordered polytomous items. On the other hand, little studies have used this model to assess measurement invariance with dichotomous items. Thus, this study demonstrates the assessment of measurement invariance with dichotomous data. More importantly, it contributes to the discussion on the invariance testing with categorical/ordinal data using factor analysis. Specifically, the study presents a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items. This discussion follows a review on the contemporary research for conditions under which application of factor analysis become problematic.

After that, reviews methods for model identification and specifications for conducting measurement invariance testing with categorical data. Besides, the estimator choices for conducting measurement invariance tests with categorical data are addressed together with fit indices. For the analysis purposes, EGMA data from the Zambian sample is used as an example to demonstrate invariance testing with dichotomous items. The study was conducted in R using the packages lavaan, semTools and semPlot. These packages contain all functions needed to efficiently run Multiple Group Confirmatory Factor Analysis (MG-CFA) in R (Hirschfeld and Brachel, 2014).

This section of the paper provides a solid overview of the problem statement, the significance of the research, purpose and research questions, and a description of study assumptions and delimitations.

1.1. Problem Statement

Measurement invariance plays a crucial role in test construction and validation for the reason that is merely administering the same test to diverse groups or at multiple time points does not ensure the measurement of the same underlying construct. That is, if measurement invariance fails to hold, it questions the validity of such scores as a measure of the underlying construct of interest. Thus, assessing measurement invariance is crucial in psychological assessment. Karakoç and Bökeoğlu, made the following statement that “any differences that are

generated by the measurement tool” that utilise the same measurement to different groups may cause shortcomings in comparing between groups and in making relevant interpretations (2018, p.12). Hence, based on this statement, comparisons become only meaningful if measurement invariance is achieved. Failure to achieve it, the obtained results will have no similar meaning since the reliability and validity of the results have not been obtained (Başusta, 2010). Hence, one can say measurement invariance study is vital not only for a newly developed measurement tool but also allow for meaningful comparisons between different groups (Cheung and Rensvold, 2000).

Furthermore, Molenaar and Borsboom (2013), adds that measurement invariance provides a way for us to determine to the extent to which the measurement models can be generalised between groups or against time. Thus, with any measurement tool developed for implementation on a heterogeneous group, it is necessary to prove that the measurement characteristics of the test are equal at the subgroup level. Besides, measurement invariance is about the measurement tool; that is, it does not focus on the characteristics of the individuals included in the universe to which measurement procedures are applied (Ayvalli and Biçak, 2018). Birnbaum (1968) pointed out that, “research results are no more valid than the measures used to collect the data” (p.15). On the other hand, the assessment of measurement invariance continues to be a methodological challenge in invariance studies due to different data type collected through different measurement instruments. That is, different instruments collect different data type; for instance, we find instruments collecting examinees’ data in dichotomous form (EGMA) and also those collecting in a continuous form (PISA, TIMMS). This fact presents a challenge to a new researcher interested in invariance studies.

However, the issue of measurement invariance is not a straightforward thing since the procedure for assessing it is dependent on the type of data. That is, the assessment of measurement invariance with dichotomous items or categorical data partly differ from those of invariance testing with continuous items. According to Muthén and Asparouhov (2002), the steps for assessing measurement invariance with categorical data differ from those of invariance testing with continuous variables. According to Millsap (2011), testing for measurement invariance with categorical data encompasses three steps; but before assessing measurement invariance, model fit is first checked in the initial analysis while allowing model parameters (factor loadings, thresholds, residual variances) to be freely estimated in both conditions preferably the group under study. Subsequently, the invariance analysis then is done at each of the three steps progressively; (1) configural invariance (checking for the invariance of model structure across the groups); (2) metric invariance (checking for the invariant of factor loadings or thresholds while holding them equal across the groups); and (3) strict invariance (checking for

the invariance of loadings, thresholds or error terms while holding them equal across the groups) (Bollen, 1989).

Before 1995, children's math knowledge has been measured at the fourth-grade level through easy-to-administer written tests (Sitabkhan and Platas, 2018). However, these fourth grades and later skills are reported to be significantly influenced by earlier skills that are less amenable to written assessment. Claessens and Engel (2013) investigations found that early mathematical development has a strong influence on children's academic achievement across domains, including literacy. Because of the magnitude of these findings and the implications at both individual and societal levels, measuring of informal mathematical skills has become a priority for governments in low and middle-income countries. One such instrument in use in low- and middle-income countries is the EGMA. The EGMA is used by governmental and non-governmental organisations to measure baseline mathematical competencies of students in low- and middle-income countries whose findings form as a basis for interventions (RTI International, 2014).

As documented by USAID (2012), EGMA items were designed after extensive consultation with experts in mathematics education and cognition and with researchers interested in early mathematics learning and assessment. Besides, the conceptual framework for EGMA builds on the assumption that before formal schooling, children possess informal mathematical skills that are critical to the development of more advanced skills taught in formal education, e.g., mathematical skills acquired outside a formal school context or before a child enters school (RTI International, 2014). Sitabkhan and Platas (2018) put it that all societies use mathematical knowledge to some extent, that is, before one starts school, he or she has some mathematics skills upon which they build on more advanced skills taught in the formal curriculum. Hence, if we are to measure students' progress in mathematics skills successfully, we need to take into account of these skills. According to Booth and Siegler, informal skills play a vital role in everyone's life. For instance, these skills are at the centre of everything a child does before starting formal education, and their use includes that of keeping track of things, engaging in small businesses to that of describing direction or distance or unit of measurements (2008).

Students' performance in mathematics varies across countries. Many factors affect this variation in mathematics. Of interest in this study is the difference between female and male students' performances and the condition termed as a gender gap. While EGMA results are useful for many developing countries to assess their educational systems and identify educational policies, EGMA also allows comparison across many groups and countries. According to Ayvalli and Biçak (2018), the measurement between groups involves the assessment of psychological constructs like personality, intelligent or other constructs. Thus, the measurement tool used to assess them should be meaningful for the target group or cultures for meaningful and comparable

results (Ayvalli and Biçak, 2018). On the other hand, Quirk and others indicated that reliability and validity of the measurement tool could only be possible through the interpretation of the measured variables in the same way across different groups (Quirk, et al, 2015).

Furthermore, the developer also claims that although other cross-national mathematics studies like the TIMSS provide valuable information useful for comparing progress in mathematics performance between developing countries and developed countries, these instruments fail to address the reason behind poor performance in mathematics or where it starts (RTI International, 2014). Thus, EGMA according to the developer offers an opportunity to ascertain whether children are developing the required critical skills in mathematical and, if not, helps to track areas of concerns where efforts may best be directed (RTI International, 2014).

What differentiates EGMA from other assessment tools is based on the following questions as asked by the developer: how do we measure whether children are achieving critical mathematical skills? Besides, it also answers the question of whether educational systems at national or subnational levels are creating opportunities for students to achieve mathematical skills (RTI International, 2014). Although these questions appear to be similar in goals, a significant difference is that using the instrument one is able both to measure a child's progress along a trajectory and determine whether the extent to which an educational system is empowering a student with maths skills required to function in the twenty-first century. Hence, according to RTI International, the primary purpose of EGMA is to answer the latter question (RTI International, 2014). That is, the tool assesses, on a large scale, the level to which students have acquired specific critical skills in maths for today and tomorrow. EGMA has the following subtest; number identification, number discrimination, recognition of number patterns, and addition and subtraction that were arrived after cross-cutting research in mathematics achievement in developing nations by experts (USAID, 2016).

Besides, there is a universal agreement among researchers that any thoughtful instruction in mathematics must help students to understand numbers, ways of representing them and express relationships among numbers and number systems (Hedges and Nowell, 1995). TIMSS is built on a similar understanding. Hence, the content of EGMA is in line with those topics that TIMSS assess. Booth and Siegler give a summary of mathematical skills that should be assessed; (1) what the society intends to teach its students (the intended curriculum); 2. what the schools teach (the implemented curriculum); and 3. what the students acquire after instruction (the attained curriculum) (2008).

Similar to TIMSS, EGMA assesses mathematical skills that the intended curriculum aims to empower its students with (RTI International, 2014). Therefore, the instrument consists of foundational skills in mathematics upon which the later or more complicated skills in upper grades, for instance, those that TIMSS measure at fourth grade and above builds on. Additionally,

EGMA measure what the students ultimately learn; that is, it ties test items to what students learn in classrooms (USAID, 2016). Hence, the instrument is connected (1) to the state since it designs the curriculum, (2) to teachers since they implement the curriculum and (3) to students since they are the target of to be empowered with mathematical skills (RTI International, 2014).

However, we say that developing an assessment tool is not a mean and an end in itself. Nevertheless, it is a process that involves examining the psychometric properties of the instrument. As Murphy and Davidshofer (2004) put it after development, scrutiny of an assessments tool is required to understand the skills we are measuring and determine how accurate we were in capturing them (that is, checking for reliability and validity of the tool). Given the broader acceptance that this instrument across African countries and among others in Asia, it is worth to investigate what the instrument measures across different groups of students. Thus, this study assesses the measurement invariance of EGMA for the Zambian sample by students' gender. Assessing measurement invariance across groups of students is of fundamental importance both to the literature and to the researchers using the assessment tool.

1.1.1. Theoretical Framework of Measurement Invariance

As all test involves assessing several students from different sub-groups in a population, thus to allow for meaningful comparisons, measurement invariance is assumed. According to Hansson and Gustafsson (2013), invariance assumes that the individuals responding to a test are conceiving the test items in the same way on the construct of interest. Therefore, failure to meet this assumption would imply biased items against one or more groups in a test (Wendt, Kasper and Trendtel, 2017). Thus, measurement invariance is vital, especially when the tool involves testing different groups on the construct of interest. Let us examine this simple illustration by Glockner-Rist and Hoijsink (2003) on a measure of depression about crying. For instance, assuming that female cry more often than males, observed items about crying would be approved less likely by males than females. Hence, females would get high scores on this item and eventually, higher scores on the depression scale than males even though all the groups are equally depressed. With this difference in the way the item is capturing the construct of depression, any observed mean differences between males and females on this scale might be deceitful, for they may not be correctly representing the real gender differences on the observed construct. According to Millsap (2007) observed that scores obtained after assessment should not be related to the background characteristics of the test takers such as demographic variables for the instrument to achieve the invariance of measurement.

Below is the mathematical definition of invariance of measurement:

$$p(x_i|\xi_i, v_i) = p(x_i|\xi_i) \quad (1)$$

where $X_i = (X_{ij})$ reflects the observed scores of the i th person of the population, $\xi_i = (\xi_{ij})$ is a factor scores for that person and $V_i = (V_{ij})$ is the vector for the viewed items irrespective of

the individual characteristics irrelevant to X_i , taking into account of ξ_i . As summarised by (Millsap, 2007), therefore, for X concerning ξ and V to be invariant $P(X)$ subject to ξ and V should be equal to the $P(X)$ subject to ξ only.

Therefore, for any tool devised to assess the observed variable ξ , it is essential to ascertain the invariant of the observed item relating to the underlying construct variable ξ and the variable V , as the latter reflects the test taker's background characteristics that are unrelated to the underlying construct variable were are assessing (Wirth and Edwards, 2009). Wendt, Kasper and Trendtel noted that measurement invariance is achieved only when individual's scores on; (1) the items, (2) levels of ξ and (3) levels of V meet the following rule: the probability of getting a specific score given the level of ξ and V equals the probability of getting a similar score given the level of ξ (2017).

1.1.2. Measurement Invariance Models

Two models are used to investigate the invariant of an instrument. These models follow different assumptions; that is, we have those that work well with categorical-ordered items and those that work well with continuous items (see Mellenbergh, 1994). The most commonly used models in invariance assessment include the linear factor model (Confirmatory Factor Analysis (CFA)) (Spearman, 1904; Thurstone, 1947), which links continuously observed items to their latent variable (Mellenbergh, 1994), and the Item Response Theory (IRT) (e.g., two-parameter logistic model) which links dichotomously scored items to their latent variable (Ayala, 2009).

According to Kim and Yoon (2011), we can assess measurement invariance with the CFA and IRT models. However, many research studies exploring the relationship between the two have noted down some weakness associated with them when testing for measurement invariance with ordered-categorical items. For instance, Cimetta, Marx, Yaden, Alkhadim, and Cutshaw (2018) noted that testing measurement invariance under CFA with categorical items is not adequately addressed because it does not examine scalar invariance which is applicable only with continuous items. Besides, with respect to IRT also, testing measurement invariance is not adequately addressed because assessing for invariant is identical to assessing for differential item functioning (DIF) (Hambleton, Swaminathan, and Rogers, 1991). Thus, the CFA is considered an optional analysis for ordered-categorical data also, even though it does not adequately address the third step in the invariance testing.

On the other hand, Kim and Yoon (2011, p.12) stated that "in some cases, [IRT] can provide different and potentially more useful information for the establishment of measurement invariance". However, IRT models work well when checking for differential item functioning and item bias. The paper would have given more comparative information between the two; however, comparing the differences between them is not the aim of this study. For more information on the differences see (Kim and Yoon, 2011).

1.1.3. Measurement Invariance with Confirmatory Factor Analysis

According to Brown and Moore (2012), confirmatory factor analysis (CFA) is one tool that can allow us to investigate measurement invariance. Moreover, today it is commonly used in the literature of invariance studies. As cited by Bentler (1990), with the CFA model, the observed variable X_i follows;

$$x_i = \tau + A\xi_i + \varepsilon_i \quad (2)$$

which is a representation of the expected value and covariance matrix, respectively:

$$E(X_i) = \tau + Ak \quad (3)$$

$$\text{var}(x_i) = A\phi A' + \theta \quad (4)$$

where $\varepsilon = (\varepsilon_{ij})$ is the error term for the i th person; $\Lambda = (\lambda_{jj'})$ is the matrix of the regression coefficients; $\tau = (\tau_j)$ is the intercept and $\kappa = (\kappa_j)$ is the mean of the latent variable. $\Phi = (\phi_{jj'})$ corresponds to the covariance matrix for the unobserved variable and $\Theta = (\theta_{jj'})$ is the covariance matrix for the residuals.

Furthermore, maximum likelihood is the frequently used method of estimation in CFA, which assumes a multinormal distribution of the observed variables with parameters μ and Σ representing the means and covariances for the observed variables (Kaplan, 2009). Therefore, $P(X_i = x | \xi_i)$ can be calculated based on the means and the covariance matrix of X_i given ξ_i , which is as shown below:

$$E(x_i | \xi_i) = \tau + A\xi_i \quad (5)$$

$$\text{Var}(x_i | \xi_i) = \theta \quad (6)$$

According to Rojas, Rojas and Brizuela (2018), assessing for measurement invariance across groups, it is fundamental to make sure that the loadings (Λ), the intercepts (τ) and the residuals (Θ) are constant within the designated groups of V ; considering that for any value of V it must be true that;

$$E(x_i | \xi_i, v_i) = E(x_i | \xi_i) = \tau + A\xi_i \quad (7)$$

$$\text{Var}(x_i | \xi_i, v_i) = \text{Var}(x_i | \xi_i) = \theta \quad (8)$$

Hence, defining V for any K groups in the population, measurement invariance is assessed as follows:

$$E(x_{ik} | \xi_{ik}) = \tau_k + A_k \xi_{ik} = \tau + A \xi_{ik} \quad (9)$$

$$\text{Var}(x_{ik} | \xi_{ik}) = \theta_k = \theta \quad (10)$$

where $X_{ik} = (X_{ijk})$ equals the observed variables for the i th individual in the k th group of the population, with k in $\{1, 2, \dots, K\}$; $\xi_{ik} = (\xi_{ijk})$ are factor scores and $\tau_k = (\tau_{jk})$, $A_k = (\lambda_{jj'k})$ and $\theta_k = (\theta_{jj'k})$ are the intercepts, loadings and the matrix of residuals for the k th group, respectively (Rojas, Rojas and Brizuela (2018)).

Meredith (1993), gives a summary of the four primary levels of invariance analysis as involving; configural invariance, metric invariance, scalar invariance and strict invariance. The details of these follow in the sections below.

1.1.3.1. Configural Invariance

According to Brown and Moore (2012), configural invariance represents the fit of the presumed model for the designated groups, inconsiderate of the model's fit for each group in the population. Thus, we can treat configural as a baseline model where the interest is testing whether or not the same items are measuring our construct of interest in the same way across groups or times. Widaman and Reise (1997) further adds that configural invariance shows that the unobserved variables in the defined groups are identical, though not the same. In other words, it shows that members in different groups conceive a given construct in the same way (Milfont and Fischer, 2010). Therefore, testing for configural invariance is checking for the overall fit of the model.

1.1.3.2. Metric Invariance

Metric invariance assesses the similarity of the regression coefficients between groups ($\Lambda_1 = \Lambda_2 = \dots = \Lambda_k$). It builds upon configural invariance by checking whether factor loadings are equal across groups or times. Factor loadings reflect the degree to which differences among participant's responses to the items are due to differences among their levels of the underlying constructs of interest. According to Steenkamp and Baumgartner (1998), metric invariance implies that for any given value of j , the unobserved variables $\xi_{j1}, \xi_{j2}, \dots, \xi_{jk}$ have the same unit of measurement. In other words, achieving metric invariance means that observed variables and latent variables linearly correlate in every group (Chen, 2007; Hortensius, 2012). Hence, attaining metric invariance suggest that the construct has the same meaning to participants across time or groups.

1.1.3.3. Scalar Invariance

Scalar Invariance investigates the extent to which regression coefficients and intercepts are equal between groups ($\Lambda_1 = \Lambda_2 = \dots = \Lambda_k$ and $\tau_1 = \tau_2 = \dots = \tau_k$). It builds upon metric invariance by requiring that the item intercepts are also equal across groups or time. Scalar invariance gives the ability to justify mean comparisons across time or groups. Intercepts are the origin or starting value of the scale on which the factors are based. Thus, participants with the same value on the underlying construct should have equal values on the viewed items.

Hence, obtaining scalar invariance, denotes that for any observed value of j , the underlying variables $\xi_{j1}, \xi_{j2}, \dots, \xi_{jk}$ have the same origin and unit of measurement across groups (Meredith and Teresi, 2006). According to Cheung and Rensvold (2002), people who get similar scores on the latent variables tend to get similar scores in the observed variables, independent of a group membership.

Thus, this particular level of invariance is vital in psychological measurement since it indicates that groups do not bias test scores. Additionally, achieving this level of invariance implies that we can reliably compare scores of people belonging to different groups.

1.1.3.4. *Strict or Residual Invariance*

Strict or residual invariance is about the extent to which error terms are equal across the groups. Rojas, Rojas and Brizuela (2018) write that this level of invariance is achieved when the factor loadings, intercepts and errors are equal in all the groups ($\Lambda_1 = \Lambda_2 = \dots = \Lambda_k$, $\tau_1 = \tau_2 = \dots = \tau_k$ and $\Theta_1 = \Theta_2 = \dots = \Theta_k$). According to Chen (2007), strict invariance implies that variations in the observed variables exist due to variations in the underlying variables. However, Kaplan notes that the evaluation of strict invariance is not straightforward for the reasons of variation of error variance given different constructs under investigation, contrary to the assumption of independence between latent variables and errors when using confirmatory factor analysis (Kaplan, 2009).

Hence, for others like Widaman and Reise (1997), recommends the removal of strict invariance from the steps of invariance testings considering that the details it provides are usually not vital to the analysis of measurement invariance.

1.1.4. *Support for Using CFA with Dichotomous Items*

In the literature, there are currently two methods for establishing measurement equivalence. We find models based on structural equation modelling (e.g., CFA) and those based on IRT (Millsap and Everson, 1991). Many testing instruments measuring attitudes or personality traits usually uses ordered-categorical response format such as those on Likert scale but analysed using confirmatory factor analysis model while assuming the continuous character of those variables (Meade and Lautenschlager, 2004). Multiple studies are describing those models; among them are few studies are giving details about using CFA to test for measurement invariance with dichotomous data (Millsap and Everson, 1993). Hence, this study focuses on the usage of confirmatory factor analysis to assess measurement invariance with dichotomous items.

Even though various studies have analysed the effectiveness of the continuous linear models in measurement invariance, less attention has been paid on the usage of these models with ordered categorical data (French and Finch, 2006). On the other hand, Lubke and Muthén have warned about using the continuous model in categorical data (2004). Different models demand to consider some things before using them because the continuous and ordered models are not of the same measure type. According to Elosua, the latter includes thresholds, which are inherent to ordered dichotomous items; thus, using a continuous linear model to assess the measurement invariance across groups becomes complicated (2011). On the other hand, Larry (2013) notes that limitations are only theoretical, that is the usage of a linear model with ordered categorical data to assess measurement invariance is possible. Hence, today we can find some

studies using CFA with dichotomous items in the applied context to assess for measurement invariance. Besides, the use of continuous linear models to analyse ordered categorical data while violating the condition of multivariate normality has become common in the literature.

Despite the differences and complications, several papers have investigated the power and type I error associated with the CFA model in assessing measurement invariance with ordered data (Elosua and Wells, 2008; Stark, Chernyshenko, and Drasgow, 2006). For instance, Elosua and Wells (2008) compared the power related to the procedures that are designed explicitly for ordered items, e.g., IRT, to that obtained using the maximum likelihood estimation under CFA model. There were variations in the finding of the studies. For instance, Meade and Lautenschlager (2004) and Elosua and Wells (2008) reported more power associated with IRT models, mainly when invariance or differential item functioning (DIF) generated in the form of thresholds fails to hold. On the other hand, Stark, Chernyshenko, and Drasgow (2006) found that more power was associated with the CFA model when using the maximum likelihood estimation. One of the reasons for the differences in the findings was attributed to the model used to generate data. For example, the first two studies generated data using the IRT logistic models, and the last study generated data using the continuous linear model.

Further studies have examined the power in determining measurement invariance when using the CFA model with dichotomous items. For instance, French and Finch (2006) under a simulation study, examined the power and type I error associated with detecting measurement invariance on factors loadings with dichotomous items. While using the chi-square tests statistic as criteria for determining measurement invariance, the two found meagre power in detecting differences in factor loadings (that is, in the assessed conditions the power failed to reach 20%) (French and Finch, 2006). On the other hand, Finch and French used the IRT model to investigate factorial invariance in the unidimensional dichotomously scored test by generating DIF on the discrimination parameter (2007). Even though the model recommended for binary items was used the power did not reach 60%.

Thus, this study uses the model for continuous data to evaluate measurement invariance with dichotomous items. Using this model in our study contributes to other studies that have evaluated the applicability and possible deficiencies of using a linear model with categorical outcomes. The study contributes to the literature about the usage CFA model with dichotomous items and shows the utility of this model in assessing measurement invariance. In their simulation study, Elosua and Wells (2008) confirmed the adequacy of the CFA model with ordered categorical data in assessing measurement invariance. In summary, the two found that the linear model was good and promising (Elosua and Wells, 2008). Even though the researchers conducted the study using the simulated data, we say that the standard factor model can still successfully be used with ordered responses to analyse invariance between groups.

Further in the literature, observed is the fact that most of the studies on measurement invariance with dichotomous items are very technical and of a theoretic baseline. Besides, most of these studies (Muthén, and Asparouhov, 2002; Millsap and Yun-Tein, 2004; Meade and Lautenschlager, 2004), analyses were conducted on simulative data. Thus, making this study unique for it is conducted with the data from the real-world application collected using EGMA.

1.1.5. Measurement Invariance with Dichotomous Items

When using binary items, the threshold structures are considered for embedding the ordered-categorical variables in invariance testing with the linear model (Kim and Yoon, 2011). Although the ordered-categorical variables are discrete, researchers assume the assumption of continuity concerning the underlying latent variables (Molenaar and Borsboom, 2013). Following this assumption, one can use linear models to linearly relate the underlying latent variables to their respective observed scores (through factor loadings).

Additionally, we assume the assumption of the multivariate normal distribution on the latent variables. Hence, taking this assumption, we are also able to compute the correlations among factor scores and use them for confirmatory factor analysis with dichotomous items (Kim and Cohen, 1995). Muthén, and Asparouhov (2002) adds that responses on the underlying variables are then manifested in binary scores through thresholds. Cimetta et al. (2018), summarise it that we determine the distribution of observed binary responses on a particular variable from the distribution of latent response concerning a set of thresholds (v) parameters. Hence, it can be said that;

$$X_{ij} = c, \text{ if } V_{jc} < X^*_{ij} \leq V_j (c + 1) \quad (11)$$

Where, V_{jc} equals to the ordered binary responses of the j th item ($V_{j0} = -\infty$; $V_j (c + 1) = \infty$; and $C = 0, 1, \dots, C - 1$, where thresholds of an item equal the number of categories on an item minus one ($C - 1$). Hence, a dichotomous item has a single threshold if V_{j1} equals -1. Therefore, a score on a latent variable less than -1 equals to the score of zero and individual's scores greater than -1 equals to the score of one (Rojas, Rojas, and Brizuela, 2018). Hence, measurement invariance with binary items follows the invariance of thresholds, because the distribution of responses of binary items is determined by thresholds, ($V_{jk1}, V_{jk2}, V_{jkc}$) as well as of other factor model parameters ($\Lambda_k, \theta_{\varepsilon_k}, t_k$) as in continuous variable cases (Kim and Yoon, 2011).

With dichotomous items the basic equation for CFA is;

$$X^*_i = \tau + A\xi_i + \varepsilon_i \quad (12)$$

where all model parameters are according to the general formula of CFA (Rojas, Rojas and Brizuela, 2018). The only additional feature to the equation is a score for the underlying observed dichotomous variables for the i th individual ($X^*_i = (X^*_{ij})$) (Rojas, Rojas, and Brizuela, 2018). Where X^*_i represents a multivariate normal distribution for the variable with μ^* and Σ^* as

means and covariance matrixes respectively. X_j^* equals to the underlying variable to X_j defined as follows;

$$x_{ij} = \begin{cases} 1, & \text{if } X_{ij}^* \geq v_j \\ 0, & \text{if } X_{ij}^* < v_j \end{cases} \quad (13)$$

where v_j is a real number called a *threshold*. Cimetta et al. (2018) write that with the dichotomous item, the probability of success of individuals on each observed item is usually 1 ($X_{ij} = 1$). Therefore, using the CFA model assumptions, we can get the probability of success of X_{ij} given ξ_i :

$$P(x_{ij} = 1 | \xi_i) = \Phi(\tau_j + \Lambda_j \xi_i - v_j) \quad (14)$$

where Λ_j is the j th row of Λ and Φ are probability function derived from the standard normal distribution table (Rojas, Rojas, and Brizuela, 2018). Hence, one can get the probit of success probability by using the inverse of Φ in both sides of an equation, which is linearly associated $\tau_j + \Lambda_j \xi_i - v_j$. Thus, following this pragmatic approach, we can interpret the coefficients of the CFA model with ordinal variables (Sideridis, Tsaousis and Al-Harbi, 2015).

The three levels given by (Milfont and Fischer, 2010) are (i) configural invariance, (ii) metric invariance, and (iii) scalar invariance. Configural invariance measures the invariance of the structure of factors present on an instrument across different groups. Metric invariance, on the other hand, determines whether loadings of each indicator on factors vary across groups (Muthén and Muthén, 2009). Scalar invariance, as the third one examines whether various groups are conceiving and reacting to test items in the same way (Molenaar and Borsboom, 2013). However, Elwin et al. (2019) indicate that we cannot determine this level of invariance on measurement instruments and items that have more than two response options, that is, “exception of binary items that have an inadequate range of scores to assess for intercept or threshold differences meaningfully” (p.5). Hence, with dichotomous items, three invariance levels are meaningfully analysed. Even though the other level, strict invariance (e.g., the equivalence of residuals, factor loadings or thresholds), is often considered as optional (Elwin et al., 2019). Thus, leaving us with only two levels of configural and metric invariances with dichotomous items.

However, in Muthén and Muthén (2009), we find another view, the two aptly put it that with dichotomous, investigating measurement invariance requires that factor loadings and thresholds are same since both parameters follow the same item distribution curves. Hence, according to them, metric invariance is skipped when testing invariance with dichotomous items (Muthén and Muthén, 2009). That is, measurement invariance testing with dichotomous items involves three steps: configural invariance, strong invariance, and strict invariance. However, in this study, we adopt the view by Elwin et al. (2019). Therefore, we have assessed three measurement invariances in this paper: (1) configural invariance (model fit); (2) metric

invariance (equality of factor loadings or thresholds) and (3) strict invariance (equality of loadings or threshold or residuals).

1.1.6. Procedures for Measurement Invariance and Fit Indices

According to Elosua (2011), to evaluate measurement invariance, multiple group factor analysis is usually used, followed by computing chi-square tests for previous models and succeeding models, respectively. Then, we compute the differences between the chi-square value and degrees of freedom for the less previous models and the chi-square value and degrees of freedom for the succeeding models. Last but not least, we make conclusions for the invariance or non-invariance of the model across groups depending on the significance of difference values.

In addition to the model's parameters of regression intercepts, factor loadings, and errors produced, in case of continuous response data, thresholds are added in the case of dichotomous items (Elosua, 2011). The analysis follows the most common procedure, the forward approach of sequentially adding more model constraints (Dimitrov, 2010). In terms of model estimation, Elosua (2011) write that maximum likelihood estimation is appropriate when assessing invariance for continuous items while focusing primarily on factor loading and intercept. On the other hand, with dichotomous items, the most recommended method of estimation as recommended by Meredith (1993) is the weighted least squares mean, and variance adjusted (WLSMV) which is considered appropriate since it does not assume for multivariate normality of noncontinuous data.

Furthermore, there have several fit indices recommended to determine if measurement invariance is achieved across groups. Meade and Lautenschlager (2004), gave a list of fit indices that we can use like chis-square, CFI, and RMSEA. Meredith (1993) pointed out that the chi-square statistic is valuable since it allows us to draw inferences regarding model differences in the population. On the other hand, Barbara and Linda (2013) cautioned about using chi-square, especially in small sample size. The decision about which indices to use remains a subjective issue solely dependent on the researchers. Although there is no universal consensus concerning fit indices for evaluating model fit in invariance testing, a typical approach as recommended by Meade and Lautenschlager is to consider the (1) statistical significance difference of the chi-square tests, (2) change in comparative fit statistics, and (3) the degree of variations among model parameters (2004).

The above fit indices are most used when assessing measurement invariance irrespective of the data type. That is, both with continuous and dichotomous items, the above indices are traditionally used to evaluate the invariance of the models (Brown, 2006). After that, the next procedure involves comparing the fit indices of previous models to that of succeeding models for the assessment of invariance. In his simulation study, Chen (2007) recommended computing the differences between the model's respective root mean square error of approximation (RMSEA)

and the comparative fit index (CFI), respectively. Based on the differences calculated between the fit indices of the previous model and succeeding model, one can determine the invariance. Thus, to decide whether invariance or not, the following cut-off values are commended; sample sizes at least in all the groups ($n > 300$): $CFI > .90$ and $RMSEA < .05$ (Chen, 2007). These are the commonly used fit indices in providing researchers with an objective and evidence-based information of whether invariance exists.

1.2. Significance of the research

This study aims to assess measurement invariance with dichotomous. Today, methodologists have continued to address and clarify issues associated with invariance testing using different procedures and techniques. Thus, this study is of significance, for it demonstrates the assessment of measurement invariance with dichotomous items using factor analysis. Additionally, the study contributes to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items. Today, methodologists have continued to address and clarify the issues associated with invariance testing using different procedures and techniques. The current study adds to this ongoing endeavour by contributing to an important issue that has not adequately addressed in the literature, that is, assessing measurement invariance with dichotomous items.

Examination of the literature reveals that measurement invariance studies are specifically conducted to explore the generalizability of psychological constructs across groups and to determine the comparability of these constructs. Also, measurement invariance is tested in groups of the same culture for comparisons purposes across the subgroups of gender, age or ethnicity or different times (Hansson and Gustafsson, 2013). Measurement invariance is vital for construct validity in international assessment. Many studies have established the construct validity for many international assessment instruments like PISA, TIMSS. However, no studies have assessed the construct validity of EGMA. EGMA is a vital tool whose validity needs to be also established, just like other international assessment tools like PISA, TIMSS.

Furthermore, few studies are giving details about using CFA to test for measurement invariance with dichotomous data (Millsap and Everson, 1993). Most of the studies on measurement invariance are those focusing on instrument measuring psychological constructs using continuous or polytomous items (Uzun and Ogretmen, 2010; Uzun, 2010; Bofah and Hannula, 2015; Wendt, Kasper and Trendtel, 2017). For those focusing on the dichotomous items, most of them are very technical and of a theoretic baseline, mostly conducted on simulative data to demonstrate measurement invariance using factor analysis.

Different from the literature reviewed, this study assessed measurement invariance with dichotomous items using EGMA as an example. Since the inception of this instrument in 2008, no studies have investigated measurement invariance of this instrument. Unlike previous studies for

other international testing instruments that concentrated so much on continuous variables, this study focuses on dichotomous items using the ordinary linear CFA for a continuously measured variable. This study is also unique for it was conducted with the data from the real-world application collected using EGMA. In this context, the study is related to testing measurement invariance of EGMA 2018 across students' gender subgroups.

Furthermore, the current study contributes to this ongoing endeavour by contributing to an essential issue of assessing measurement invariance with dichotomous items using EGMA for the Zambian sample by students' gender. The research is of significant importance for it is set out to provide answers to the central question of whether the EGMA, as administered to a sample of Zambian students, assess the same latent constructs across groups of students. Addressing early maths ability is a critical component of the school readiness assessment literature, thus, by assessing measurement invariance of EGMA helps to answer whether the pattern of latent constructs remains the same when measuring invariance across gender subgroups. The construct of maths skills in Zambia is essential; as such, we need to understand how both male and female conceptualize them. Hence, the study not only contributes to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items but also help to establish the measurement invariance for the EGMA thereby allowing program evaluators, and researchers to assume that the interpretation of mathematics scores of students from different gender groups are appropriate across various groups.

1.3. Purpose of the research

The purpose of this study was to assess measurement invariance with dichotomous items using EGMA from the Zambian sample as evidence of construct validity.

1.4. Research Objectives

- A. To demonstrate the assessment of measurement invariance using Confirmatory Factor Analysis with dichotomous items.
- B. To establish the measurement invariance for the EGMA test for the Zambian sample by students' gender as evidence of construct validity.
- C. To assess measurement invariance for each factor of EGMA model for the Zambian sample by students' gender.
- D. To contribute to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items; make recommendations to program evaluators and researchers using EGMA in measuring students early mathematical ability.

1.5. Research Questions

- A. How may we implement confirmatory factor analysis to assess measurement invariance with dichotomous items?

B. Is there measurement invariance for the EGMA test for the Zambian sample by students' gender?

C. Is there measurement invariance for each factor of the EGMA test for the Zambian sample by students' gender?

1.6. Assumptions

It was assumed that all the samples for the national assessment survey were sampled as appropriately. It was further assumed that all the students sampled were those in grades 2 and 3 of primary schools in Zambia and that the sample consisted of all gender groups. It was also assumed that the data in the EGMA maintained by the Examinations Council of Zambia contains accurate and valid data on students that responded to EGMA during National Assessment Survey of 2018.

1.7. Delimitations

This research was only limited to primary school students from Zambia that were sampled during the 2018 National Assessment survey, as it used data from a student database consisting of grades 2 and 3 students. Also, the study focused on the EGMA in the National Assessment survey conducted by the Examinations Council of Zambia. Finally, the study performed an analysis based on students' performance in EGMA according to gender.

1.8. Chapter Summary

This study aims to assess measurement invariance with dichotomous. Today, methodologists have continued to address and clarify issues associated with invariance testing using different procedures and techniques. Thus, this study is of significance, for it demonstrates the assessment of measurement invariance with dichotomous items using factor analysis. Additionally, the study contributes to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items. This study is unique for it was conducted with the data from the real-world application collected using EGMA. In this context, the study is related to testing measurement invariance of EGMA 2018 across students' gender subgroups.

2. REVIEW OF LITERATURE

2.1. Introduction

The literature review section reviews studies related to measurement invariance for different large-scale assessments used in international testing such as PISA and TIMSS. Through the usage, the keywords like measurement invariance, assessment in education with large scale testing programs, structural equation modelling, and differential item functioning among others, the study was able to generate the potential literature for review that served as a guide for assessing measurement invariance with dichotomous items using the Early Grade Mathematics Assessment (EGMA) for the Zambian Sample as a case study.

2.2. Measurement Invariance Related Studies

In the literature, there are many studies related to measurement invariance for different groups in large-scale instruments used in international assessments such as PISA and TIMSS. The following includes some of the studies related to or have assessed measurement invariance for these measurement instruments by groups on different constructs, depending on the researcher's interest.

2.2.1. PISA Measurement Invariance Related Studies

Karakoç and Bökeoğlu (2018) examined the measurement invariance of mathematical and scientific literacy tests using the subtests of reading skills on PISA 2012 dataset. The two used MGCFA to assess measurement invariance across the constructs of mathematical and scientific literacy. Based on the MGCFA results, PISA 2012 mathematical and scientific literacy and subtests of reading skills demonstrated structural invariance while there was no metric invariance. Using a similar approach, Demir (2017) tested the measurement invariances between gender sub-groups hierarchically using PISA 2012 data for Turkish student. The study consisted of a total sample of 4,848 fifteen-year-old students from 170 schools and 12 territories in Turkey. Similarly, the study used the multi-group confirmatory factor analysis to assess the invariance between gender sub-groups Demir (2017). The study found a strong invariance between gender subgroups.

Furthermore, Kibrislioğlu and Arikan (2018) investigated measurement invariance of Science Self-Efficacy Scale in PISA using Multi-group Confirmatory Factor Analysis across year and gender. The two found that the science self-efficacy scale satisfied all stages of invariance by gender both in 2006 and 2015. From the comparative model, perspective is Scherer, Nilsen, and Jansen (2016), who examined different latent variable using the linear modelling approaches like confirmatory factor analysis, exploratory structural equation modelling, and bifactor modelling. The three used the linear approaches to describe individual perceptions concerning their factor structure, measurement invariance, and the relations to selected educational outcomes (achievement, self-concept, and motivation in mathematics). In their study, the three used PISA

2012 scale datasets for Australia, Canada, and the USA consisting of a total sample of (N=26,746 students). The study found support for the distinction between three factors of individual students' perceptions and evidence for full measurement invariance across countries for all modelling approaches (Scherer, Nilsen, and Jansen, 2016).

While the studies given above assessed measurement invariance for different constructs, what is common among them is that Multi-group Confirmatory Factor Analysis (MGCFA) was the primary model used to investigate measurement invariance. Further, the gender variable was the standard variable on which measurement invariance was tested. The study provides further support for our current study in investigating measurement invariance for gender variable on EGMA, which has not been explored elsewhere.

2.2.2. TIMSS Measurement Invariance Related Studies

Wendt, Kasper and Trendtel (2017) while investigating the extent to which the assumption of measurement invariance is affected by latent background variables, reported that equating procedures could be used for cross-national comparisons of the underlying indices when country-specific measurement models fit the data. The three concluded that cross-national comparisons of the coefficients of the GLMM should take into account the applied measurement model for scaling the HRL indices. This process could be achieved by, for example, adjusting the standard errors of the coefficients (Wendt, Kasper and Trendtel, 2017).

Karakoc, Ayan, Polatdemir and Uzun (2018) examined measurement invariance of TIMSS 2011 mathematics test in terms of different cultures. The study sample was composed of 1,987 fourth graders from Turkey, England, Japan and the USA. The three found support for the proposed model on all the countries under study; configural invariance was found in the first step but failed in the second step (Karakoç et al., 2018). Uzun (2010) investigated the factors that are related to students' science achievement while assessing the measurement invariance of these factors across gender. The data for this study came from TIMSS datasets for Turkish students with the sample size of 7841, and Structural Equation Modelling (SEM) was used as a model of analysis. The study found invariance problems across gender in TIMSS for Turkish students.

Furthermore, Bofah and Hannula (2015) while examining the relationships among motivational constructs using (TIMSS 2011 dataset) for eight-graders from five participating African countries, found empirical support for the multidimensionality of the construct and invariance for TIMSS 2011 motivational construct across cultures. Furthermore, a series of confirmatory factor analyses as cited in Bofah, and Hannula, (2015) revealed that there is a need to control method effects associated with negatively worded items in the measurement model. Bofah and Hannula further found support suggesting that in many cultures, responses to negatively worded items are systematically different (2015). For instance, they reported that factor structures and reliabilities (i.e., confidence and mathematics scales) were affected by

negatively worded items. Secondly, they found a strong relationship between the constructs, achievements and background variables such as parental education, gender and students' educational aspirations (Bofah and Hannula, 2015).

Different from the literature reviewed above, this study assessed measurement invariance with dichotomous items using EGMA as an example. However, since the inception of this instrument in 2008, no studies have investigated measurement invariance of this instrument. Unlike previous studies for other international testing instruments that concentrated so much on continuous variables, this study focuses on dichotomous items using the ordinary linear CFA for a continuously measured variable.

2.2.3. Measurement Invariance Studies with Dichotomous Items

Few studies have applied CFA for ordered-categorical variables to assess measurement invariance of the measurement instrument. Among them are Schroeders and Wilhelm (2011). The two in their study "Equivalence of Reading and Listening Comprehension Across Test Media" demonstrated how different levels of measurement invariance with dichotomous items might be assessed with Multiple Group CFA for ordered-categorical data (Schroeders and Wilhelm, 2011). Schroeders and Wilhelm used Mplus software to investigate measurement invariance while using weighted least squares mean, and variance adjusted (WLSMV) estimator for estimation (2011). The study reported the CFA provides a pleasant and effective way to test whether a particular measure is consistent across test media. Finally, they noted one advantage with an MGCFA approach and was "its ability to reflect more complex structures such as higher-order or nested factor models" (Schroeders and Wilhelm, 2011, p.866).

Similarly, is Rojas, Rojas and Brizuela (2018) in their study "The Use of Measurement Invariance with Dichotomous Variables as Evidence of Validity". Additionally, the trio performed their analysis using confirmatory factor analysis while using WLSMV as a method of estimation (Rojas, Rojas and Brizuela, 2018). The study performed all its analyses in R with the Lavaan package. The three noted one methodological limitation, and that was the impossibility of analysing the scalar invariance for dichotomous items when using Factor Analysis. Additionally, the absence of studies focusing on the criteria for investigating measurement invariance with dichotomous items using CFA was also echoed by (Rojas, Rojas and Brizuela, 2018). Hence, the present study fills this gap by assessing measurement invariance with dichotomous items using factor analysis. Additionally, the study contributes to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items. Thus, the study by (Rojas, Rojas and Brizuela, 2018) provided useful guidance for our study as we have performed all our analyses in R. We followed Rosseel (2012) steps for assessing measurement invariance in R.

2.1. Causes of Measurement Invariance in Dichotomous Items

Chen notes that scalar invariance signifies that subjects with the same underlying factor score will have similar responses on average and the same-origin for an item (2008). However, when testing for invariance with dichotomous items, observed items are assumed to be continuous an assumption which is inconceivable with binary items. Thus, this becomes a complication when using linear models to investigate for invariance. On the other hand, when the intercepts are noninvariant in the case of continuous items, it becomes the issue of *differential item functioning* commonly associated with the IRT model (Millsap and Meredith, 2007). Even though one is unable to check for scalar invariance with dichotomous items, intercept invariance is treated as threshold invariance (Millsap and Yun-Tein, 2004) with binary observed variables.

Therefore, as Givens and Jaccard (2002) put it that intercept or threshold invariance could arise due to the following reasons; (1) social desirability reasons or social norms, (2) different response styles and preferences to an item among the groups despite having the same ability level, and (3) different self-identify among groups, e.g., some students make different statements about themselves regarding their interest in maths (Chen, 2008). Furthermore, concerning configural invariance, Vandenberg and Lance (2000) note that some factors could also contribute to the lack of configural invariance; hence, they recommended to carefully scrutinise the content of the item to ascertain the reason for lack of invariance. According to Wirth and Edwards (2009), sometimes lack of invariance are due to differences that may be distinctively unique to a specific item or set of items. Hence, it is vital for the test developer and researchers to carefully scrutinise all items during analysis to identify the cause of poor fit.

Furthermore, Kline in their study revealed that the statistical conclusions made by comparing mean differences might differ or be invalid, depending on the type of analysis conducted (e.g., ANOVA vs multilevel CFA [MCFA]) and whether or not measurement invariance is achieved (2011). The trio study cautions against using means comparisons through ANOVA tests to assess measurement invariance (Kline, 2011). Thus, with more advanced models like CFA the following cut-off values ($n > 300$): CFI $> .010$ and RMSEA $< .015$ are recommended by (Chen, 2007). These are the commonly used fit indices in providing researchers with an objective and evidence-based data of whether invariance exists.

The other source for lack of invariance may be in the loadings, intercepts or thresholds themselves for they make different contributions to means, hence, inhibiting valid and comparability of test scores. This fact is in line with Millsap and Meredith (2007), who noted that for noninvariant measures, invariance occurs when item contribution to the overall score is equal across groups. Widaman, Ferrer, and Conger (2010) adds that differences could occur due to the following reasons if ; (1) different groups conceptualise items differently, (2) specific items are

more suitable to one group than the another, (3) test items were not well written, and (4) some groups respond to difficulty items more quickly than others.

2.2. Gender Differences in Mathematics Performance

In today's world high achievement in mathematics plays a vital role at all levels of students' development, e.g., schooling level, career stage, and at a professional level, hence, making gender differences in mathematics achievement a subject of extensive research (Hedges and Nowell, 1995). Research on gender differences, however, have not been consistent due to methodological problems, mainly used in studying gender difference in cognitive measures. Our study, while adopting a different methodology of structural equation modelling, focuses on the cognitive measures concerning equality of student performance on EGMA. Using this methodology is not the first and last, several studies have used and continue to use linear models to examine gender differences. For instance, several studies have tested measurement invariance across gender subgroups, like Hirschfeld and Brown (2009) investigated the structural relationships to achievement in reading comprehension across student sex, year level, and ethnicity in a sample of 3,506 students from New Zealand with Multiple Group Confirmatory Factor Analysis. Furthermore, Uzun and Ogretmen (2010) who investigated factors related to the fifteen-year-old students' science achievement using confirmatory analysis method to assess measurement invariance across gender in a Trend of International Mathematics and Science Study (TIMSS) 2009 sample from Turkey. Though the study met partial invariance across gender subgroups, none of the factors met the strict invariance.

As can be observed from the above literature, gender difference has been at the centre of research in educational measurement. Caporrimo grouped these studies as (a) those focusing on factors associated with gender differences in math achievement such as cognitive, affective, and social variables; (b) those studies using nonrepresentative samples to examine for the difference in performance; (c) those studies using a reasonably representative sample; and (d) meta-analytic studies (Caporrimo, 1990). Furthermore, some researchers have examined the potential determinants of gender differences in math achievement. An example, based on the first group is Battista, who examined the relationship between math achievement and cognitive abilities (Battista, 1990). His findings reveal that male students have more exceptional visual-spatial ability than female students do, and he concluded that this differences in the ability had been used by others to explain for gender differences in mathematics scores (Battista, 1990).

Concerning the affective and social factors, (Kaiser, 1993) examined the attitudes of male and females toward math by focusing on the social stereotyping items and other negative social expectancies that different students have towards mathematics and lastly examined the influence these social expectations have on female students' participation in maths-related courses. His study found that male students have more positive attitudes toward math and participate more

actively in math-related courses (Kaiser, 1993). Others like Cannor and Serbin have argued that even though gender differences in maths may be real, they may not be constant across age groups (Cannor and Serbin, 1985). That is, young girls typically show a slight advantage in math performance, but this advantage declines as they get older compared to boys.

Dickerson, McIntosh and Valente (2013), summarised two main types of arguments that give account for the gender gap in mathematics test scores. According to them, a first potential explanation is the biological one that “boys are genetically more able mathematically” (Dickerson, McIntosh and Valente, 2013, p.25). On the other hand, in her recent critical review, while refuting this claim, Spelke (2005) (in Dickerson, McIntosh and Valente, 2013, p.25) summarised three most prominent genetic arguments for genders differences as (i) male infants’ inherent more immense interest in objects versus people, (ii) gender-specific profiles in spatial and numerical abilities leading to a more excellent aptitude for mathematics, and (iii) higher dispersion of male than female performance in quantitative and spatial ability (Dickerson, McIntosh and Valente, 2013), as some of the reasons why more massive numbers of men have unusually high scores in mathematics.

However, Friedman and Romero (1989) focused on the second group of explanations which are linked (namely cultural or societal explanations) to the gender gap. Here Friedman and Romero (1989) tested two possibilities. The first possibility focused on the observable factors by focusing on those factors that could help to explain why mathematics tests scores vary systematically across the gender, with boys having observable characteristics that lead to higher test scores. For instance, he observed that some parents send their sons to better schools than their daughters and this could lead to differences in performance, or also he noted that teachers discriminate against girls in the allocation of schooling inputs (Friedman and Romero, 1989). The second possibility focused on the influence that the family, school and societal has on students, Friedman observed that girls do not have observable characteristics that are less conducive to performing well in maths tests than boys. However, he noted that family, school and societal influences affect both boys and girls differently in ways that we cannot directly observe (Friedman and Romero, 1989). Spencer et al., (1999) noted that some of these influences result in genuine or perceived gender-specific returns to maths skills or in stereotype threats about one’s group that are not captured by our regressors. Thus, these possibilities had implications on our study concerning how the interpretation of measurement invariance results after testing for invariance on EGMA.

Furthermore, researchers like (Howes, 2002; Sinnes, 2006) have added that since mathematics is a science subject, females in principle will also produce the same scientific knowledge as males if they put sufficient rigour in scientific inquiry. According to Ajai and Imoko (2015), Gender inequality in science, technology, and mathematics education have become

global; hence, fighting gender inequality in education has become a global topic. Besides, bridging the gender gap is seen as a significant way of achieving egalitarianism and enhancing human development. Therefore, there is need to give boys and girls the same opportunities and challenges. It is in light of the above that this study used structural equation modelling to assess measurement invariance of EGMA for the Zambian sample by students' gender. They were answering the question of whether measurement invariance of mathematical ability based on gender holds for the data set obtained as a result of EGMA 2018 for the Zambian sample. These findings make an essential contribution to the current debate on the gender difference in mathematic performance.

2.3. Chapter Summary

The literature examined in this study has shown that measurement invariance studies are explicitly conducted to assess the generalizability of psychological constructs between cultures and to determine the comparability of these constructs across groups or times. Besides, measurement invariance is also tested in groups with the same culture for comparisons across subgroups of gender, age or ethnic origin. However, explicit in the related literature, is that most of the studies on measurement invariance are those focusing on instrument measuring psychological constructs using continuous items. For those focusing on the dichotomous items, most of them are very technical and of a theoretic baseline, mostly conducted on simulative data to demonstrate measurement invariance using factor analysis. Also, most of them aimed to improve the technical framework of the concept of invariance and to define the best conditions to implement this technique. Thus, making this study unique for it was conducted with the data from the real-world application collected using EGMA. In this context, the study is related to testing measurement invariance of EGMA 2018 across students' gender subgroups.

3. METHODOLOGY

3.1. Introduction

This chapter explains the methodology that was followed by the author to carry out the research. In this chapter, the reader finds an explanation of the design of the research, research universe and sample, instrument and data analysis that were employed to conduct this research.

3.2. Research Design

The research is quantitative in methodology with descriptive and comparative design methods. The descriptive and comparative designs are well suited for this study since the researcher thought to assess whether the EGMA, as administered to a sample of Zambian students, assess the same latent constructs across student's gender subgroups. The study uses the descriptive and comparative designs, and methods to (1) assess measurement invariance using Confirmatory Factor Analysis with dichotomous items; (2) establish the measurement invariance for the EGMA model for the Zambian sample by students' gender; and (3) to assess measurement invariance for each factor of EGMA model for the Zambian sample by students' gender.

3.2. Participants

The participants for this instrument were primary school students in grade 2 (88%, n = 4122) and grade 3 (12%, n = 576), who answered the EGMA from different provinces of Zambia in 2018. The mean age of the students that participated in EGMA during the 2018 National Assessment Survey was 9.88 years. The total sample consisted of 4698 individuals (49% male, n = 2305; 51% female, n = 2393). The sample for this study came from an equally distributed sample representing most of the geographical areas of Zambia. All this was due to the efforts and design methodologies used by the ECZ team when administering the exam.

3.3. Instrument

We requested for access to the EGMA data before research see appendix 2. As pointed out in the introduction, EGMA is an assessment tool that measures early mathematics skill with an emphasis on number and operations, and it is composed of six subtests which captured children's mathematics competencies considered fundamental in early grade mathematics (RTI International, 2014). The test's subtests include number identification, number discrimination, recognition of number patterns, addition and subtraction and word problems. The reliabilities of these subtests ranged between 0.78 for oral counting and 0.86 for number discrimination (USAID, 2016). Furthermore, this research also investigated the reliabilities of the instrument based on the Zambian sample, answered by primary school students in 2018 National Assessment Survey. All the six subtests were timed during the assessment; however, some unexplained improvements were noted with three subtests due to the way the examiners administered the tests hence these were dropped from (3 subtests of addition, subtraction and word problems) the

2018 survey data test. Thus, the three subtests that remained in the entire data set for 2018 survey are number discrimination, recognition of number patterns and number identification, and were the basis for our analysis.

Table 1.
Summary Description of Subtests

Subtests	Number of items	Reliability
Number Identification	20	0.91
Number discrimination	10	0.80
Recognition of number patterns	10	0.78

As shown in Table 1 below the participants responded to 20 items of number identification, and ten items of recognition of number patterns and number discrimination, respectively. Before analysis, the study investigated the reliability of the three subtests, and their reliabilities were KR-20 0.91, 0.80 and 0.78, for number identification, number discrimination, and recognition of number patterns, respectively. Hence, we conducted our analysis analyzed on all three subtests in this study.

3.4. Data Analysis

Study analysis and model estimations were all based on a matrix of tetrachoric correlations since the items were dichotomous. The method of estimation with the CFA was WLSMV that was used throughout MCCFA analysis levels while using theta parameterization, recommended for use with dichotomous items (Millsap and Yun- Tein, 2004). WLSMV is idea because it uses polychoric correlations once data is specified as ordered-ordinal, and produces correlations between two dichotomous indicators to their latent variables that are assumed to have an underlying continuous latent distribution (Kosiol, 2010). Furthermore, WLS-based estimators are considered to be “distribution-free” and do not assume multivariate normality. We performed all the analyses with the lavaan package (Rosseel, 2012) using R. The study used conditions stated by Millsap and Yun-Tein (2004) to identify the fit of the underlying model. Furthermore, we tested the following three levels of measurement invariance recommended for dichotomous data cited by Elwin et al. (2019) in this study; (1) configural invariance; (2) metric invariance and (3) strict invariance.

For configural invariance and model fit, the study relied on the following criteria: comparative fit index (CFI) greater than or equal to .95, root mean square error of approximation (RMSEA) less than or equal to .05 and normed fit index (NFI) (high values greater than .95 as indicators of a sound- fitting model (Barbara and Linda, 2013). Furthermore, we also relied on the fit indices criterion as suggested by Brown (2006) to assess the overall fit of the model: the chi-square test, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI). He suggested that the RMSEA values of less than 0.05 represent a “close fit,” and

models with values above 0.08 should be rejected, and for CFI, a value of at least 0.95 is required to accept a model (Brown, 2006). The following levels of measurement invariance (metric and strict) were determined based on the criteria suggested by Chen (2007). He recommends an adequate sample size of (total $N > 300$) and a change of (≥ -0.01) in CFI, which would indicate noninvariance in the measurement (Chen, 2007).

Figure 1 below shows the CFA model to be estimated. The model factor structures are based on the confirmatory factor analysis for the model fits defined by three highly correlated latent variables: number identification, recognition and discrimination of the EGMA model in the mathematical context. The given model was first estimated, in samples defined by the gender of the student and then in both groups concurrently to check for configural invariance of the model. After that, the model was assessed for metric invariance and, latterly, strict invariance after meeting the restrictions of model fit required before the assessment of the subsequent levels of invariances. The study met the required sample size of, as Chen (2007) recommended ($N > 300$). Thus, our sample sizes were adequate in all the groups under study. The dataset contains responses from a total sample that consisted of 4698 primary school students. There were complete data for all 4698 participants with no missing values on all variables.

Table 2.
CFA Model Fit by Gender Subgroups

	df	χ^2	NFI	CFI	SEA
All Groups	737	20991.376	0.967	0.967	0.076
Male	737	9325.249	0.968	0.970	0.071
Female	737	12284.088	0.963	0.965	0.081

First, we computed MGCFA to assess whether there was a model fit for all three subtests in all gender group. Table 2 above shows the fit indices for all steps of the MGCFA analysis for all gender subgroups and each group, respectively. As shown in Table 2 above, fit indices indicated a better fit to the observed data, $\chi^2 (737) = 20991.376$, $NFI = 0.967$, $CFI = 0.967$, $RMSEA = 0.076$. Hence, we concluded that the model by all gender subgroups was a better fit and served as a basis for more stringent tests of measurement invariance.

In the second step, the study ran a separate MGCFA on both males and females, and overall fit indices indicated that each model fit better in both male ($\chi^2 (737) = 9325.249$, $NFI = 0.968$, $CFI = .970$, $RMSEA = .071$) and female ($\chi^2 (737) = 12284.088$, $NFI = 0.963$, $CFI = .965$, $RMSEA = .081$) respectively. Thus, figure 1 below by all gender subgroups formed the final model for our measurement invariance analysis. The model's items loadings were statistically different from zero ($p < .05$), and their completely standardised values were greater than .30 for a greater part of items as shown in Figure 1, 2 and 3 and Table 3 below. These results suggest that the observed

items were related to their corresponding underlying variables of number identification, number discrimination, and recognition of number patterns, respectively.

Table 3.

Summary of Model Fit with Estimates and Standardised Values

Subfactors	Items	All Subgroups		Male		Female	
		R^2	Std values	R^2	Std values	R^2	Std values
Number Identification	num1	0.50	0.71	0.62	0.79	0.41	0.64
	num2	0.60	0.77	0.60	0.78	0.60	0.77
	num3	0.30	0.55	0.37	0.61	0.25	0.50
	num4	0.71	0.84	0.74	0.86	0.68	0.82
	num5	0.82	0.90	0.84	0.92	0.80	0.89
	num6	0.77	0.87	0.78	0.89	0.75	0.87
	num7	0.80	0.89	0.79	0.89	0.81	0.90
	num8	0.82	0.90	0.82	0.91	0.81	0.90
	num9	0.82	0.91	0.80	0.89	0.85	0.92
	num10	0.90	0.95	0.89	0.94	0.91	0.95
	num11	0.91	0.96	0.91	0.96	0.92	0.96
	num12	0.60	0.78	0.62	0.79	0.59	0.77
	num13	0.91	0.95	0.91	0.95	0.91	0.95
	num14	0.87	0.93	0.85	0.92	0.90	0.95
	num15	0.92	0.96	0.91	0.96	0.93	0.96
	num16	0.37	0.61	0.40	0.63	0.35	0.59
	num17	0.26	0.51	0.29	0.54	0.23	0.48
	num18	0.57	0.76	0.53	0.73	0.62	0.78
	Num19	0.60	0.78	0.64	0.80	0.57	0.76
	Num20	0.54	0.74	0.56	0.75	0.53	0.73
Number Discrimination	quant1	0.48	0.69	0.47	0.69	0.48	0.70
	quant2	0.71	0.84	0.70	0.84	0.72	0.85
	quant3	0.67	0.82	0.73	0.85	0.62	0.79
	quant4	0.72	0.85	0.72	0.85	0.73	0.85
	quant5	0.66	0.81	0.67	0.82	0.65	0.80
	quant6	0.55	0.74	0.55	0.74	0.55	0.74
	quant7	0.47	0.68	0.48	0.70	0.45	0.67
	quant8	0.37	0.61	0.36	0.60	0.39	0.62
	quant9	0.48	0.69	0.52	0.72	0.44	0.66
	quant10	0.47	0.69	0.47	0.68	0.48	0.69
Recognition of Number Patterns	misnum1	0.51	0.71	0.49	0.70	0.53	0.73
	misnum2	0.88	0.94	0.90	0.95	0.87	0.93
	misnum3	0.57	0.75	0.56	0.75	0.57	0.76
	misnum4	0.86	0.93	0.89	0.94	0.83	0.91
	misnum5	0.46	0.68	0.47	0.69	0.44	0.67
	misnum6	0.66	0.81	0.67	0.82	0.64	0.80
	misnum7	0.48	0.70	0.47	0.68	0.51	0.71
	misnum8	0.55	0.74	0.55	0.74	0.55	0.74
	misnum9	0.38	0.61	0.33	0.58	0.43	0.66
	misnum10	0.26	0.51	0.34	0.58	0.20	0.45

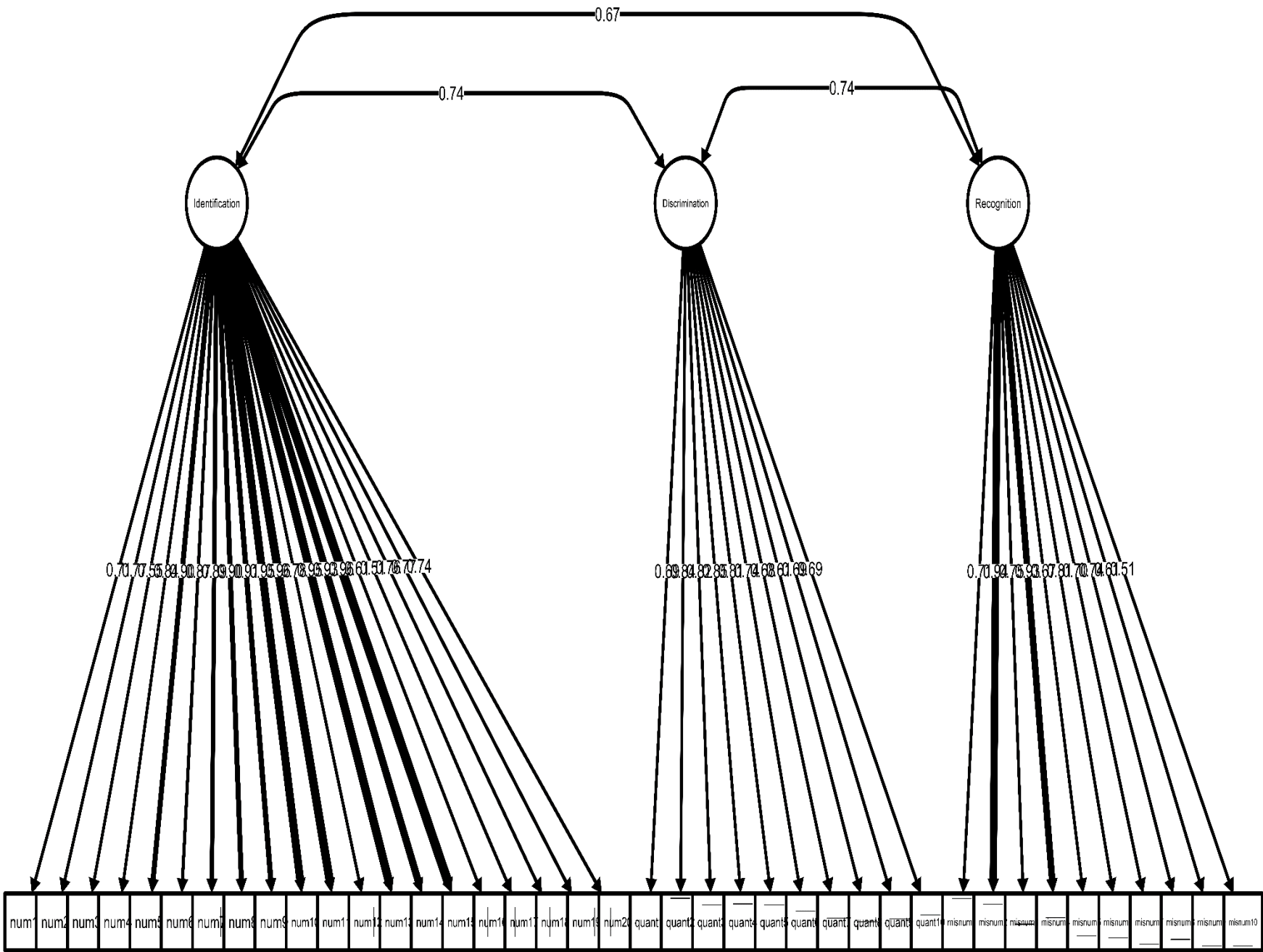


Figure 1. CFA Model with Model Estimates and Standardised Parameters by All Gender Group



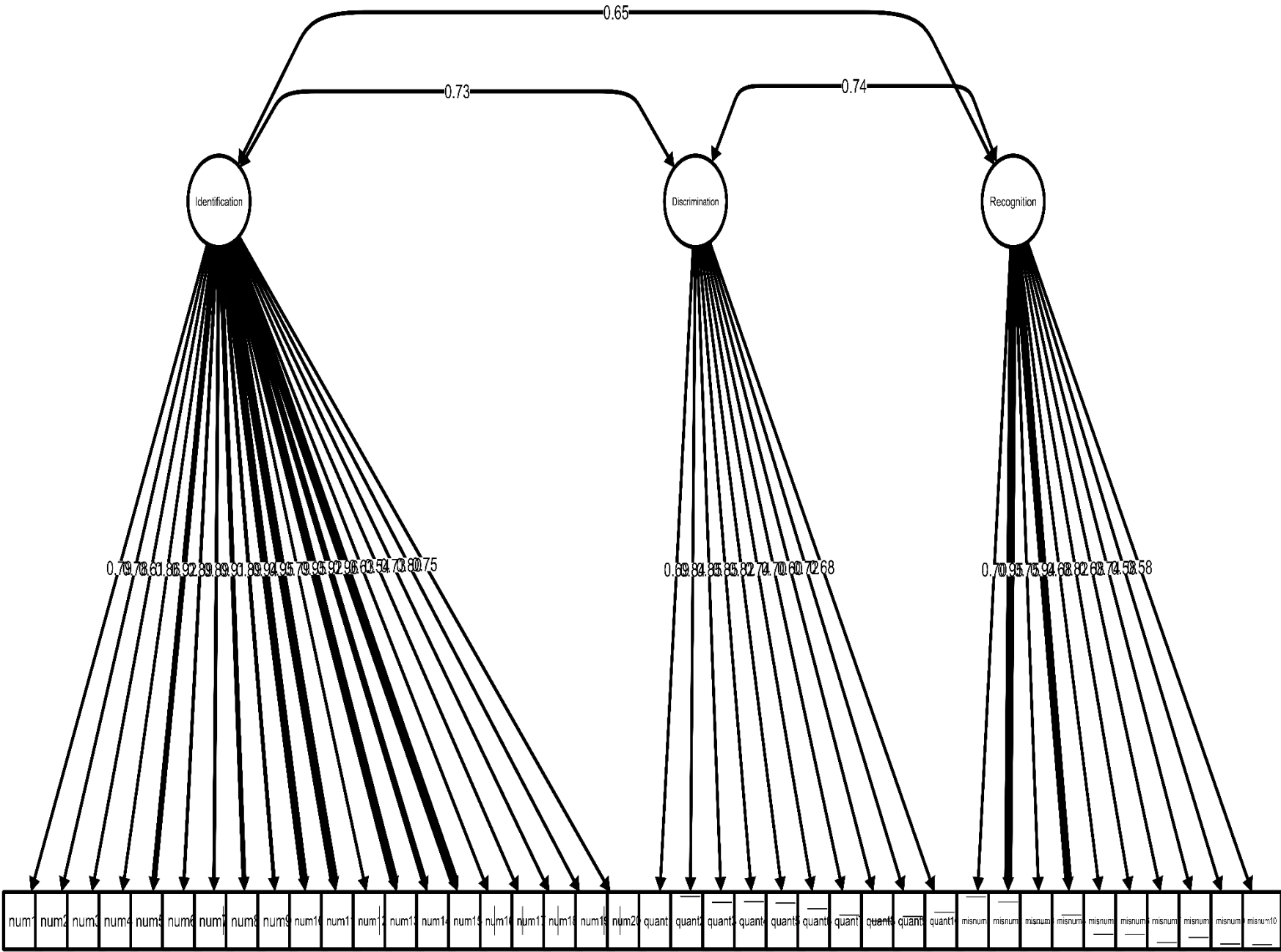


Figure 2. CFA Model with Model Estimates and Standardised Parameters by Male Group



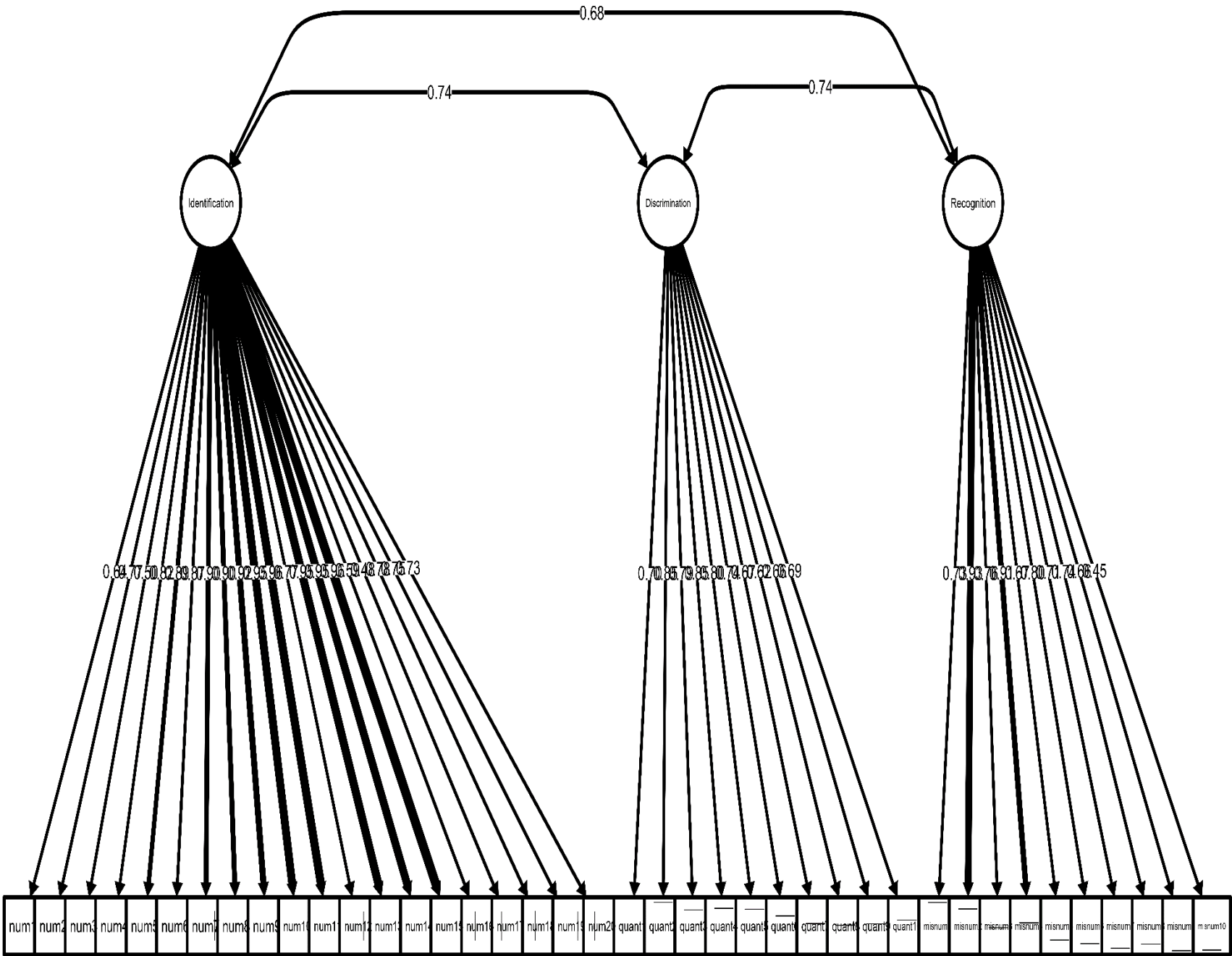


Figure 3. CFA Model with Model Estimates and Standardised Parameters by Female Group



4. RESULT

4.1. Introduction

This chapter presents the results of the study in line with the primary purpose of the research of assessing measurement invariance with dichotomous items using EGMA from the Zambian sample as a study case. The study demonstrated and assessed measurement invariance of EGMA for the Zambian sample by students' gender. The findings reported below are about invariances across gender subgroups. The study analysed all the measurement invariance tests with a three-step multi-group confirmatory factor analysis. In each step, the goodness of fit statistics obtained from analyses, are in Tables 2, 3, 4 and 5 for all model, and for each factor (that is number identification, number discrimination and recognition of number patterns), respectively. Last but not least is a summary table showing the presence or the absence of measurement invariance following the analysis conducted. The results are presented in line with the main objectives of the study of (1) to assess measurement invariance with dichotomous items using Confirmatory Factor Analysis, (2) to establish the measurement invariance for the EGMA model for the Zambian sample by students' gender, (3) to assess measurement invariance for each factor of EGMA model for the Zambian sample by students' gender, and (4) to contribute to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items; and make recommendations to program evaluators and researchers using EGMA in measuring students early mathematical ability.

4.2. Assessment of Measurement Invariance with Dichotomous Items

The first objective of the study was to demonstrate how measurement invariance with dichotomous items may be achieved with dichotomous items while using confirmatory factor analysis. Thus, overall, the study has demonstrated and assessed measurement invariance of EGMA for the Zambian sample by students' gender using multiple-group confirmatory factor analysis. According to literature reviewed, three measurement invariances are analysed with dichotomous items while using the linear model: (1) configural invariance, (2) metric invariance and (3) strict invariance. The analysis, as conducted in R using the syntax given in the appendix, confirms the findings of other studies. The first step involved testing for configural invariance. This step served as a basis for more stringent tests of measurement invariance (metric invariance and strict invariance). The results are shown below in the respective sections.

4.3. Measurement Invariance for the EGMA Model

The second objective of the study was to establish the measurement invariance for the EGMA model for the Zambian sample by students' gender. Table 4 below gives the results for the three-structure model of EGMA as defined by three factors of number identification, number discrimination, and recognition of number patterns.

Table 4.**Measurement Invariance for the Three-Factor Structure Model**

	df	χ^2	CFI	Δ CFI	RMSEA
Configural Invariance	1474	21609.337	0.967		0.076
Metric Invariance	1508	21718.867	0.967	0.00	0.076
Strict Invariance	1548	22062.054	0.967	0.00	0.075

In the first step, the structure model was fit for both males and females at the same time to assess configural invariance. As shown in Table 4 above, fit indices indicated a better fit to the observed data, χ^2 (1474) = 21609.337, CFI = 0.967 and RMSEA = 0.076. This served as a basis for the subsequent tests of measurement invariance.

Results of the second step, metric invariance (loadings or thresholds held equal across gender subgroups), indicated that there were no significant differences in the model in comparison to the configural invariance model when using a change of ($\geq$ - 0.01) in CFI as the indicator of significant change (Δ CFI = .00). The nonsignificant change suggested that the loadings or thresholds are invariant for gender subgroups across the three subtests of EGMA.

Last but not least, the study checked for strict invariance for all model (while holding loadings or thresholds or errors equal across gender subgroups), and the results indicated no significant differences in the model in comparison to the metric invariance model (Δ CFI = .00). These findings provided us with sufficient evidence to conclude that the all model structure, as shown in figure 1 above was consistent across males and females, and that configural, metric and strict invariances were achieved across gender group. These findings point out that the students' early literacy mathematical abilities in number identification, number discrimination and recognition of number patterns are significantly available in each gender sub-groups. However, these findings were considered not enough to conclude that the model was invariant conclusively; thus, further invariance analyses were conducted for each subtest as demonstrated below.

4.4. Measurement Invariance by each Factor of the EGMA Model

The third objective of the study was to assess measurement invariance by each factor of EGMA model for the Zambian sample by students' gender. The study defined three factors of number identification, number discrimination, and recognition of number patterns on the EGMA model. Thus, the study conducted further invariance analyses for each subtest, as demonstrated in the following tables results.

4.3.1. Measurement Invariance for Number Identification

The study assessed invariance for the subtest of number identification. Table 5 below gives the results of the analysis.

Table 5.**Measurement invariance for number identification**

	df	χ^2	CFI	Δ CFI	RMSEA
Configural Invariance	340	13244.417	0.969		0.127

Firstly, the study computed MGCFA to assess whether mathematic ability in Number Identification was invariant across males and females. Following similar steps as in the previous analysis, in the first step, the model was evaluated by fitting it to both male and female at the same time to determine for configural invariance. The fit indices together indicated a poor fit for the underlying construct of number identification to the viewed items across males and females (RMSEA = .127, CFI = .969). Even though the model met a CFI threshold of at least 0.95 required to accept a model, it failed to reach the other threshold of RMSEA of less than 0.08. Given this poor fit of the model of configural invariance, we concluded that all other levels of invariance failed to hold across all gender groups for number identification. Thus, the subtest of number identification failed to achieve configural, metric and strict invariances.

4.3.2. Measurement Invariance for Number Discrimination

The study further assessed invariance for the subtest of number discrimination. Table 6 below gives the results of the analysis.

Table 6.**Measurement Invariance for Number Discrimination**

	df	χ^2	CFI	Δ CFI	RMSEA
Configural Invariance	70	856.041	0.973		0.069
Metric Invariance	78	872.216	0.973	0.00	0.066
Strict Invariance	88	884.525	0.973	0.00	0.062

Secondly, we conducted a multiple-group CFA to check whether mathematic ability in Number Discrimination was invariant across males and females. Following similar steps as in the previous analysis, in the first step, the model was checked for configural invariance by fitting it to all gender sub-groups simultaneously. The fit indices in combination indicated a moderate fit of number discrimination model to the observed items across males and females (RMSEA = .069, CFI = .973). Given that the model met both thresholds of CFI of at least 0.95 and RMSEA of less than 0.08, required to accept a model, we observed that the model was indicative of a moderate optimal fit model. Thus, we can conclude that the number of discrimination construct achieved

this level of invariance in comparison by gender. Also, the differences in the fit indices between the configural invariance model and the metric model indicated that metric invariance is achieved ($\Delta\text{CFI} = .00$). Further, the results indicated that strict invariance was also achieved ($\Delta\text{CFI} = .00$).

4.3.3. Measurement Invariance for Recognition of Number Patterns

The study further assessed invariance for the subtest of recognition of number patterns. Table 7 below gives the results of the analysis.

Table 7.

Measurement Invariance for recognition of number patterns

	df	χ^2	CFI	ΔCFI	RMSEA
Configural Invariance	70	760.842	0.976		0.065
Metric Invariance	78	776.324	0.975	0.00	0.062
Strict Invariance	88	813.178	0.974	0.00	0.059

Thirdly, MGCFA was conducted to investigate whether mathematic ability for Recognition of Number Patterns was invariant across males and females. Following similar steps as in the previous analysis, in the first step, configural invariance was investigated by fitting a model that included all the gender sub-groups at the same time. The fit indices together indicated a slightly better fit for recognition of number patterns model to the observed items across males and females (RMSEA = .065, CFI = .976). Similarly, as to number discrimination model, the thresholds of CFI of at least 0.95 and RMSEA of less than 0.08, required to accept a model were met for this model. Thus, we concluded that the construct of recognition of number patterns achieved this level of invariance in comparison by gender. Also, the differences in the fit indices between the configural invariance model and the metric model pointed out that metric invariance was achieved ($\Delta\text{CFI} = -.001$). Further, the results indicated that strict invariance was also achieved ($\Delta\text{CFI} = -.001$).

Table 8.

The study's summary table for measurement invariance analysis

	Configural Invariance	Metric Invariance	Strict Invariance
Three Structure Model	Yes	Yes	Yes
Number Identification	No	No	No
Number Discrimination	Yes	Yes	Yes
Recognition of Number Patterns	Yes	Yes	Yes

Table 8 gives a summary of the results of all our invariance analyses conducted in this paper. As shown in the above, it appears that the three-structure model was a good fit in all the gender subgroups and passed all measurement invariance tests. However, results from additional analyses conducted revealed a poor fit to the observed items across males and females for

number identification model, as such, we concluded that this model failed to achieve configural, metric and strict invariances across gender subgroups. On the other hand, both number discrimination and recognition of number patterns indicated a better optimal fit to the observed data, and consequently, they passed all measurement invariance tests.

4.4. Implications of using MGCFA in Measurement Invariance with Dichotomous Items

The results presented in this paper are limited only to three levels of invariance testing (configural, metric and strict invariances). That is, scalar invariance was not analysed. According to the literature reviewed, e.g., Elwin and others write that scalar invariance we can only meaningfully determine with measurement tool having observed items with more than two responses (Elwin et al., 2019). By implication, this means using MGCFA in measurement invariance with dichotomous items scalar invariance cannot be analysed. Hence, in this study, it was not possible to determine scalar invariance with EGMA because it contains dichotomously scored questions that do not have adequate range of scores required to assess for invariance on intercept or thresholds across different groups. Despite this, the MGCFA provided us with a convenient and effective way to test whether EGMA was invariant across gender groups. Therefore, the study has produced results for three levels of invariances (configural, metric and strict).

4.5. Summary of the Findings

In sum, it appears that the three-structure model was a good fit in all the groups; the three-structure model passed all measurement invariance tests. Additionally, we performed specific analyses for each factor of the three-structure model. Following this analysis, we that number identification model represent a poor fit to the observed items across males and females; as such, the study concluded that it failed to achieve configural, metric and strict invariances. On the other hand, both number discrimination and recognition of number patterns indicated a better optimal fit to the observed data, and consequently, they passed all measurement invariance tests. These results show that a full invariance of the EGMA model as defined by three subtests (number identification, number discrimination and recognition of number patterns), is achieved at all levels analysed, that is, configural, metric and strict levels. By implication, this indicates that there is consistency in the way that these two sub-groups of gender conceptualize the psychological construct of number identification, number discrimination and recognition of number patterns.

However, individual analyses of each subtest revealed group differences at configural invariance level for the construct of number identification. The above findings suggested a lack of consistency in the factor structure for this construct across gender subgroups. On the other hand, on two subtests of number discrimination and recognition of number patterns, findings revealed

no statistically significant difference in the models for configural, metric and strict invariances. That is, at both levels of configural and metric invariance, measurement invariance was achieved. Following this finding, it was possible to conclude that there was consistency in the way that these two sample groups conceptualized the underlying constructs of number discrimination and recognition of number patterns.



5. DISCUSSION

5.1. Introduction

This chapter presents a summary of the study, purpose, and procedure. The study is a demonstration of how Confirmatory Factor Analysis with dichotomous items may be used to assess measurement invariance. Hence, to do this, the study used multiple-group confirmatory factor analysis to evaluate measurement invariance of EGMA for the Zambian sample by students' gender. This study has established measurement invariance for the EGMA and made some recommendations to program evaluators and researchers using EGMA in measuring students' early mathematical ability. All in all, the study contributes to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items. We have also recommended for future studies.

5.2. Study Discussion

Before 1995, children's math knowledge has been measured at the fourth-grade level through easy-to-administer written tests (Sitabkhan and Platas, 2018). However, these fourth grades and later skills are reported to be significantly influenced by earlier skills that are less amenable to written assessment. Claessens and Engel (2013) investigations found that early mathematical development has a strong influence on children's academic achievement across domains, including literacy. Because of the magnitude of these findings and the implications at both individual and societal levels, measuring of informal mathematical skills has become a priority for governments in low and middle-income countries. One such instrument in use in low- and middle-income countries is the EGMA. EGMA results are useful for many developing countries to assess their educational systems and identify educational policies. The instrument also allows for comparisons across many groups and nations.

Despite this full acceptance, there has been a lack of sub-group measurement invariance assessment regarding this instrument. According to Gomez and Stavropoulos (2018), lack of measurement invariance information for any measurement instrument, and in this case EGMA, is problematic because users cannot reliably compare for differences in performance for various groups on the test. Thus, to help establish the invariance of EGMA and contribute to its comparability, the present work applied multiple groups confirmatory factor with dichotomous items to investigate the invariance of EGMA for the Zambian sample by students' gender.

First of all, we analysed measurement invariance across students' gender sub-groups with a three-step multi-group confirmatory factor analysis. In each step, the study obtained the goodness of fit statistics. The three steps are in line with many previous studies that have applied multiple groups confirmatory factor analysis with dichotomous items (Elwin et al., 2019; Rojas, Rojas and Brizuela, 2018); Schroeders and Wilhelm, 2011). The three measurement invariances assessed in this paper are (1) configural invariance; (2) metric invariance and (3) strict

invariance. The study did not compute scalar invariance because according to Elwin et al. (2019), scalar invariance we can only meaningfully assess it in measurement tools that have items capturing more than two response options from test-takers (e.g., continuous items). That is, it was not possible to determine scalar invariance with EGMA because it contains dichotomous scored items that do not provide an adequate range to scores required to assess for invariance of intercept or threshold across different groups. Furthermore, one of the requirements in the identification of invariance is that of allowing some thresholds to vary across populations, but this is not possible when the measured variables are dichotomous (Millsap, 2011).

However, an interesting observation in literature is whether metric or scalar invariance to use when applying multiple-group CFAs on dichotomous items. For instance, Muthén and Muthén (2009) used scalar in place of metric invariance. We found this usage confusing because scalar invariance uses continuous items to assess equality of factors loadings and intercepts. However, in the case of dichotomous items, the analysis of scalar invariance usually involves evaluation of equality of thresholds, rather than that of intercepts (Kosiol, 2010). Assessing invariance requires holding both loadings and thresholds equal because both parameters are on the same item distribution curves. Besides, scalar invariance builds on metric invariance. However, its evaluation is impossible since we use thresholds in place of intercepts. Hence, we say that what we evaluate in the second step (metric invariance) is the same thing that we would evaluate in the third step (scalar invariance). Thus, to avoid confusion, metric invariance should be used.

In summary, the findings of the confirmatory factor analysis confirmed a three-structure model defined by three highly correlated latent variables: number identification, number discrimination and recognition of number patterns of the EGMA model in the mathematical context. The presented model revealed no group differences at any level of invariance testing. Here in the study, full invariance of the EGMA model as defined by three subtests (number identification, number discrimination and recognition of number patterns), was observed at all levels analysed, that is, configural, metric and strict levels. By implication, this shows that there was consistency in the way male and female conceptualized the psychological construct of number identification, number discrimination and recognition of number patterns, as well as the way that all observed items associated well with their respective latent constructs.

On the contrary, following individual analyses of each subtest, the findings revealed group differences at configural invariance level for the construct of number identification. The findings above suggested a lack of consistency in the factor structure for this construct across gender subgroups. Specifically, this suggests that there are differences in the way male and females conceptualise number identification. Thus, the findings present a caution regarding making comparisons with number identification subtest across different gender groups. Additionally,

these findings are particularly imperative given the increasing research on gender differences in educational measurement. Researchers like (Sinnes, 2006) have examined the potential determinants of gender differences in mathematics among men and females. For instance, he investigated the relationship between math achievement and cognitive abilities, and his findings revealed that male students have more exceptional visual-spatial ability than female students do (Sinnes, 2006). Battista (1990), further adds that this disparity in the ability can account for the observed gender differences in math scores favouring male students.

Interestingly, with the other two subtests (number discrimination and recognition of number patterns), findings revealed no statistically significant difference in the models for configural, metric and strict invariances. That is, the study found the two subtests to be invariant across the groups, implying that there was consistency in the way that the two groups conceptualized the underlying constructs of number discrimination and recognition of number patterns, as well as the way that all items associate to the latent construct. The findings have an important implication for the construct of maths in Zambia for they have shown that there no much differences in how young female and male learners conceptualize maths items. Hence, it might relatively be safer to compare scores of EGMA for matched samples of gender sub-group from primary school students in Zambia on number discrimination and recognition of number patterns. Additionally, the study results build on the suggestions by Cannor and Serbin, who reported that gender differences in mathematics achievement might not be constant across gender subgroups (1985).

Furthermore, the study revealed some interesting findings regarding model fit, the fit indices indicated a better fit model in male group (RMSEA = .07, CFI = .97) than in female group (RMSEA = .08, CFI = .97). Considering these differences, we can say there could be some differences in the meaning of some test scores across gender groups. Thus, if we are to consider scalar invariance, these two gender sub-groups present some differences that could be affecting item scoring. However, this was impossible to identify without the analysis of scalar invariance that enables one to examine the source for items showing different levels of ability. Hence, given the dichotomous nature of EGMA items, it was not possible to assess for scalar invariance in the present study.

Following the findings of the study, we found the instrument to be invariant on both the factor structure and the item loadings across gender sub-groups for the three-structure model and individual factors of number discrimination and recognition of number patterns. Thus, by implication, the results suggest that it might relatively be safer to compare scores of EGMA for matched samples of gender sub-group from primary school students in Zambia, even though screening of the items is required primarily for number identification. However, we have made these conclusions in exception to the construct of number identification which the study found to

be noninvariant. Hence, we recommend for extreme cautiousness when comparing group differences in number identification items.

5.3. Strength and Limitations of the study

The present study has not only contributed to a conceptual discussion on the application of confirmatory factor analysis on measurement invariance with dichotomous items. However, it has also established the measurement invariance for the EGMA thereby allowing program evaluators, and researchers to assume that the interpretation of mathematic scores of students from different demographic backgrounds and genders are appropriate across various groups.

Thus, checking for invariance and acknowledging the sources of noninvariance contributes significantly to future research in the assessment of early mathematical literacies. Secondly, these findings provide evidence for the comparability of binary scores of EGMA across gender sub-groups on two subtests of number discrimination and recognition of number patterns. However, users of EGMA should consider making a significant improvement to the construct of number identification where the study did not find configural invariance. Finally, more studies with large samples are required to enable the generalizability of the findings to the broader population of primary school students in Zambia across gender sub-groups and give way to other future studies to explore invariance of this instrument similarly.

Despite the uniqueness of the study and the soundness mentioned above, the present study is not without some limitations. Since the study uses student participants from Zambia, the results may not be generalizable to similar populations in other countries. Besides, the variable of interest in the study was gender. However, this variable alone may not be fully comprehensive as other demographic variables may exist on which student's performance varies in EGMA. Furthermore, another limitation is concerning data; EGMA consists of 6 subtests: number identification, number discrimination, recognition of number patterns, addition and subtraction and word problems (RTI International, 2014). During the 2018 assessment survey, students were assessed on all the six subtests. However, some unexplained improvements were noted with three subtests due to the way the examiners administered the tests. Hence, these were dropped from (3 subtests of addition, subtraction and word problems) the 2018 survey data test.

Thus, the study was limited only to the three subtests that remained in the entire data set for 2018 survey; number identification, number discrimination, and recognition of number patterns. In turn, this had an influence on our measurement invariance testing for we relied just on three subtests instead of 6. Consequently, measurement invariance testing did not apply to the subtests of addition, subtraction and word problems. Hence, this invites further research to explore the invariance of the other three subtests.

5.4. Recommendations

The study recommends for careful screening of the test items, especially for number identification. Further, the study invites future studies to aim at remedying the limitations of this study by screening for invariance using other models like IRT. Additionally, future studies should test for the invariance of the instrument by including other demographic variables such as grade, region, and language of instruction on which student's performance may vary and determine whether these variables influence the equivalence of the testing instrument. Investigating measurement invariance of EGMA by these demographic variables would not only provide more understanding of the applicability of EGMA concerning these variables but also illustrate the extent at which variations on these variables impact the testing of EGMA. Furthermore, future studies should be done with all the six subtests of EGMA. Future studies should also use simulative polytomous data in comparison to dichotomous items to check for scalar invariance. We also recommend checking for the similarities of fit indices those produced by R with those produced by AMOS and Lisrel.

5.5. Conclusion

Measurement invariance is a vital part of construct validity. However, it continues to be a methodological challenge in invariance studies due to different data type. Different instruments collect different data type; that is, we find instruments collecting examines data in dichotomous form (EGMA) and also those collecting in a continuous form (PISA, TIMMS). This fact makes it challenging for a researcher interested in invariance studies. The methodologists have continued to address and clarify on issues associated with invariance testing using different procedures and techniques. The present study contributes to this ongoing endeavour, that is, it has assessed and established measurement invariance for EGMA test for the Zambian sample by students' gender as evidence of construct validity. Last but not least, the study results, analysis and discussions contribute to other invariance studies that have used Multiple Group Confirmatory Factor Analysis with dichotomous items.

REFERENCES

- [1] Acar, T. (2008). *Maddenin farklı fonksiyonlaşmasını belirlemede kullanılan genelleştirilmiş aşamalı doğrusal modelleme, lojistik regresyon ve olabilirlik oranı tekniklerinin karşılaştırılması (Yayımlanmamış doktora tezi)*. Hacettepe Üniversitesi, Ankara.
- [2] Ajai, J.T. and Imoko, I.I. (2015). Gender differences in mathematics achievement and retention scores: A case of the problem-based learning method. *International Journal of Research in Education and Science (IJRES)*, 1(1), 45- 50.
- [3] Alatlı, B.K., and Bökeoğlu, Ö.Ç. (2018). Investigation of Measurement Invariance of Literacy Tests In The Programme For International Student Assessment (PISA-2012). *Elementary Education Online*, 2018; 17(2): pp. 1096-1115.
- [4] Ayala, R. J. de. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.
- [5] Ayvalli, M., and Biçak, B. (2018). An Investigation into the Measurement Invariance of Pisa 2012 Mathematical Literacy Test. *European Journal of Education Studies*, 6(4), 311-360.
- [6] Barbara, G.T., and Linda, S.F. (2013). *Using Multivariate Statistics (6th Ed.)*. New York: Pearson.
- [7] Başusta, N. B. U. (2010). Ölçme eşdeğerliği. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 1(2), 58-64.
- [8] Battista, M. T. (1990). Spatial Visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education*, 21, 47-60.
- [9] Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107(2), 238-246.
- [10] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In E. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores (pp. 397-479)*. Reading, MA: Addison Wesley.
- [11] Bofah, and Hannula, M. S. (2015). TIMSS data in an African comparative perspective: Investigating the factors influencing achievement in mathematics and their psychometric properties. *Large-Scale Assessments in Education*, 3(1), 300-340.
- [12] Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford, England: John Wiley.
- [13] Booth, J.L., and Siegler, R.S. (2008). Numerical magnitude representation influence arithmetic learning. *Child Dev.* 79, 1016-1031.
- [14] Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.
- [15] Brown, T. A., and Moore, M. T. (2012). "Confirmatory factor analysis," in *Handbook of Structural Equation Modelling*, ed. R. H. Hoyle. New York, NY: The Guilford Press.
- [16] Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., and Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology*, 40(5), 415-436.
- [17] Caporrimo, R. (1990). *Gender, confidence, math: why aren't girls "where the boys are"?* Boston, MA: Annual Meeting of the American Psychological Association. ERIC Document Reproduction Service No. ED 334 074.
- [18] Chen, F. F. (2007). The sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modelling*, 14, 464-504.
- [19] Cheung, G. W. and Rensvold, R. B. (2000). Assessing final and acquiescence response sets in cross-cultural research using structural equations modelling. *Journal of Cross-Cultural Psychology*, 31(2) 187-212.
- [20] Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling*, 9(2), 233-255.
- [21] Cimetta, A. D., Marx, R. W., Yaden, D. B., Alkhadim, G. S., and Cutshaw, C. A. (2018). Latent variable structure and measurement invariance of a modified early literacy assessment. *International Journal of School and Educational Psychology*, 6(3), 208-222.

- [22] Claessens, A., and Engel, M. (2013). *How Important is Where You Start? Early Mathematics Knowledge and Later School Success*. Teachers College Record.
- [23] Demir, E. (2017). Testing measurement invariance of the students' affective characteristics model across gender subgroups. *Educational Sciences: Theory & Practice*, 17(1), 5–20.
- [24] Dickerson, A., McIntosh, S., and Valente, C. (2013). *Do the Maths: An Analysis of the Gender Gap in Mathematics in Africa*. UK: University of Bristol.
- [25] Dimitrov, D. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development*, 43(2), 121-149.
- [26] Drasgow, F. (2006). Polychoric and Polyserial Correlations. *Applied Psychological Measurement*, 19, 143–165.
- [27] Edwards, M. C., and Wirth, R. J. (2009). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58–79.
- [28] Elosua, P. (2011). Assessing Measurement Equivalence in Ordered-Categorical Data. *Psychological*, 32, 403-421.
- [29] Elosua, P., and Wells, C. (July 2008). *A Comparison of MACS and the IRT Likelihood Ratio Test for Identifying DIF*. Paper presented at III European Congress of Methodology, Oviedo.
- [30] Ergül, D. (2016). Testing Measurement Invariance of the Students Affective Characteristics Model across Gender Sub-Groups. *Educational Sciences: Theory & Practice*, 17(1), 47–62.
- [31] French, B. F., and Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modelling – A Multidisciplinary Journal*, 13(3), 378–402.
- [32] Friedman, A., and Romero, L.A. (1989). Functional differential equations for the determination of the viscosity function in a rheometer. *Archive for Rational Mechanics and Analysis*, 107, 85–97.
- [33] Gerrit Hirschfeld, G., and Brachel, R.V. (2014). Multiple-Group Confirmatory Factor Analysis in R. A Tutorial in Measurement Invariance with Continuous and Ordinal Indicators. *The Journal of applied psychology*, 8, 230-245.
- [34] Glockner-Rist, A., and Hoijsink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modelling. *Structural Equation Modeling*, 10, 544–565.
- [35] Gomez, R., and Stavropoulos, V. (2018). Malaysian parent ratings of the strengths and difficulties questionnaire factor structure and measurement invariance across language versions, gender, informants, and race. *Assessment DOI: 10.1177/ 1073191118787284*
- [36] Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- [37] Hansson, Å., and Gustafsson, J. E. (2013). Measurement Invariance of Socioeconomic Status Across Migrational Background. *Scandinavian Journal of Educational Research*, 57(2), 148–166.
- [38] Hedges, L.V., and Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *The British Journal of Educational Psychology*, 269, 41-5.
- [39] Hirschfeld, G. H. F., and Brown, G. T. L. (2009). Students' conceptions of assessment factorial and structural invariance of the SCoA across sex, age, and ethnicity. *European Journal of Psychological Assessment*, 25(1), 30–38.
- [40] Hortensius, L. (2012). *Project for Introduction to Multivariate Statistics: Measurement Invariance*. Retrieved on 4th December 2019 from <https://pdfs.semanticscholar.org/>.
- [41] Kaiser, A. P. (1993). Parent-implemented language intervention: An environmental system perspective. Kaiser, A. P., and Gray, D. B. (Eds.). *Enhancing children's communication: Research foundations for intervention*, 2, 63-84. Baltimore, MD: Brookes.
- [42] Kaplan, D. (2009). *Structural Equation Modelling: Foundations and Extensions* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

- [43] Karakoç Alatlı, B., and Çokluk Bökeoğlu, Ö. (2018). Investigation of Measurement Invariance of Literacy Tests in the Programme for International Student Assessment (PISA-2012). *Elementary Education Online*, 17(2), 1096–1115.
- [44] Karakoc Alatli, B., Ayan, C., Polatdemir, B., and Uzun, G. (2017). Examination of the TIMSS 2011 Fourth Grade Mathematics Test in Terms of Cross-Cultural Measurement Invariance. *Eurasian Journal of Educational Research*, 16(66), 1–30.
- [45] Kibrislioğlu Uysal, N., and Akin Arikan, Ç. (2018). Measurement Invariance of Science Self-Efficacy Scale in PISA. *International Journal of Assessment Tools in Education*, 5(2), 325–338.
- [46] Kim, E. S., and Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modelling*, 18(2), 212–228.
- [47] Kim, S.-H., and Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on the detection of differential item functioning. *Applied Measurement in Education*, 8, 291–312.
- [48] Kline, R. B. (2011). *Principles and practices of structural equation modelling*. New York, NY: The Guilford Press.
- [49] Koziol, N., and Bovaird, J. A. (2009). *Using experimental designs to evoke construct validity evidence*. Poster presented at the annual meeting of the American Psychological Association, Toronto, Canada
- [50] Koziol, N.A. (2010). *Evaluating Measurement Invariance with Censored Ordinal Data: A Monte Carlo Comparison of Alternative Model Estimators and Scales of Measurement*. Theses and Dissertations from the College of Education and Human Sciences. University of Nebraska – Lincoln
- [51] Larry, R. P. (2013). *Psychometric Methods: Theory into Practice*. New York: The Guilford Press
- [52] Levine, A. (2005). *Educating School Leaders*. The Education Schools Project. Retrieved on 20th January 2020 from <http://www.edschools.org/pdf/Final313.pdf>.
- [53] Lubke, G. H., and Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modelling*, 11, 514–534.
- [54] Meade, A. W., and Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modelling*, 11(1), 60–72.
- [55] Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, 115, 300–307.
- [56] Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrical*, 58(4), 525–543.
- [57] Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, 44(11), S69–S77.
- [58] Milfont, T. L., and Fischer, R. (2010). Testing measurement invariance across groups: applications in cross-cultural research. *International Journal of Psychological Research*, 3, 111–121.
- [59] Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, 72(4), 461–473
- [60] Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. R. Cudeck and R. C. MacCallum (Eds.). *Factor analysis at 100: Historical developments and future directions* (p. 131–152). Lawrence Erlbaum Associates Publishers.
- [61] Millsap, R. E., and Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research*, 26(3), 479–497.
- [62] Millsap, R. E., and Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17(4), 297–334.
- [63] Millsap, R. E., and Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9(1), 93–115
- [64] Millsap, R. E., and Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39, 479–515.

- [65] Millsap, R.E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Taylor and Francis Group
- [66] Molenaar, D., and Borsboom, D. (2013). The formalization of fairness: Issues in testing for measurement invariance using subtest scores. *Educational Research and Evaluation*, 19(2-3), 223-244.
- [67] Molennar, D., and Borsboom, D. (2015). *Psychometrics*. Amsterdam: University of Amsterdam.
- [68] Murphy, R. K., and Davidshofer, O. C. (2004). *Psychological Testing- Principles and Applications (6th Ed.)*. New Jersey: Pearson Education.
- [69] Muthén, B. O., and Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modelling in Mplus (Mplus Web Notes: No. 4)*. Retrieved 4th May 2020 from <http://www.statmodel.com/examples/webnote.shtml>
- [70] Muthén, L. K., and Muthén, B. O. (2009). *Mplus (Version 5.21) [Computer software]*. Los Angeles, CA: Muthén & Muthén
- [71] Quirk, M., Mayworm, A., Furlong, M. J., Grimm, R., and Rebelez, J. (2015). Dimensionality and Measurement Invariance of a School Readiness Screener by Gender and Parent Education Levels. *International Journal of School and Educational Psychology*, 3(3), 167-177
- [72] R Core Team. (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna Austria. Retrieved 4th May 2020 from <http://www.R-project.org/>
- [73] Rojas, L., Rojas, G., and Brizuela, A. (2018). *The Use of Measurement Invariance with Dichotomous Variables as Evidence of Validity*. Instituto de Investigaciones Psicológicas, Universidad de Costa Rica.
- [74] Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modelling. *Journal of Statistical Software*, Volume 48, 78-90.
- [75] RTI International. (2014). *Early Grade Mathematics Assessment (EGMA) Toolkit, (March)*, 1-11.
- [76] Sax, L. J. (1994). Predicting gender and major-field differences in mathematical self-concept during college. *Journal of Women and Minorities in Science and Engineering*, 1, 291-307.
- [77] Scherer, R., Nilsen, T., and Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, 7, 110.
- [78] Schroeders, U., and Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, 71, 849-869.
- [79] Sideridis1, G.D., Tsaousis, I., and Khaleel A. Al-harbi. (2015). Multi-Population Invariance with Dichotomous Measures: Combining Multi-Group and MIMIC Methodologies in Evaluating the General Aptitude Test in the Arabic Language. *Journal of Psychoeducational Assessment*, 1-7.
- [80] Sinnes, A. (2014). Understanding Student Participation and Choice in Science and Technology Education. *Springer*, 17-19.
- [81] Sitabkhan, Y., and Platas, L. M. (2018). Capturing Children's Mathematical Knowledge: An Assessment Framework, 5(3), 106-124.
- [82] Spearman, C. (1904). "General intelligence" objectively determined and measured—*American Journal of Psychology*, 15, 201-293.
- [83] Spearman, C. E. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.
- [84] Spelke, E. S. (2005). Sex Differences in Intrinsic Aptitude for Mathematics and Science? A Critical Review. *American Psychologist*, 60(9), 950-958.
- [85] Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35(1), 4-28.
- [86] Stark, S., Chernyshenko, O. S., Drasgow, F., and Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, 91(1), 25-39.
- [87] Steenkamp, J.B., and Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research*, 25:78-90.

- [88] Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press.
- [89] Ulrich Schroeders, U., and Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension Across Test Media. *Educational and Psychological Measurement*, 5, 849-869.
- [90] USAID. (2016). *Early Grade Reading and Mathematics Assessment in the Republic of Macedonia: Study Report*.
- [91] Uzun, B., and Ogretmen, T. (2010). Assessing the measurement invariance of factors that are related to students' science achievement across gender in TIMSS-R Turkey. *Education and Science*, 35(155), 26-35.
- [92] Van de Vijver, F. J. R. and Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Mahwah, NJ: Lawrence Erlbaum Associates.
- [93] Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-69.
- [94] Wendt, H., Kasper, D., and Trendtel, M. (2017). Assuming measurement invariance of background indicators in international comparative educational achievement studies: a challenge for the interpretation of achievement differences. *Large-Scale Assessments in Education*, 5(1).
- [95] Widaman, K. F., and Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *Educational and Psychological Measurement*, 43, 729-740.
- [96] Widaman, K. F., Ferrer, E., and Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, 4(1), 10-18.

APPENDIXES

Appendix 1: Timetable

Activity	2019					2020					
	Jun	Aug	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun
Write a research proposal											
Thesis Research Proposal Defence											
Request for Project Datasets											
Final research proposal submission											
Extended Literature Review											
Data preparation											
Data analysis and its interpretation											
Report writing											
First draft report submission											
The final draft of the thesis											
Thesis defence											
Final thesis submission											

Appendix 2. Request Letter for Dataset



T.C.
MERSİN ÜNİVERSİTESİ REKTÖRLÜĞÜ
MERSIN UNIVERSITY RECTORATE
DEPARTMENT OF EDUCATION
DIVISION OF MEASUREMENT AND
EVALUATION IN EDUCATION



The Director,
Examinations Council of Zambia,
P O Box 50432,
Lusaka.

Dear the Director,

I am a supervisor for Brian Mumba a master's student at Mersin University in Turkey.

We are planning for his master's thesis entitled "Assessing the Measurement Invariance of Early Grade Mathematic Assessment (EGMA) Tool by Different Demographic Variables for the Zambian Sample". So, we request for National Assessment Dataset for grades 2 and 3.

I sincerely request as his academic supervisor that he be granted access to EGRA/EGMA/SSME data sets for the following years 2014 and 2018.

The named datasets will be purely used for academic purposes. Your permission to access the datasets will be greatly appreciated. Thank you for your time and consideration in this matter.

Sincerely,

Assoc. Prof. Devrim ALICI
Head of Measurement and Evaluation
in Education Department
Mersin University, TURKEY

10 / 07 / 2019

Adres : Mersin University Çiftlikköy Campus, Faculty of Education
Department of Measurement and Evaluation in Education, PO Box 33343,
Mersin/ TURKEY
Ayrıntılı bilgi için irtibat:
Tel : 324 3610001/11214 Fax :
e-posta : devrimozdemir@mersin.edu.tr Elektronik Ağ : www.mersin.edu.tr



MEÜ.ID.FR-001/03

Sayfa ..J..

Yayın Tarihi : 22 / 11 / 2013

Appendix 3: R Syntax Measurement Invariance Analysis

```
# Measurement Invariance Analysis in R
getwd()
setwd("C:/Users/HP/Google Drive/My projects")
library(pacman)
pacman::p_load(pacman, dplyr, psych, lavaan, semPlot, polycor)
mydata1 <- readRDS("EGMA_2")
names(mydata1)
summary(mydata1$age)
# Subsetting only the items
mydata2 = subset(mydata1, select = num1:misnum10)
head(mydata2)
nrow(mydata2)
# Reliability.....
# Using sapply to convert factors items to numeric
must_convert<-sapply(mydata2,is.factor)
M2<-sapply(mydata2[,must_convert],unclass)
Factor_items <-cbind(mydata2[,!must_convert],M2)
str(Factor_items)
factor1= c(1:20)
factor2 = c(21:30)
factor3 = c(31:40)
psych::alpha(Factor_items[, factor1])
psych::alpha(Factor_items[, factor2])
psych::alpha(Factor_items[, factor3])
# Model Fit
mydata = mydata1
names(mydata)
# ordereing items
for(i in 1:42) {mydata [,i]= ordered(mydata[,i])}
class(mydata$num1)
# Testing model fit by Gender
Gender_data <- select(mydata, -c(grade,age))
# Fit measures
order = c("df", "chisq", "cfi", "RMSEA", "SRMR")
Egma_model=
"
```

```

Identification                                     =~
num1+num2+num3+num4+num5+num6+num7+num8+num9+num10+num11+num12
+num13+num14+num15+num16+num17+num18+num19+num20

Discrimination                                     =~
quant1+quant2+quant3+quant4+quant5+quant6+quant7+quant8+quant9+quant10

Recognition                                       =~
misnum1+misnum2+misnum3+misnum4+misnum5+misnum6+misnum7+misnum8+misnum9
+misnum10
"

Gender_fit = cfa(Egma_model, data=Gender_data, estimator="WLSMV")
summary(Gender_fit, standardized=T, rsquare=T)
fitmeasures(Gender_fit, fit.measures = order)
semPaths(Gender_fit, "standardized", layout = "tree2")
fitGender= fitmeasures(Gender_fit, fit.measures = order)
write.csv(fitGender, file = "Fitmeasures1.csv")
# Three model fit per group
#male
table (Gender_data$Gender)
Male_fit<-cfa(Egma_model,data=subset(Gender_data,Gender=="Male"),
estimator="WLSMV")
fitmeasures(Male_fit, fit.measures=order)
fitmale= fitmeasures(Male_fit, fit.measures = order)
write.csv(fitmale, file = "Fitmeasures2.csv")
#Female
Female_fit<-cfa(Egma_model,data=subset(Gender_data,Gender=="Female"),
estimator="WLSMV")
fitmeasures(Female_fit, fit.measures=order)
fitfemale= fitmeasures(Female_fit, fit.measures = order)
write.csv(fitfemale, file = "Fitmeasures3.csv")
# Assessing Measurement Invariance; Three structure Model
# Gender group
# Configural Invariance
Model_Conf_fit      <-cfa(Egma_model,data=Gender_data,      group="Gender",
estimator="WLSMV")
fitmeasures(Model_Conf_fit, fit.measures=order)
layout(t(1:2))

```

```

semPaths(Model_Conf_fit, "mod", "est", ask=FALSE, layout = "tree",
         levels = c(1,2,4,5), edge.color = "black", mar = c(2,2,4,2))

# Weak Invariance
Model_Weak_fit      <-cfa(Egma_model,data=Gender_data,          group="Gender",
group.equal=c("loadings"), estimator="WLSMV", parameterization ="theta")
fitmeasures(Model_Weak_fit, fit.measures=order)
Weak_fit= fitmeasures(Model_Weak_fit, fit.measures = order)
write.csv(Weak_fit, file = "Fitmeasures5.csv")

# Metric Invariance
Model_scalar_fit    <-cfa(Egma_model,data=Gender_data,          group="Gender",
group.equal=c("loadings","thresholds"), estimator="WLSMV", parameterization ="theta")
fitmeasures(Model_scalar_fit, fit.measures=order)
scalar_fit= fitmeasures(Model_scalar_fit, fit.measures = order)
write.csv(scalar_fit, file = "Fitmeasures6.csv")

# Strict Invariance
Model_Strict_fit     <-cfa(Egma_model,data=Gender_data,          group="Gender",
group.equal=c("loadings", "thresholds", "residuals"), estimator="WLSMV", parameterization
="theta")
fitmeasures(Model_Strict_fit, fit.measures=order)
Strict_fit= fitmeasures(Model_Strict_fit, fit.measures = order)
write.csv(Strict_fit, file = "Fitmeasures7.csv")

# Assessing Measurement Invariance Per Factor Model According to Gender
# Factor one: Number identification Model
Factor_1_model=
"
Identification
num1+num2+num3+num4+num5+num6+num7+num8+num9+num10+num11+num12
+num13+num14+num15+num16+num17+num18+num19+num20
"

# Configural Invariance
F1_Conf_fit         <-cfa(Factor_1_model,data=Gender_data,          group="Gender",
estimator="WLSMV")
fitmeasures(F1_Conf_fit, fit.measures=order)
conf_fit1= fitmeasures(F1_Conf_fit, fit.measures = order)
write.csv(conf_fit1, file = "Fitmeasures8.csv")

# Weak Invariance

```

```

F1_Metric_fit      <-cfa(Factor_1_model,data=Gender_data,      group="Gender",
group.equal=c("loadings","thresholds"), estimator="WLSMV", parameterization ="theta")
fitmeasures(F1_Metric_fit, fit.measures=order)
Weak_fit1= fitmeasures(F1_Metric_fit, fit.measures = order)
write.csv(Weak_fit1, file = "Fitmeasures9.csv")

# Strict Invariance
F1_Strict_fit      <-cfa(Factor_1_model,data=Gender_data,      group="Gender",
group.equal=c("loadings", "thresholds", "residuals"), estimator="WLSMV", parameterization
="theta")
fitmeasures(F1_Strict_fit, fit.measures=order)
Strict_fit1= fitmeasures(F1_Strict_fit, fit.measures = order)

# Factor Two: Number discrimination Model
Factor_2_model=
"
Discrimination
quant1+quant2+quant3+quant4+quant5+quant6+quant7+quant8+quant9+quant10
"

# Configural Invariance
F2_Conf_fit      <-cfa(Factor_2_model,data=Gender_data,      group="Gender",
estimator="WLSMV")
fitmeasures(F2_Conf_fit, fit.measures=order)
conf_fit2= fitmeasures(F2_Conf_fit, fit.measures = order)
write.csv(conf_fit2, file = "Fitmeasures11.csv")

# Metric Invariance
F2_Metric_fit      <-cfa(Factor_2_model,data=Gender_data,      group="Gender",
group.equal=c("loadings", "thresholds"), estimator="WLSMV", parameterization ="theta")
fitmeasures(F2_Metric_fit, fit.measures=order)
Weak_fit2= fitmeasures(F2_Weak_fit, fit.measures = order)

# Strict Invariance
F2_Strict_fit      <-cfa(Factor_2_model,data=Gender_data,      group="Gender",
group.equal=c("loadings", "thresholds", "residuals"), estimator="WLSMV", parameterization
="theta")
fitmeasures(F2_Strict_fit, fit.measures=order)
Strict_fit2= fitmeasures(F2_Strict_fit, fit.measures = order)

#Factor three: Recognition of Number Patterns Model
Factor_3_model=

```

```
"
Recognition
misnum1+misnum2+misnum3+misnum4+misnum5+misnum6+misnum7+misnum8+misnum9
+misnum10
"

# Configural Invariance
F3_Conf_fit      <-cfa(Factor_3_model,data=Gender_data,      group="Gender",
estimator="WLSMV")
fitmeasures(F3_Conf_fit, fit.measures=order)

# Metric Invariance
F3_Metric_fit      <-cfa(Factor_3_model,data=Gender_data,      group="Gender",
group.equal=c("loadings", "thresholds"), estimator="WLSMV", parameterization ="theta")
fitmeasures(F3_Metric_fit, fit.measures=order)

# Strict Invariance
F3_Strict_fit      <-cfa(Factor_3_model,data=Gender_data,      group="Gender",
group.equal=c("loadings", "thresholds", "residuals"), estimator="WLSMV", parameterization
="theta")
fitmeasures(F3_Strict_fit, fit.measures=order)
```

CURRICULUM VITAE

Name ve Surname : Brian Mumba

Date of Birth : 23 January 1990

E-mail : brianmumba0@gmail.com

Education Background:

Levels	Program	University	Year
Undergraduate	Adult Education	The University of Zambia	2011-2015
Graduate	Educational Measurement and Evaluation	Mersin University	2017-2020

Professional Background:

Service	Place of Work	Year
Research Assistant	Renavatio Business Service (Lusaka-Zambia)	2019-present
English Teacher	Mavikent College, Mersin Turkey	2019-2020
Social Protection Ambassador	UNDP Mersin - Turkey	2019-2019
Research Assistant (internship)	International Youth Federation (Mersin-Turkey)	2018-2019

Publications

1. Mumba, B., and Alici, D. (2019). Understanding Quantitative versus Qualitative Dichotomies. Social Sciences Education eJournal, Volume 2, Issue 7.
2. Mumba, B. (2019). The Call for the Introduction of University Entrance Exams in Zambia: The Case of University Admission Criteria. Pedagogy journal, Volume 2, Issue 9.
3. Mumba, B. (2019). The skill of the 22nd century in the automated world: How can education prepare its students for future work? The 5th International Students Social Science Congress Conference Presentation, Karadeniz Teknik University, Trabzon.